RAD and the demographic history of a hybrid zone

new insights into the evolution of hybrid sterility

Claudius Kerth

Department of Animal and Plant Sciences

The University of Sheffield

This dissertation is submitted for the degree of

Doctor of Philosophy

Faculty of Science

March 2018

I would like to dedicate this thesis to my grandmother who has never stopped believing in me.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Claudius Kerth
March 2018

Acknowledgements

I would like to acknowledge Anja Westram, Ludovic Duvaux, Helen Hipperson, Toni Gossmann, Stefan Prost, Simon Baxter, John Davey, Paul Etter, Eric Johnson and Günter Köhler for insightful discussions, lending me a helpful hand once in a while, openly sharing their knowledge or supporting me during field work. I would also like to thank Jamie Hutchison and Roger Butlin for sampling the grasshoppers that I used in this study. Roger Butlin has acquired funding for a large part of this project that included genomic data acquisition and the purchase of necessary computer hardware. He has also suggested the RAD marker method to me. I am grateful to him for his generosity. My special thanks go to Martina Ottenbruch, Christian Kerth and Käthe Kerth for being my hosts during the last one and a half years. I couldn't have done it without them.

Abstract

Restriction Site associated DNA (RAD) is a molecular method involving restriction digestion and high throughput DNA sequencing. It promises the systematic subsampling of the genome and highly repeatable scoring of genetic variation in hundreds of individuals at current sequencing costs. However, it comes with its own problems. De novo assembly of RAD sequence data usually creates many putative reference tags that are only found in one or a few individuals leaving only relatively few markers for population genomic analyses. In the first chapter, I investigate three potential reasons for this outcome – incomplete digestion, genomic religation and insufficient DNA template amount – by looking at the occurrence of restriction enzyme recognition sequences within the resultant sequencing data of two different types of RAD libraries.

Analysis of sequence clusters as well as the proportion of concordantly mapping read pairs against a *Locusta* reference sequence suggest that incomplete digestion has affected one of the restriction enzymes used and thereby the number of loci that could be sequenced at sufficient depth across individuals. The other restriction enzyme is found to be much less affected by incomplete digestion and instead random religation of restriction fragments indicates an inefficient adapter ligation step that also leads to low read depths across individuals. Finally, qPCR and read mapping against four newly reconstructed paired-end (PE) contig pair reference sequences suggests that low amount of starting DNA and/or high loss of DNA during the library preparation are a major cause for the locus drop-out observed in the de novo assembled read data.

In the second part of this thesis, I use RAD sequence data to make inferences about several aspects of the demographic history of two grasshopper subspecies that form a hybrid zone in the Pyrenees between France and Spain. Sequence data was generated from 36 individuals sampled at the two opposite ends of a hybrid zone that is characterised by hybrid male sterility. I use a state-of-the-art de novo assembly strategy that utilises the shotgun-type PE reads from standard RAD to distinguish alleles from paralogs. I then conduct several population genomic analyses with the programme `ANGSD` that incorporates uncertainty in genotypes by using genotype likelihoods instead of called genotypes. Results based on more than 1 million filtered sites confirm the high genetic differentiation of the two subspecies found in pre-genomic studies and a surprisingly high genetic diversity in the subspecies that is thought to be derived from a very distant glacial refuge. Further, demographic modelling

with the programme $\delta a\delta i$ reveals a robust signal of low but significant gene flow during the divergence of the two subspecies ($Nm \simeq 0.471$, until about 25 thousand years ago (kya)). Allowing for gene flow roughly doubles the divergence time estimate from about 0.5 to 1.1 million years ago (mya). The divergence time estimate without allowing for gene flow is highly consistent with previous estimates from a mitochondrial sequence marker. A history of divergence with gene flow also indicates that alleles causing Dobzhansky-Muller incompatibilitys (DMIs) are unlikely to have risen in frequency by genetic drift alone. The gene flow is clearly asymmetric between the two subspecies in line with many previous studies of the hybrid zone that indicated asymmetric introgression in the same direction. There is no signal of recent (postglacial) gene flow in the data set. However, this may well be due to a lack of power. Further analysis of this data set promises to yield more insights, e.g. loci potentially under divergent selection between the two subspecies.

Table of contents

[1] Let the adapter oligos anneal slowly over a couple of hours in the heat block or thermal cycler. The adapter can be tested through ligation to a Taq PCR product and subsequent test PCR with the following P2Y primer: 5' – TCTCGGCATTCCTGCTGAAC – 3' (Kang-Wook Kim)

[2] optional because of the cost and effort involved with cloning, but recommended before spending a lot of money on Solexa sequencing

[3] optional because of the cost and effort involved with cloning, but recommended before spending a lot of money on Solexa sequencing

Glossary

all pairs all the pairs of sequences below a given Levenshtein distance are identified during the graph construction phase. 66

barcode short DNA sequence incorporated into adapter oligonucleotides that becomes part of the sequence read. Barcodes are used in order to be able to pool the DNA of different individuals, populations, treatments, etc. into one library that can be sequenced on one lane of an illumina flow cell. 9, 46

C$_t$ PCR cycle when a certain fluorescent threshold is reached. 54

concordant A read pair is called concordant if it aligns with the expected relative mate orientation (here: forward–reverse or reverse–forward) and within the expected range of distances between mates. This is also called a proper pair. The complement of discordant. xvi, 32, 39, 47, 62, 63

connected component All nodes (here sequence reads) after all–pairs search (and before clustering!), that are directly connected by an edge or indirectly connected via several nodes, belong to the same connected component. 66

contig longer consensus sequence derived from assembling smaller overlapping sequence reads. 25, 26, 50, 69

discordant A read pair is called discordant if it aligns without the expected relative mate orientation (here: forward–reverse or reverse–forward) or outside the expected range of distances between mates. Note that `bowtie2` only calls discordant read pair mappings if both reads map *uniquely*. Here, I am NOT adopting this requirement. xv, 32

edit distance minimum number of operations (one symbol insertion, deletion or substitution) required to change one string of symbols into another. Also known as *Levenshtein distance*. 17, 52, 66, 67

Expect (E) value The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. 50

fragment not a PCR duplicate. With paired reads from standard RAD (i. e. including random shearing of restriction fragments) typically identified by having different PE read sequences or different insert sizes after read mapping against a reference. 26–29, 56, 67

heterochromatin Chromatin that remains in a highly condensed state throughout the cell cycle. 44

index similar to barcode and serves the same purpose; generally incorporated into the centre of the adapter so that special sequencing run for the index is required. 10

kmer subsequence with a specified length (k) of a longer sequence. 25, 48, 49, 71, 107

linked RAD tag site position in the reference sequence with at least one concordant read pair on each side of a putative restriction site and the SE reads overlapping each other as expected from the restriction enzyme. 25, 27, 47, 51

mapping quality score The mapping quality score Q is the Phred transformation of the estimate of the probability p that the reported mapping position does not correspond to the read's true point of origin: $Q = -10\log_{10} p$. The way p is estimated is different for each mapping programme, but in any case a mapping quality score Q of 3 roughly corresponds to a mis-mapping probability p of 0.5, i. e. the read has an estimated 50% chance to have derived from a location other than the one reported. 46

population minor allele frequency The population minor allele frequency is the (unknown) frequency of the minor allele in the entire population (as opposed to the sample).. 76, 78

proper pair read pair from illumina paired-end sequencing that got mapped to a reference in the correct orientation within a maximum expected distance from each other that is determined by the fragment size selection during the sequencing library preparation. Also called a concordantly mapping pair. xv, 27, 32, 46, 47, 62–64, 69

RAD tag genetic marker from RAD sequencing; the sequence up or downstream of a restriction site. 13, 23, 25–27, 40, 41, 45, 49, 51, 52, 56, 69

read any sequence that comes out of the sequencer. 9

sample allele frequency The sample allele frequency is the frequency of the allele among the individuals in a specific sample. 76, 80, 117

SbfI restriction enzyme with the recognition sequence CCTGCA↓GG. 22–24, 29, 42–45

site frequency spectrum Also known as allele frequency spectrum (AFS). It is constructed by computing the sample frequency (i. e. an integer ≥0) of the ancestral (unfolded) or minor allele (folded) at each nucleotide site. The SFS is then the histogram of the number of sites at each frequency. 11, 69, 76, 77, 79–83, 86, 87, 91–96, 100, 109, 117, 127, 129

XhoI restriction enzyme with the recognition sequence C↓TCGAG. 23, 45

Acronyms

bp base pair. 22, 23, 43, 45, 48, 50, 54, 62

CI confidence interval. 73, 80, 82, 84, 85, 88, 94, 97–100, 103, 125, 127, 128

ddRAD double digest RAD. 9, 11

DEM digital elevation model. 42

DMI Dobzhansky-Muller incompatibility. x, 125–127

EM Expectation Maximisation. 80

EST expressed sequence tag. 46

GWAS genome wide association study. 13, 14

HWE Hardy Weinberg equilibrium. 17, 76, 93

indel small sequence insertion or deletion polymorphism. 9, 16, 27, 53, 67, 76, 107, 109

kya thousand years ago. x, 6, 8, 100, 102, 104

LD linkage disequilibrium. 11

LRT likelihood ratio test. 65, 70, 82, 83, 96, 97, 99, 103, 105, 109, 111

MAF minor allele frequency. 78, 79, 86–90, 109–111

Mbp mega base pairs. 68–70

ML maximum likelihood. 76, 87, 91, 97–100, 103

mya million years ago. x, 106, 125

PE paired-end. ix, 9, 10, 13, 23–25, 34, 43–45, 47, 56, 67–69

RAD Restriction Site associated DNA. ix, 9–11, 21, 29, 66

SAM Sequence Alignment/Map format. 46, 62, 63

SE single-end. 10, 13, 23, 24, 34, 42, 43, 45, 47, 56, 66–69, 107

SNP single nucleotide polymorphism. 9, 48, 49, 69, 70, 91, 127

Chapter 1

General Introduction

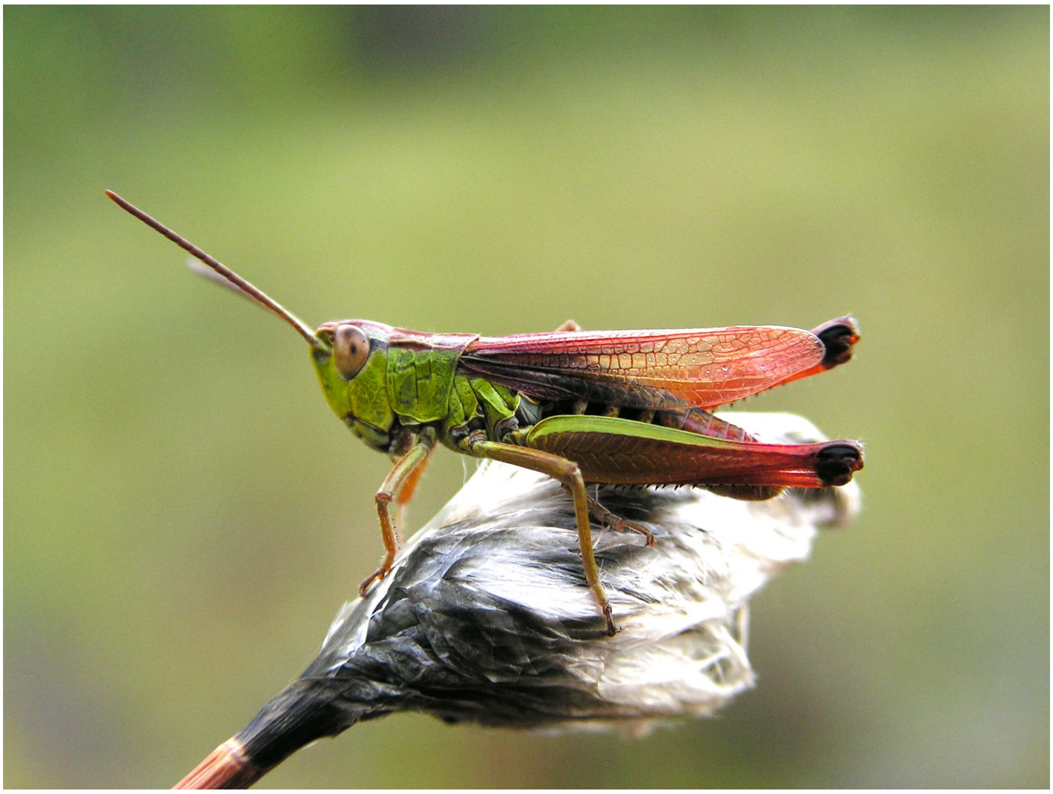


Fig. 1.1 *Chorthippus parallelus parallelus*. Male individual from Finland.

1.1 Hybrid Zone in the Pyrenees

A hybrid zone of the grasshopper *Chorthippus parallelus* (Orthoptera: Acrididae) occurs along the Pyrenees mountains between France and Spain where an Iberian subspecies, *erythropus*, meets the nominate subspecies *parallelus* (Butlin and Hewitt, 1985a,b). In Europe, the subspecies *parallelus* is distributed from the Balkans in the southeast to southern France, northwards up to England (but not Ireland) and to southern Finland and eastwards at least until the Ural and Caucasus mountains. On the Iberian peninsula, it is replaced by the subspecies *erythropus*. The two subspecies meet and hybridise in transverse mountain passes (cols) of the Pyrenees. The hybrid zone is not confined to the High Pyrenees. The hybrid zone extends into the eastern (Catalan region) and western foothills (Basque region) of the Pyrenees (Buno et al., 1994; Hewitt, 1993). The hybrid zones in the Basque region and probably also the Catalan region appear to be much wider than in the High Pyrenees (Buno et al., 1994; Vazquez et al., 1994). This is likely due to greater population densities and fewer barriers to dispersal (e. g. high mountains) allowing a greater degree of mixing.

In what characters have subspecies diverged?

The two subspecies have diverged in many characters. They include morphological characters (Butlin and Hewitt, 1985b; Butlin et al., 1991), mating behaviour, i.e. male calling and courtship song, cuticular hydrocarbons and female preference (Buckley et al., 2003; Butlin and Hewitt, 1985a; Ritchie, 1990; Ritchie et al., 1989), chromosomal characters (Bella et al., 2007; Gosálvez et al., 1988) and a neuropeptide (Roth et al., 2007). The two subspecies also have different patterns of infection by the bacterial endosymbiont *Wolbachia* (Zabal-Aguirre et al., 2010, 2014) with a generally lower infection frequency in *parallelus* than in *erythropus*.

Phylogeographic studies of *Chorthippus parallelus* in Europe

The historical biogeography and population structure of *C. parallelus* in Europe has been extensively studied by Cooper et al. (1995) and Lunt et al. (1998). Based on these studies, the current hybrid zone is believed to have formed by secondary contact following range expansion after the last glaciation from two different refugia as the ice retreated in the Pyrenees some 9-12,000 years ago. The absence of shared haplotypes between Spanish and French populations at an anonymous nuclear sequence marker (cpnl-1) indicates extensive lineage sorting, which can only be caused by a substantial time of divergence with no or very low gene flow (Cooper and Hewitt, 1993). Mitochondrial sequence divergence between *erythropus* and *parallelus* indicates a total divergence time of around 500,000 years (Lunt et al., 1998). This calculation assumed a mutation rate of 2% per million years and no gene flow during divergence.

It seems very likely that large fluctuations in climate, like ice ages, and the concomitant habitat changes have played a major role in repeatedly separating the ranges of the two subspecies for thousands of generations (Hewitt, 1990). Secondary contact during interglacial periods would then have allowed some gene exchange between the two subspecies. This would have reduced genetic divergence at least for those populations close to the contact zone. However, those populations would go extinct at the onset of the next ice age. So any gene flow that did not reach populations in glacial refuges would not have affected the longterm divergence of the two subspecies.

Phylogeographic studies strongly suggest that the *parallelus* populations on the French side of the Pyrenees are derived from populations in the Balkans (southeastern Europe) and that *erythropus* populations on the Spanish side of the Pyrenees are derived from a refuge in the south of the Iberian peninsula. Strong molecular genetic and chromosomal differentiation within Spain suggests that current *erythropus* populations are derived from several independent refugia in southern Spain (Bella et al., 2007; Cooper and Hewitt, 1993). In particular, there seems to be strong divergence between Pyrenean and northern Spanish populations on the one hand and central and southern Spanish populations on the other hand.

Hybrid zone as a natural laboratory

With an average haplotype divergence of 1.7% at a single nuclear locus (Cooper and Hewitt, 1993), the two subspecies may still be at an intermediate stage in the process of speciation (Mallet, 2008; Roux et al., 2016). This would make this system suitable for the study of reproductive isolation mechanisms that evolve during the early stages of speciation which do not necessarily have to be similar to those that evolve toward the end of speciation or after speciation is complete.

In the laboratory, F_1 hybrid males produced from crosses between 'pure' *parallelus* and 'pure' *erythropus* taken from either side of the hybrid zone are almost completely sterile with degenerate testes and a severely disrupted meiosis (Bella et al., 1990; Hewitt et al., 1987; Virdee and Hewitt, 1992). The fact that F_1 males are affected by sterility, not only F_2 or backcross males, indicates that at least one locus must be (co-)dominant in its expression of the genetic incompatibility (the other locus may be recessive if located on the X chromosome). Hybrid females on the other hand are fully fertile. This is therefore an example of Haldane's rule (Turelli and Orr, 1995). Males collected from the hybrid zone, however, are almost completely fertile (Ritchie et al., 1992). The absence of sterile males in the hybrid zone is likely due to the reconstruction of compatible (presumably ancestral) genotypes caused by selection against incompatible combinations of derived alleles (Gavrilets, 1997). The two subspecies are polymorphic for these negative epistatic interactions (roughly 10% of F_1 hybrid males from *pure parents* are fully fertile, Shuker et al., 2005) and thus compatible (i.e. ancestral) alleles can reach the centre of the hybrid zone. This also means that genetic markers, even when in complete linkage disequilibrium with a causative variant underlying sterility, are not necessarily expected to be divergently fixed between the two subspecies.

Within a transect of the hybrid zone, clines for different characters vary greatly in width suggesting that the underlying genes have introgressed to different degrees since the secondary contact of the two subspecies. Selection in hybrids will restrict the introgression of those parts of the genome that are responsible for Dobzhansky-Muller incompatibilities between the genomes of the two subspecies, whereas neutral loci should introgress to a large extent into the other subspecies' genomic background by recombination. The clines of several characters are also much wider than would be expected for neutral introgression given the life time dispersal distance of these grasshoppers of less than 30 m (Virdee and Hewitt, 1990) and the number of generations since secondary contact of around 9,000 years (Hewitt, 1990). Those broad clines are a likely relict of the colonisation of the hybrid zone by rare long distance dispersal (Nichols and Hewitt, 1994). Initially broad clines for characters under selection in hybrids should on the other hand have converged to equilibrium narrow clines in just a few hundred generations (Barton and Hewitt, 1985). This should increase the power to detect genetic markers under selection in hybrids in a hybrid zone analysis with a dense genome-wide marker set (Gompert and Buerkle, 2011a). In addition, the bimodal distribution of testis follicle lengths (a good proxy for sterility) of backcross hybrids suggests that only a few loci of large effect contribute to hybrid sterility [Llewellyn, 2008, fig. 4-3 and Shuker et al., 2005].

The two subspecies also have divergent infection patterns with the endosymbiont *Wolbachia* (Zabal-Aguirre et al., 2010). That is a low infection frequency in the *parallelus* populations where the B strain of *Wolbachia* is predominant versus a high infection frequency in *erythropus* populations where the F strain is predominant. Unidirectional cytoplasmic incompatibility causes an average reduction of 32.5% in the proportion of eggs with embryos as compared to crosses between uninfected individuals (Zabal-Aguirre et al., 2014). This, in addition with the slightly less severe bidirectional cytoplasmic incompatibility, should lead to considerable reproductive isolation between the two subspecies. So, male hybrid sterility and divergent mating systems are not the only reproductive barriers between the two subspecies. The occurrence of several reproductive barriers also speaks for a rather old divergence of the two subspecies.

Another subspecies of *C. parallelus* has been detected in Italy from nuclear sequence (Cooper et al., 1995) and chromosomal differentiation and a hybrid zone between the Italian subspecies and the northern European and Balkan subspecies (*parallelus*) based on cytogenetic markers has been described in the Alps (Flanagan et al., 1999). Interestingly, the Italian subspecies seems to be more closely related to the Iberian subspecies (*erythropus*) than to *parallelus* (Cooper et al., 1995) and, in contrast to the Pyrenean hybrid zone, natural male hybrids have been found that do show meiotic abnormalities and abnormally developed sperm. This may indicate that the hybrid zones in the Alps are much younger than those in the Pyrenees.

Evidence for asymmetric introgression across the hybrid zone

Clines of several morphological as well as molecular markers have been shown to be very wide in different transects through the hybrid zone (up to 40 km) (Butlin and Hewitt, 1985a,b; Butlin et al., 1991; Vazquez et al., 1994) and their centres are often displaced from each other, which indicates substantial gene flow across the hybrid zone.

Bella et al. (1992) have found preferential homogamy (i.e. production of pure rather than hybrid offspring) in *parallelus* females sequentially mated with males from both subspecies. In a similar study, Hutchison (2013) found that *parallelus* males outcompete *erythropus* males in siring offspring regardless of the females' subspecies status. This indicates some mechanism(s) of postmating isolation. A very good candidate for this is unidirectional cytoplasmic incompatibility caused by divergent frequencies of *Wolbachia* infection in the populations used for those crossing experiments (Zabal-Aguirre et al., 2010, 2014). The fact that *parallelus* males sire more offspring in both types of matings indicates a greater potential for introgression of the *parallelus* genome (including the X chromosome) into the *erythropus* genome than vice versa. This is particularly true when males contribute more to gene flow than females due to their higher dispersal distances as has been estimated in a closely related grasshopper species (Bailey et al., 2003).

Lunt et al. (1998) have found that a population from Pyrenean Spain (Escarilla) is more similar to French populations at a mitochondrial sequence marker (COI) than to all other Spanish populations. If this pattern is a result of introgression of mitochondrial haplotypes from *parallelus* into Pyrenean *erythropus*, then it could be driven by *Wolbachia* (Gompert et al., 2008; Hurst and Jiggins, 2005; Raychoudhury et al., 2010). This is supported by the finding of Zabal-Aguirre et al. (2010) that, in the hybrid zone, the shift of infection from *parallelus* type to *erythropus* type occurs on the Spanish side of the Pyrenees. Mitochondrial and *Wolbachia* genomes should be in strong linkage disequilibrium, since both are maternally inherited. The spread of a *Wolbachia* strain (or the uninfected state) should therefore also lead to the spread of its associated mitochondrium.

Ferris et al. (1993) have shown that an X-chromosomal nucleolar organiser region (NOR, a cluster of ribosomal DNA) specific to *parallelus* has introgressed deeply into *erythropus* genomic background in the Pyrenean hybrid zone. At the transect across the Col de la Quillane the cline centre for this marker is some 15 km south of the centre of most other clines. This indicates substantial introgression of X-chromosomal sequences from *parallelus* into *erythropus*. A much more extreme indication of introgression has been found by Bella et al. (2007). They have shown that an X-chromosomal interstitial heterochromatic band that is closely associated with the NOR characteristic of *parallelus* has introgressed about 400 km into the northwest of the Iberian peninsula into the south of the Cantabrian mountains. Interestingly, the X-NOR did not introgress or is not active in *erythropus* genomic background.

There is not just evidence of introgression from *parallelus* into *erythropus*. Butlin et al. (1991) have shown that at a transect in the western Pyrenees (Col du Pourtalet) many morphological characters have cline centres that are greatly displaced towards the north into *parallelus* territory. However, this is the only study that I am aware of that indicates substantial introgression from *erythropus* into *parallelus*. So there seems to be ample direct and indirect evidence for substantial introgression from *parallelus* into *erythropus* since the formation of the Pyrenean hybrid zone some 10 kya (and potentially also in previous interglacials) as detailed above and only scant evidence for gene flow in the other direction.

Dispersal and serial founder effects

Most species on Earth have experienced range shifts and expansions into their current habitat following the changes in environmental conditions due to climate warming at the end of the last ice age between 18,000 and 10,000 years ago. Many species are geographically subdivided into distinct subspecies which form hybrid zones where they meet. Many of these hybrid zones are secondary, having formed after allopatric divergence in different ice age refugia (Hewitt, 1999; Taberlet et al., 1998).

The recolonisation of previously inhospitable habitat during range expansion should have led to a series of founder events (Peter and Slatkin, 2015). These are severe population bottlenecks caused by the founding of new populations by very few colonists. This is followed by initially exponential (census) population size increase. Later gene flow should have little impact on gene frequencies in established populations (Hewitt, 1996; Waters et al., 2013). A series of founder events should produce a strong negative correlation between genetic diversity and geographical distance from the source population. Similarly, it should produce a strong positive correlation of genetic divergence with geographical distance from the origin of expansion (Peter and Slatkin, 2013). Both correlations are strongest right after the time of range expansion and should decay over time depending on the dispersal ability of the species.

In many species it has indeed been found that genetic diversity decreases the further away a population is from the origin of range expansion (e. g. in humans, Luca et al., 2011). In particular, populations from the north of a distribution range often have lower genetic diversity than southern populations, which are closer to glacial refugia (Hewitt, 1996). There are, however, also exceptions to this overall pattern. For example, Petit et al. (1999) have shown a lack of decreasing genetic diversity of mitochondrial sequences with distance to putative glacial refugia in the bat species *Nyctalus noctula*. It is very likely that the exceptional dispersal capability of this migratory species can at least partly explain this observation.

In *Chorthippus parallelus*, the spatial distribution of genetic variation across Europe at neither a nuclear nor a mitochondrial locus had clearly shown a reduction of genetic diversity with distance from the putative glacial refuge in the Balkans or Greece (see intraregionial K_S values in table 2 of Cooper et al., 1995 and table 1 of Lunt et al., 1998). There is also no increase of genetic distance with geographic distance to putative ancestral populations in the southern Balkans in those two data sets (see their pairwise K_{ST} values). The general reduction in genetic diversity from southern to northern European populations found by Cooper et al. (1995) does not necessarily need to be a signal from postglacial range expansion either. A possible alternative explanation would be smaller effective population sizes in the north due to greater forest cover. For instance, the dryer climate in the south could have allowed more wildfires to create the open grassland that is the suitable habitat for *C. parallelus*.

Serial founder models are based on stepping-stone dispersal that is over discrete locations that are colonised one by one and only by individuals from neighbouring locations. However, it is obvious that most species do not disperse that way. Individual based simulation studies of recolonisation by Ibrahim et al. (1996) have shown that the current spatial distribution of genotypes is very dependent on the *distribution* of individual lifetime dispersal distances of the species (i.e. not just the average individual dispersal distance) as well as on the dispersal of other species it depends on, e. g. food plant species. They included leptokurtic dispersal functions with fat tails that allow rare long distance dispersal. The simulations showed that colonists founding new populations can initially reproduce exponentially, while later migrants will arrive when the population size is closer to its carrying capacity and therefore contribute relatively little to the gene pool of the population. The result are large patches of reduced genetic diversity and those patches persist for many hundreds of generations. According to this study, recolonisation by rare long distance dispersal seems to produce a smaller increase of genetic distance with geographic distance than stepping-stone dispersal and this signal also decays more quickly over time. It also shows that rare long distance dispersal leads to a greater reduction in genetic diversity than stepping-stone dispersal.

However, when looking at a slightly broader spatial scale than individual subpopulations, the reduction in genetic diversity in a newly colonised area with respect to refugial areas due to serial founder events can be reversed if long distance dispersal is common enough (Bialozyt et al., 2006). In order for long distance dispersal to speed up recolonisation appreciably (e. g. reducing colonisation time by one half) as compared to diffusion dispersal, it's rate needs to be so high that genetic diversity on a regional scale is hardly diminished during colonisation. Founder populations do not have time to grow into large patches and admixture between adjacent founder populations as well as ongoing migration from the source of expansion could erode the small expansion signal of reduced local genetic diversity very quickly (Peter and Slatkin, 2013, fig. 3).

While Virdee and Hewitt (1990) have estimated the average dispersal distance for the flightless *Chorthippus parallelus* at only 30 m per generation, its postglacial range expansion must have proceeded at a much greater pace. Assuming the distance between the putative Balkan glacial refugia and Britain is about 2,000 km, with a dispersal distance of 30 m it would have taken about 67,000 years to reach Britain. For *C. parallelus* to reach Britain between the Younger Dryas cold spell at around 11 kya and the flooding of the English Channel at around 8 kya (Sturt et al., 2013), the average dispersal distance would have needed to be 600–700 m per generation (year) (Cooper et al., 1995). This indicates that long distance dispersal during postglacial range expansion of *C. parallelus* was common enough to allow preservation of genetic diversity at a regional scale over its whole current distribution range. The initially reduced genetic diversity at a local scale may have recovered since the end of recolonisation of northern Europe due to gene flow between adjacent founder populations.

Recently, a more powerful measure to detect range expansions has been proposed by Peter and Slatkin (2013). It is based on the expected increase of frequencies of shared alleles between source population and newly founded population due to genetic drift, a phenomenon also known as "allele surfing".

Of particular interest is whether the different processes associated with the postglacial recolonisation of northern Europe (including allele surfing and adaptation to different environments) accelerated the divergence that results in hybrid sterility between the two subspecies *parallelus* and *erythropus*. Tregenza et al. (2002) have investigated this and have found no significant effect associated with recolonisation during postglacial range expansion on testis follicle length (a good proxy for sterility) in crosses between populations across Europe, i.e. a cross between France and Spain does not produce more sterile offspring than a cross between Greece and Spain.

1.2 RAD

What is RAD?

RAD, aka *RADseq* and *sRAD*, is a restriction enzyme-based genotyping by sequencing (GBS) technique. It targets regions of the genome for sequencing that flank the recognition sites of restriction enzymes (fig. 1.2). Altshuler et al. (2000) were the first to use reduced-representation sequencing libraries from genomic restriction fragments for SNP discovery. After the introduction of massively parallel sequencing technologies, these were further developed for population allele frequency and genotype estimation (Andolfatto et al., 2011; Baird et al., 2008; Davey et al., 2011; Elshire et al., 2011; Van Tassell et al., 2008). By sequencing only the ends of restriction fragments, RAD creates sequencing libraries with reduced complexity as compared to whole-genome shotgun sequencing, thus subsampling the genome at homologous locations to identify and type single nucleotide polymorphisms (SNPs) and small sequence insertion or deletion polymorphisms (indels) more or less randomly throughout the genome and increasing read coverage for the enriched sites. It is therefore a cost-effective alternative to whole genome shotgun sequencing. If template molecules for sequencing are labelled with different barcodes from ligated adapter oligonucleotides, then individual or population samples can be pooled during library preparation and sequenced together in the same parallel sequencing run. Different reads from the same RAD tag are fully overlapping, which focuses coverage at those restriction sites and facilitates variant discovery and genotype calling for RAD tags as compared to shotgun sequencing. When standard RADseq according to Baird et al. (2008) is done with PE sequencing, the forward reads begin at the cut site but the second (reverse) reads are coming from the sheared end and are only partially overlapping each other (shotgun type reads). Those paired-end reads that originate from the same side of the same restriction site can be used to identify PCR duplicates and can be assembled into a few hundred base pair long contigs (fig. 1.2). double digest RAD (ddRAD) together with paired-end sequencing, i.e. sequencing from each end of the restriction fragment, produces two RAD tags per fragment, i.e. one on each end (fig. 1.3).

As opposed to microsatellite markers or target capture sequencing, RADseq can be used with study organisms that are currently lacking genome sequence information and allows genome-wide discovery of co-dominant genetic markers and their genotyping in one go (Davey et al., 2011). Only genome size and GC content are required to estimate the required number of reads for a certain target coverage. Previously, microsatellites had been the marker of choice for many population genetic studies. However, developing hundreds or even thousands of polymorphic microsatellite markers is costly and time-consuming for species without a close relative with a sequenced genome (DAWSON et al., 2010). Once

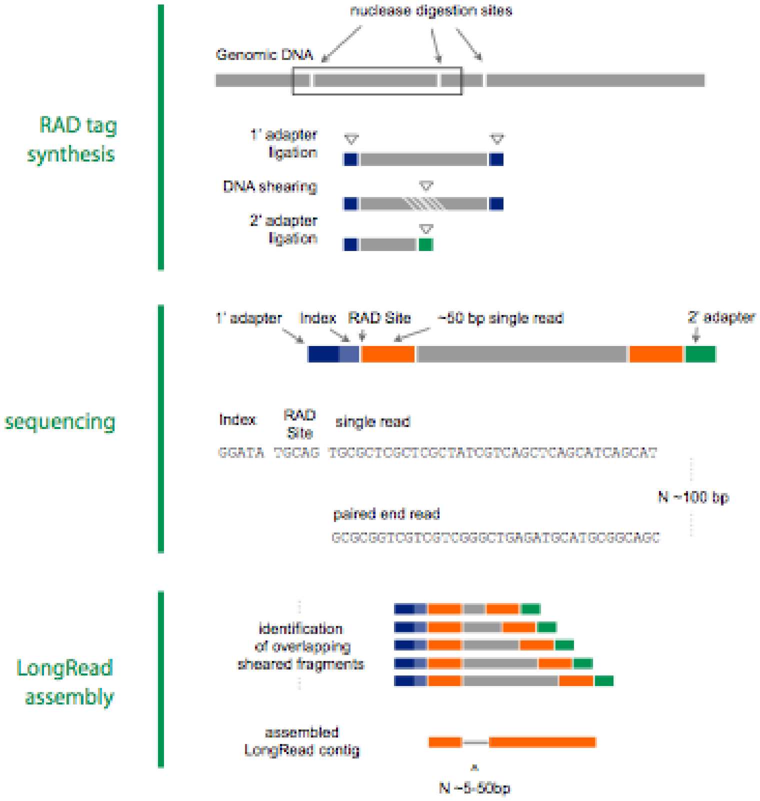


Fig. 1.2 Overview of the standard RAD marker technique according to Baird et al. (2008). (1) One restriction enzyme is used to digest genomic DNA. The first illumina adapter (P1), containing a different barcode (here called index) sequence for each individual, is then ligated to the restriction fragments. The restriction fragments are then sheared, usually by high-frequency sonication, into a fragment size range that is suitable for illumina sequencing, which is selected on an agarose gel. After ligation of the second illumina adapter (P2), fragments with at least one P1 adapter are enriched by selective PCR. (2) The single-end (SE) reads start with a barcode sequence, followed by the remainder of the restriction site. Only relatively short sequences (tags) are generated from the ends of the fragments. (3) Due to random shearing of restriction fragments, the PE reads start at variable genomic distances from the restriction site (unless they are PCR duplicates) and thus can be assembled into short PE contigs, depending on the size range selected on the gel. Taken from Atwood et al. (2011).

developed, the genotyping of many microsatellite markers in many individuals is again costly and time-consuming. In addition, mutation mode and mutation rates are generally not known for microsatellites. Like AFLP, RADseq does not require prior sequence information and selection of polymorphic markers for genotyping of population samples. However, instead of dominant restriction fragment length polymorphisms, RADseq provides sequence information from those fragments that makes determination of homology much easier (Hohenlohe et al., 2013) and allows the application of site frequency spectrum and haplotype-based population genomic analyses (Pool et al., 2010).

While many questions in evolutionary genetics can already be addressed with a fairly moderate number of markers (from 10's to 100's of loci) – e. g. inference of population structure, phylogeny, phylogeography, historical demography and gene flow as well as QTL mapping – some other questions greatly benefit from a dense genome-wide marker set (10's of thousands to several million markers depending on genome size). Genome scans, with samples from large, outbred populations in particular, greatly increase their chances of signal detection with a dense genome-wide marker set (Catchen et al., 2017; Lowry et al., 2016; McKinney et al., 2017). For instance, while association analysis of F2 or backcross individuals from experimental crosses or linkage analysis from wild pedigrees capture only few recombinations that allow for a disassociation of weakly linked markers to a causative allele, the same study within wild populations effectively uses many historical recombinations and thus allows in principle a much finer scale of mapping of loci affecting variation in the focal trait. Thus, genome-wide association studies can benefit from a high density of genetic markers across the genome, and more so for species with low genome-wide levels of linkage disequilibrium (LD). The same is true in principle for genome-wide scans for signals of selection, e. g. hybrid zone analyses (Gompert et al., 2012b).

Since *standard* version of RAD according to Baird et al. (2008) generates tags around every restriction site – one upstream and one downstream of it (figure 1.2), it allows for only limited complexity reduction. Currently available rare cutter restriction enzymes have an up to 8 bp long recognition sequence (ignoring ambiguous positions). Further complexity reduction can be achieved by skipping the step of shearing the restriction fragments and instead size selecting a range of restriction fragments that have the right size for the sequencing platform (Andolfatto et al., 2011; Elshire et al., 2011). This produces RAD tags only at cut sites which have a second cut site within a short distance from the first cut site. Further fine tuning of library complexity can be achieved by adding a second restriction enzyme to the protocol and sequence only fragments cut by both enzymes, so-called ddRAD (see fig. 1.3) (Peterson et al., 2012).

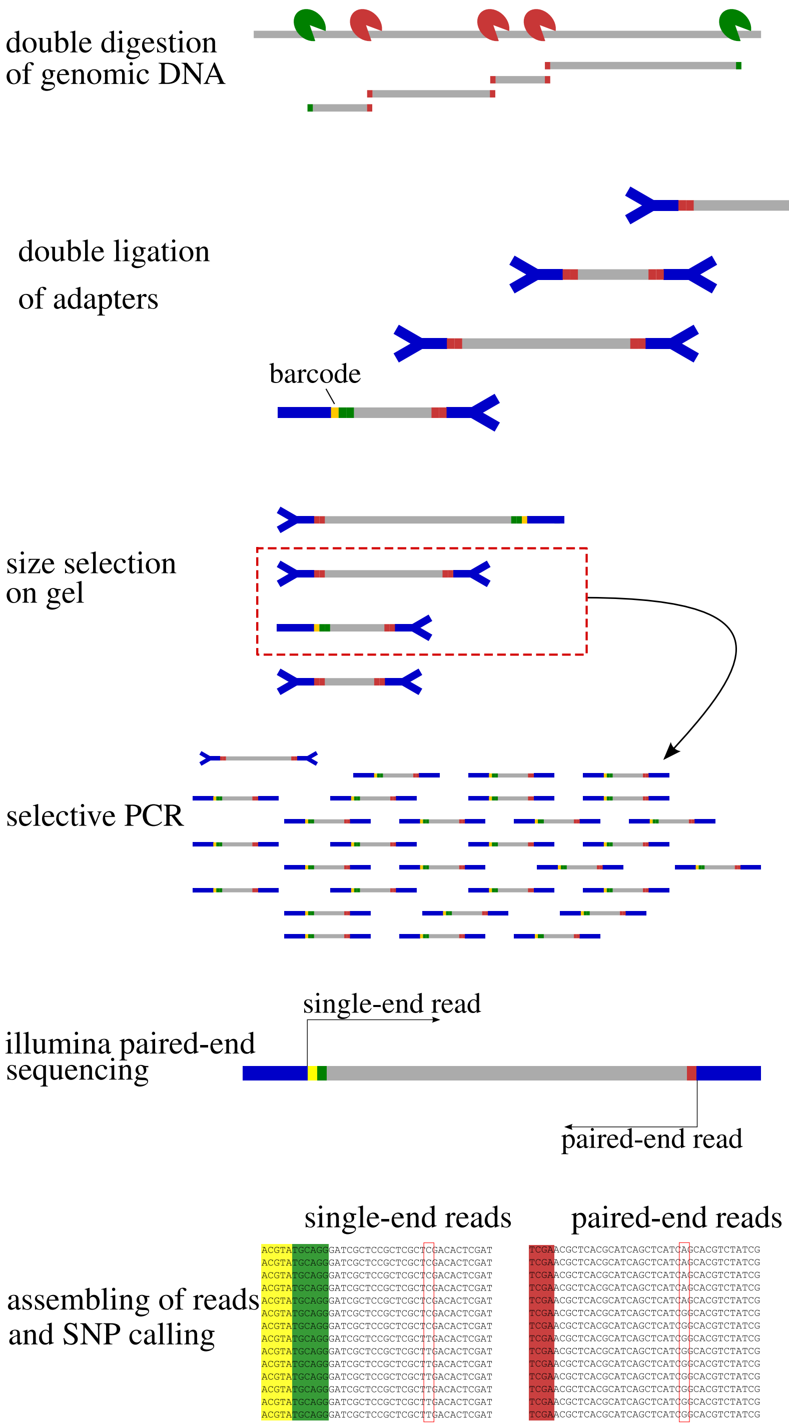

double digestion
of genomic DNA
double ligation
of adapters
barcode
P1 adapter
P2-Y adapter
size selection
on gel
selective PCR
illumina paired-end
sequencing
single-end read
paired-end read
single-end reads
paired-end reads
assembling of reads
and SNP calling
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCAGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCAGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCAGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCAGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCAGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTCGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG
ACGTATGCAGGGATCGCTCCGCTCGCTTGACACTCGAT TCGAACGCTCACGCATCAGCTCATCGGCACGTCTATCG

Fig. 1.3 Double-digest RAD protocol overview. Genomic DNA is first digested by two different restriction enzymes (red and green). illumina adapters (P1 and P2) are then ligated to the restriction fragments. The P2 adapter is a so-called divergent-Y adapter that, if anything, only contains the reverse-complement of the backward PCR primer binding site needed during the selective PCR step (see figure 2.14 on page 42). Restriction fragments are then size-selected on a gel. Here, in contrast to the standard RAD protocol (see fig. 1.2), gel size selection selects which markers get into the final sequencing library. The selective PCR step enriches the library for fragments with at least one P1 adapter ligated to it. Bridge-amplification on the illumina flow cell requires a P1 and a P2 adapter. illumina paired-end sequencing results in two RAD tags per fragment that can be assembled and used for SNP and indel calling.

When making the decision between whether to use standard RAD or ddRAD several things need to be considered. The first is surely whether standard RAD can provide enough complexity reduction. If that is the case, then the next consideration will be the availability of a sonicator for the random shearing of restriction fragments to a size suitable for the sequencing platform. The availability of a sonicator may not be critical though, since there are now enzymes that can perform random DNA fragmentation such as those provided by the NEBNext Ultra II kit from NEB or the Nextera DNA Library Preparation Kit from illumina. One advantage built into standard RAD if the RAD library is sequenced with PE sequencing (i.e. a forward read from the RAD tag and a reverse read from the opposite end created by shearing) is the easy detection of PCR duplicates. Due to random shearing, the SE and PE reads of two pieces of DNA after PCR should only be identical if at least one of them is a clone from PCR. When doing PE sequencing on a ddRAD library, both forward and reverse reads are fixed at restriction sites and PCR duplicates are not detectable unless at least a few degenerate bases are added to the sequencing adapter and sequenced (see Casbon et al., 2011; Schweyen et al., 2014; Tin et al., 2015). Another great feature of standard RAD is the possibility to assemble an a few hundred base pair long contig from the partially overlapping PE reads of each RAD tag (see fig. 1.2). This is very useful for paralog detection and functional annotation (Chong et al., 2012; Etter et al., 2011; Hohenlohe et al., 2013).

Even when standard RAD could provide enough complexity reduction, ddRAD may be still preferable. One reason is the slightly easier library preparation protocol. It has fewer steps, fewer clean-ups with the associated loss of DNA and it doesn't require a sonicator (Puritz et al., 2014b). With PE sequencing it also provides two RAD tags from each restriction fragment that is being sequenced (see fig. 1.3), a SE RAD tag and a PE RAD tag which can both be used for SNP and indel genotyping. One major disadvantage that comes with ddRAD is the fact that size selection of restriction fragments selects markers. So reproducibility of markers can be an issue. A good overlap in recovered markers has however been reported with precise size selection tools such as a Pippin Prep (Sage Science) (DaCosta and Sorenson, 2014; Peterson et al., 2012). Another downside is the greater susceptibility to allele-drop-out due to polymorphisms in the restriction sites (Arnold et al., 2013).

What can it be used for?

RADseq (either standard or ddRAD) has been used for virtually any kind of evolutionary genetic analysis including genome wide association study (GWAS) (Nadeau et al., 2014; Parchman et al., 2012), genomic clines analysis (Gompert et al., 2012a; Nosil et al., 2012), genetic mapping (Andolfatto et al., 2011; Baxter et al., 2011; Chutimanitsakun et al., 2011; Pfender et al., 2011; Richards et al., 2013), phylogenomics (Merz et al., 2013; Nadeau et al., 2014; Wagner et al., 2012), genome scans for selection (Andersen et al., 2012; Gompert et al., 2012a; Hohenlohe et al., 2010; Pujolar et al., 2014; Stölting et al., 2013) and inference of population demographic history (Evans et al., 2014; Lozier, 2014; Luca et al., 2011).

Assuming the majority of a genome evolves neutrally or nearly neutrally, RAD markers provide a lot of neutral sequence variation that can be used to infer fine-scale population structure, phylogeography as well as parameters for models of demographic history, like effective population size and gene flow. This information is useful for the planning and interpretation of genome scans for selection, GWASs and genomic clines analyses. However, some RAD markers will fall into functional regions of the genome and, therefore, cannot be assumed to be neutral genetic markers.

Problems with RAD

RAD markers, however, pose several challenges to unbiased estimation of population genetic parameters. One challenge are null alleles from polymorphisms in the restriction recognition sequence (Arnold et al., 2013; Gautier et al., 2012; Luca et al., 2011). Allele-drop-out leads to a systematic underestimate of genetic polymorphism and it also tends to increase estimates of genetic differentiation between populations. The biasing effect is stronger the greater the genetic diversity of the species. Luca et al. (2011) have provided formulas for an approximate correction of the effect of allele-drop-out on these measures based on equilibrium assumptions. However, filtering loci by across individual coverage can already greatly reduce its effect. With a frequency of the restriction site mutation of 0.5, 25% of individuals would not have reads at the locus, assuming HWE. With a frequency of 0.75, 56% of sampled individuals would not have reads from the locus. So requiring a minimum proportion of individuals to have reads from the locus could mitigate the effect of allele-drop-out but not completely remove it, even if only loci with full coverage across individuals are included (Arnold et al., 2013). That is because the set of loci not affected by allele-drop-out are not a random subset of all loci, but tend to have lower genetic diversity (and therefore also a smaller chance of a mutation in the restriction site).

RADseq is also sensitive to the quality of input DNA (Graham et al., 2015). Shearing of genomic DNA results in smaller fragments that may be too short for efficient sonication in standard RADseq (Davey et al., 2012) or disconnect two restriction sites in ddRAD, preventing their PCR and bridge amplification required for sequencing. However, only highly degraded DNA (i.e. containing only low molecular weight fragments) leads to a marked reduction in RAD tags recovered per individual.

Most RAD-like protocols include a selective PCR step that is intended to increase the fraction of fragments with correctly ligated adapters in the DNA library before sequencing.

PCR duplicates in the sequence data set, however, should be removed before genotype calling. This has two reasons. First, misincorporation events that occur and are propagated by PCR amplification during the library preparation would appear like genuine alleles to all genotype callers due to their high read coverage. Second, PCR drift can produce more copies of one allele at a heterozygous locus in the sequencing library. Since most genotype callers assume binomial read sampling from a bi-allelic locus, non-random read sampling from a heterozygous locus will lead to increased false-homozygote genotype calls. Both biases also apply when doing population genetic analyses in a probabilistic framework from genotype likelihoods, as done in the chapter Investigation into the demographic history of the hybrid zone.

In general, RAD cannot be expected to provide binomially distributed read depths of alleles at bi-allelic diploid loci nor can it be expected to lead to Poisson distributed read depths across loci. Observed read depth distributions are always overdispersed with respect to these theoretical distributions. For this reason detection of paralogous loci should not be based on Poisson quantile thresholds of locus read depth. There are several reasons for overdispersed read depths. Davey et al. (2012) have shown that in standard RAD libraries smaller restriction fragments have lower read coverage than longer ones. This is probably caused by a less efficient shearing through sonication of small restriction fragments (Davey et al., 2012). Additional variation in coverage between RAD tags is caused by a positive GC bias from PCR amplification (at least with the commonly used Phusion DNA polymerase). Variation in DNA methylation of restriction sites is a further reason for overdispersed read depths when using methylation sensitive restriction enzymes. There is also evidence for methylation at CpG sites in another acridian grasshopper related to *C. parallelus* (Keller et al., 2007). Due to random shearing, each locus in a standard RAD sequencing library is represented by DNA copies of different lengths. With all other RAD-like protocols, each locus is represented by DNA copies of only one length. PCR amplification as well as bridge amplification on the flow cell of the illumina sequencer are biased toward smaller DNA fragments. This leads to a greater read coverage of loci from smaller restriction fragments (Andrews et al., 2014).

RAD de novo assembly pipelines

Since RAD produces fully overlapping sequences from RAD tags, de novo assembly algorithms for RAD data have mostly been based on distances between read sequences rather than length of sequence overlaps, which is how shotgun-type reads are assembled into contigs longer than an individual read length. However, assembly programmes designed for shotgun-type reads have been successfully employed for assembling RAD tags (e. g.

Parchman et al., 2013). One of the most widely used programmes for RAD de novo assembly, stacks, has only recently introduced gapped alignment (via Needleman-Wunsch algorithm) into its de novo assembly procedure (with version 1.38 from 18 April 2016). This allows it to assemble RAD tags with indels. Eaton (2014) has shown convincingly with simulated and empirical data the advantage of using pyRAD over stacks (previous to version 1.38) when the sequences contain realistic numbers of indels. It recovers loci shared among a greater number of individuals. Like pyRAD, the de novo assembly method of dDocent (Puritz et al., 2014a) allows for indels during the clustering step and therefore its assembled loci are shared among a greater number of individuals than those assembled by the indel-unaware clustering algorithm of stacks (previous to version 1.38).

A common problem with most RAD analysis pipelines published to date is their lack of modularisation and documentation. For instance, a good de novo assembly strategy can be combined with a suboptimal SNP detection or paralogy detection algorithm without clear separation of both functionalities in the code and appropriate documentation. This makes it difficult for users to recombine the best functionality from different programmes or pipelines and results in a lot of underused computer code. Two pipelines that exemplify this problem shall be mentioned here: PRGmatic from Hird et al. (2011) and rtd from Peterson et al. (2012). Writing feature-rich, complex, monolithic and sparsely documented software is obviously a common issue in software development and has been addressed by providing basic rules in the Unix philosophy. A noteworthy positive exception to this general lack of modularisation and thorough documentation is dDocent (Puritz et al., 2014a).

The de novo assembly of RAD loci without a reference sequence poses a particular challenge to the application of RAD to non-model organisms. With short read lengths and repetitive genome sequences, the distinction of homologous from paralogous read sequences can become impossible for RAD tags in repetitive parts of the genome. Importantly, mis-assembled clusters of sequence reads from repetitive RAD tags need to be detected and removed from downstream population genetic analyses because they violate the assumption of orthology underlying most of these analyses. Otherwise, diversity estimates will be inflated, for instance.

There have been several attempts to optimise RAD de novo assembly (e. g. Ilut et al., 2014; Mastretta-Yanes et al., 2015). Ilut et al. (2014) describe an empirical method to determine the optimal within cluster distance for RAD tag de novo assembly. Using digital digests of genome sequences with one restriction enzyme (but without random shearing, i.e. equivalent to double-digest), they create a clustering threshold series. For each clustering threshold they determine the number of clusters that contain one haplotype, two haplotypes or ≥3 haplotypes. They then propose that the optimal cluster distance threshold is where the

1–haplotype and 2–haplotype clusters start to plateau in frequency and the ≥3–haplotype clusters are still low in proportion. Higher thresholds would not increase the detection of heterozygous loci much while increasing the proportion of paralogous sequences within the 2–haplotype cluster category. However, I tested their method on the standard RAD data set used in Investigation into the demographic history of the hybrid zone and could not find a plateau in the cluster frequency curves and therefore also no optimal cluster distance threshold (fig. 1.4).

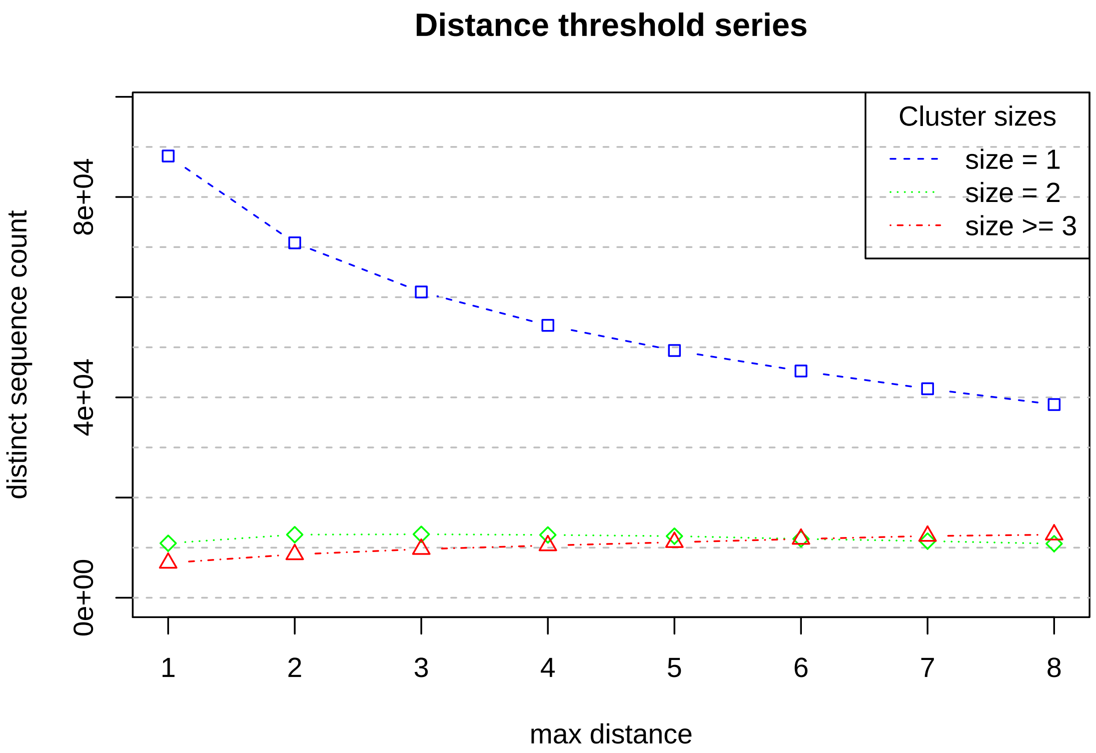


Fig. 1.4 Distance threshold series from the Ilut pipeline. The number of distinct sequences in clusters of size 1, 2 and ≥3 are plotted for each maximum edit distance threshold for transitive clustering.

Many paralogy filters have been proposed. One commonly used filter is based on an excess of heterozygotes with respect to Hardy Weinberg equilibrium (HWE) (HOHENLOHE et al., 2011). Unless a locus is affected by strong overdominant selection (which is rare), an excess of heterozygous genotypes in a population sample is most likely an artefact of clustering similar sequence reads from repetitive parts in the genome.

Alternatives to RAD

RAD is an untargeted reduced representation sequencing method (with the exception of GC and methylation bias dependent on restriction enzyme and genome studied) which has been criticised recently by Lowry et al. (2016) and Tiffin and Ross-Ibarra (2014) when used in genome scans for regions under selection. Their main point is that many such genome scans had used RAD data with too few markers to have at least one polymorphic marker in strong enough LD with a locus under selection and therefore have probably missed many such loci, particularly those with incomplete selective sweeps or starting from standing genetic variation. The extent of LD (i.e. average size of linkage blocks) varies greatly among species and seems to be fairly unpredictable (Lowry et al., 2016, see their table 1). So the scale of this problem very much depends on the average LD and the size of the genome of the species studied. This lack of complete marker coverage in regions under selection has led Lowry et al. (2016) to suggest exome capture sequencing for genomic scans for selection, i.e. enriching for functionally relevant sequences at the risk of missing those distant from genes (Good, 2011; Jones and Good, 2016; Mamanova et al., 2010; Yi et al., 2010).

Targeted capture sequencing is only slowly being adopted for evolutionary and ecological studies of non-model organisms (Bi et al., 2012; Nadeau et al., 2012). There are probably two main reasons for that. First, probe design requires prior sequence information. So, unless a genome reference sequence is available from at least a fairly closely related species (Good, 2011) a transcriptome needs to be assembled first (Bi et al., 2012). Second, probe synthesis has greatly added to the costs of sequencing. Both issues have been addressed successfully by Puritz and Lotterhos (2017). They show that probes can be generated directly from cDNA. Given that already Maricic et al. (2010) have reported custom in-house generation of capture probes, it seems surprising that this approach has not already been adopted more widely. Also note that not only protein coding sequences would be captured when using cDNA for probe generation but also many long non-coding RNA's most of which have a poly-A tail (Ulitsky, 2016).

There is now good experimental support for the argument that the inference of neutral processes, like demographic history and gene flow, from exome capture sequence data should suffer from severe biases due to pervasive effects of linked selection even when inference is restricted to putatively neutral four-fold degenerate sites (e.g. Andolfatto, 2007; Elyashiv et al., 2016; McVicker et al., 2009; Sella et al., 2009). Since RADseq samples sequences more or less randomly from across the genome, it should suffer much less from this bias but can also not be assumed to provide fully neutrally evolving sequences, especially in species with large effective population sizes and low recombination rates (Corbett-Detig et al., 2015). However, current data still suggests that fluctuations in population sizes are the main contributor to variation among species of genetic diversity at four-fold degenerate sites and that therefore useful information about neutral demographic processes can be retrieved from such data (Coop, 2016; Hsieh et al., 2017). In addition, 4-fold degenerate sites have been repeatedly shown to be the most polymorphic among all sites examined in whole-genome resequencing studies of diverse organisms (see Begun et al., 2007, table 1 and Small et al. 2007) Furthermore, powerful tests of selection based on the ratio of non-synonymous to synonymous substitutions can be performed with protein coding sequences from exome capture but not from RAD data.

1.3 What to expect in the following chapters

In this study, I address the following questions:

- Can a genome-wide set of sequence markers be generated from the genome of *C. parallelus*?
- Can such a data set be assembled and genetic variation be analysed that allows inferences to be drawn about forces that affect the majority of markers, i.e. can genome-wide signals be detected?

The chapter 2 on page 21 documents an investigation into the possible reasons for a lack of across-individual sequence coverage at almost all de novo assembled RAD tags. In the second chapter, I then apply a new de novo assembly strategy (adapted from `dDocent`) to one standard RAD data set (Baird et al., 2008) from two population samples of *C. p. erythropus* and *C. p. parallelus*. While avoiding SNP and genotype calling from this low coverage data set, I derive genome-wide population genetic summary statistics and use two different programmes to fit demographic models for the two subspecies using their one-dimensional as well as joint site frequency spectra.

Chapter 2

Testing incomplete digestion

> Bonzai: Are you as successful as you would like to be?
> Zappa: I would say that the basic characteristic of my life is failure. If there is one thing that I excel at, it's failure – I manage to fail at 100 percent of the things that I do. Since most of the things that I set out to do are theoretically impossible, it's very easy to fail. I've learned to live with it. In terms of machinery and personnel, there never seems to be enough to get things done exactly right.
>
> interview with Frank Zappa, 1985

2.1 Introduction

In this chapter, I use two different types of RAD data sets[1] to investigate three issues that can occur during the preparation of RAD libraries and that can lead to unusually low overall sequence read coverage and an extreme variation in coverage among loci and individuals: incomplete digestion, genomic re-ligation and low genomic template amount. I show bioinformatic analyses that can detect or at least distinguish among these issues post-sequencing and I suggest suitable measures for the library preparation that can mitigate their impact.

[1] that were initially created with the intention of doing a hybrid zone analysis

2.1.1 The Problem

A standard RAD library with the restriction enzyme SbfI was prepared according to the protocol of Paul Etter, University of Oregon (Baird et al., 2008) (see Appendix section 5.1 on page 152). The RAD library contained DNA from 36 grasshoppers sampled from the two distal populations ("Aunat" and "Greixer") of a transect through the *Chorthippus parallelus / erythropus* hybrid zone in the Pyrenees between France and Spain (fig. 2.1). The RAD library was sequenced on an illumina GAIIx at GenePool in Edinburgh and the resulting 46 base pair (bp) long reads[2] assembled with the programme suite `stacks` (Catchen et al., 2011).

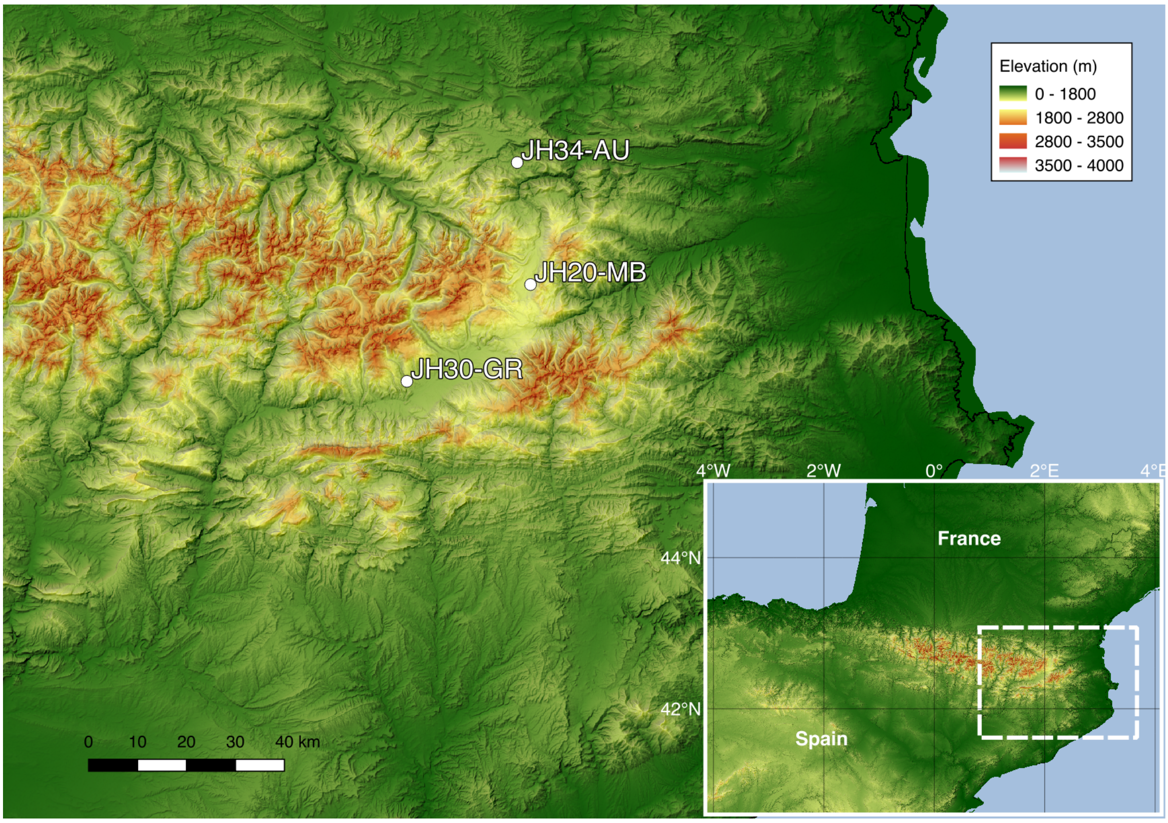
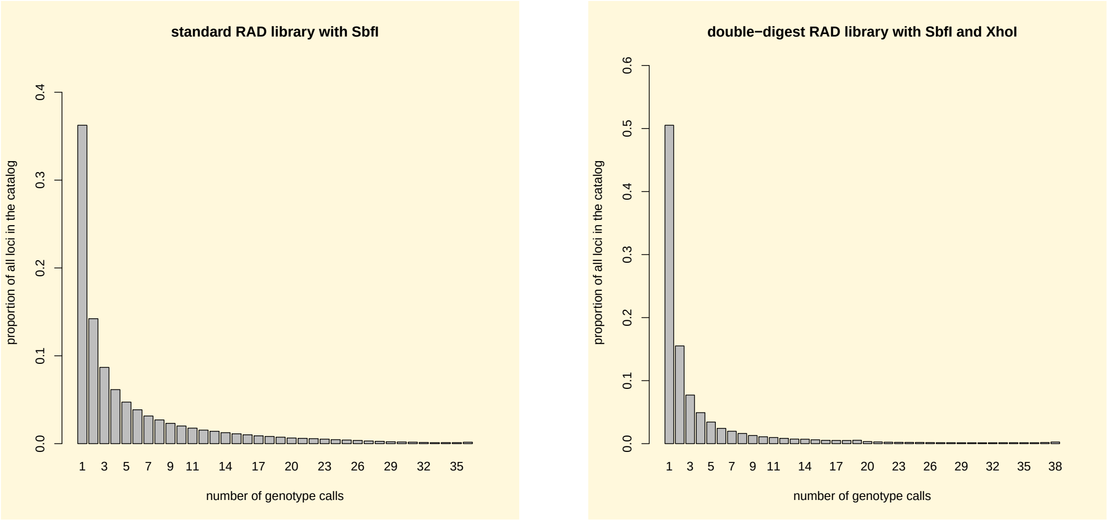


Fig. 2.1 Map of sampling locations. JH34-AU: *parallelus*; JH30-GR: *erythropus*. JH20-MB: marks the centre of the cline of sterility as determined by Shuker et al. (2005). Details about the creation of this map are provided in the section 2.3.1 on page 42.

Figure 2.2a shows the frequency distribution of loci – reconstructed by `stacks` version 0.998 – over the number of individuals that have a genotype called for that locus. `Stacks` was run with a minimum allele read depth of 3 per individual and a maximum number of mismatches between alleles of 2 for merging alleles into loci within individuals as well as

[2] after removing the barcode sequence



(a) (b)

Fig. 2.2 Frequency distribution of loci, reconstructed by stacks (i.e. so-called "catalog stacks"), over the number of individuals in which they have a genotype call.

assembling a catalog of loci for the whole sample of 36 grasshoppers (further details can be looked up on the stacks home page). About 50% of the 379,720 unfiltered reconstructed loci have a genotype call in only one or two of the 36 individuals. About 170,000 RAD markers were expected from this library (see section 2.3.5) assuming a genome size of 12 giga bps (see section 2.3.4).

Figure 2.2b shows the frequency distribution of reconstructed loci over the number of individuals for which they have a genotype called for the data of an SbfI+XhoI double-digest RAD library. This ddRAD library was prepared from the same individuals but according to the protocol in Appendix section section 5.2 on page 152. The SE and PE RAD tags have been merged into a 196 bp long tag[3] before the assembly with stacks. Stacks was run with a minimum allele depth of 3 and maximum mismatch distance of 6 for merging alleles into read stacks within individuals as well as assembling a catalog of stacks (i.e. putative loci) from the individual stacks of all 36 grasshoppers (plus 2 technical replicates). Of the 156,532 putative loci that stacks had assembled, 51% can only be found in one individual and only 3.5% can be found in 20 or more individuals. Note that these numbers are from the raw output of stacks and do not include credibility filtering of putative loci and genotype calls. Around 16,000 RAD markers were expected from this double-digest RAD library (see equation 2.3 in section 2.3.5).

[3] including 11 bp remainders of restriction sites

In the standard SbfI RAD data set, there are 5,014 SE reads and 25,268 PE reads that apparently contain SbfI recognition sites within them (30,282 in total)[4]. I searched in 88,734,712 quality filtered reads altogether. That is, 0.034% of quality filtered reads contain an SbfI recognition site. Figure 2.3 shows the frequency distributions of SbfI sites in SE and PE reads for all 36 individuals separately (for further description see section 2.3.11). Obviously, if SbfI restriction and following P1 adapter ligation were 100% efficient, there should be no SbfI recognition sequences in either SE or PE reads. Could this pattern be an indication that SbfI restriction during the library preparation was incomplete? If so, could there be a systematic variation between individuals in the completeness of restriction at individual SbfI sites that could lead to many loci only being detected in a few individuals as shown in figure 2.2? In the following I will investigate the potential role of incomplete restriction enzyme digestion during the sequencing library preparation on the distribution

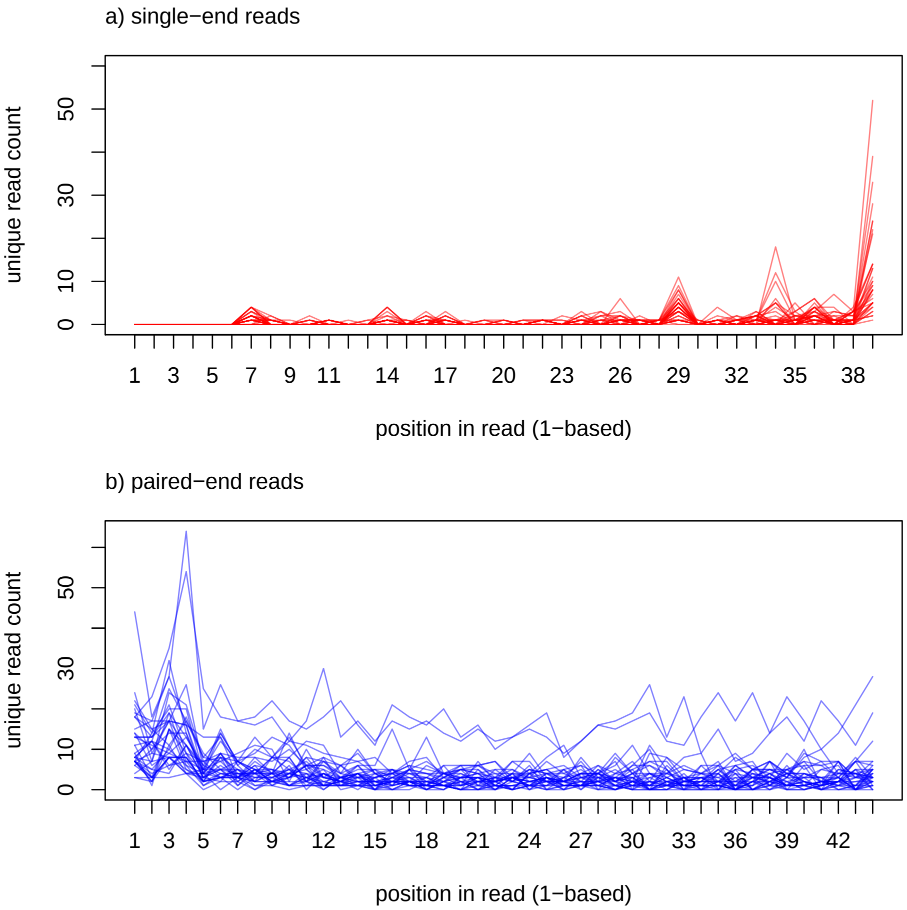


Fig. 2.3 SbfI site frequency distributions across (a) SE and (b) PE reads for each of the 36 individuals in the standard RAD data set.

[4] counted with `grep -c`

of read coverage over RAD tags and on the resulting detection probability of RAD tags as shown in figure 2.2.

2.2 Results and Discussion

2.2.1 Assembling pairs of PE read contigs

Incomplete digestion for a specific SbfI site could in principle be tested with a PCR (or even quantified with *q*PCR) across this site before and after digestion (Luca et al., 2011). However, no close reference sequence was available for PCR primer design at the time of this analysis[5]. For a given RAD tag from a standard RAD library, the PE reads can be used to assemble them into a contig that could be long enough for primer design (see figure 1.2 on page 10). Still, for PCR primer design, *pairs* of PE contigs need to be created. This requires some reference sequence to provide the information for which RAD tags come from the same SbfI site in the *Chorthippus parallelus* genome.

I decided to use the transcriptome of the desert locust *Schistocerca gregaria* (*Cyrtacanthacridinae*) as a reference sequence (Badisco et al., 2011) . The transcriptome of another grasshopper – *Locusta migratoria* (*Oedipodinae*) – was also available (Kang et al., 2004), but Liu et al. (2008) have shown that the subfamily *Cyrtacanthacridinae* is more closely related to *Gomphocerinae* – the subfamily that *Chorthippus parallelus* belongs to – than the subfamily *Oedipodinae*.

I have mapped all standard SbfI RAD reads[6] of all 36 individuals against the *Schistocerca* transcriptome with `stampy` (Lunter and Goodson, 2011) (see 2.3.6 on page 46). From the mapping output files I first extracted all read pairs where at least one read from a pair mapped and then merged them into one big file. I then ran my custom script `find_linked_RADtags.pl` on this file. This script collected from all individuals all PE reads that belong to the same RAD tag of a linked RAD tag site that was detected in as little as one individual. For each detected linked RAD tag site this script collected the PE reads upstream and downstream of the SbfI restriction site in separate files (for further description see section 2.3.6). It thus collected PE reads from 77 *Schistocerca* reference contigs with linked RAD tag sites .

I then used the programme `SSAKE` (Warren et al., 2007) together with my wrapper script `SSAKEoptimiser.pl` for the de novo assembly of PE reads into contigs (see section 2.3.6). `SSAKEoptimiser.pl` finds the optimal kmer length for each individual assembly, optimising

[5] April 2014

[6] informal name "Big Data"

for the length of the longest contig assembled. There are 64 *Schistocerca* reference contigs with a RAD tag site for which at least one upstream and one downstream SSAKE contig could be assembled.

The SSAKEoptimiser output for each assembly of PE reads generally contains several contigs of similar length and with similar read counts. It is therefore not straightforward to pick those SSAKE contigs upstream and downstream of the restriction site that genuinely belong together, i.e. come from the same locus. Using a heuristic that uses contig number, contig length and blast hits against the putative *Schistocerca* reference contig, I could assemble and confidently pick 20 *pairs* of PE read contigs for PCR primer design (further details on page 49).

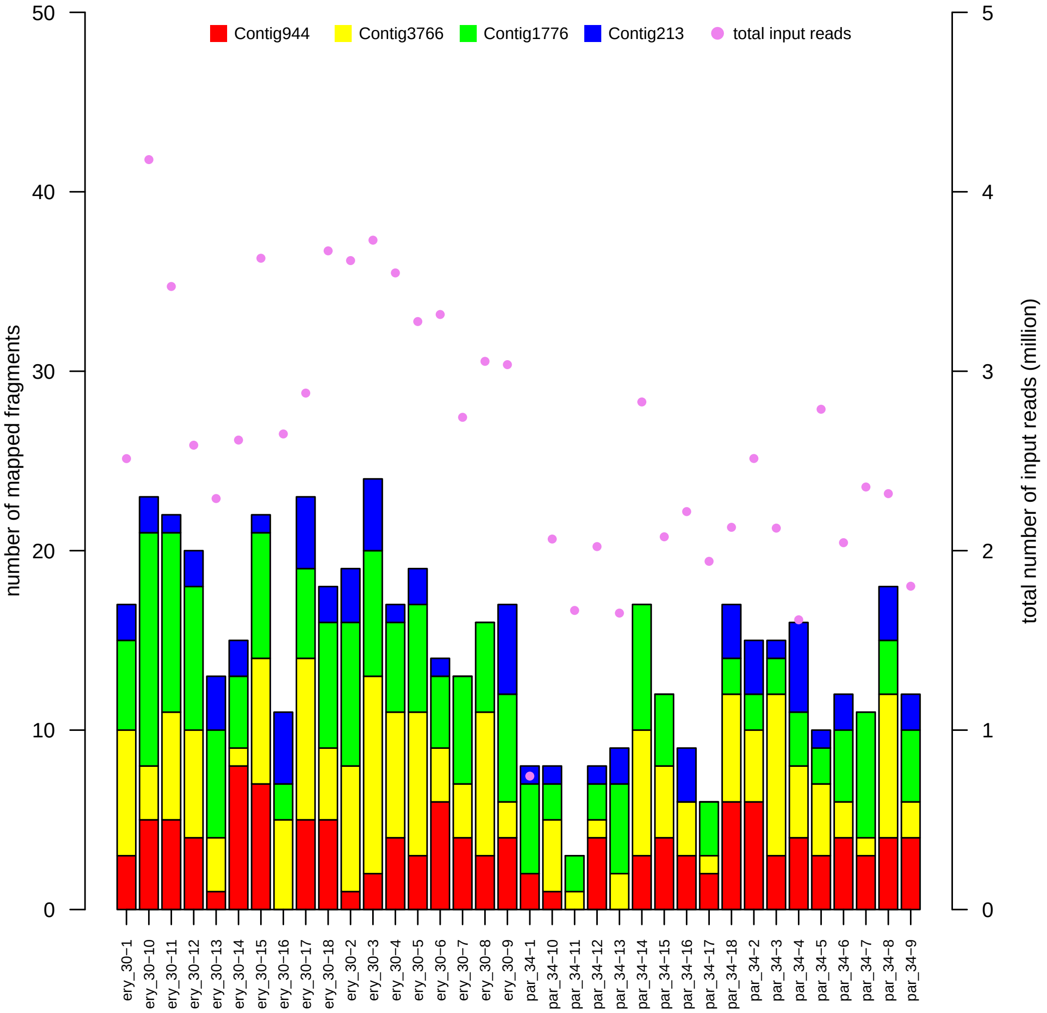


Fig. 2.4 Distribution of RAD fragment numbers from the 36 individuals in the library mapped against 4 *PE contig pair* reference sequences.

I combined pairs of PE contigs by aligning them to their putative *Schistocerca* reference contig and filled the gap between them with N's. I then used these 20 newly created *C. parallelus* consensus sequences as a reference against which to map all standard SbfI RAD reads (further details in section 2.3.7). That is because before setting out to do PCR I wanted to find out how many of the 36 individuals get reads mapped to these linked RAD tag sites and whether individuals actually have reads mapped to both RAD tags at an SbfI restriction site. This could be important information for prioritising some of the 20 loci over others for the analysis of between individual variation in the completeness of digestion with PCR.

Among the 20 *C. parallelus* PE contig pair reference sequences, there are 5 which:

- do not show very high or very low number of reads mapped
- do not show large numbers of very divergent reads mostly containing indels
- do not show other signs of repetitiveness, e. g. SE reads mapping to PE contigs instead of the RAD tags

One of these 5 reference sequences has only reads mapped from *C. p. parallelus* (12 individuals) and none from *C. p. erythropus*. This could be caused by a polymorphism in the restriction site sequence. For each individual, I counted the number of fragments mapped towards the remaining four reference sequences by counting only SE reads from proper pairs and after removing PCR duplicates by collapsing multiple occurrences of the same insert size into one. If two read pairs on a RAD site from an individual have the same insert size, they constitute only one fragment, i.e. one read pair is likely to be a PCR duplicate.

Figure 2.4 shows the distribution of these counts over all 36 individuals for the four PE contig pair reference sequences. It suggests that none of the 4 loci would be a good candidate to test possible variation in restriction enzyme digestion with PCR. That is because the distribution of coverage at the 4 loci is rather even across the 36 individuals. Even though between individual variation in the completeness of digestion cannot be ruled out by this data yet, a different pattern would be expected if it was common. That is, more individuals would have no fragments mapping while others would have many. Given this data, a systematic variation in the completeness of restriction enzyme digestion between individuals is now less likely to be the reason for the dominance of singleton loci in the `stacks` assembly (see figure 2.2). Instead, the variation in coverage among individuals in figure 2.4 can be largely explained by variation in the number of input reads (figure 2.5).

The fragment count for the four loci is generally not very high (see table 2.1), indicating that low unique template amount for sequencing prevented any individual from having many fragments mapped. I started the standard SbfI RAD library preparation with about 130 ng of

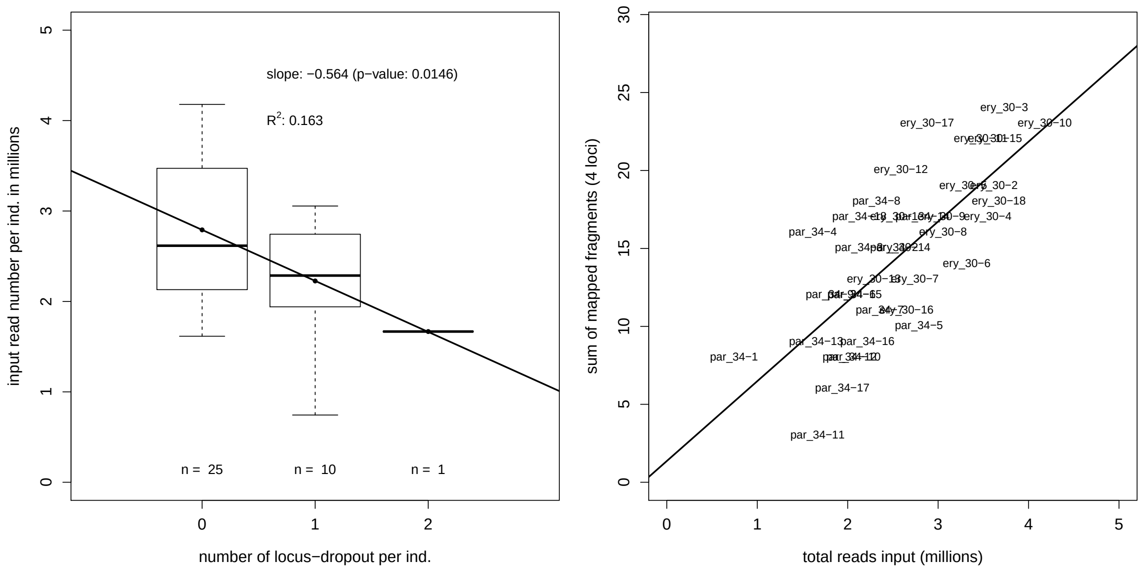


Fig. 2.5 Correlation of locus dropout (left) and fragment counts (right) with number of input reads. Left: Locus dropout is the number of loci (which are the same as in figure 2.4) for which an individual had no fragment mapped. Right: the sum of mapped fragments over the four loci for each individual versus the number of input reads for the read mapping.

DNA from each grasshopper (see sRAD protocol on page 152). Assuming that the genome is 12 Gbp long (see section 2.3.4), this would only correspond to around 10,000 copies of the genome (equation 2.1).

Table 2.1 Mean and coefficient of variation of fragment counts per individual for the 4 loci shown in figure 2.4.

	mean	CV
Contig944	3.5	0.5
Contig3766	4.5	0.6
Contig1776	4.8	0.5
Contig213	1.9	0.8

For the SbfI+XhoI double-digest RAD library I estimated the template amount that went into the selective PCR during library preparation with *q*PCR. This also indicated a very low template amount of on average 1.26 (± 0.37) template molecules per locus and individual (see section 2.3.8).

$$
\begin{aligned}
\text{molar amount of template} &= \frac{\text{amount of DNA}}{\text{MW of bp} \times \text{genome size}} = \frac{130 \times 10^{-9}\text{g}}{660\frac{\text{g}}{\text{mol}\times\text{bp}} \times 12 \times 10^{9}\text{bp}} \qquad (2.1)\\
&= 1.26 \times 10^{-20}\text{mol}\\
\text{number of template molecules} &= 1.26 \times 10^{-20}\text{mol} \times \text{Avogadro's number}\\
&= 1.26 \times 10^{-20}\text{mol} \times 6.0221413 \times 10^{23}\\
&= 9884
\end{aligned}
$$

The problem of low fragment count is unlikely to be alleviated much by creating more sequence reads from the same RAD library (figure 2.6). Instead, the template amount from each individual for the selective PCR during the library preparation needs to be increased. One obvious way would be to increase the total DNA input from each individual, but that reduces the number of individuals that can be pooled during the library preparation since the capacity of spin columns and the agarose gel (for size selection) is then reached with fewer individuals. Another option could be to postpone the gel fragment size selection until after a selective PCR, thus reducing the loss of template amount before the PCR (see Parchman et al. 2012 and the table on page 54).

If the dominance of singleton loci in the `stacks` assembly was not caused by a systematic variation in the completeness of digestion between individuals but simply by random dropout due to low template amount, could this in turn again be caused by incomplete restriction enzyme digestion of genomic DNA as indicated by the full SbfI recognition sequences found in the RAD sequence data (figure 2.3)? Or can genomic religation of restriction fragments[7] account for the observed restriction enzyme recognition sequences in the RAD data? The rest of this chapter will be investigating this question.

[7] instead of ligation to Illumina adapters

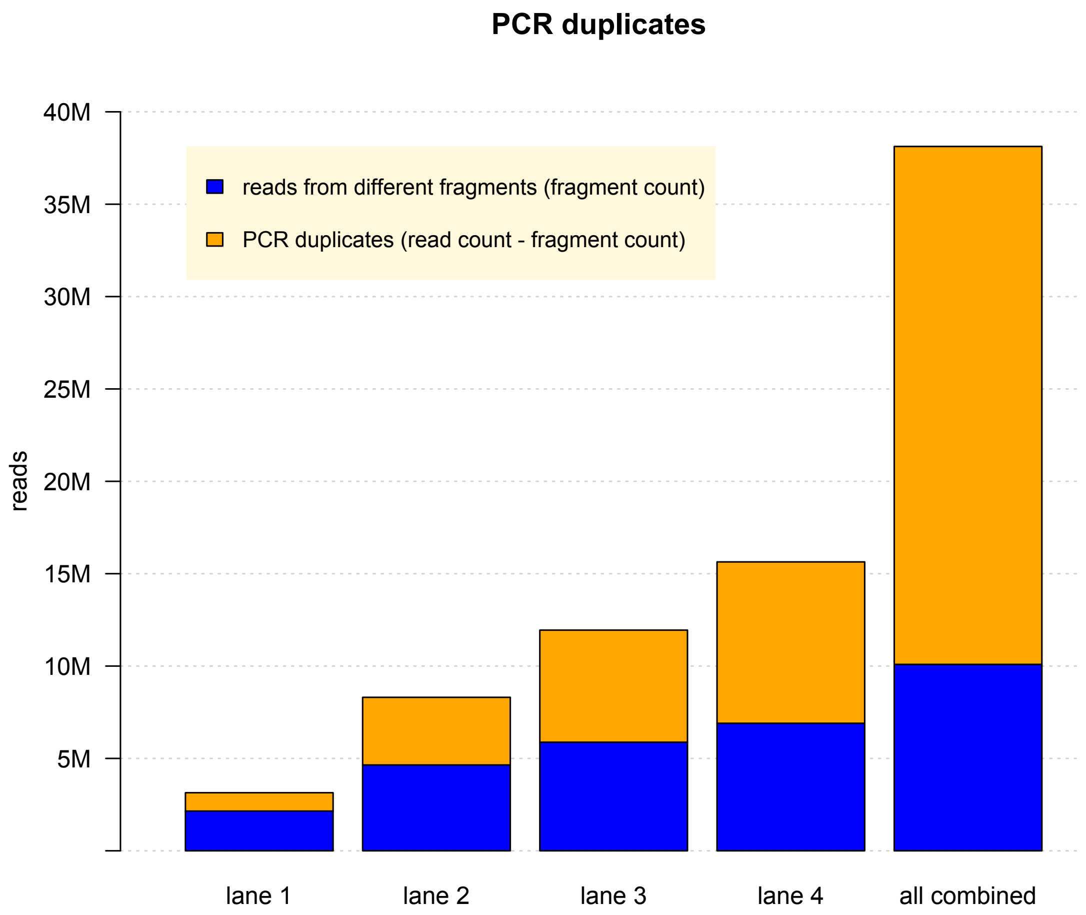


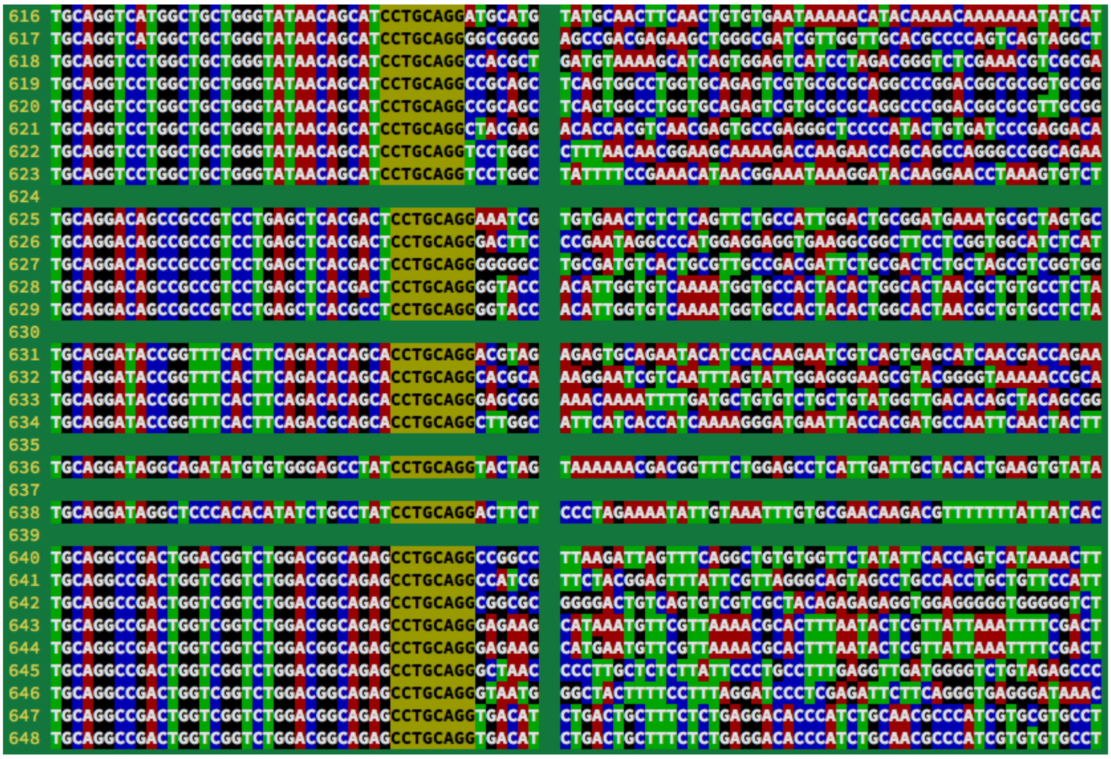
Fig. 2.6 The standard SbfI RAD library was sequenced on four different Illumina GAIIx flow cell lanes with increased sequence template amount resulting in an increased read yield. However, this increased sequencing effort for the same RAD library generated an ever higher proportion of PCR duplicates.

2.2.2 Incomplete digestion or genomic religation in the standard RAD library?

Cluster analysis

If non-homologous SbfI fragments were ligated during the adapter ligation step in the library preparation, then the subsequences right (i.e. downstream) of the SbfI site in the SE reads should be very divergent within clusters of similar reads, except when the clusters contain PCR duplicates which can be recognised by (almost) identical PE reads[8]. I have therefore collected all read pairs containing an SbfI site from each individual and, after collapsing all identical read pairs into one, clustered these reads by the subsequence left (5') of the SbfI

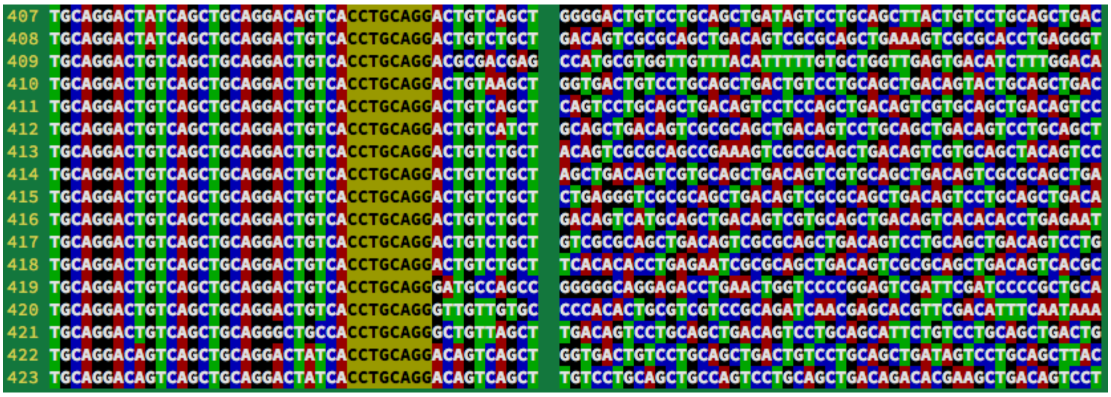
[8] they should only differ by sequencing errors



```
616 TGCAGGTCATGGCTGCTGGGTATAACAGCATCCTGCAGGATGCATG  TATGCAACTTCAACTGTGTGAATAAAAACATACAAAACAAAAAAATATCAT
617 TGCAGGTCATGGCTGCTGGGTATAACAGCATCCTGCAGGGGCGGGG  AGCCGACGAGAAGCTGGGCGATCGTTGGTTGCACGCCCCAGTCAGTAGGCT
618 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGCCACGCT  GATGTAAAAGCATCAGTGGAGTCATCCTAGACGGGTCTCGAAACGTCGCGA
619 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGCCGCAGC  TCAGTGGCCTGGTGCAGAGTCGTGCGCGCAGGCCCGGACGGCGCGGTGCGG
620 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGCCGCAGC  TCAGTGGCCTGGTGCAGAGTCGTGCGCGCAGGCCCGGACGGCGCGTTGCGG
621 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGCTACGAG  ACACCACGTCAACGAGTGCCGAGGGCTCCCCATACTGTGATCCCGAGGACA
622 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGTCCTGGC  CTTTAACAACGGAAGCAAAAGACCAAGAACCAGCAGCCAGGGCCGGCAGAA
623 TGCAGGTCCTGGCTGCTGGGTATAACAGCATCCTGCAGGTCCTGGC  TATTTTCCGAAACATAACGGAAATAAAGGATACAAGGAACCTAAAGTGTCT
624
625 TGCAGGACAGCCGCCGTCCTGAGCTCACGACTCCTGCAGGAAATCG  TGTGAACTCTCTCAGTTCTGCCATTGGACTGCGGATGAAATGCGCTAGTGC
626 TGCAGGACAGCCGCCGTCCTGAGCTCACGACTCCTGCAGGGACTTC  CCGAATAGGCCCATGGAGGAGGTGAAGGCGGCTTCCTCGGTGGCATCTCAT
627 TGCAGGACAGCCGCCGTCCTGAGCTCACGACTCCTGCAGGGGGGGC  TGCGATGTCACTGCGTTGCCGACGATTCTGCGACTCTGCTAGCGTCGGTGG
628 TGCAGGACAGCCGCCGTCCTGAGCTCACGACTCCTGCAGGGGTACC  ACATTGGTGTCAAAATGGTGCCACTACACTGGCACTAACGCTGTGCCTCTA
629 TGCAGGACAGCCGCCGTCCTGAGCTCACGCCTCCTGCAGGGGTACC  ACATTGGTGTCAAAATGGTGCCACTACACTGGCACTAACGCTGTGCCTCTA
630
631 TGCAGGATACCGGTTTCACTTCAGACACAGCACCTGCAGGACGTAG  AGAGTGCAGAATACATCCACAAGAATCGTCAGTGAGCATCAACGACCAGAA
632 TGCAGGATACCGGTTTCACTTCAGACACAGCACCTGCAGGCACGCA  AAGGAATCGTCAATTTAGTATTGGAGGGAAGCGTACGGGGTAAAAACCGCA
633 TGCAGGATACCGGTTTCACTTCAGACACAGCACCTGCAGGGAGCGG  AAACAAAATTTTGATGCTGTGTCTGCTGTATGGTTGACACAGCTACAGCGG
634 TGCAGGATACCGGTTTCACTTCAGACGCAGCACCTGCAGGCTTGGC  ATTCATCACCATCAAAAGGGATGAATTACCACGATGCCAATTCAACTACTT
635
636 TGCAGGATAGGCAGATATGTGTGGGAGCCTATCCTGCAGGTACTAG  TAAAAAACGACGGTTTCTGGAGCCTCATTGATTGCTACACTGAAGTGTATA
637
638 TGCAGGATAGGCTCCCACACATATCTGCCTATCCTGCAGGACTTCT  CCCTAGAAAATATTGTAAATTTGTGCGAACAAGACGTTTTTTTATTATCAC
639
640 TGCAGGCCGACTGGACGGTCTGGACGGCAGAGCCTGCAGGCCGGCC  TTAAGATTAGTTTCAGGCTGTGTGGTTCTATATTCACCAGTCATAAAACTT
641 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGCCATCG  TTCTACGGAGTTTATTCGTTAGGGCAGTAGCCTGCCACCTGCTGTTCCATT
642 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGCGGCGC  GGGGACTGTCAGTGTCGTCGCTACAGAGAGAGGTGGAGGGGGTGGGGGTCT
643 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGGAGAAG  CATAAATGTTCGTTAAAACGCACTTTAATACTCGTTATTAAATTTTCGACT
644 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGGAGAAG  CATGAATGTTCGTTAAAACGCACTTTAATACTCGTTATTAAATTTTCGACT
645 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGGCTAAC  CCCTTGCTCTCTTATTCCCTGCCTTTGAGGTTGATGGGGTCTGTAGAGCCC
646 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGGTAATG  GGCTACTTTTCCTTTAGGATCCCTCGAGATTCTTCAGGGTGAGGGATAAAC
647 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGTGACAT  CTGACTGCTTTCTCTGAGGACACCCATCTGCAACGCCCATCGTGCGTGCCT
648 TGCAGGCCGACTGGTCGGTCTGGACGGCAGAGCCTGCAGGTGACAT  CTGACTGCTTTCTCTGAGGACACCCATCTGCAACGCCCATCGTGTGTGCCT
```

Fig. 2.7 Snapshot of clusters of unique read pairs from the output of command 2.14. Each line shows a read pair. Left column: SE reads, right column: PE reads. The SbfI recognition sequence is highlighted in yellow. The SE reads in the top cluster are quite diverse right of the SbfI site, although in the lines 619 and 620 as well as 622 and 623 they are identical. In the first case this is clearly due to PCR duplication indicated by the almost identical PE reads (they only differ by sequencing errors). In the second case, however, PCR duplication can be ruled out and only incomplete digestion of the SbfI site at this putative locus seems plausible.

site, thus ignoring the potentially non-homologous subsequence right of the SbfI site (for details see section 2.3.11). Figure 2.7 shows a snapshot from the output. Even though there clearly is some indication of incomplete digestion (see figure 2.8), most clusters are largely consistent with genomic religation.



```
407 TGCAGGACTATCAGCTGCAGGACAGTCACCTGCAGGACTGTCAGCT  GGGGACTGTCCTGCAGCTGATAGTCCTGCAGCTTACTGTCCTGCAGCTGAC
408 TGCAGGACTATCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  GACAGTCGCGCAGCTGACAGTCGCGCAGCTGAAAGTCGCGCACCTGAGGGT
409 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACGCGACGAG  CCATGCGTGGTTGTTTACATTTTTGTGCTGGTTGAGTGACATCTTTGGACA
410 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTAAGCT  GGTGACTGTCCTGCAGCTGACTGTCCTGCAGCTGACAGTACTGCAGCTGAC
411 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCAGCT  CAGTCCTGCAGCTGACAGTCCTCCAGCTGACAGTCGTGCAGCTGACAGTCC
412 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCATCT  GCAGCTGACAGTCGCGCAGCTGACAGTCCTGCAGCTGACAGTCCTGCAGCT
413 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  ACAGTCGCGCAGCCGAAAGTCGCGCAGCTGACAGTCGTGCAGCTACAGTCC
414 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  AGCTGACAGTCGTGCAGCTGACAGTCGTGCAGCTGACAGTCGCGCAGCTGA
415 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  CTGAGGGTCGCGCAGCTGACAGTCGCGCAGCTGACAGTCCTGCAGCTGACA
416 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  GACAGTCATGCAGCTGACAGTCGTGCAGCTGACAGTCACACACCTGAGAAT
417 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  GTCGCGCAGCTGACAGTCGCGCAGCTGACAGTCCTGCAGCTGACAGTCCTG
418 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGACTGTCTGCT  TCACACACCTGAGAATCGCGCAGCTGACAGTCGCGCAGCTGACAGTCACGC
419 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGGATGCCAGCC  GGGGGCAGGAGACCTGAACTGGTCCCCGGAGTCGATTCGATCCCCGCTGCA
420 TGCAGGACTGTCAGCTGCAGGACTGTCACCTGCAGGGTTGTTGTGC  CCCACACTGCGTCGTCCGCAGATCAACGAGCACGTTCGACATTTCAATAAA
421 TGCAGGACTGTCAGCTGCAGGGCTGCCACCTGCAGGGCTGTTAGCT  TGACAGTCCTGCAGCTGACAGTCCTGCAGCATTCTGTCCTGCAGCTGACTG
422 TGCAGGACAGTCAGCTGCAGGACTATCACCTGCAGGACAGTCAGCT  GGTGACTGTCCTGCAGCTGACTGTCCTGCAGCTGATAGTCCTGCAGCTTAC
423 TGCAGGACAGTCAGCTGCAGGACTATCACCTGCAGGACAGTCAGCT  TGTCCTGCAGCTGCCAGTCCTGCAGCTGACAGACACGAAGCTGACAGTCCT
```

Fig. 2.8 Snapshot of a cluster of unique read pairs from the output of command 2.14. Each line shows a read pair. Left column: SE reads, right column: PE reads. The SbfI recognition sequence is highlighted in yellow. The sequences right of the SbfI site are not very diverse. In fact, they are so similar that they could be derived from a repetitive genomic element. The read pairs in the lines 413–418, in particular, seem to suggest incomplete digestion of the SbfI site.

Read pair mapping analysis

Random genomic restriction fragment religation should create chimeras. If mapped to a reference sequence, read pairs from such chimeras should generally not map as proper pairs. I therefore mapped all SbfI containing read pairs against the *Locusta migratoria* genome (Wang et al., 2014) with bowtie2 (Langmead and Salzberg, 2012). Proper pairs should be a clear indication of incomplete digestion whereas discordantly mapped read pairs could be the result of genomic religation or lack of synteny between the *Chorthippus* and *Locusta* genome or simply the small size of the contigs in the *Locusta* genome assembly.

I therefore tried to estimate the proportion of concordantly mapping read pairs from a random sample of read pairs where, for the vast majority of read pairs, the two reads in a pair should come from the same genomic location as in the case of incomplete digestion. Any reduction in the proportion of concordantly mapping read pairs among the SbfI site containing read pairs with respect to this expectation should be caused by random genomic religation of SbfI restriction fragments.

As a quasi random sample, I have taken the 200,001st to 300,000th read pair from each individual. After mapping against the *Locusta* genome reference sequences, I first extracted all read pairs where both reads mapped, concordantly or not and disregarding mapping quality. The total number of these read pairs is 846,583. Among these read pairs there are 440,015 that map concordantly (further details from page 62 onwards). So 52% of mapping read pairs mapped concordantly. That means that a random read pair – where both reads generally come from the same genomic location in *C. parallelus* as with incomplete digestion – is about as likely to map discordantly as concordantly against the *Locusta* genome reference.

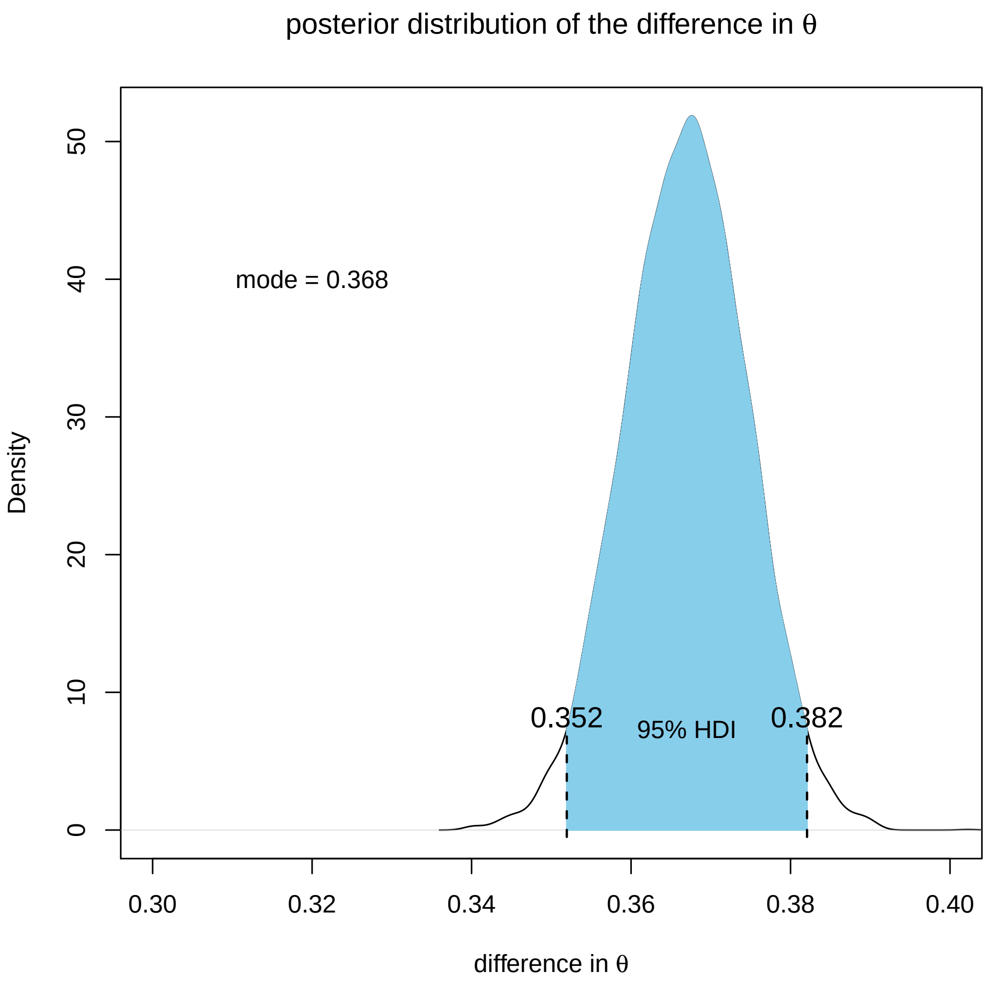


Fig. 2.9 Density plot from 10,000 samples from the posterior credibility distribution of the differences in θ between randomly selected reads and SbfI site containing reads. θ stands for the probability of a read pair to map concordantly.

Using the same command lines as above for read pairs containing SbfI sites, I counted among a total of 2,184 mapping read pairs 333 which mapped concordantly, or 15.5%.

Figure 2.9 shows the posterior credibility distribution of the difference in θ – the probability that a read maps concordantly – between the randomly selected read pairs and the SbfI site containing read pairs. The difference in the probability to map concordantly between a randomly selected read pair and a read pair containing an SbfI site is 0.367 (95% HDI: 0.352, 0.382). This is a strong difference that says that 37% *more* reads map concordantly among the randomly selected read pairs than among the SbfI site containing read pairs.

This result is consistent with random genomic religation, which recreates SbfI sites and disrupts concordant mapping of read pairs. If the SbfI sites observed in some reads would only be caused by incomplete digestion, this large difference would not be expected. On the other hand, the fact that 15% of SbfI site containing reads do map concordantly indicates that genomic religation cannot account for all the observed SbfI sites in the reads.

2.2.3 Incomplete digestion or genomic religation in the double–digest RAD library

There are also reads from the SbfI+XhoI double-digest library that contain full restriction enzyme recognition sequences. I have counted the SbfI and XhoI site positions in PCR-deduplicated SE and PE reads from each individual. The figures 2.10 and 2.11 show the relative site frequency distributions across SE and PE reads for SbfI and XhoI, respectively (further details in section 2.3.11).

Cluster Analysis

As with the SbfI site containing reads from the standard SbfI RAD library, if random genomic religation were responsible for the the SbfI and XhoI sites in the double-digest RAD reads, then the subsequences left and right of the SbfI or XhoI sites in the reads should come from different loci. When clustering reads by similarity of the subsequences left of the SbfI or XhoI sites, the subsequences right of the SbfI or XhoI sites should be very divers within clusters.

After collapsing identical SbfI and XhoI site containing SE reads into one, I have therefore clustered them by the subsequence left (5') of the recognition sequence (further details on page 60). Looking into these clusters of uniqued SE reads with XhoI sites does for the vast majority show clusters that are consistent with incomplete digestion (see fig 2.12)! The clusters from SE reads with an SbfI site, on the other hand, for the vast majority support genomic religation (see figure 2.13).

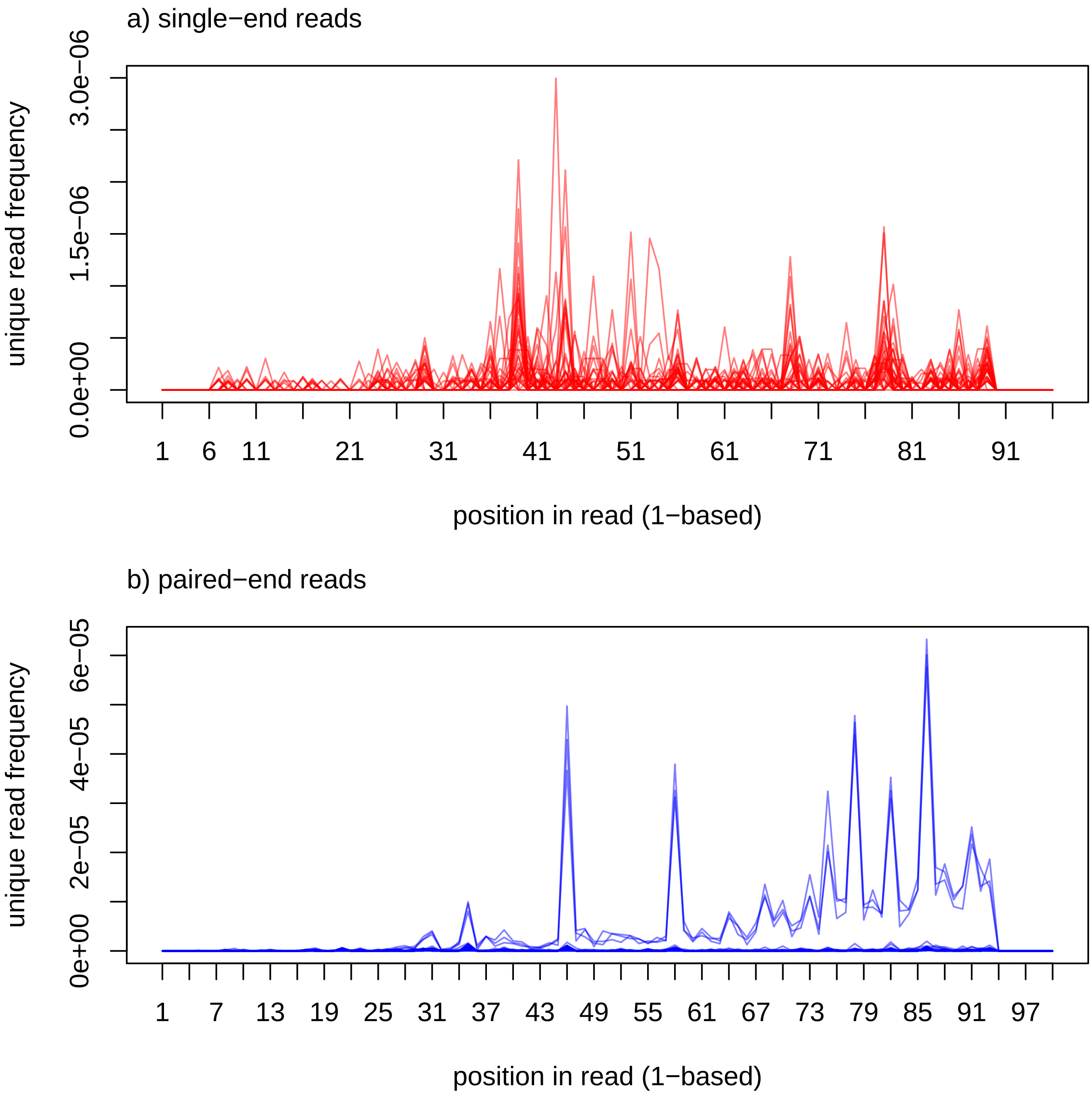


Fig. 2.10 Relative SbfI site frequency distributions for the SbfI-XhoI double-digest RAD data set with per individual uniqued reads relative to individual read count. Note that the graph for the PE reads has a 20 times larger y-axis than the graph for SE reads. SE reads are 96 base pairs long, PE reads are 100 base pairs long.

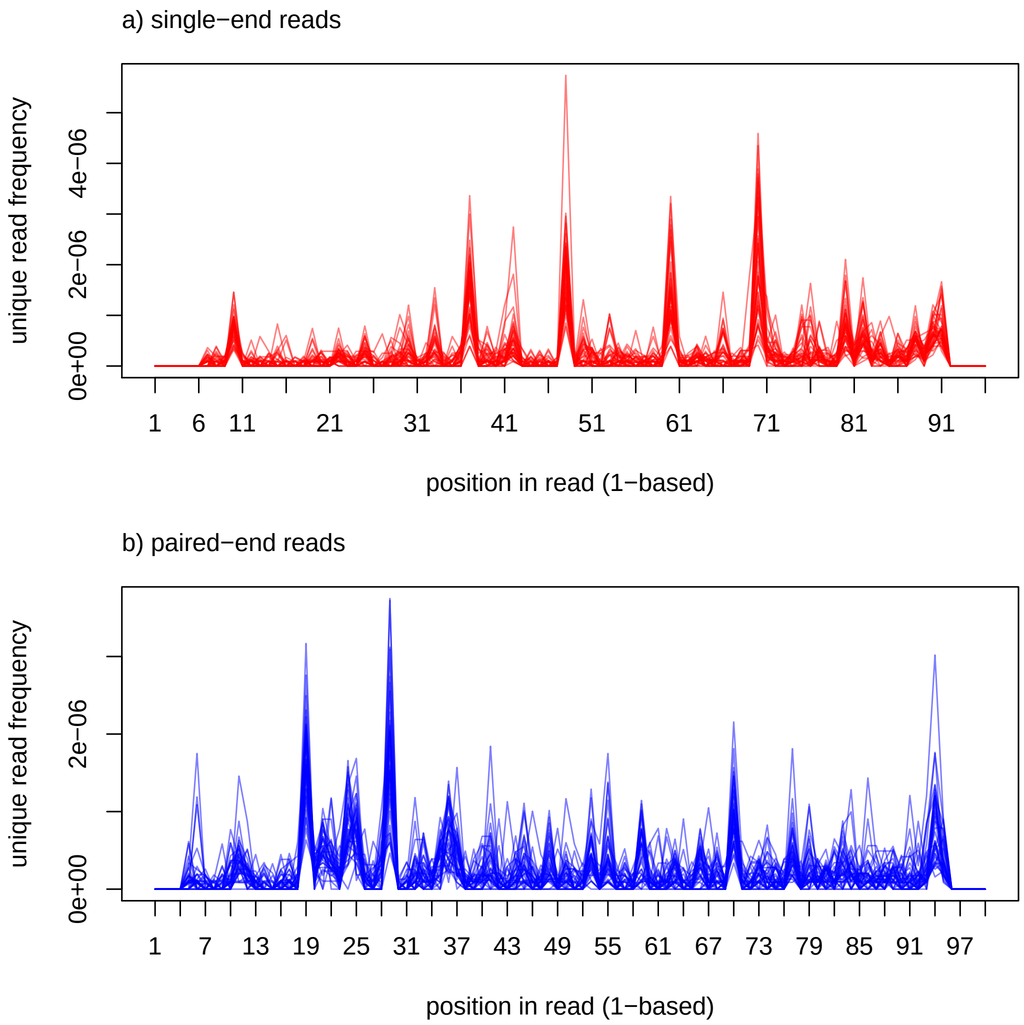


Fig. 2.11 Relative XhoI site frequency distributions for all 38 individuals (including two technical replicates) with per individual uniqued reads relative to individual read count.

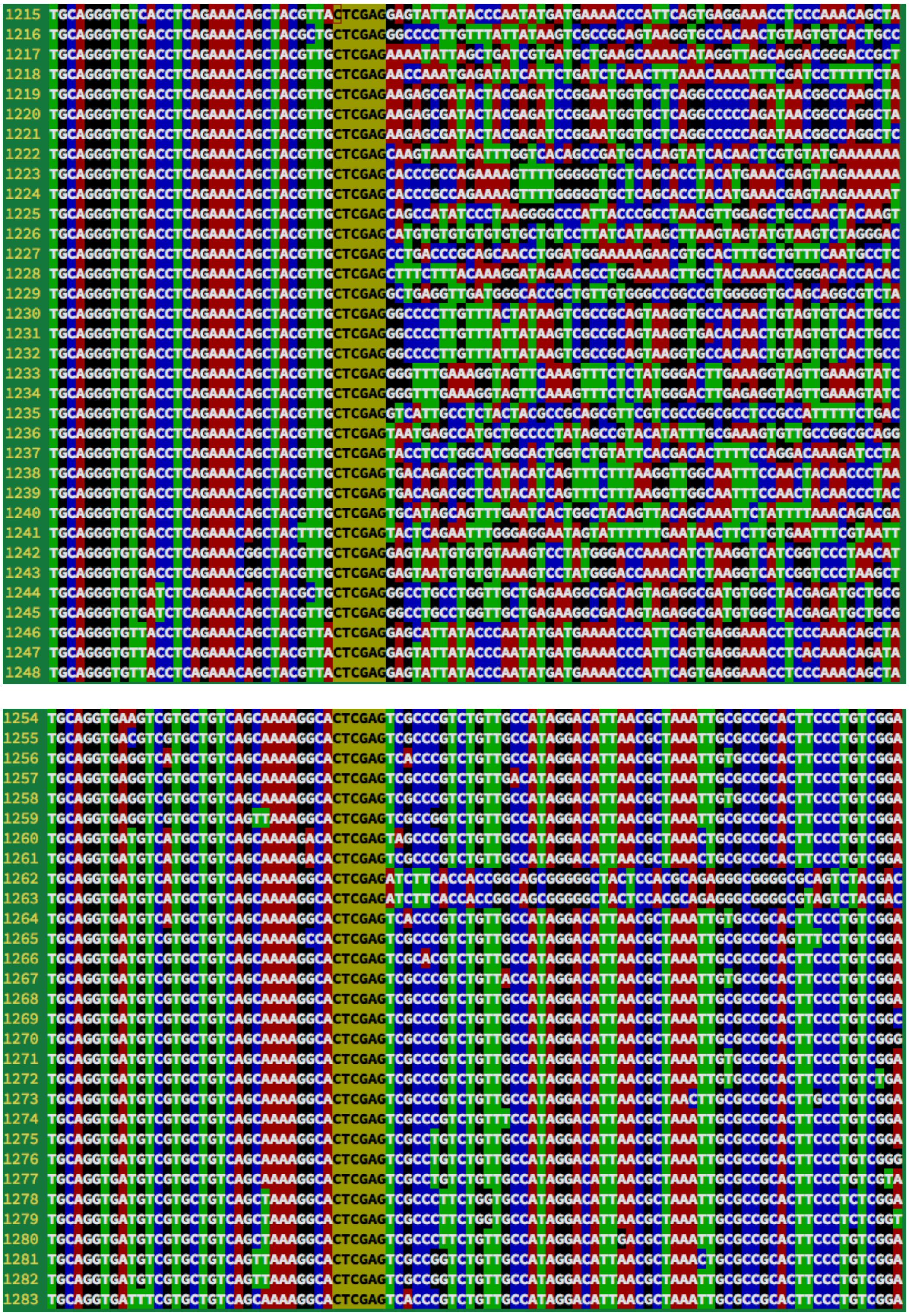
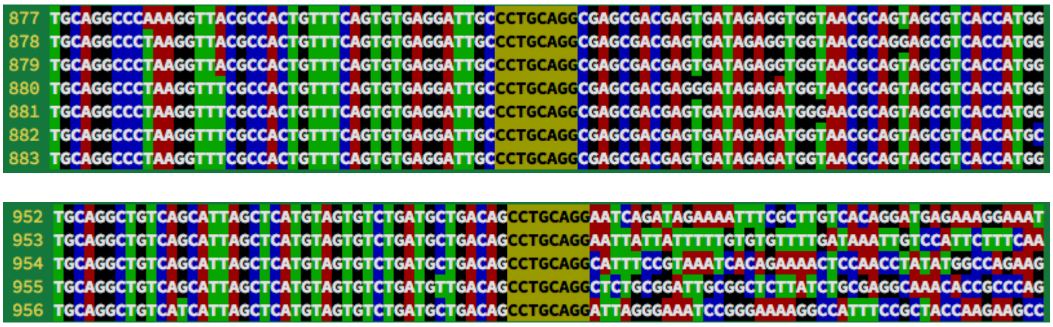


Fig. 2.12 Two clusters produced by `cluster.pl` on all uniqued SE reads containing an XhoI site. The top cluster is one of only very few that is indicative of genomic religation given its sequence diversity downstream of the XhoI site. The bottom cluster is one of the vast majority (including the many small ones) that are only consistent with incomplete digestion given the lack of sequence divergence downstream of the XhoI site.



```
877 TGCAGGCCCAAAGGTTACGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGGTGGTAACGCAGTAGCGTCACCATGG
878 TGCAGGCCCTAAGGTTACGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGGTGGTAACGCAGGAGCGTCACCATGG
879 TGCAGGCCCTAAGGTTACGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGGTGGTAACGCAGTAGCGTCACCATGG
880 TGCAGGCCCTAAGGTTTCGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGGGATAGAGATGGTAACGCAGTAGCGTCACCATGG
881 TGCAGGCCCTAAGGTTTCGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGATGGGAACGCAGTAGCGTCACCATGG
882 TGCAGGCCCTAAGGTTTCGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGATGGTAACGCAGTAGCGTCACCATGC
883 TGCAGGCCCTAAGGTTTCGCCACTGTTTCAGTGTGAGGATTGCCCTGCAGGCGAGCGACGAGTGATAGAGATGGTAACGCAGTAGCGTCACCATGG
```

```
952 TGCAGGCTGTCAGCATTAGCTCATGTAGTGTCTGATGCTGACAGCCTGCAGGAATCAGATAGAAAATTTCGCTTGTCACAGGATGAGAAAGGAAAT
953 TGCAGGCTGTCAGCATTAGCTCATGTAGTGTCTGATGCTGACAGCCTGCAGGAATTATTATTTTTGTGTGTTTTGATAAATTGTCCATTCTTTCAA
954 TGCAGGCTGTCAGCATTAGCTCATGTAGTGTCTGATGCTGACAGCCTGCAGGCATTTCCGTAAATCACAGAAAACTCCAACCTATATGGCCAGAAG
955 TGCAGGCTGTCAGCATTAGCTCATGTAGTGTCTGATGTTGACAGCCTGCAGGCTCTGCGGATTGCGGCTCTTATCTGCGAGGCAAACACCGCCCAG
956 TGCAGGCTGTCATCATTAGCTCATGTAGTGTCTGATGCTGACAGCCTGCAGGATTAGGGAAATCCGGGAAAAGGCCATTTCCGCTACCAAGAAGCC
```

Fig. 2.13 Two examples of clusters produced by `cluster.pl` on all uniqued SE reads containing an SbfI site. The upper cluster is indicative of incomplete digestion. The lower cluster is indicative of genomic religation.

Read pair mapping analysis

As for the standard RAD library, if the SbfI or XhoI sites in the reads of the double-digest RAD library are due to random restriction fragment religation, then they should be chimeras. The sequences left and right of the restriction site should map to different parts of a set of genome reference sequences. That's why I wrote a script called `digidig.pl` that takes a FASTQ SE read file, looks whether there are SbfI (or XhoI) sites in the reads at a position that leaves at least 30 bp to the left and right of the recognition sequence, splits the reads at the SbfI (or XhoI) site and writes the upstream part to a new SE file and the reverse complement of the downstream part to a new PE file. Next, I mapped these individual paired read files against the *Locusta* reference genome (Wang et al., 2014) with `bowtie2`. I have then extracted all reads from read pairs where both reads mapped, properly or not and disregarding mapping quality (for further details see page section Read pair mapping analysis on page 62).

Using the same command lines as above for the standard RAD read pairs, but specifying an accepted insert size range of 30 till 120 (instead of 50 till 900), I counted 116 concordantly mapped read pairs (indicating incomplete digestion) among a total of 932 mapped read pairs (12.4%).

I have also digitally digested the SE reads with XhoI using `digidig.pl` and mapped the digested reads against the *Locusta* genome reference. Using the same methods as for SbfI site containing SE reads, I counted 949 read pairs mapping concordantly (indicating incomplete digestion) among a total of 2250 mapped read pairs (42.2%).

In order to estimate the expected proportion of concordantly mapping read pairs for reads where both read subsequences are derived from the same location in the *C. parallelus* genome (as with incomplete digestion), I modified `digidig.pl` to also allow *random digital digestion*. It picks a random cut position while making sure that each new read in the resulting read pair is at least 30 bp long.

As a quasi random sample of reads I took the 200,001st – 300,000th SE reads from each individual. I then digitally digested these random reads randomly with `digidig.pl` creating new read pairs with variable read lengths. I then mapped these read pairs against the *Locusta* genome reference with `bowtie2` with the same settings as before (see page 62). I used the same set of command lines as above for the SbfI and XhoI digitally digested SE reads. I thus counted 689,319 concordantly mapping read pairs among a total of 1,041,305 mapping read pairs, or 66.2%.

Table 2.2 Summary of results from read pair mapping analyses. The percentage of concordantly mapping read pairs is shown for each data set together with the expected percentage from random read pairs.

data	% concordant	
	observed	expected
SbfI in standard RAD	15.5	52.0
SbfI in ddRAD	12.4	66.2
XhoI in ddRAD	42.2	66.2

2.2.4 Summary

Table 2.2 summarises the mapping results for the two different RAD libraries and the two restriction enzymes used and shows that reads with SbfI sites map concordantly much less often than would be expected if they were mainly caused by incomplete digestion. Reads containing XhoI sites, on the other hand, map concordantly much more frequently, although still far fewer than expected if incomplete digestion were the sole reason for XhoI sites in these reads. These results are consistent with the results from the analysis of clusters of cut site containing reads.

So incomplete digestion seems to have affected XhoI sites much more than SbfI sites in both types of RAD libraries. In order to detect an XhoI site that was not cut within SE reads of the double-digest library, it (i) needs to be within less than 90 base pairs of the SbfI site and (ii) have a second XhoI site within the next few hundred base pairs that *was* cut, so that a P2 adapter could be ligated. Given these highly restrictive conditions, it seems likely that incomplete digestion of XhoI sites has significantly reduced the template amount for the SbfI+XhoI double-digest RAD library.

Incomplete digestion is a much more severe problem with the library preparation than genomic religation. When assembling RAD tags only from SE reads as in standard RAD, genomic religation should only interfere with clustering when the second restriction site is within a read length of the first. Other chimeras should not be a problem for clustering. By contrast, incomplete digestion does not so much interfere with RAD tag assembly – two SbfI sites within a read length of each other should be rare anyway – but it indicates that an unknown (and probably unknowable) fraction of restriction sites were not cut. That is because an uncut site can only be detected if there was a restriction site close by that *did* get cut. So, the indication that the SbfI restriction sites in the reads of the two RAD libraries are probably mostly due to genomic religation is a good sign, although it indicates an inefficient ligation of Illumina adapters, which also reduces PCR template amount. The general lack of signs for genomic religation in the SE reads containing an XhoI site, on the other hand, is reason for concern, since it suggests incomplete digestion.

Given the rarity of signs for incomplete digestion of SbfI restriction sites, it seems unlikely that it alone could have reduced the template amount enough to cause the extreme skew in the distribution of putative RAD tag loci over genotype calls (i.e. individuals) as observed in figure 2.2a. Although a high-fidelity version of SbfI was used, remaining star activity of SbfI may be an explanation for the large number of singleton loci in figure 2.2a.

2.3 Methods

All commands and programmes shown here have been executed in the command processor BASH on the Debian-based Linux operating system called Ubuntu.

2.3.1 Creation of figure 2.1

The base map was created from two digital elevation model (DEM) tiles of the SRTM 90m Digital Elevation Database v4.1 provided by CGIAR–CSI (Jarvis et al., 2008). I used QGIS to extract the polygons corresponding to the country borders of France and Spain from a shapefile containing all world-wide country borders (downloaded from Natural Earth). Raster manipulations (projection, merging, clipping to polygon outline, color relief, hill shade and slope shade) were carried out with utilities from the Geospatial Data Abstraction Library (GDAL). The exact GDAL command lines can be provided upon request. Raster overlays and map design was performed with QGIS (2017) (version 2.18).

2.3.2 Adapter sequences for SbfI+XhoI ddRAD

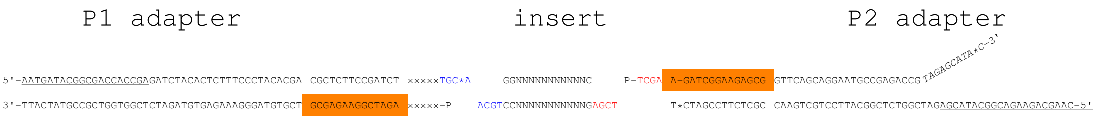

P1 adapter
insert
P2 adapter
5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA CGCTCTTCCGATCT xxxxxTGC*A
3'-TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCT GCGAGAAGGCTAGA xxxxx-P
GGNNNNNNNNNNNC
ACGTCCNNNNNNNNNNNGAGCT
P-TCGA A-GATCGGAAGAGCG GTTCAGCAGGAATGCCGAGACCG TAGAGCATA*C-3'
T*CTAGCCTTCTCGC CAAGTCGTCCTTACGGCTCTGGCTAGAGCATACGGCAGAAGACGAAC-5'

Fig. 2.14 Outline of P1 and P2 adapters for double-digest RAD. Underlined sequences are selective PCR primer sequences. An asterisk * stands for a phosphorothioate bond. A "P" stands for a phosphate group. Sequences in *italics* are non-complementary (wavy). Sequences with an orange background are identical to each other. An "x" stands for a base in a barcode sequence.

2.3.3 Estimating genome wide GC content

Using the SE reads from the SbfI RAD library (excluding the SbfI recognition sequence part) and command 2.2 I have determined the GC content of the SE reads: 49.5%. So, it seems that sequences close to SbfI sites are more GC rich than further distant sequences.

Command 2.1 For each individual, this command takes the PEs reads, removes exact duplicates and determines their overall GC content. Finally, the average of the individual GC contents is taken. Note the brute-force parallelisation by sending each iteration of the `for` loop into the background with `(...)&`.

```
for i in *fq_2.gz;
do
( awk '(NR-2)%4==0' <(zcat $i) | perl -ne' $h{$_}=1;END{foreach $s (keys %h)
{$gc += $s =~ tr/GC/GC/;} print $gc/(51 * scalar keys %h), "\n";} ' )&
done | \
perl -ne'$sum+=$_; END{print $sum/$., "\n";}'

0.4607
```

Command 2.2 This command is a different version of command 2.1. It is used here to determine the GC content of all SE reads from the standard SbfI RAD library. It first creates and exports two functions, `gc` and `mean`, and then uses the programme `parallel` in order to parallelise the determination of GC content over 10 cores. After stripping barcode and the remainder of the restriction site, the reads are 40 base pairs long. Note, the space between { and `awk` (line 1) as well as { and `perl` (line 7) is required.

```
1   gc(){ awk '(NR-2)%4==0' | sed 's/^......//' | \
2   perl -ne'$h{$_}=1;END{foreach $s (keys %h){$gc+=$s=~tr/GC/GC/;}
3   print $gc/(40*scalar keys %h), "\n";}' ;}
4
5   export -f gc
6
7   mean(){ perl -ne'$sum+=$_; END{print $sum/$., "\n";}'; }
8
9   export -f mean
10
11  parallel -j 10 'zcat {} | gc' ::: *fq_1.gz | mean
12
13  0.4956
```

2.3.4 Genome size of *Chorthippus parallelus*

Chorthippus parallelus has a chromosome complement of 2n = 16 + X. Males have one X-chromosomes, females have two. Table 2.3 shows four studies that provide genome size estimates for *Chorthippus parallelus*. Note that all studies are measuring the DNA content of spermatids. However, none of the studies explicitly deal with the issue that half of their measurements are missing the contribution from the X chromosome.

Table 2.3 mentions the country of origin of the grasshoppers used for genome size estimates. Note, however, that only Belda et al. (1991) provide sampling locations. For the rest it is assumed that individuals were sampled close to the authors institutes. So two studies provide genome size estimates for *C. parallelus parallelus* and two for *C. parallelus erythropus*. The *parallelus* subspecies seems to have a genome 2–4 giga bps larger than the one of subspecies *erythropus*. Apart from possible systematic differences in methodology, this apparent difference in genome size could be real, since several studies have shown

Table 2.3 Studies that provide genome size estimates for *Chorthippus parallelus*

study	sample source	C-value $[10^{-12}g]$	Method	tissue type	standard species
John and Hewitt (1966)[a]	UK	12.37 (± 0.75)[b]	Feulgen staining, microdensitometry	spermatid	*Locusta migratoria*[c]
Wilmore and Brown (1975)[d]	UK	13.36 (± 0.04)	Feulgen staining, microdensitometry	spermatid	mouse spermatid
Gosalvez et al. (1980)[e]	Spain	8.58 (± 0.47)[f]	Feulgen staining, microdensitometry	spermatid	*Allium cepa*, root tissue
Belda et al. (1991)	Spain	10.73 (± 0.97)[g]	Feulgen staining, microdensitometry	spermatid	chick erythrocytes

[a] see their table 3
[b] 5 individuals
[c] assuming 6.4 on relative scale corresponds to C-value of 5.5 pg
[d] see their table 2
[e] *C. longicornis* syn. of *C. parallelus*
[f] 3 individuals
[g] 3 individuals

chromosomal differentiation between the two subspecies, in particular on the X chromosome: an active nucleolar organiser region (NOR) distally on the X in *C. p. parallelus* but not in *C. p. erythropus* (Gosálvez et al., 1988). This NOR on X lies in or near a distinctive distal C-band[9]. In addition, Pyrenean *C. p. erythropus* also show an interstitial C-band on X that does not occur in pure *C. p. parallelus* (Bella et al., 2007). Further chromosomal differences are listed in table 1 of Ferris et al. (1993).

Gosálvez et al. (1988) showed that all the heterochromatin present in both subspecies is particularly rich in GC DNA base pairs.

2.3.5 Expected RAD marker number

Using PE reads from the standard RAD library – PCR-deduplicated per individual – as a proxy for the whole genome, I estimate the GC content of the *Chorthippus parallelus* genome to be around 46% (see Estimating genome wide GC content on page 42). However, Wilmore and Brown (1975) have determined the GC content of the *C. p. parallelus* genome from thermal dissociation profiles (41.2%) and sedimentation in CsCl and Cs_2SO_4 density gradients (41.7% and 42.0%)[10]. I think that PE reads from SbfI standard RAD are still a biased sample towards GC rich regions of the genome due to the GC rich SbfI recognition sequence. Assuming a genome size of 12 giga base pairs, the expected number of RAD tag

[9] heterochromatin stained with Giemsa
[10] see their table 1

loci from a standard RAD library with SbfI in the grasshopper genome is ∼170,000 (equation 2.2).

$$\begin{aligned}\text{expected number of RAD tags} &= \underbrace{12\times 10^{9}}_{\text{genome size}} \times \underbrace{\left(\frac{0.42}{2}\right)^{6} \times \left(\frac{(1-0.42)}{2}\right)^{2}}_{\text{SbfI site probability}} \times \underbrace{2}_{\text{tags per SbfI site}} \\ &= 173,110 \end{aligned} \tag{2.2}$$

The number (and identity) of markers in a double-digest RAD library depends very much on the size selection of restriction fragments. I selected fragments roughly between 300 and 800 bp length. The P1 adapter is 63 bp long (excluding 4 bp overhang), the P2 adapter is 61 bp long (excluding 4 bp overhang). The SbfI remainder after the cut is 6 bp long and the XhoI remainder is 5 bp long. If XhoI cuts a fragment at a distance less than about $300-63-61-6-5=165$ bp away from the SbfI cut site, then this fragment would not be size selected because it would be shorter than the lower bound of size selection (in this example). The SE sequences (excluding the SbfI recognition sequence parts) have a mean GC content of 49.5% (see command 2.2). The following formula requires the GC content of sequences of length 168 bp adjacent to SbfI sites. I will use the average of SE and PE GC contents – 48% (see section 2.3.3 on page 42) – for calculating the probability of no XhoI cut within the first 168 bp after the SbfI restriction site. In the next 500 bp then needs to be at least one XhoI site to make the SbfI fragment a marker. The expected number of RAD markers per genome with an SbfI-XhoI double-digest and a selected size range of 300–800bp is:

$$\begin{aligned}\text{RAD markers per genome} &\simeq 12\times 10^{9} && \text{(genome size in bp)} \\ &\times \left(\frac{0.42}{2}\right)^{6} \times \left(\frac{(1-0.42)}{2}\right)^{2} && \text{(SbfI cut prob. per bp)} \\ &\times 2 && \text{(each cut creates two potential RAD tags)} \\ &\times \left[1-\left(\frac{0.48}{2}\right)^{4} \times \left(\frac{(1-0.48)}{2}\right)^{2}\right]^{165} && \text{(prob. of no XhoI cut in the first 165 bp after SbfI site)} \\ &\times \left(1-\left[1-\left(\frac{0.46}{4}\right)^{2} \times \left(\frac{(1-0.46)}{4}\right)^{4}\right]^{500}\right) && \text{(prob. of at least one XhoI cut in the following 500bp)} \\ &\simeq 16,000 \end{aligned} \tag{2.3}$$

I have created an Excel file called `ComplexityReduction.xls` that implements equation 2.3 and that allows easy modification of variables.

2.3.6 Assembling pairs of PE read contigs

READ MAPPING The 49 million standard SbfI RAD reads used here were base call quality filtered with process_radtags[11] and a quality score threshold of 20 in a 20 bp sliding window. process_radtags also made sure that the remainder of the SbfI restriction site was present at the 6th position in the SE reads and that it was preceded by one of the 36 5 bp long barcodes that I used, for each case allowing 1 bp mismatch. I mapped these reads against the *Schistocerca* transcriptome (Badisco et al., 2011) with stampy (Lunter and Goodson, 2011) and setting the switches -noautosense, to turn off inference of insert size distribution, and -insertsd=400, to specify a very wide insert size distribution. This is to ensure that the proper pair bit in the SAM flag is set correctly by stampy. I then used command 2.3 in order to extract all Sequence Alignment/Map format (SAM) records, where at least one read from a pair got mapped, with subsequent position sorting. After this filtering, I merged all individual mapping output files into one big file with samtools merge. Note, that I did not filter for mapping quality scores, so that reads with ambiguous mapping position[12] were also retained. When stampy identifies several equally good mapping locations for a read or read pair, it reports one of these at random. Also note that the *Schistocerca* expressed sequence tags (ESTs) were assembled with the programme phrap and the authors do not report any attempt to merge different transcripts from the same gene into so-called unigenes. It can therefore happen that reads that are derived from the same position in the genome map to different parts of the *Schistocerca* reference.

Command 2.3 Command line that uses samtools and awk to create position sorted bam files in parallel that only contain records of paired reads where at least one read of the pair got mapped (i.e. skipping records with flag 77 and 141). Note the brackets around the command line and the skipping of ";" between "&" and "done".

```
for i in *sam.gz; \
do (samtools view -hS $i | gawk '/^\@/ || and($2, 4)==0 || and($2, 8)==0' | \
samtools view -uhS - | samtools sort - `basename $i .fq_1.sam.gz`) & done
```

DETECTING LINKED RAD TAG SITES The programme IGV among others can be used to visualise the alignment of many reads against a set of reference sequences. However, visually inspecting all *Schistocerca* reference contigs for whether they have read pairs mapped to both sides of one SbfI restriction site is very tedious and time consuming. That is why I wrote the script called find_linked_RADtags.pl which reports reference contigs where at least two read pairs map to opposite sites of an SbfI restriction site (or any cut site leaving a 4 bp overhang). This script also detects the contig shown in figure 2.15. With this script

[11] from the stacks pipeline

[12] with quality score < 3, see stampy README section 11.5

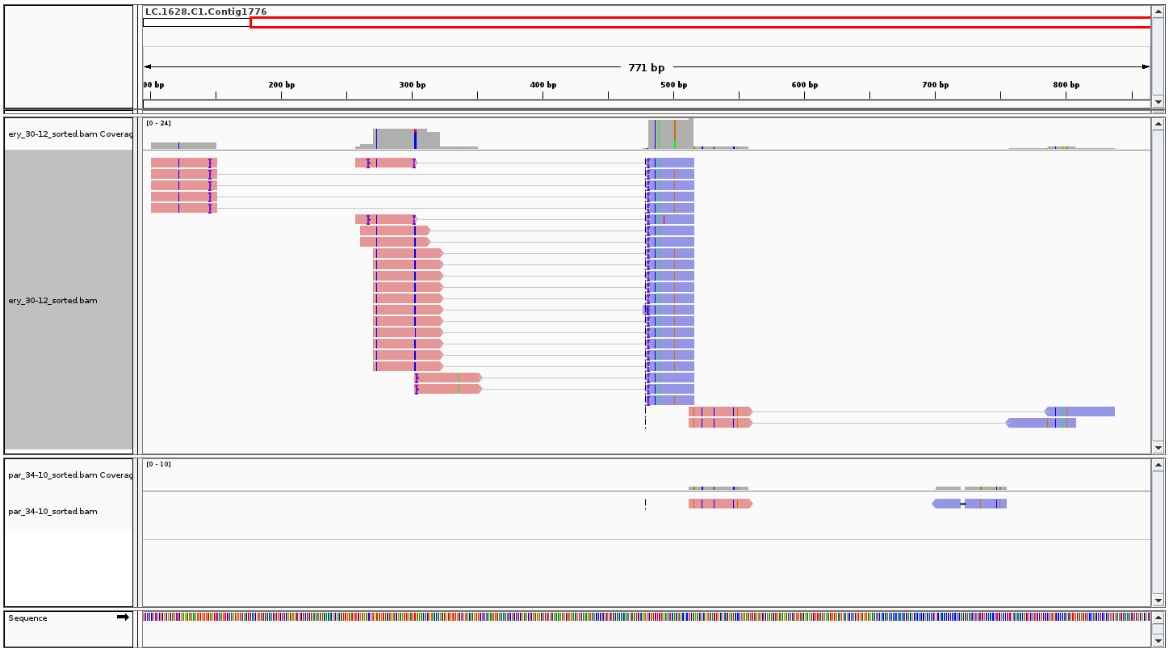


Fig. 2.15 Alignment of standard RAD read pairs to both sides of an SbfI restriction site. Read pairs are connected by a line. Forward mapping reads are pink, backward mapping reads are blue. The upper individual has 7 unique read pairs, i.e. with different paired-end reads.

the detection of a reference contig requires one concordantly mapped read pair (also called proper pair) on both sides of an SbfI restriction site. So SE as well as PE reads need to map on both sides of the restriction site. This is stringent and will obviously miss contigs with genuine SbfI RAD tag sites, but it is necessary to remove many false positive detections. The purpose of the script is not to detect as many contigs as possible, but only to detect several contigs with genuine SbfI RAD tag sites. The script `find_linked_RADtags.pl` collects all PE reads that are mates of SE reads that mapped to a detected linked RAD tag site. For each detected reference contig it prints reads upstream or downstream of that site to a separate file. Note, that at this stage `find_linked_RADtags.pl` will only detect one linked RAD tag site per reference contig. However, due to the small sizes of the transcriptome contigs, this should not be major shortcoming.

PE READ ASSEMBLY I attempted to use `VelvetOptimiser.pl` to assemble the collected PE reads into PE contigs (Zerbino and Birney, 2008). However, the programme fails to assemble three PE read contigs with low read coverage – Contig1776_downstream (see figure 2.15), Contig4139_upstream and LC03019A1F03.f1_upstream – despite my diligent attempts to provide the necessary settings (Davey et al., 2012; Etter et al., 2011; Zerbino, 2010).

`SSAKE` (Warren et al., 2007) is a simpler but also less heuristic and more tunable assembly programme than `Velvet`. It does not take base call quality scores into account and takes only multi-fasta files as input. It searches for *perfect* kmer matches between reads. i.e. does not allow for sequencing error or SNPs. SSAKE by default does not allow setting a minimum overlap (-m) of less than 16 bp. This could be too stringent for some of the low coverage PE contigs that I wanted to assemble. I, therefore, modified SSAKE to allow a minimum overlap of as small as 10 bp. When calling SSAKE with

Command 2.4 Command line that turns fastq files into multi-fasta files. It takes all fastq files in the parent directory, extracts the header and sequence part (while skipping the quality string), replaces the "@" at the beginning of the fastq headers with a required ">" sign and writes the stream to a new file with the same base name.

```
for i in ../*fq; do awk '(NR-1)%4==0 || (NR-2)%4==0' $i | \
sed 's/^@/>/' > `basename $i .fq`.fa; done
```

-w 1 Minimum depth of coverage allowed for contigs

-o 1 Minimum number of reads needed to call a base during an extension

... on the "Contig1776_downstream" reads from one individual for PE assembly (just two overlapping reads, see fig. 2.15), it is able to assemble a full length contig of 81 bp.

TagCle Any non-genomic sequence, i.e. adapter sequence, in the reads should interfere with de novo assembly. The Perl script TagCle by Kang-Wook Kim (Sheffield University) detects overlap between paired reads by Smith-Waterman local alignment and clips off read segments downstream of the end of the local alignment, i.e. generally adapter sequence. That way the script can also detect a *single* adapter (or barcode) base at the end of a read. I used command 2.5 to remove adapter sequences from the reads. TagCle clipped 159 SE and 216 PE reads of a total of 1,584,732 reads (0.02%). It did not discard any sequence.

Command 2.5 This is the command line that I used in order to run the script TagCle on all 154 pairs of input files in parallel. The -me switch turns off any direct search for adapter sequences.

```
for i in ../input/*fq_1;
do
(j=`echo $i | sed 's/1$/2/'`; TagCle_0.70.pl -me -i1 $i -i2 $j > `basename $i .fq_1`.log) &
done
```

Kmer size optimisation All de novo assemblers require optimisation of kmer length (Davey et al., 2012), which is mainly what VelvetOptimiser.pl does with Velvet. So I wrote a script called SSAKEoptimiser.pl which for each set of PE reads iterates through kmer lengths from 11 to 33 and keeps the assembly which produces the longest contig. This script exists in several parallelised versions (using different parallelisation modules) that

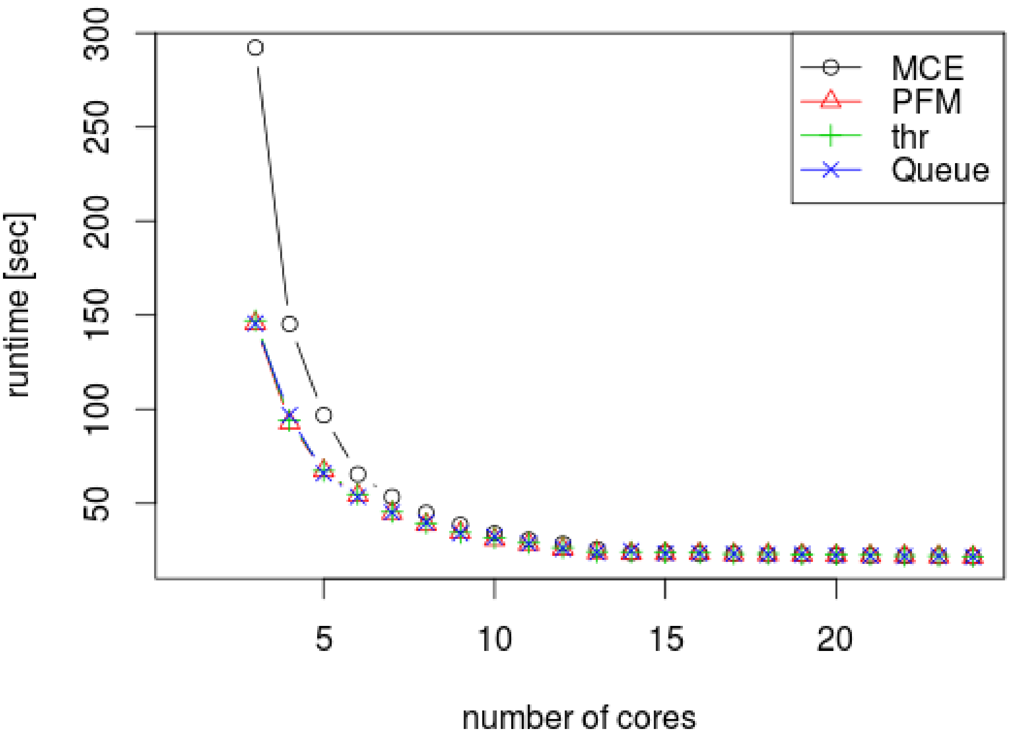
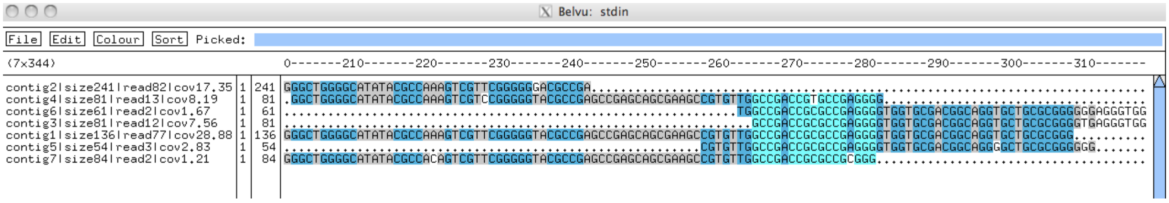


Fig. 2.16 The run times of the four parallelised versions of the `SSAKEoptimiser` script on 11 input files.

parallelise not only over input files but also over iterations through kmer lengths. This full parallelisation greatly speeds up execution on a multi-core machine (figure 2.16).

PICKING THE RIGHT SSAKE CONTIGS `SSAKE` generally assembles many contigs for a region. In some cases this is due to different SE RAD tags mapping to the same position in the *Schistocerca* transcriptome. In other cases, this is clearly due to insufficient merging of contigs by `SSAKE` due to low coverage and SNPs (see figure 2.17) That's why I created multiple alignments of `SSAKE` contigs from each assembly with `Muscle` (Edgar, 2004) in order to manually merge them in the alignment editor of `MEGA` (Tamura et al., 2013).



```
(7x344)                                   0-------210-------220-------230-------240-------250-------260-------270-------280-------290-------300-------310-------
contig2|size241|read82|cov17.35  1 241 GGGCTGGGGCATATACGCCAAAGTCGTTCGGGGGACGCCGA.............................................................................
contig4|size81|read13|cov8.19    1  81 .GGCTGGGGCATATACGCCAAAGTCGTCCGGGGGTACGCCGAGCCGAGCAGCGAAGCCGTGTTGGCCGACCGTGCCGAGGGG....................................
contig6|size61|read2|cov1.67     1  61 ..............................................................TGGCCGACCGCGCCGAGGGGTGGTGCGACGGCAGGTGCTGCGCGGGGGAGGGTGG
contig3|size81|read12|cov7.56    1  81 ................................................................GCCGACCGCGCCGAGGGGTGGTGCGACGGCAGGTGCTGCGCGGGGTGAGGGTGG
contig1|size136|read77|cov28.88  1 136 GGGCTGGGGCATATACGCCAAAGTCGTTCGGGGGTACGCCGAGCCGAGCAGCGAAGCCGTGTTGGCCGACCGCGCCGAGGGGTGGTGCGACGGCAGGTGCTGCGCGGG..........
contig5|size54|read3|cov2.83     1  54 .........................................................CGTGTTGGCCGACCGCGCCGAGGGGTGGTGCGACGGCAGGGGCTGCGCGGGGGG.......
contig7|size84|read2|cov1.21     1  84 GGGCTGGGGCATATACGCCACAGTCGTTCGGGGGTACGCCGAGCCGAGCAGCGAAGCCGTGTTGGCCGACCGCGCCGCGGG.....................................
```

Fig. 2.17 Multiple sequence alignment of `SSAKE` contigs assembled from reads collected from the downstream side of the RAD tag site in the *Schistocerca* reference contig "LC.1628.C1.Contig1776". The aligment view was created with the command: `muscle -in *LC.1628.C1.Contig1776_downstream*contigs -msf | belvu -` . The 7 contigs can clearly be merged into one big contig if allowing for a few SNPs.

Since the SSAKEoptimiser output generally contains several contigs of similar length and with similar read counts, it is difficult to pick those contigs upstream and downstream that genuinely belong together, i.e. come from the same locus. Visual inspection of a pairwise alignment showed that it is by no means always the longest contig assembled that aligns significantly to the *Schistocerca* reference.

In order to help me pick the right SSAKE contigs, I started by determining for each PE read assembly

- the length of the longest contig assembled
- the total number of contigs assembled and
- the number of contigs with length > 100, > 200 and > 300 bp

Another important piece of information would be a significant blast hit of a SSAKE contig against its putative *Schistocerca* reference contig. I therefore extracted all 64 *Schistocerca* reference contig sequences from the *Schistocerca* transcriptome reference file with command 2.6. With my script blast2seq.pl – calling NCBI blastn 2.2.28+ (Camacho et al., 2009)

Command 2.6 Example of a command line that extracts FASTA sequences from an indexed multi-FASTA file using a file listing FASTA headers.

```
for i in `cat reference_contig_names_with_up-downstream_contig`; \
do samtools faidx LC_unique.seq  $i > $i.fa; \
done
```

– I then blasted all 128 relevant SSAKE contig files against their putative *Schistocerca* reference contig sequence and recorded the number of blast hits as well as the fasta headers of the 10 best hitting SSAKE contigs together with their Expect (E) value (further explanation) in a .SSAKE_contig_stats file. I only recorded blastn hits with an Expect (E) value of less than 10^{-10}. I then used this table as a guide for picking and possibly merging SSAKE contigs in MEGA. I used command lines similar to 2.7 in order to find overlapping SSAKE contigs that haven't been merged yet.

Command 2.7 This command line example is a very quick way to find out which sequences in a multi fasta file are similar to each other. It prints out hits of an all by all blastn of the sequences in a file. Note that query and subject get the same file. The first awk command removes hits against itself, the sort part brings reciprocal hits together and the second awk command keeps only one line for each pair of matching sequences.

```
blastn -query *LC03012A1D06.f1_downstream.fa.ssake*.contigs \
-subject *LC03012A1D06.f1_downstream.fa.ssake*.contigs -task blastn \
-evalue 1e-10 -outfmt 6 | awk '$1 != $2' | sort -k3 -nk11 | awk 'NR%2' | less -S
```

I only called pairs of PE contigs if on each side of the restriction site I could unambiguously pick a `SSAKE` contig. This either required a much better blast hit than the second best `SSAKE` contig or, if no blast hits could be obtained, a small number of `SSAKE` contigs, one of them being much longer than the others. I required at least one of the two PE contigs from a restriction site to have a significant blast hit to the *Schistocerca* reference contig.

After picking and potentially merging `SSAKE` contigs, I aligned upstream and reverse complemented downstream contigs against their putative *Schistocerca* reference contig (if possible, i.e. significant blast hit) and filled the gap between them with N's. I thus created a new *C. parallelus* PE contig pair consensus sequence for each *Schistocerca* reference contig.

2.3.7 Backmapping

After the creation of 20 *C. parallelus* PE contig pair consensus sequences, I wanted to find out how many of the 36 individuals get reads mapped to these 20 linked RAD tag sites and whether individuals actually have reads mapped to both sides of an SbfI restriction site.

INCLUDING SE RAD TAG SEQUENCES INTO THE NEW REFERENCE Before mapping all standard SbfI RAD reads back against the newly created PE contig pair reference, I wanted to include the SE RAD tag sequences into the PE contig pair consensus sequences. For the determination of the consensus RAD tag sequences I obviously only want to use reads whose PE mate was used for the assembly of the `SSAKE` PE contig that I finally picked (see section 2.3.6). That's because the script `find_linked_Radtags.pl` had printed out all read pairs that mapped to a detected linked RAD tag site in the *Schistocerca* reference, but after the `SSAKE` assembly I mostly only picked PE contigs that got a `blast` hit to their *Schistocerca* reference contig. Other `SSAKE` contigs are much more likely to derive from similar but non-homologous loci to the *Schistcerca* reference contig. I started by using command 2.8 in order to extract the FASTQ headers from those PE reads that get a blast hit to their inferred PE contig. I then used the output files from this command, containing headers of the required SE sequences, as pattern files for a `grep` filter of the SE read files that

Command 2.8 Using `blastn` to find PE reads that map to the inferred PE contig (see section 2.3.6 on page 49). The `for` loop iterates over all 40 PE read files. The first part of the loop converts fastq to fasta format. The second line feeds that into `blastn` (using megablast by default) and uses the corresponding PE contig (from section 2.3.6) as subject. The third line takes the first column with the query headers from the blast output table and writes it to an output file.

```
for i in *fq_2; \
do awk '(NR-2)%4==0 || (NR-1)%4==0' $i | sed 's/@/>/' | sed 's/_pp//' | \
blastn -subject `basename $i .fq_2`_consensus.fas -evalue 1e-10 -outfmt 6 | \
cut -f1 | sed 's/2$/1/' > `basename $i .fq_2`_blast_mapped.ids; \
done
```

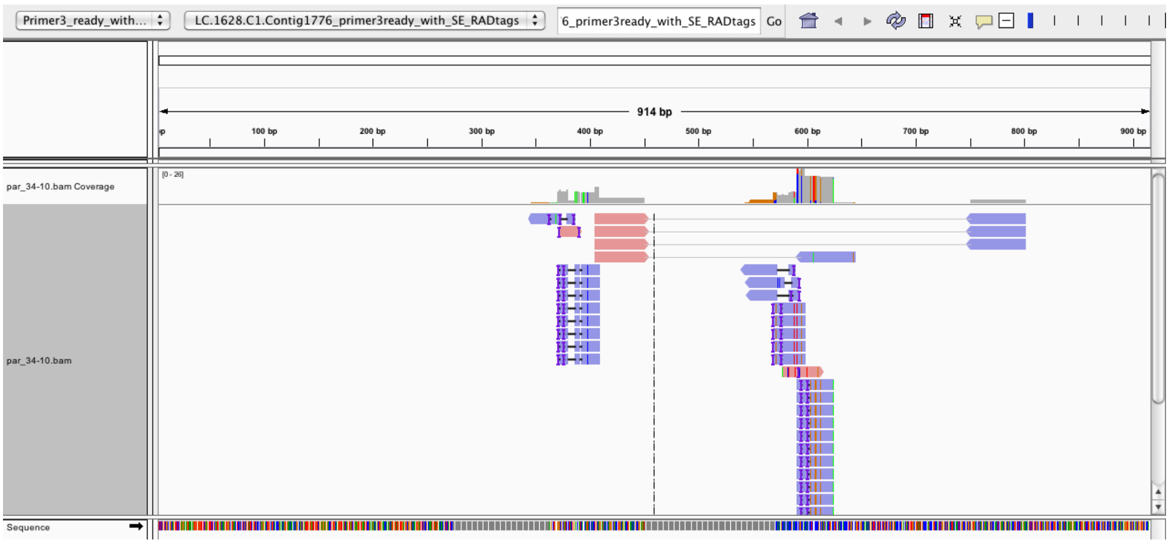


Fig. 2.18 Example read alignment of all standard RAD reads of individual par_34-10 against one PE contig pair reference sequence.

`find_linked_RADtags.pl` has put out (command 2.9). Having extracted these SE reads

Command 2.9 Using the header files created by the previous command (2.8) to extract corresponding SE reads from `find_linked_RADtags.pl` SE read files.

```
for i in *ids; \
do grep -A1 -f $i ../all_Big_Data_`basename $i _blast_mapped.ids`.fq_1 | \
egrep -v "\-\-" | sed 's/@/>/' > `basename $i .ids`_SE.fas; \
done
```

for each RAD tag, I created multiple sequence alignments of them with `muscle` in `.msf` format, which I could then use for the `consambig` programme from the `emboss` package in order to create SE RAD tag consensus sequences. Finally, I included these sequences into the 20 PE contig pair consensus sequences manually in `MEGA`.

TARGETED ALIGNMENT AND CLEAN UP OF MAPPING RESULT I used the programme `stampy` to align all standard SbfI RAD reads from all 36 individuals against this new set of reference sequences. Figure 2.18 shows an example of an alignment of this `stampy` mapping. There are many low quality mappings which are very likely wrong (e. g. SE reads mapped to PE contig without SbfI site). However, here I have been using reads derived from a much larger source than is represented in the small reference of 20 pairs of PE contigs. Therefore, `stampy` finds unambiguous mapping locations[13] for reads that have an edit distance of 17 to the reference sequence. `stampy` does not have an option for maximum allowed distance to the reference. Kosugi et al. (2013) have developed a few Perl scripts that

[13] indicated by a mapping quality >3

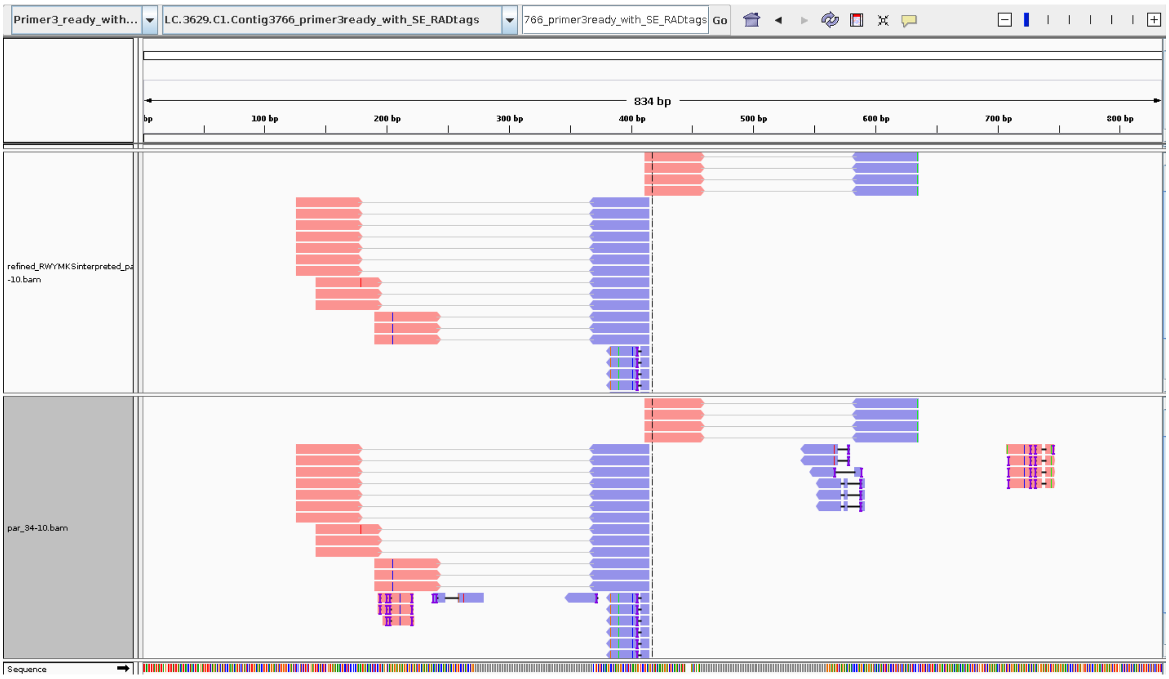

Primer3_ready_with...
LC.3629.C1.Contig3766_primer3ready_with_SE_RADtags
766_primer3ready_with_SE_RADtags
Go
834 bp
bp
100 bp
200 bp
300 bp
400 bp
500 bp
600 bp
700 bp
800 bp
refined_RWYMKSinterpreted_pa
-10.bam
par_34-10.bam
Sequence

Fig. 2.19 Read alignment of all standard RAD reads of individual par_34-10 against one PE contig pair consensus sequence. Upper track after `coval-refine` filtering. Lower track is without `coval-refine` filtering for comparison.

deal with the problem of false positive alignments especially when doing a so-called *targeted alignment*[14] by removing reads that map with too many mismatches. They show that filtering by mapping quality score is rather ineffective when trying to improve a targeted alignment. Their programme `coval-refine` by default removes all reads with more than 2 indels, 1 indel and 1 soft-clipped end and 2 soft-clipped ends. I have changed `coval-refine` so that indels count like two mismatches. I have set the maximum proportion of mismatches to 0.1. Thus SE reads are 46 base pairs long and are allowed to have up to 5 mismatches. The 51 base pair long PE reads are also allowed up to 5 mismatches. By default, `coval-refine` counts ambiguous positions in the reference as mismatches. I therefore changed `coval-refine` to take correct account of the dual ambiguity codes RWYMKS. Figure 2.19 shows the mapping result after `coval-refine` treatment for one individual and one reference sequence.

2.3.8 Estimation of template amount for selective PCR

I estimated the amount of template that went into the selective PCR for the SbfI+XhoI double-digest library (see figure 1.3) with quantitative PCR (Rutledge and Côté, 2003).

[14] when the reference sequence is much smaller than the source of the reads to be mapped

I took an aliquot of the selective PCR product and determined its DNA concentration with picoGreen and a fluorometer. I made a serial dilution of 8 times 1:10 from this PCR product and used this to produce a standard curve. I created three replicates of the standard curve and three replicates of the target, i.e. the solution before selective PCR. The C_t threshold was automatically set by the *q*PCR machine and, except for the lowest template concentration of the standard dilution, the melting temperature of the PCR products were always very close to 79°C. Assuming a mean template molecule length of 550 bp and 16,000 loci from this double-digest library (see equation 2.3), the *q*PCR results indicate that on average only 1.26 (± 0.37) template molecules per locus and individual went into the selective PCR[15].

2.3.9 comparison of different double-digest RAD protocols

Please see the table on the following page.

[15] see DD_RAD_LIB_101111_data.xls

Table 2.4 comparison of different protocols for non-standard RAD[a] (see figure 1.3)

protocol	adapters	DNA isolation	digestion	ligation	gel size selection	PCR	purification
Peterson et al. (2012)	P1 and P2 adapter at stock conc. of 40μM, short adapters and long PCR primers	NA	digest for 3h, no heat-inact. – instead bead puri.	30min ligation at RT, 2-10 fold excess of adapters to sticky ends	before PCR, automated DNA size selection with Pippin-Prep (Sage Science)	20ng size-selected library per PCR reaction, 2μM end conc. of each PCR primer, only 8-12 cycles (?!)	AMPure beads
Andolfatto et al. (2011)	10μM stock conc., short adapters + long PCR primers	Puregene (Qiagen)	10ng DNA per sample, with 3.3U of 4bp cutter MseI (i.e. no double-digerst), 3h at 37°C followed by heat-inact.	5nmole adapters, only 1 U of T4 DNA ligase in a volume of 50μl, 1h at 16°C	before PCR, ladder mixed into library, 2% gel	only 15 cycles, Phusion	AMPure beads
Parchman et al. (2012)	1μM stock conc. for P1 (EcoRI) adapter, 10μM stock conc. for P2 (MseI) adapter, annealing in pure water (?!), short adapters + long PCR primers	NA	6- and 4bp cutter, 10U EcoRI, only 1U MseI, digestion in T4 buffer, NaCl added to ∼50mM end conc., volume 9μl, 8h, heat-inact.	1pmole EcoRI adapter, 10pmole MseI adapter, 67 NEB units T4 ligase, 6h at 16°C, ligation in only 11.4μl	after PCR, 2.5% gel, low electric field gel runs, EtBr gels, many lanes	individual PCR before pooling samples and before gel size selection, 30 cycles (?!), only 0.08μM end conc. of each primer in PCR (?!), BioRad Iproof High Fidelity DNA polymerase	QiaQuick spin columns
this study	long adapters short PCR primers (fig. 2.14)	silica spin columns (Qiagen)	132 ng per sample, 10 U/sample SbfI-HF, 20 U/sample XhoI , NEBuffer 4, 3 h at 37°C followed by heat-inact.	∼22 fold excess of adapters to sticky ends, 400 NEB U/sample ligase, 2 h at RT, then over night at 4°C followed by heat-inact.	before PCR, 1% agarose gel with EtBr, whole pooled library in one lane, 13 V/cm	20-24 cycles, Phusion Mastermix, 1.0μM of each thiol-protected primer	Qiagen MinElute spin columns

[a] without a random shearing step

2.3.10 PCR duplicates in the standard SbfI RAD data

The standard SbfI RAD library had been sequenced on 4 lanes of an Illumina GAIIx sequencer with increasing template amount. I ran the programmes `RADpools` and `RADtags` from the `RADtools` package version 1.2.1 on the raw data of these four lanes separately. `RADpools` demultiplexes reads by barcode, filters them according to base call quality scores and finally sorts read sequences alphabetically in individual read files. The programme `RADtags` then takes these sorted read files and calls RAD tags based on maximum pairwise mismatch counts. `RADtools` therefore uses alphabetical sorting of reads (within individuals) in order to reduce the number of pairwise mismatch counts it has to perform to call clusters of reads into RAD tags. This is obviously a suboptimal clustering strategy since it does not guarantee to cluster all similar sequences together. For each RAD tag it reports the read count as well as fragment count. Different fragments are recognised by different PE reads, allowing for sequencing errors. PCR duplicates have the same PE read sequences and thus don't increase fragment count but do increase read count. I then ran my script `investigate_PCR_duplicates.pl` on the individual output files of `RADtags` from each lane in order to determine the total number of PCR duplicates from the read and fragment counts that it had put out for each cluster of SE reads. I then repeated the same analysis for all reads combined.

2.3.11 Investigation of restriction enzyme recognition sequences within RAD reads

SbfI and XhoI site frequency distributions

For the figures 2.3, 2.10 and 2.11 I used my script `position.pl` to tally – for each individual separately – the positions in the *unique* read sequences where an SbfI or XhoI recognition sequence was found. Each line in these plots connects the site frequencies of one individual. SE and PE reads were analysed separately. The reason for collapsing the reads into unique sequences is to remove the effect of PCR duplicates. Any peaks in the distributions can thus not be caused by PCR duplicates.

In figure 2.3 (a), there are three peaks in SbfI frequency at position 29, 34 and 39. Are these peaks caused by repetitive genome sequences that just happened to have an SbfI site at those positions? In order to investigate this, I used command line 2.10 in order to cluster all SbfI site containing reads from from the standard SbfI-RAD data set by the subsequence left (5') of their SbfI site. If the peaks at read position 29, 34 and 39 in figure 2.3 (a) were due to repetitive genome sequences, then the reads with SbfI sites at those positions should find themselves in bigger clusters than other reads. Large cluster sizes are indeed found for reads

Command 2.10 This command line clusters `uniq`-ed SE reads of all individuals that contain an SbfI site and then prints out foreach SbfI site position in the SE read length the cluster sizes that have been found. My custom script `cluster.pl` first groups reads by SbfI position and then clusters them within groups by mismatch count on the subsequence left of the SbfI site, thus ignoring the potentially non-homologous genomic sequence (due to religation) downstream of the SbfI site.

```
zcat *fq_1.gz | awk '(NR-2)%4==0' | grep "CCTGCAGG" | \
sort | uniq | cluster.pl > all_ind_pre_SbfI_cl_size_by_pos.cl
```

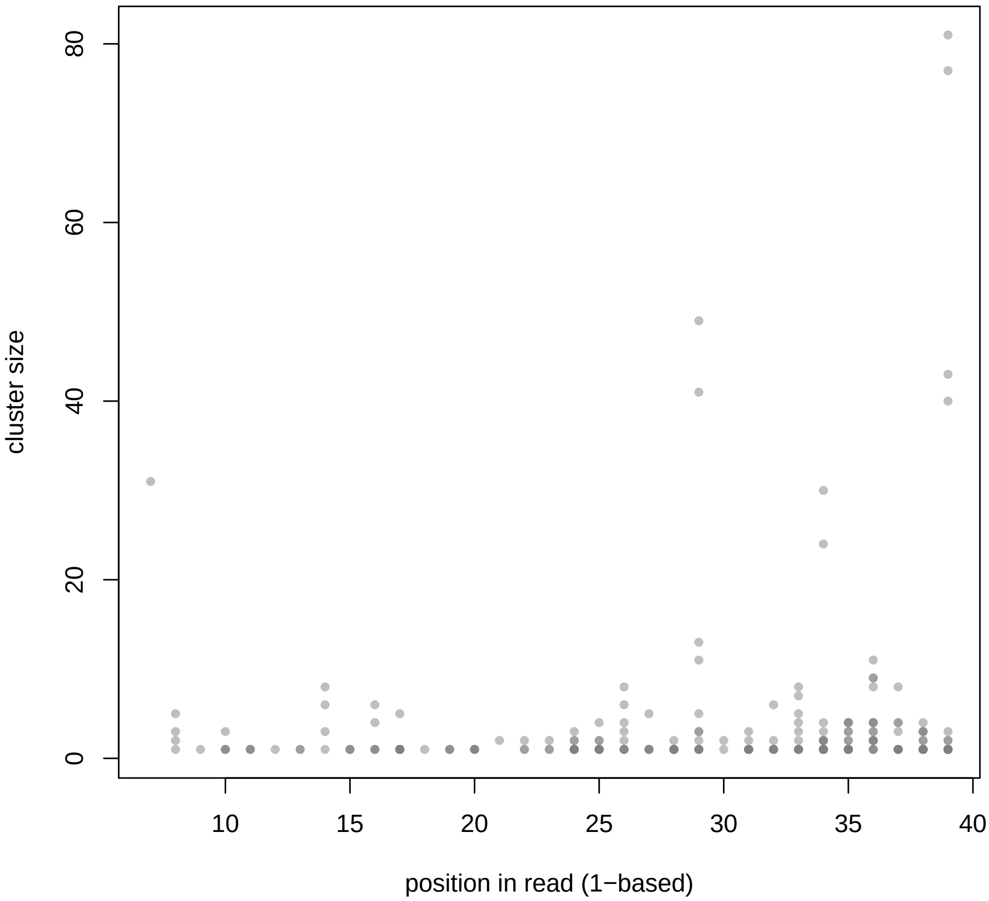
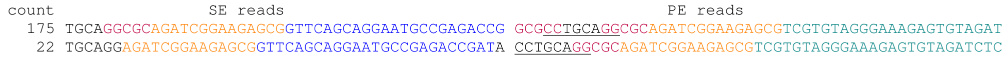


Fig. 2.20 Cluster size for reads with different position of SbfI site in the SE reads of the standard SbfI-RAD library.

with SbfI site at read position 29, 34 and 39 (figure 2.20). This supports the idea, that the peaks of SbfI site frequency are caused by repetitive genome sequences that happen to have an SbfI site at these positions.



```
count                   SE reads                                                              PE reads
  175 TGCAGGCGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG  GCGCCTGCAGGCGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGAT
   22 TGCAGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATA  CCTGCAGGCGCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTC
```

Fig. 2.21 PE reads with SbfI site from ind. ery_30-8. *underlined*: SbfI recognition sequence; `GCGCC`: barcode sequence; `AGA...GCG`: sequence common to P1 and P2 adapter; `TCG...`: P1 adapter sequence; `GTT...`: P2 adapter sequence. This figure was created from the output of command 2.11.

There are two individuals in figure 2.3 (b) which have exceptionally high SbfI site frequencies in PE reads (individual "ery_30-8" and individual "ery_30-17"). The exceptionally high number of SbfI sites at the beginning of PE reads from these individuals can be explained by P1 adapter dimers and the fact that the barcode sequences for these two individuals (`GCGCC` and `TGACC`) end with `CC` and that these barcode sequences (or at least the `CC` part) appear at the beginning of PE reads into P1 adapters, thus recreating an SbfI recognition sequence (fig. 2.21). An equivalent but not as frequent pattern can be found in the reads of individual "par_34-3", which has the barcode `AACCC`. The fact that the barcode sequence (or part of it) can be found at the beginning of PE reads can be explained by P1-P1 dimers where one of the adapters gets sheared off after (or within) the barcode sequence followed by ligation of the P2 adapter, which is necessary for illumina sequencing. The overall higher frequency of SbfI sites in PE reads of these three individuals (figure 2.22) can be explained by reads into the P1 adapter, i.e. sequencing the recreated SbfI site (figure 2.23).

Command 2.11 This command takes the SE and PE reads from one individual and pastes them side by side separated by a tab character. It then extracts those lines where the PE read contains an Sbf recognition sequence, counts the number of occurrences of identical lines and presents the most common lines at the top.

```
paste <(zcat ery_30-8.fq_1.gz | awk '(NR-2)%4==0') <(zcat ery_30-8.fq_2.gz | \
awk '(NR-2)%4==0') | grep " .*CCTGCAGG" | sort | uniq -c | sort -nrk 1 | less -S
```

Command 2.12 This command is similar to command 2.11. However, in addition to the SbfI recognition sequence it extracts lines that contain an 8 base pair sequence from the beginning of the illumina adapters (see figure 2.14).

```
paste <(zcat par_34-3.fq_1.gz | awk '(NR-2)%4==0') <(zcat par_34-3.fq_2.gz | \
awk '(NR-2)%4==0') | grep "CCTGCAGG.*AGATCGGA" | sort -k 2 | uniq | vim -
```

The reads for figure 2.10 and 2.11 are from reads of the SbfI+XhoI double-digest RAD library and were quality filtered with `stacks`' `process_radtags` and my custom script `grep_true_RADtag.pl`. In contrast to figure 2.3, the SbfI and XhoI site frequencies in these two figures are expressed in relation to the individual's read count, thus removing the effect of inter-individual variation in read count on the site frequency distributions.

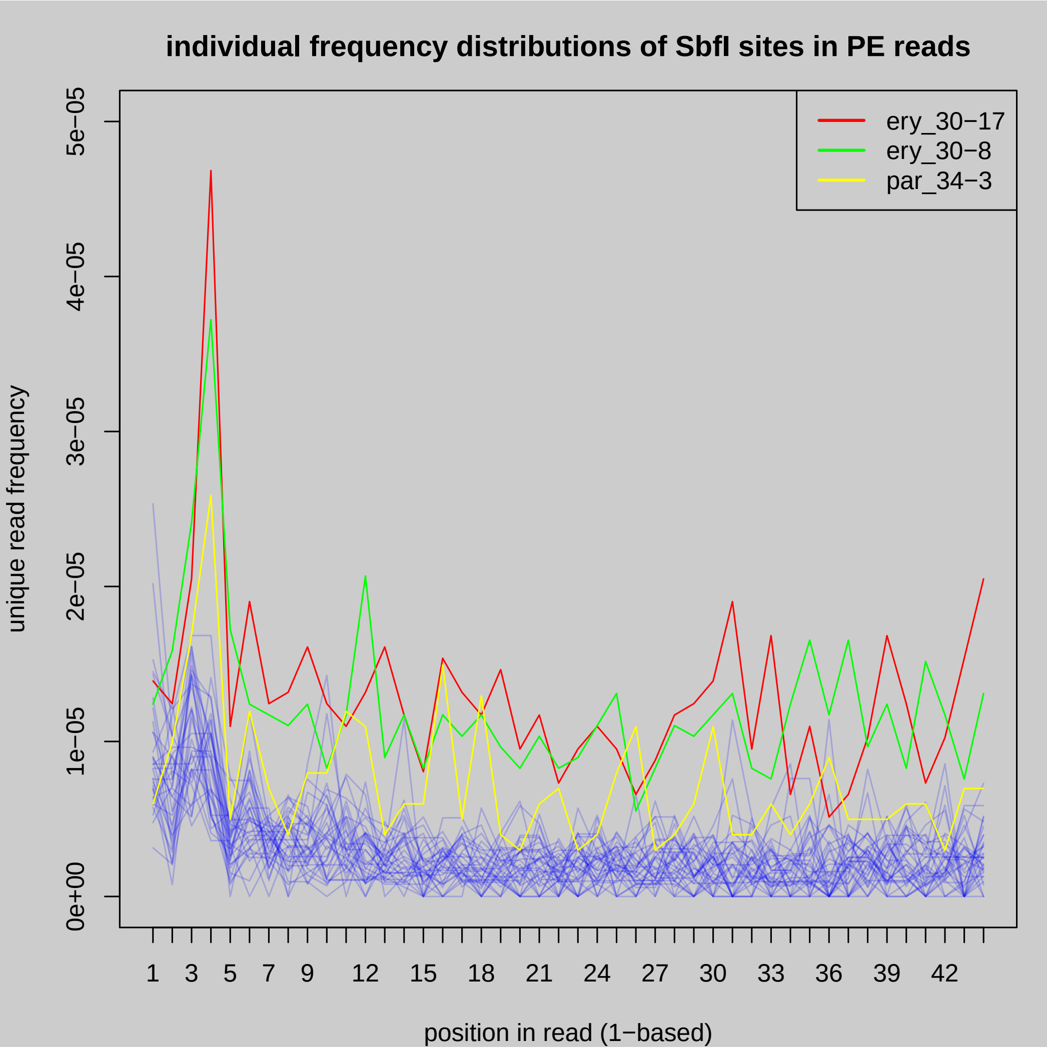
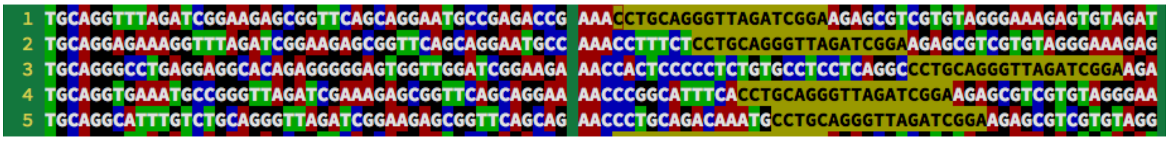


Fig. 2.22 SbfI site frequency distributions in uniqued PE reads for each individual relative to its read count (after quality filtering with `process_radtags`). The three individuals with barcodes ending in `CC` are highlighted with different colours.



```
1 TGCAGGTTTAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG AAACCTGCAGGGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGAT
2 TGCAGGAGAAAGGTTTAGATCGGAAGAGCGGTTCAGCAGGAATGCC AAACCTTTCTCCTGCAGGGTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAG
3 TGCAGGGCCTGAGGAGGCACAGAGGGGGAGTGGTTGGATCGGAAGA AACCACTCCCCCTCTGTGCCTCCTCAGGCCCTGCAGGGTTAGATCGGAAGA
4 TGCAGGTGAAATGCCGGGTTAGATCGAAAGAGCGGTTCAGCAGGAA AACCCGGCATTTCACCTGCAGGGTTAGATCGGAAGAGCGTCGTGTAGGGAA
5 TGCAGGCATTTGTCTGCAGGGTTAGATCGGAAGAGCGGTTCAGCAG AACCCTGCAGACAAATGCCTGCAGGGTTAGATCGGAAGAGCGTCGTGTAGG
```

Fig. 2.23 Snapshot from the output of command 2.12. The highlighted subsequences contain the SbfI recognition sequence overlapping with the reverse complement of the barcode sequence `AACCC` and the beginning of the P1 adapter sequence. The subsequences left of the SbfI sequence in the PE reads are genomic. Note that these genomic sequences can also be found in the corresponding SE reads (after reverse-complementing) right after the remainder of the SbfI sites at the beginning of the SE reads.

The most striking pattern are the three outlier individuals for the PE reads containing SbfI sites in figure 2.10. The outlier individuals are "par_34-3", "ery_30-18" and "par_34-14". This high relative frequency is not due to low read counts.

sample ID	barcode
34-3	AACCC
34-14	GCGCC
30-18	TGACC

However, these three individuals are the only three whose barcode ends with CC. When looking at the output of the following command line, it becomes clear that the high frequencies of reads with SbfI sites from these three individuals is caused by reads into the P1 adapter, their barcode regenerating the SbfI recognition sequence.

```
awk '(NR-2)%4==0' par_34-3_cleaned_trueTags.fq_2 | \
grep CCTGCAGG | sort | uniq | cluster.pl > par_34-3_SbfI_PE.cl
```

The extreme similarity of the curves for these three individuals is also striking. This might be caused by common repetitive sequences containing two SbfI sites in close proximity.

Figure 2.11 shows the XhoI site frequency distribution over SE and PE reads. There are no outlier individuals but clear common peaks of XhoI frequency at certain positions in the reads. I clustered all XhoI site containing SE reads by the subsequence left (5') of their XhoI site with my script `cluster.pl` and command 2.13. Figure 2.24 shows for the SE reads that most of these peaks are the result of repetitive sequences since they coincide with large cluster sizes.

Command 2.13 This command is analogous to command 2.10. It clusters `uniq`-ued SE reads of all individuals that contain an XhoI site and then foreach XhoI site position in the SE read length prints out the cluster sizes that have been found. My custom script `cluster.pl` first groups reads by XhoI position and then clusters them within groups by mismatch count on the subsequence left of the XhoI site, thus ignoring the potentially non-homologous genomic sequence (due to religation) downstream of the XhoI site.

```
cat *cleaned_trueTags.fq_1 | awk '(NR-2)%4==0' | grep "CTCGAG" | \
sort | uniq | cluster.pl > all_ind_XhoI_in_SE_cl_by_pos.csv &
```

Cluster Analysis

For the cluster analysis of genomic religation versus incomplete digestion, I have first collected all read pairs containing a restriction site from each individual with my script `grep_fq_read_pairs.pl`. I have then used command 2.14 with my script `cluster.pl` in order to cluster these reads by similarity of the subsequence left (5') of the restriction site. I

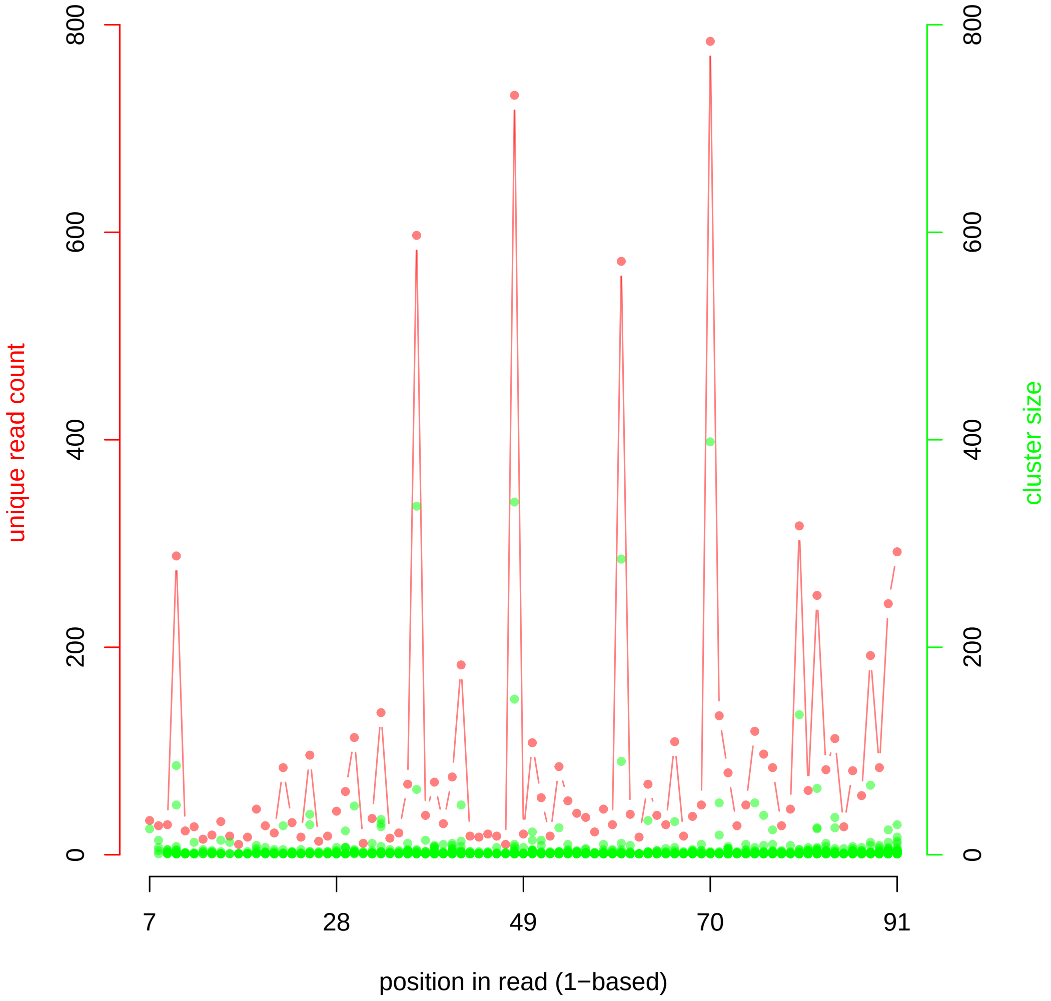


Fig. 2.24 XhoI site frequency distribution for all uniqued SE reads together (red) and the corresponding sizes of clusters in which these reads find themselves (green).

have used command 2.13 to cluster SE reads with SbfI or XhoI sites from the double-digest RAD library. In addition to cluster sizes, the script `cluster.pl` can also be set to print out the clusters themselves. The output file with clusters is then opened in the text editor VIM and the DNA sequences are highlighted with colours by a VIM plugin of Johan Nylander.

Command 2.14 This command first pastes read pairs that contain SbfI sites sites side by side (SE left, PE right), then removes exact duplicate lines before clustering read pairs by the subsequence left of the SbfI site.

```
paste <(awk '(NR-2)%4==0' *SbfI_reads.fq_1) <(awk '(NR-2)%4==0' *SbfI_reads.fq_2) | \
grep "CCTGCAGG" | sort | uniq | cluster.pl > all_uniq_SbfI_reads_clustered.cl
```

Read pair mapping analysis

QUASI-RANDOM READ PAIRS With command 2.15 I have taken the 200,001st to 300,000th read pair from each individual in the standard RAD data set. For the SbfI+XhoI double-digest RAD data set I have used a very similar command, but extracted only the SE reads. All reads had passed my quality filtering steps.

Command 2.15 This command line extracts the third set of 100,000 FASTQ records from each read file.

```
for i in ../*fq_[12].gz; do (head -n 1200000 <(zcat $i) | \
tail -n 400000 > `basename $i .gz`.random)& done
```

BOWTIE2 For the standard RAD library, I ran `bowtie2` with a maximum allowed fragment/insert size for a "valid paired alignment"[16] (`-X`) of 900. The upper end of gel size selection was at around 800 base pair fragment length (including Illumina adapters). For the read pair data from digital digestion, I have set the maximum fragment/insert size to 120. The length of the original SbfI+XhoI SE reads was 96 bps. For both types of read pairs I ran `bowtie2` in `-very-sensitive-local` mode.

COUNTING CONCORDANT READ PAIRS I then extracted all reads from read pairs where both reads mapped, concordantly or not and disregarding mapping quality, sorted on names (both done as in command 2.16) and concatenated the individual sorted BAM files into one with `samtools cat`. I then used command 2.17 in order to count the total number of read pairs. Although, `bowtie2` reports a number of concordantly mapped read pairs in its mapping summary report, I have found that the proper pair bit in the SAM flag of each read record is not always set correctly. This seems to be linked to the inference of fragment/insert size, i.e. the length of the genomic fragment from which the SE read and the PE read is

Command 2.16 This command line extracts from each individual read mapping output file (in SAM format) those read pairs where both reads mapped (`-F12`) and then sorts these SAM records by read name, which is necessary for concatenating the individual files into one with `samtools cat`.

```
parallel -j 5 'samtools view -huF12 {} |
samtools sort -n -o {.}_F12_sort.bam -T {.} -' ::: *sam &
```

[16] presumably synonymous to proper pair

Command 2.17 This command line counts the number of SE reads (first in pair) in a mapping output file (binary SAM format).

```
samtools view -f64 all_random_F12_sort.bam | wc -l
```

produced. I have therefore applied my own filtering to count proper pairs (command 2.18). My script `Isize.pl` calculates the insert/fragment size, also known as *template length*, from

Command 2.18 This command line counts the number of genuinely concordant read pairs in a mapping output file by applying a sequence of filters. The first line extracts all reverse mapping reads (`-f16`) whose mate did not map as reverse-complement (`-F32`) and which have the proper pair SAM flag bit set (`-f2`). The second line makes sure that both reads in the pair got mapped to the same reference contig. The third line makes sure that the reverse mapping read has a higher mapping position than its forward mapping mate (no dovetailing). The fourth line uses my custom script `Isize.pl` in order to calculate the fragment/insert size and the fifth line makes sure that the insert size is within the bounds of 50 and 900.

```
1 samtools view -f2 -f16 -F32 all_random_F12_sort.bam | \
2 awk '$7 ~ "="' | \
3 awk '$4 > $8' | \
4 Isize.pl | \
5 awk '$10 < 900 && $10 > 50' | \
6 wc -l
```

the reported mapping positions of the read pair and the CIGAR string of the reverse mapping read in the pair. It does this by adding up the matching bases reported in the CIGAR string of the reverse mapping read in order to add this to its reported mapping position[17] and then subtracts from the result the reported mapping position of its forward mapping mate. Thus soft-clipped bases are not included in the calculated insert size.

> A mapped base is a base in the read that corresponds one-to-one to a base in the reference. So soft-clipped, hard-clipped, inserted and deleted bases are not mapped bases in that sense.
>
> Tim Fennell on the `samtools` mailing list on SourceForge

> If all segments [reads] are mapped to the same reference, the unsigned [absolute] observed template length equals the number of bases from the leftmost mapped base to the rightmost mapped base.
>
> Li and Durbin (2011)

I have also searched among the read pairs, which did not get the proper pair bit set by `bowtie2`, for read pairs which still fulfill the criteria of proper pairs (command 2.19). For the read pairs from digital digestion I used the same three commands (2.17 till 2.19) as for the standard SbfI RAD read pairs, except for applying an accepted insert size range of 30 to 120.

[17] this results in the reference coordinate of the position right of the rightmost mapping base with respect to the reference sequence

Command 2.19 This command line applies the same filters as command 2.18 to read pairs which did not get the proper pair SAM flag bit set (-F2).

```
samtools view -F2 -f16 -F32 all_mapped_read_pairs_F12.bam | \
awk '$7 ~ "="' | \
awk '$4 > $8' | \
Isize.pl | \
awk '$10 < 900 && $10 > 50' | \
wc -l
```

Chapter 3

Investigation into the demographic history of the hybrid zone

> There's no such thing as a data analysis pipeline.
>
> John McCutcheon

3.1 Introduction

Traditionally, population genetic analyses relied on a genotype calling step. However, with low to medium coverage ($< 10\times$) next generation sequencing data there can be substantial uncertainty in genotype inferences. The major sources of error are assembly and mapping errors (Li, 2011b) as well as base call errors and the random sampling of allele copies at hererozygous loci. If this uncertainty is ignored during downstream analyses by using the most likely genotype given the sequencing read data at a locus from an individual – and usually some hard filtering of putatively variant sites based on quality scores or likelihood ratio test (LRT) p-values – this can lead to errors or biases in population genetic inferences (Crawford and Lazzaro, 2012; Han et al., 2014; Johnson and Slatkin, 2008). Methods that avoid genotype calling by incorporating genotype uncertainty in downstream analyses can largely avoid these biases (Li, 2011b; Nielsen et al., 2012). In addition, by avoiding SNP calling and instead incorporating uncertainty in polymorphic sites, these methods can also avoid ascertainment bias in downstream analyses (Albrechtsen et al., 2010).

The following study is an attempt to infer some details of the demographic history of the two subspecies *erythropus* and *parallelus* from patterns of genetic variation observed in next-generation sequencing data from a standard RAD library. In this chapter, *population* *size* always stands for *effective population size*, not census population size, unless stated otherwise.

3.2 Materials & Methods

3.2.1 Sampling and library preparation

Individuals from each of the two distal populations of a transect through the eastern part of the Pyrenees Mountains (Col de la Quillane, fig. 2.1) were sampled by Jamie Hutchison and Roger Butlin in 2008 and preserved in ethanol until DNA extraction in 2009. A RAD library was prepared according to the protocol of Baird et al. (2008) (see Appendix section 5.1 on page 152 for details). It was then sequenced on 4 lanes of an Illumina GAIIx with 51 base pair paired-end sequencing which yielded a total number of 52,872,783 read pairs.

3.2.2 Read filtering

Raw sequencing reads were split by individual barcodes and filtered for base call quality with `process_radtags` from `Stacks` (Catchen et al., 2011). Since all used barcode sequences differed from each other by at least 3 mismatches, SE reads were allowed to start with a 5 base pair sequence that differed from one of the barcodes used by 1 mismatch. Read pairs were discarded if they did not contain the remainder of the SbfI restriction site with at most one mismatch following the barcode in the SE read. Base call quality filtering was done with a sliding window across the reads. A read pair was discarded if the average Phred-scaled base call quality in a window dropped below 20. This read pair filtering discarded 8,505,427 raw read pairs (16%) leaving 44,367,356 read pairs for the analysis. After removal of the 5 bp barcode sequence, the SE reads were 46 bp long.

3.2.3 De novo assembly

In order to create a reference sequence for the RAD tags in this library, I used an assembly strategy very similar to the one implemented in `dDocent` (Puritz et al., 2014a) (see figure 3.1 for an overview). I used only non-redundant sequence reads for de novo assembly of a "RADome". I used `starcode` (commit 1034408ca6) (Zorita et al., 2015) to collapse *read pairs* from each individual separately into unique representatives allowing for an edit distance of up to 2. These unique representatives are the canonical sequences from all connected components of the graph from an all pairs search. Canonical sequences are chosen by highest read count, then by highest number of connections with other reads. This reduced the number of quality filtered read pairs for de novo assembly to 9,179,521 (21% of the original read pair number). Due to the shotgun-type nature of PE reads, read pairs should only be identical (up to 2 mismatches) if they come from the same fragment from random shearing, i.e. if they are clonal reads from PCR.

After combining all SE sequences from these unique sequence pairs from all individuals into one fasta file, I have used `Vsearch` (commit 1116d6167b) (Rognes et al., 2016) with its subcommand `cluster_fast` for heuristic clustering with a pairwise identity threshold of 0.8 of a query sequence with the centroid of a growing cluster. Identity is defined as 1 minus edit distance as a proportion of alignment length, but excluding terminal gaps. Thus a indel between two reads will only contribute once towards their edit distance.

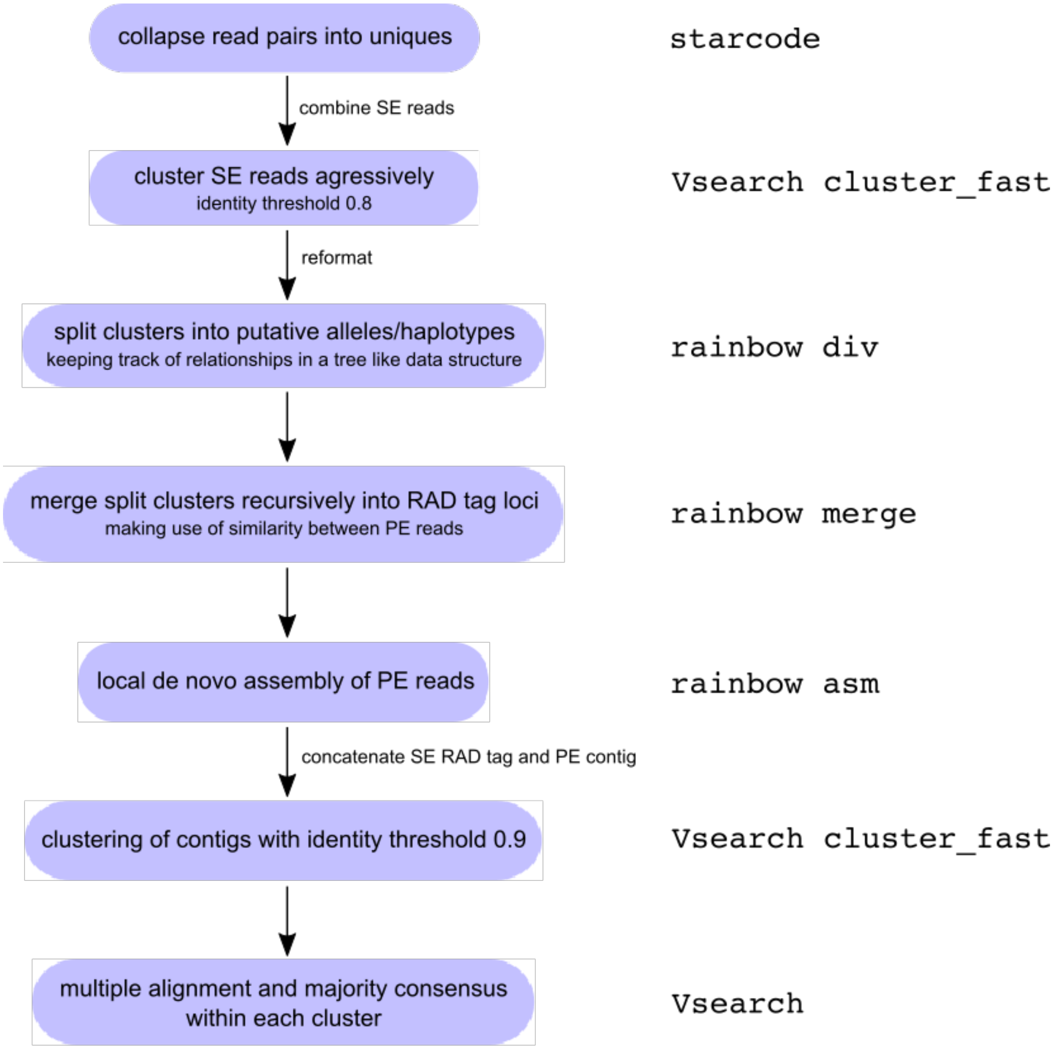

collapse read pairs into uniques
starcode
combine SE reads
cluster SE reads agressively
identity threshold 0.8
Vsearch cluster_fast
reformat
split clusters into putative alleles/haplotypes
keeping track of relationships in a tree like data structure
rainbow div
merge split clusters recursively into RAD tag loci
making use of similarity between PE reads
rainbow merge
local de novo assembly of PE reads
rainbow asm
concatenate SE RAD tag and PE contig
clustering of contigs with identity threshold 0.9
Vsearch cluster_fast
multiple alignment and majority consensus
within each cluster
Vsearch

Fig. 3.1 Overview of the de novo assembly strategy.

Next, I have used the uclust format output of Vsearch together with the collapsed read pairs from starcode to create a clustering output in the format of the output from rainbow cluster (Chong et al., 2012). I then ran rainbow div on this cluster file. This programme recursively splits the initial clusters into putative alleles/haplotypes while keeping track of the relationships between split clusters in a tree–like data structure. These split clusters are next merged back recursively along this tree structure with rainbow merge that uses the similarity between PE reads to determine whether two split clusters should be merged. Remember that the standard RAD protocol (Baird et al., 2008) includes random shearing of restriction fragments. This produces PE reads of variable distances from the same RAD tag (Etter et al., 2011). The unique feature of rainbow is to be able to utilise these shotgun–type PE reads to distinguish alleles from paralogs. In addition, after recursively merging split clusters into putative RAD loci, rainbow merge also performs a local de novo assembly with the sequences of each merged cluster using SE and PE reads. From the assembly output I then extracted the SE RAD tag sequences together with their longest PE read contig (separating each with a sequence of 10 N's) using gawk code from dDocent.

Finally, since the merge process of rainbow depends on overlap between PE reads from the same RAD locus, it could be incomplete when coverage is low. I therefore clustered the rainbow contigs again with the Vsearch subprogramme cluster_fast, this time with an identity threshold of 0.9 and producing a majority consensus sequence from a multiple alignment of the sequences in each cluster of rainbow contigs. The assembled RADome contains 583,312 contigs and a total length of 97 mega base pairs (Mbp). In the following I will frequently refer to it under the informal name of *Big Data* reference assembly.

An extensive documentation about the assembly procedure as well further downstream analyses can be found in the BASH script assembly.sh. This text file contains exact command lines used, together with extensive explanation. It should be the first stop for information when trying to reproduce the results presented here. Note, not all steps of the analysis could be documented directly in this text file, but when Rmarkdown or IPython notebooks (Pérez and Granger, 2007) have been used for analysis, then references to those files can be found in assembly.sh.

3.2.4 Read mapping

I mapped all quality filtered reads against the newly created RADome (or Big Data reference) with the programme bowtie2 version 2.2.9 (downloaded on 29th October 2016) (Langmead and Salzberg, 2012). I ran bowtie2 in end–to–end mode, i.e. without soft–clipping of query sequences. Further, I specified "very–sensitive" search mode with a seed length of 20. I allowed gaps to be up to 10 bp long, but I disallowed gaps within 4 bp of either end of the read. In addition I specified a high penalty of 10 for alignments against ambiguous characters in the reference. In most reference contigs, SE RAD tag and PE contig were separated by 10 N's (unless `rainbow asm` assembled SE and PE reads into one contig). This high penalty therefore makes successful read alignments across the gap between SE RAD tag and PE read contig very unlikely. Further, I specified a minimum fragment length of 60 for proper pair read alignments, thus allowing some overlap between reads in a pair.

Although standard RAD data from paired-end sequencing allows the detection and removal of PCR duplicates, I have opted not to do so, since `ANGSD` could not estimate a site frequency spectrum from a de-duplicated data set (see section 3.4.4 on page 110 in the supplementary material). The following analyses all depend on the correct estimation of genotype likelihoods which assumes that reads are sampled independently from the true genotype. Due to the PCR duplicates, I expect the genotype likelihoods to be generally biased against heterozygote genotypes. Andrews et al. (2016) have claimed that "PCR should not systematically favour one allele over another at a given locus, and therefore parameters estimated from a large number of loci are unlikely to be substantially biased". However, PCR drift, i.e. the random amplification of one allele at a locus at the expense of the other, as well as allele dropout can bias the site frequency spectrum towards higher frequency counts. For instance, if there were three true heterozygotes at a locus and all other individuals homozygous for the reference allele, then the count class 3 should be incremented in the SFS. With three false homozygote calls at this locus, the count class 6 will be incremented instead.

3.2.5 Filtering of the de novo reference assembly

There are 4,575 contigs with SE reads that map to the PE part of the contig, i.e. which do not pass the *mismapping* filter. I have also filtered contigs for excessive coverage by SE reads, which map to the RAD tag part of each contig (i.e. next to the restriction site). The coverage by PE reads is much more dependent on the length of the contig. With this filter a contig has excessive coverage if in any of the individual BAM files it has SE read coverage above the 99th percentile of that individual's coverage distribution. I have thus excluded 2,282 contigs due to excessive coverage. Since many false–positive SNPs fall within low–complexity sequences (Li, 2014), I have detected those regions within the *Big Data* reference assembly with `dustmasker` (Morgulis et al., 2006) and excluded them from further analysis. The above filters have excluded 10% of the reference from further analysis, leaving 88 Mbp.

I then went on to extract those bits of the filtered reference that have sufficient data for downstream analyses. I used the `samtools depth` command to get the read count for all positions and for all 36 individuals across the filtered reference. I then extracted those positions where at least 15 individuals have each at least 3× coverage. I only counted reads with a mapping quality greater than 5. This reduces the number of sites to 2% of the original *Big Data* reference assembly or 2 Mbp, spread over 34,967 contigs. The first 6 base pairs of each contig are the remainder of the SbfI restriction site (`TGCAGG`) and therefore not variable. I have excluded those sites as well from further analysis.

In addition to the excessive coverage filter, I have also applied a HWE filter to the remaining sites in order to reduce false heterozygotes caused by reads from paralogous sequences mapping to the same position in the reference. I have used the programme `ANGSD` (version 0.915-5, `git` commit `ge6e63e5`, Nov. 2016) (Korneliussen et al., 2014) to estimate per site inbreeding coefficients (F_{site}) using genotype likelihoods (Vieira et al., 2013). I removed entire contigs that had a site with a negative estimate of inbreeding coefficient and a p-value for deviation of F_{site} from 0 of less than 0.05 from an LRT. I estimated inbreeding coefficients for each population separately as well as for both populations taken together. The latter increased sample size for the estimation of F_{site} but also increased estimates of F_{site} for SNPs with different allele frequencies in the two populations. This filter removed 217 contigs (for further details see section 3.4.3 on page 109 in the supplementary material). This has left 1,799,962 filtered sites on 34,750 contigs for further analysis.

Inspection of the global, i.e. across sample, coverage distribution revealed that there were 12,693 sites with total coverage greater than 1000×. I have therefore also determined the 99th percentile of the global *per-site*[1] coverage distribution and removed all contigs with a position that had coverage greater than the global 99th percentile. This has removed 69,438 sites on 407 contigs leaving 1,730,524 sites on 34,343 contigs for further analysis. The new across-sample (global) coverage distribution is shown in figure 3.2.

The average per site, per individual coverage across the 1.7M filtered sites is 6.3× (fig. 3.3).

The above filters have selected for contigs with more unique sequences as shown by the great reduction in the proportion of reads with a mapping quality score of 1 that map to the filtered reference as compared to the unfiltered reference sequence (see fig. 3.4). Only reads with a mapping quality score of ≥ 5 have been used for downstream analyses. A Phred score of 5 for the mapping quality should indicate a probability of about 1/3 that the read truly originated elsewhere in the *reference sequence* (mapping uncertainty does not incorporate assembly uncertainty). However, as figure 3.20 on page 108 shows, `bowtie2` generally greatly underestimates the true mapping quality, given the reference sequence, particularly for very low mapping quality scores. A more stringent filtering of read mappings is therefore not necessary. Note that in *all* following analyses mapping quality information is incorporated into the calculation of genotype likelihoods by capping base quality by the mapping quality

[1] including SE and PE reads and across the whole contig, not just RAD tag part

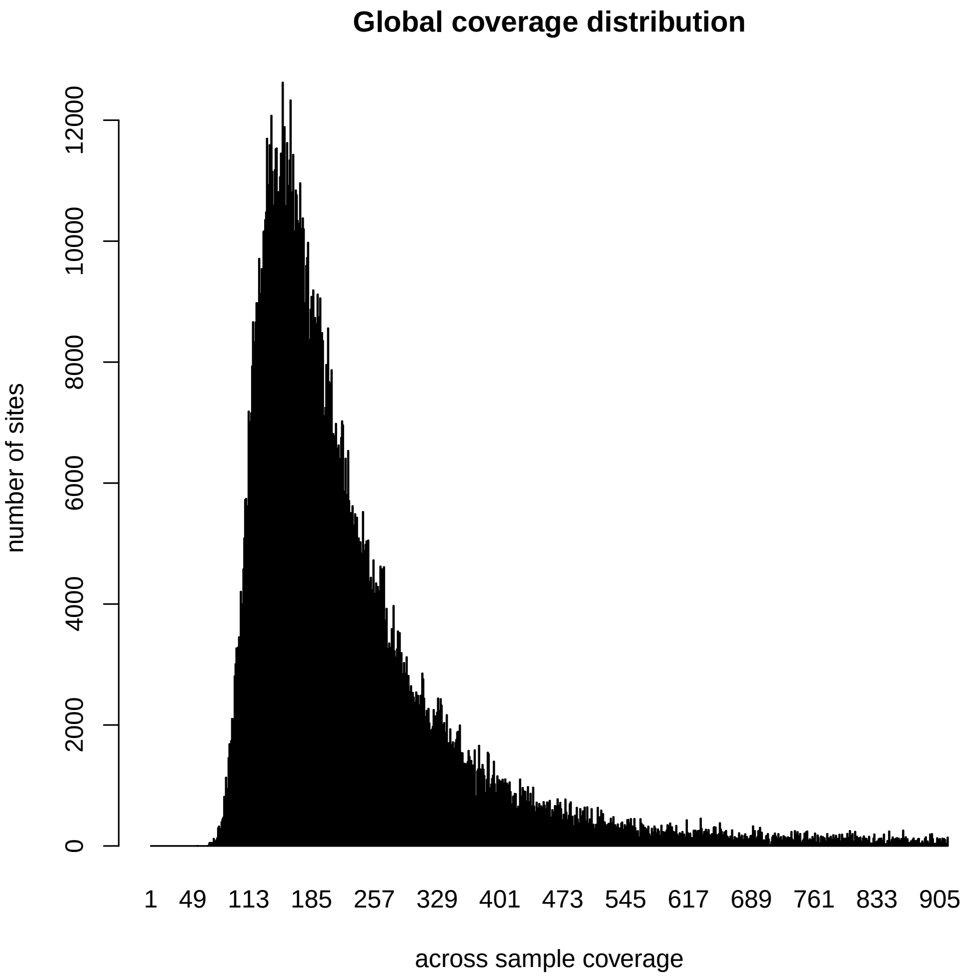


Fig. 3.2 Histogram of across sample coverage with reads of mapping quality greater than 5 from 1.7 million nucleotide sites in the *Big Data* reference after filtering.

of the read[2]. A different mapping programme will therefore result in different genotype likelihoods as mapping quality scores are computed with different heuristics by the different mapping programmes (see fig. 3.20a on page 108).

A better measure of the repetitiveness of the assembled RADome than the distribution of read mapping quality is the distribution of mappability. Mappability measures the uniqueness of a sequence of length k (kmer) sampled from a position x in the reference sequence (here Big Data reference assembly). Mappability is the inverse of how often that kmer can be found in the whole reference sequence up to a specified number of mismatches (or edits) (Derrien

[2] This claim has not been fully verified but is very likely: I have used the genotype likelihood version from `samtools` throughout (and `ANGSD` directly incorporates code from `samtools`), which should be based on `MAQ` (Li et al., 2008, section Methods: Consensus genotype calling).

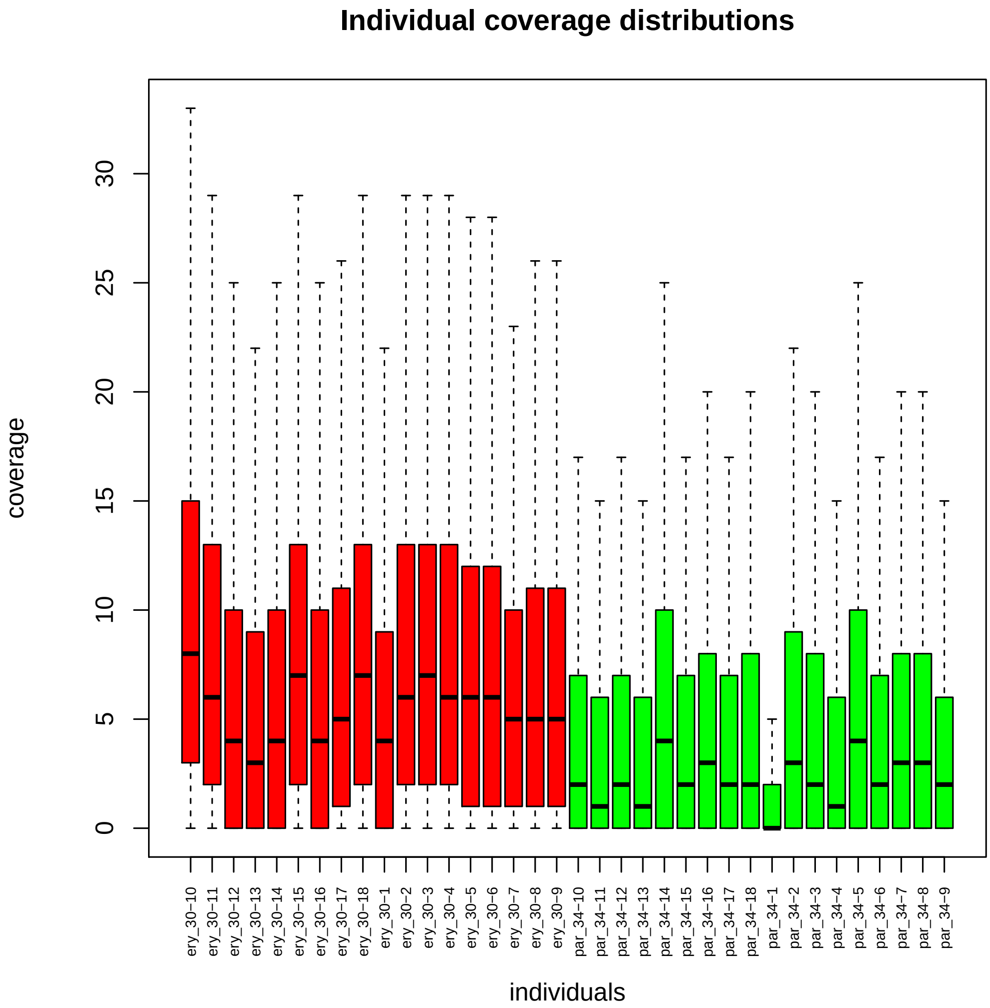


Fig. 3.3 Individual coverage distributions from reads with mapping quality greater than 5 over 1.7 million sites in the *Big Data* reference after filtering. red: *erythropus*, green: *parallelus*.

et al., 2012). Mappability has a range between 0 and 1. A position has mappability of 1 if its kmer occurs only once in the reference, 0.5 if it occurs twice and so on. Figure 3.5 shows the distribution of per contig average mappability of the raw Big Data reference assembly and of the reference assembly after applying above filters. Note that the mappability scores for the filtered reference sequence are just a subset of the mappability scores for the unfiltered reference, i.e. all mappability scores measure the uniques of kmers in the Big Data reference assembly.

Figure 3.5 shows that a large proportion of contigs have an average mappability of 1 or close to 1 and that the majority of contigs have average mappability greater than 0.5. It also shows that the above filters have *not* selected for more unique contigs in the Big Data reference assembly. In fact, the reference contigs kept after filtering for further analysis have

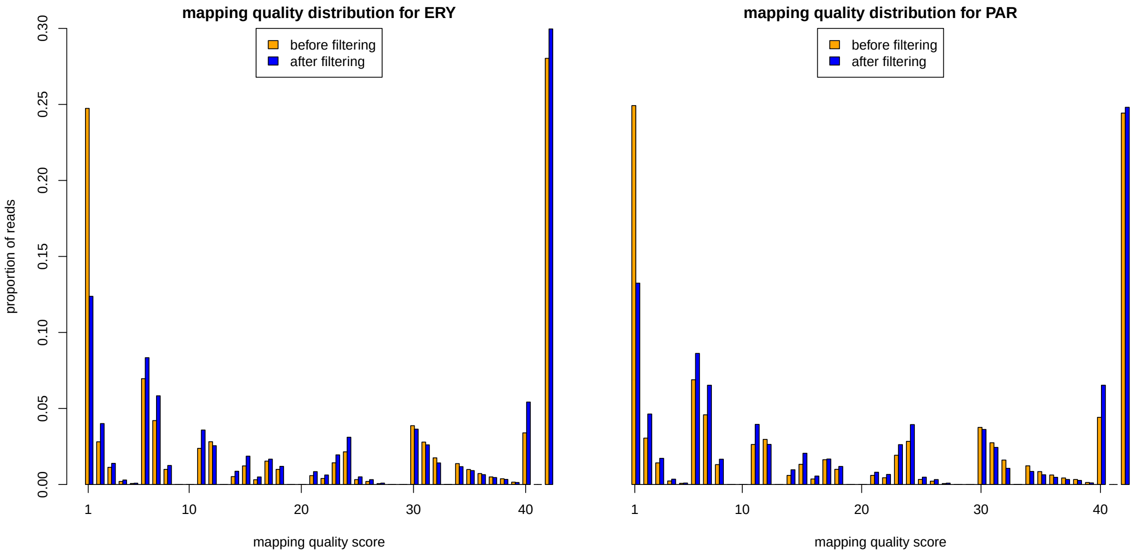


Fig. 3.4 Distribution of mapping quality scores determined by `bowtie2` for all mapped reads to the unfiltered and filtered reference sequence.

lower median mappability than before filtering. The median mappability before filtering is 0.787 whereas after filtering it is 0.737. The median of the difference in mappability between a contig before filtering and a contig after filtering is 0.053 (95% confidence interval (CI): 0.05, 0.055). Regions in the reference with lower mappability are more difficult to map reads to with data from short read next-generation sequencing. Quantitative measures of read counts as well as polymorphism detection in those regions will be compromised (Derrien et al., 2012; Lee and Schatz, 2012). However, a lower mappability is not indicative of a bad de novo assembly of the reference. In fact, collapsing repetitive sequences from the sampled genome into one reference contig will *increase* the average mappability of the assembled reference.

The above filters have also selected for contigs with lengths that are roughly consistent with the expected size distribution of sonicated restriction fragments that were size selected on an agarose gel during library preparation (see figures 3.6 and fig. 5.2 on page 156). Note that the size selected DNA fragments contained the 67 bp long P1 adapter and that PCR as well as cluster generation on the flow cell should increase the proportion of short fragments in the resulting illumina sequence data.

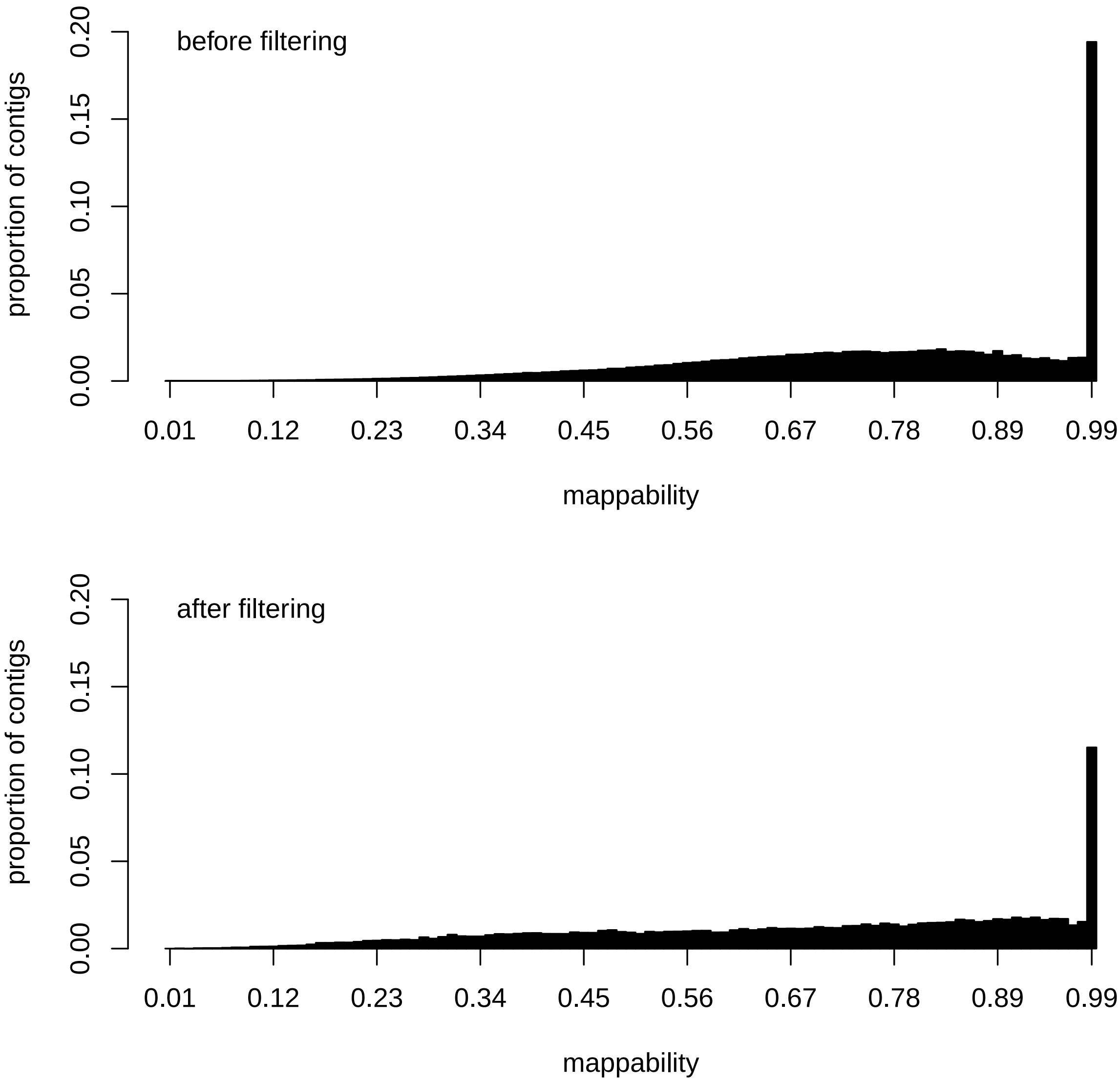
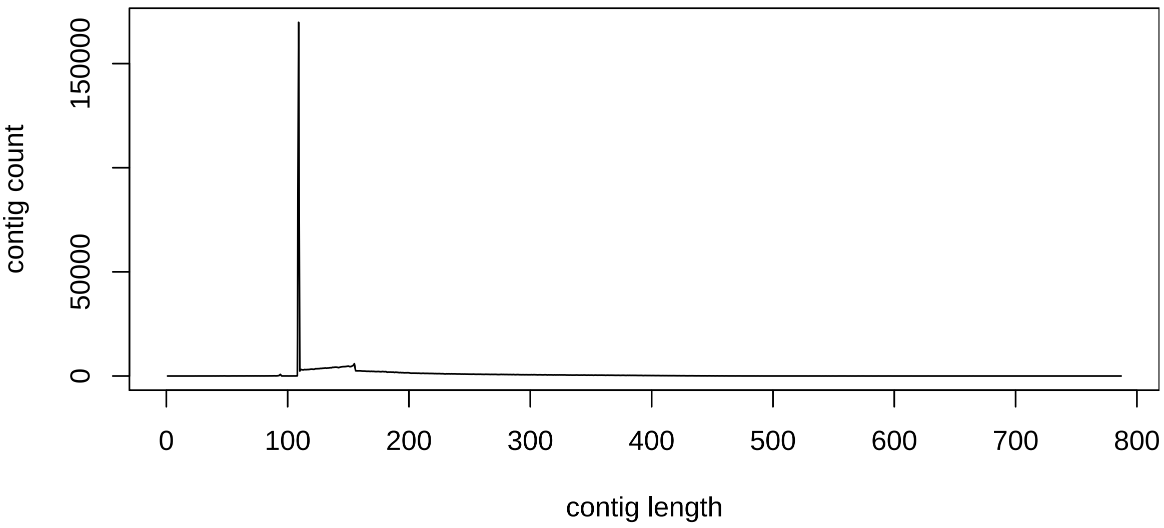


Fig. 3.5 Distribution of average mappability per contig before and after filtering the Big Data reference assembly. For details see section 3.4.2 on page 107 in the supplementary material.



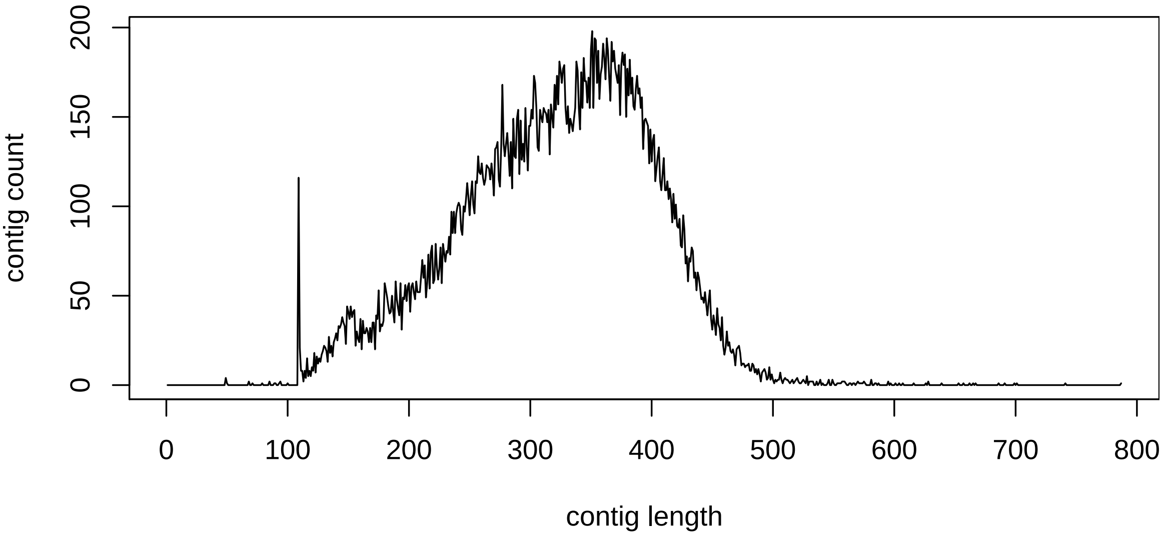


Fig. 3.6 Distribution of contig lengths. Top: raw RAD assembly (*Big Data* reference) before filtering (583,312 contigs). Bottom: RAD assembly after applying filters (34,343 contigs).

3.2.6 PCA

I have used ANGSD (version 0.915-5, git commit ge6e63e5, Nov. 2016) (Korneliussen et al., 2014) to estimate posterior genotype probabilities across all sites using as prior a maximum likelihood (ML) estimate of the population minor allele frequency (MAF) per site across all individuals from the two populations and the assumption of HWE (Kim et al., 2011). Although the assumption of HWE may well be violated, it should only make genotype probability estimates more similar between divergent individuals. Therefore, any major divergence detected between individuals based on these genotype probabilities cannot be an artefact of this prior. At low coverage, the identification of the minor allele can be uncertain. The estimation of the minor allele frequency incorporated this uncertainty (Skotte et al., 2012, suppl. mat.).

Next, I estimated the ML global unfolded site frequency spectrum based on per–site sample allele frequency likelihoods (SAF), again across all individuals, using ANGSD's subprogramme realSFS (Korneliussen et al., 2013; Nielsen et al., 2012). This global site frequency spectrum then served as prior to estimate per–site sample allele frequency posterior probabilities. Thus for each site $2N+1$ probabilities were calculated (N=36, the total number of individuals sequenced).

I have then estimated a genotype covariance matrix with ngsCovar (Fumagalli et al., 2014) between all 36 individuals. At each site the posterior expectation of the genotype covariance is computed by summing over all 9 genotype combinations and weighting each combination by the respective two posterior genotype probabilities. Each cell of the matrix contains the genotype covariance between two individuals averaged over all sites and each site weighted by its probability of being variable using the sample allele frequency posterior probabilities (Fumagalli et al., 2013, eq. 19 and 20).

During the estimation of genotype and sample allele frequency probabilities I have applied *standard* (not extended) BAQ to cap the base quality by the calculated per-base alignment quality (Li, 2011a). This is in order to reduce false-positive variant detection caused by misalignment around indels.

I then performed an eigen–decomposition on the resulting covariance matrix with the R function prcomp.

3.2.7 F_{ST}

Using ANGSD (version 0.915-5, git commit ge6e63e5, Nov. 2016) and its subprogramme realSFS (Korneliussen et al., 2014) I have determined a ML estimate of the global *unfolded* 2D–site frequency spectrum that contains the joint sample allele frequencies of the non–

reference allele from *erythropus* and *parallelus* (figure 3.7). I applied standard BAQ (Li, 2011a) and required a nucleotide site to have sequence reads from at least 9 individuals in each population. The 2D-SFS contains the joint frequencies from approximately 1,130,775 sites of which 60,573 (5%) are variable in *erythropus*, *parallelus* or both. This global unfolded 2D–site frequency spectrum was then used as (empirical) prior for posterior probabilities of all possible sample allele frequencies from which the posterior expectation of F_{ST} per site is calculated using `realSFS` from ANGSD (Fumagalli et al., 2013, eq. 3 and 16). This programme reports the numerator and denominator for either Reynolds' F_{ST} (Fumagalli et al., 2013, eq. 1–3) or Hudson/Bhatia's F_{ST} (Bhatia et al., 2013, eq. 9 and 10) for each site. I am not reporting estimates of F_{ST} per site due to their large variance (Weir et al., 2005). Instead I am reporting the global average F_{ST} across all sites. This average is always calculated by summing the numerator and denominator across sites and then taking the ratio (Bhatia et al., 2013). Note, that all estimates of F_{ST} calculated here assume within-population HWE.

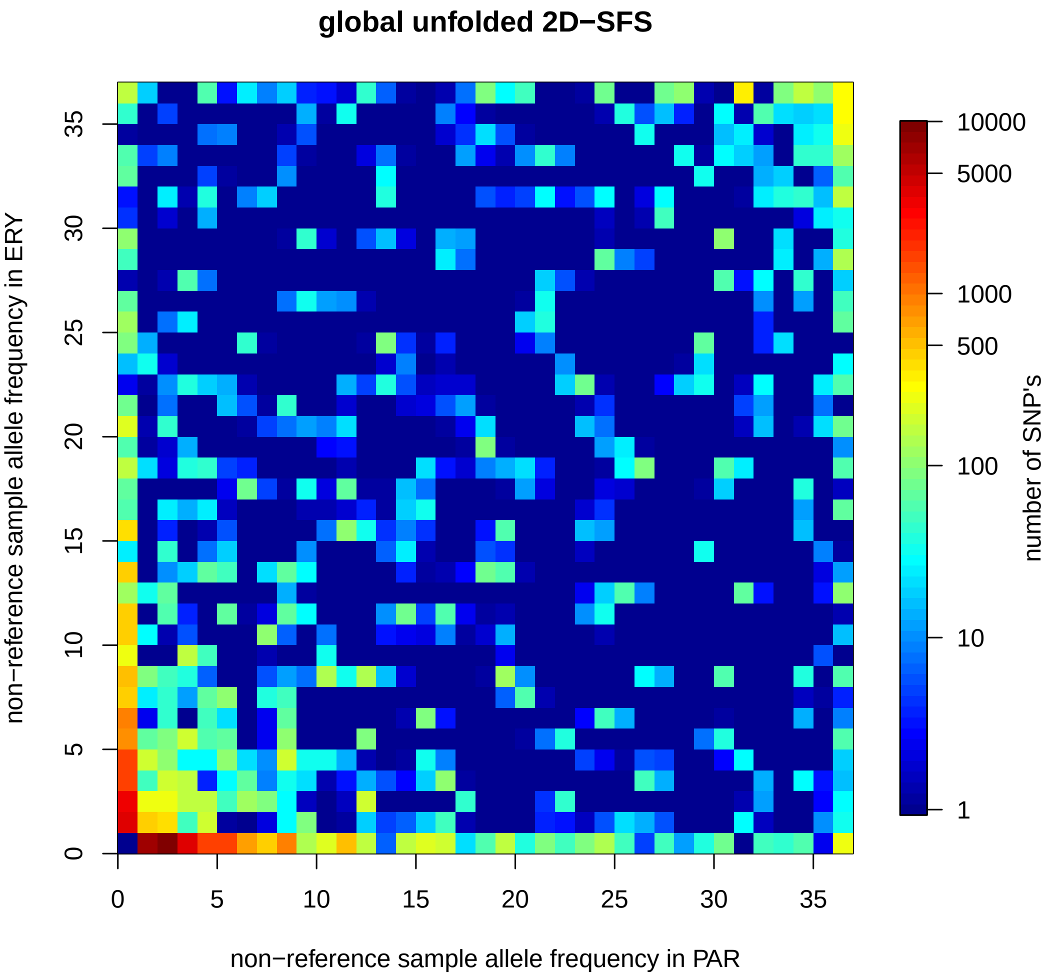


Fig. 3.7 ML estimate of the joint sample allele frequencies of the non–reference allele for *parallelus* (PAR) and *erythropus* (ERY). The count for the joint frequency class (0,0) has been set to zero for the purpose of better visualisation.

In order to investigate the dependence of F_{ST} on SNP ascertainment (i.e. simulated SNP discovery) in one or the other population, I have calculated ML estimates of population minor allele frequency (MAF) with ANGSD (Kim et al., 2011) for each population. I have then used these allele frequencies to subset sites into minor allele frequency classes and estimated average F_{ST} for each class as before. For further details of the analysis please consult `assembly.sh` and the notebook Fst.

I have also determined the distribution of F_{ST} over minor allele frequency (MAF) in either *erythropus* or *parallelus* with the function `Fst` of $\delta a\delta i$ (Gutenkunst et al., 2009) on the observed 2D SFS as shown in figure 3.7. This calculates Weir & Cockerham's F_{ST} (Weir and Cockerham, 1984, eq. for $\hat{\theta}$ at top of p. 1363). I then simulated site frequency spectra with $\delta a \delta i$ for four different demographic models and generated expected distributions of F_{ST} over MAF for these demographic scenarios. Two models included a recent bottleneck for one of the two populations, the other two included a recent population size expansion. Further details can be found in the IPython notebook 05_2D_models, section "Fst".

I have also estimated the net sequence divergence D_a between *erythropus* and *parallelus* (Nei and Kumar, 2000, see equation 12.67 on p. 256):

$$D_a = K_B - K_S \tag{3.1}$$

where K_S is the average of within-population genome-wide sequence divergences:

$$K_S = \frac{n}{n+m} \times \pi_{ery} + \frac{m}{n+m} \times \pi_{par} \tag{3.2}$$

n and m are the sample sizes (36 alleles) of *erythropus* and *parallelus*, respectively. π_{ery} and π_{par} are the average number of pairwise differences within *erythropus* and *parallelus* as calculated from the 1D site frequency spectra of each population with equation 3.7 (see section Genetic diversity estimates). K_B is:

$$K_B = \frac{1}{nm}\left[\sum_{i=0}^{m}\sum_{j=0}^{n}[i(n-j)+j(m-i)]\xi_{ij}\right] \tag{3.3}$$

where ξ_{ij} is the count in the unfolded joint allele frequency class [i, j] in the 2D-SFS (fig. 3.7). The allele frequency classes in the 2D-SFS refer to the non-reference allele. The joint reference allele frequency is given by [m-i, n-j]. K_B should be equivalent to d_{xy} as given in eq. 12.66 on p. 256 in Nei and Kumar (2000). The expectation of D_a is related to divergence time T (assuming divergence without gene flow) as follows (Nei and Kumar, 2000, eq. 12.69 an p. 256):

$$E[D_a] = 2\mu T \tag{3.4}$$

where μ is the nucleotide mutation rate. I have assumed a nucleotide mutation rate of 3×10^{-9} throughout this chapter (Liu et al., 2017).

3.2.8 Genetic diversity estimates

I have calculated ML estimates of the *unfolded* global site frequency spectrum, i.e. including all sites across contigs, for *erythropus* and *parallelus* separately with ANGSD and its subprogramme realSFS (Korneliussen et al., 2014). I used a different version of ANGSD – 0.917-142, git commit ge3dbeaa, 20 June 2017 – than for the previous analyses described above. The main difference is that this version applies *extended* instead of standard BAQ (for more details see this ANGSD issue thread). Extended BAQ is a more sensitive and less conservative algorithm than standard BAQ allowing for the detection of more polymorphic sites at the cost of slightly increased number of false positives. I folded these single-population (1D) site frequency spectra with my Python script `fold_1D_spectrum.py`.

I required at least 9 individuals to have read data in order to include a site into the estimation. For so-called *overlapping sites* I required sequence read data from at least 9 individuals in *each* population. So-called *non-overlapping sites* have sequence read data from at least 9 individuals in only one of the two populations.

Since I noticed a considerable amount of variability in the ML estimates of SFS from repeated runs of the programme `realSFS`, I turned off its accelerated Expectation Maximisation (EM) algorithm, extended the maximum number of EM iterations to 50,000 and specified a *tolerance* of 1e-06, which is the minimum difference in likelihood between two successive EM iterations and therefore a stopping criterium. These modifications completely removed the variability in SFS estimates, but also greatly extended running time, particularly for the site frequency spectrum of *parallelus*.

The programme `realSFS` allows bootstrapping the site frequency spectrum, i.e. re-estimation of the site frequency spectrum from nucleotide sites resampled with replacement. However, CI's estimated from these bootstraps will be too narrow, since it ignores that sites in the same contig are not independent and the SFS estimated is therefore a *composite* maximum likelihood estimate. Although, most of the contigs should not contain more than 1 SNP, I think it is best to resample over whole contigs instead of sites. Contigs, except for those coming from the same restriction site, should generally be unlinked. Bootstrapping over contigs is not implemented in `ANGSD/realSFS`. I have therefore created 200 bootstrap replicates of the contig id list (see bootstrap_contigs.ipynb) and specified them as so-called *regions* for the estimation of sample allele frequency files. I have then estimated site frequency spectra from these SAF files as before.

I fitted standard neutral coalescent model site frequency spectra to the observed folded single-population spectra by optimising θ in the following equation (Wakeley, 2009, eq. 4.21):

$$E[\eta_i] = \theta \frac{\frac{1}{i} + \frac{1}{n-i}}{1 + \delta_{i,n-i}} \qquad 1 \leq i \leq \lfloor n/2 \rfloor \tag{3.5}$$

This formula gives the equilibrium neutral expectation of counts (η) in each frequency class (i) in a folded spectrum. I used the `R` function `optimize` to find the value of θ that minimises the squared deviation of the above equation from all observed counts η_i.

I derived the global estimates of the number of segregating sites (S) and the average number of pairwise differences (π_{Tajima}) from the global folded site frequency spectrum according to equations 1.3 and 1.4 in Wakeley (2009), respectively:

$$S = \sum_{i=1}^{n/2} \eta_i \tag{3.6}$$

$$\pi = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n/2} i(n-i)\eta_i \tag{3.7}$$

where n is the number of alleles sampled, i.e. twice the number of sampled individuals for diploid loci. η_i is the count in the i–th frequency class of the folded SFS. The estimates of global Tajima's D are based on these estimates and equation 2.17 in Gillespie (2004):

$$D_T = \frac{\pi - \theta_W}{C} \tag{3.8}$$

C is a normalising constant that should make critical values of Tajima's D (-2 and 2 for the 5% significance level) independent of sample size (Wakeley, 2009, page 115). The formulas for its calculation can be looked up for instance in Gillespie 2004, page 45. θ_W is Watterson's θ and is calculated with:

$$\theta_W = {}^{S}\!/\sum_{i=1}^{n-1} \tfrac{1}{i} \tag{3.9}$$

Details of the above calculations are documented in the `Rmarkdown` notebook `new_1D_SFS`.

3.2.9 Inference of population demographic history

`stairway-plot`

I have used the programme `stairway-plot` version 2 beta (Liu and Fu, 2015) to infer the history of effective population sizes for *erythropus* and *parallelus*. It is a model-flexible approach, which means that it does not require the prior specification of a demographic model whose parameters shall be estimated. It can only be used on a single-population (1D) site frequency spectrum. With a user-specified mutation rate (μ) and generation time, it estimates the diploid effective population size (N_e) over time in the past by searching for the θ ($4N_e\mu$) values that maximise the composite likelihood of the given site frequency spectrum. The programme also infers the optimal number of time intervals (each with its own optimal value of θ) with different population sizes. With the number of chromosomes sampled being n, there can be from 1 up to $n-1$ time intervals with different population sizes. For the addition of another θ parameter it needs to improve the likelihood of the model significantly as determined by a LRT. The method has an intrinsically decreasing resolution going back in time. That is because the time intervals of the plot are proportional to expected coalescence time $E[T_i]$:

$$E[T_i] = \frac{4N_e}{i(i-1)}$$

with i being the number of lineages left in the genealogy and N_e being the diploid effective population size during the interval T_i between two coalescent events (Gillespie 2004, eq. 2.13).

I used a custom version of stairway-plot. I replaced the java class file Stairway_fold_training_testing5.class with Stairway_fold_random_break5.class, kindly provided by Xiaoming Liu. This was necessary to allow for the construction of bootstrap CI's for the stairway plots with version 2 and folded site frequency spectra. Usually, version 2 for folded spectra only allows the construction of pseudo CI's by random subsample of polymorphic sites and random breakpoints. I used my own bootstrap replicates of SFS (resampled over contigs), not a parametric bootstrap as implemented in version 1 of stairway-plot. I specified a mutation rate per base pair and generation of 3×10^{-9} (Liu et al., 2017) and a single number of breakpoints (steps), 34, effectively turning off random breakpoints.

Estimation of the two-dimensional joint site frequency spectrum

As for the one-dimensional site frequency spectra (see section 3.2.8), I have used ANGSD/realSFS version 0.917-142 (git commit ge3dbeaa, downloaded on 20 June 2017) for the estimation of an unfolded two-dimensional (2D) joint site frequency spectrum from only overlapping sites. I have used the same exhaustive search parameters for two-dimensional site frequency spectrum estimation with realSFS as for one-dimensional SFS estimation (see section 3.2.8 on page 79).

For the creation of 200 bootstrap resampled 2D site frequency spectra (resampled over contigs), I have used bootstrap resampled lists of contig id's and specified them as *regions* for SAF file creation with ANGSD (see bootstrap_contigs.ipynb). I used my script estimate_SAFs.py to compute SAF files in parallel. Importantly, I made sure that sample allele frequency likelihood calculation was restricted to only overlapping sites. Usually, realSFS when provided with two or more SAF files (from two or more populations) would automatically estimate the joint site frequency spectrum from only overlapping sites even if the SAF files would also include non-overlapping sites. However, the algorithm for determining overlapping sites in realSFS does not allow repetitions of the same sites (here sites from whole contigs) in the input SAF files. I have therefore modified the realSFS source file `multisafreader.hpp` by adding `return 0;` at the beginning of the function `set_intersect_pos`. This turns off determination of overlapping sites when `realSFS` is provided with more than one SAF file and instead makes it read in all sites, including repetitions of the same site. See `ANGSD` issue thread #86 for some more details.

$\delta a \delta i$

I have used $\delta a \delta i$ (Diffusion Approximations for Demographic Inference) version 1.7.0 (`git` commit b8e89915c, downloaded on 17 February 2017) for demographic modelling and inference (Gutenkunst et al., 2009). I have used its function `fold` to fold the two-dimensional unfolded spectrum from `ANGSD/realSFS`.

$\delta a \delta i$ provides a numerical solution of a diffusion equation that models the probability distribution of population allele frequencies over time. It is a continuous approximation to the bi-allelic Wright-Fisher process with a discrete number of individuals and a discrete number of populations (Kimura, 1986). It can model random genetic drift, population splits, directional migration and selection. It requires the specification of a demographic model with a set of parameters. It computes the expected population allele frequency spectrum under a certain set of parameter values and then computes the product of Poisson likelihoods over the entries in the observed site frequency spectrum with rates equal to the expected allele frequency from the optimally scaled model spectrum. The set of parameter values is optimised with one of several optimisation algorithms provided by `SciPy` (Jones et al., 2001–). For the conversion of parameters from genetic to absolute units, I have assumed one generation per year and a mutation rate per nucleotide site and generation of 3×10^{-9}. For details of the demographic model fitting with $\delta a \delta i$, see the Jupyter notebook 01_newAngsd_2D_models. For graphical representations of inferred best-fitting models, I have used functions provided by a derivative of $\delta a \delta i$ called `moments` (Jouganous et al., 2017).

When performing LRT's to test the significance of more complex models compared to simpler, nested models, I have applied an adjustment factor to the LRT test statistic (D). This adjustment is required when performing LRT's with composite likelihoods (Coffman et al., 2016). This adjustment factor can be computed with $\delta a \delta i$'s `LRT_adjust` function and a set of bootstrap data sets (i.e. site frequency spectra). I have used my 200 bootstrap resampled 2D site frequency spectra for this, generated as described above. The adjustment factor can be extremely different depending on whether it was calculated by evaluating `LRT_adjust` at the simple or complex model optimal parameterisation. Evaluating with the complex model optimal parameters is more powerful, while evaluating at the simple model optimal parameters provides a more conservative adjustment. P-values for mixtures of χ^2_x distributions have been calculated with $\delta a \delta i$'s `sum_chi2_ppf` function.

For the estimation of 95% CI's, I have used the function `GIM_uncert` in $\delta a \delta i$, which accounts for linkage in the data (Coffman et al., 2016). `GIM_uncert` calculates the Godambe Information Matrix (GIM), which requires the provision of bootstrap resampled site frequency spectra. Note that the GIM method for estimation of parameter variances and covariances is still an imperfect approximation. This is made evident by parameter variances and covariances calculated with a Fisher Information matrix (FIM), which assumes completely independent data, that are sometimes *larger* than those derived from the GIM. Also note that all CI's reported here assume a normal distribution of errors. Further details about the estimation of parameter uncertainties as well as formulas for error propagation from genetic to absolute units can be found in section 13 of the IPython notebook 01_newAngsd_2D_models.ipynb.

3.3 Results & Discussion

3.3.1 Genetic differention between *erythropus* and *parallelus*

Figure 3.8 shows a principal component analysis of genotypic covariances between all 36 individuals of *erythropus* and *parallelus* taking SNP and genotype uncertainty into account. Almost $^1/_4$ of the total genotypic variance is explained by the first principal component which clearly separates two clusters of individuals that correspond to the two subspecies. The second principal component captures only 3% of the total genotypic variance. This confirms estimates of genetic differentiation by Cooper et al. (1995) based on sequence variation at a single nuclear locus that the two subspecies are genetically distinct. Individual `par_34-1` seems to be considerably less differentiated from *erythropus* than the other *parallelus* individuals. This may be due to insufficient information because of extremely low coverage (see fig. 3.3 and Patterson et al. 2006). See section 3.4.5 on page 110 for the effect of SNP calling, genotype calling and normalisation of covariances on PCA.

I estimated posterior expectations of F_{ST} from 1.6 million sites across 32,706 contigs. Note, that this included 0.5 million sites for which there were fewer than 9 individuals with read data in one of the two populations. The global average Hudson/Bhatia's estimate of F_{ST} is 0.298 (95% bootstrap CI: 0.294 – 0.303) (see fig. 3.9). Note, that the 95% CI only captures the variance due to sampling loci in the genome (*genetic sampling* according to Weir 1999, p. 161), not the variance due to sampling individuals from the population (statistical sampling). The latter could be estimated by bootstrapping individuals from each population. Estimating CI's by bootstrapping over contigs assumes that they are independent replicates of the evolutionary process in the history of the two subspecies, i.e. that they are unlinked. However, with the standard RAD protocol (Baird et al., 2008), *two* RAD tags (here called contigs) are recovered from each restriction site. The bootstrap CI for genome-wide F_{ST} computed here may therefore slightly underestimate uncertainty. Also note that the empirical null distribution of global F_{ST}, estimated by randomly permuting the population label of individuals, does not include 0. There therefore seems to be a positive bias in the estimation of global F_{ST} of about 0.0249. See section 3.4.6 on page 117 of the supplementary material for different estimates of global F_{ST}.

Figure 3.10 shows the dependence of average F_{ST} on the minor allele frequency when only sites are included that are polymorphic in the ascertainment population. Not surprisingly, SNP ascertainment in one of the two populations lowers average F_{ST} estimates since it also excludes highly differentiated sites where the ascertainment population is fixed for one allele. When ascertaining SNP sites in *erythropus* or *parallelus*, in each case the average F_{ST} estimate tends to decrease with decreasing minor allele frequency. This would be consistent

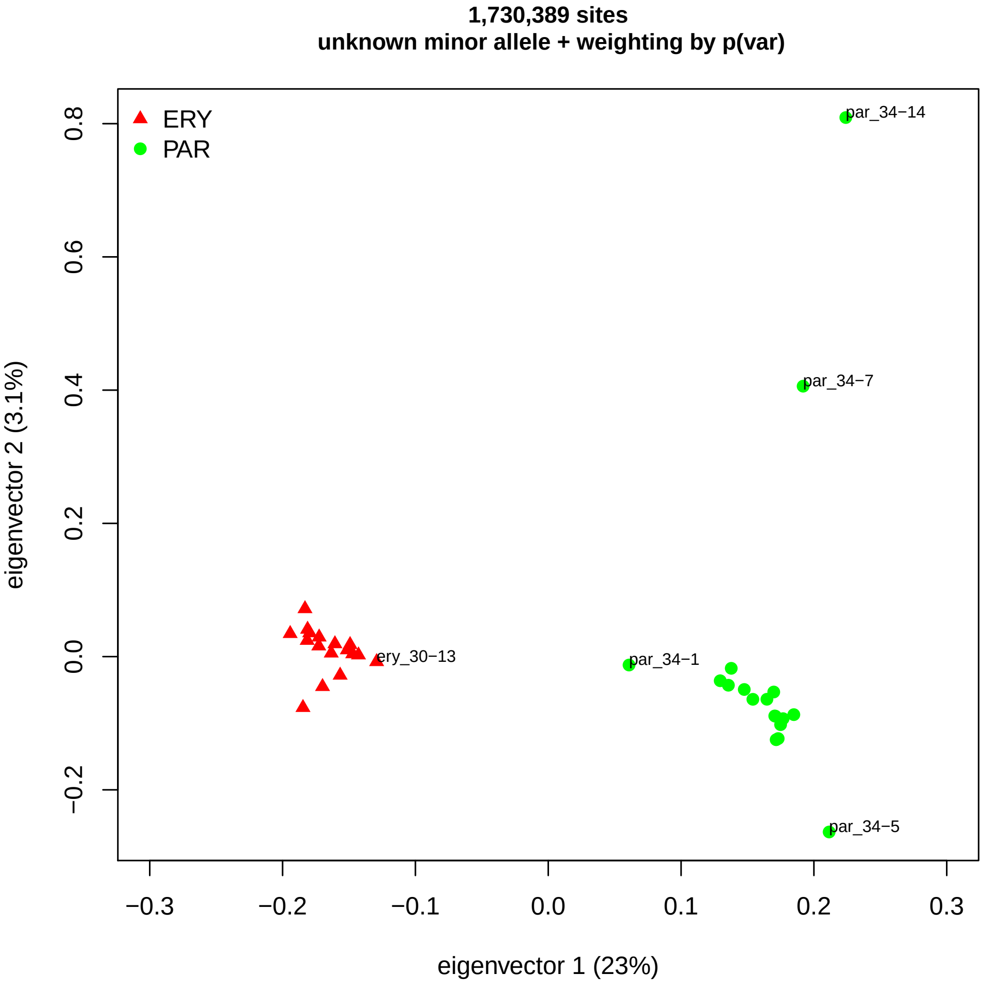


Fig. 3.8 Principal component analysis with genotype probabilities across 1.7 million sites from all 36 individuals of *erythropus* and *parallelus*. Numbers in brackets indicate the percentage of total variance explained by this axis.

with both populations having undergone a population expansion where rare alleles are more likely to be private to one population. This is made evident by the estimate of global unfolded 2D-site frequency spectrum that was used as prior for per–site F_{ST} estimates (fig. 3.7). Since these sites have a low allele frequency differentiation, they lower the average F_{ST} estimate. In contrast, if one or both populations had undergone a recent population bottleneck, variable sites would more often be ancient and polymorphic in the ancestral population. Sites with rare alleles in the bottlenecked population would therefore also include those that have drifted apart in allele frequency between the two populations, thus increasing average F_{ST} estimates (see fig. 3.12) (Bhatia et al., 2013). According to these predictions, the sudden increase in F_{ST} for the lowest MAF when ascertaining in *erythropus* (fig. 3.10) would therefore indicate

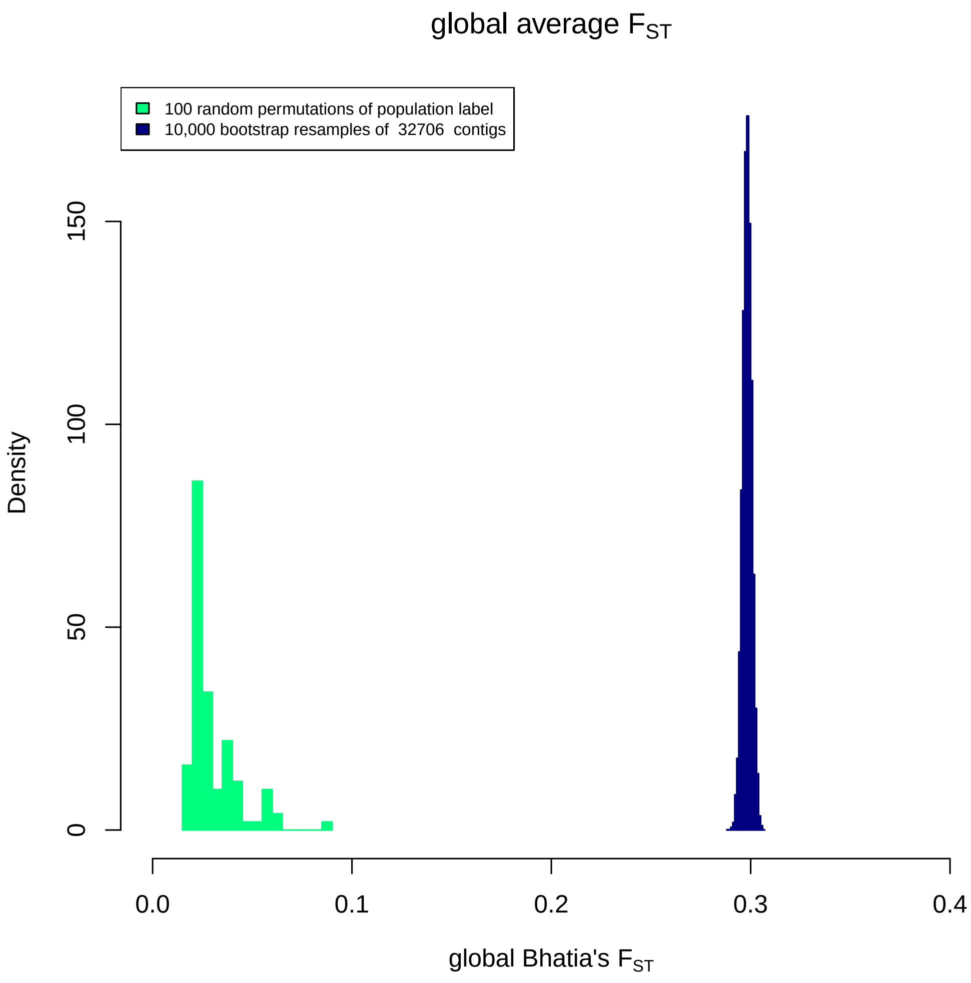


Fig. 3.9 Resampling based estimation of the variance of global F_{ST} (blue) and the empirical null distribution of global F_{ST} by randomly permuting population labels of individuals (green).

a very recent bottleneck, whereas the sudden decrease in F_{ST} when ascertaining in *parallelus* would indicate a very recent population expansion. However, the observed distribution of F_{ST} over MAF as calculated with $\delta a \delta i$ from the 2D SFS of figure 3.7 on page 78 does not show the same tendency to decrease with minor allele frequency (figure 3.11).

A ML estimate of the number of nucleotide sites with fixed differences between these two samples of 18 individuals each from *erythropus* and *parallelus* can also be extracted from the 2D-site frequency spectrum in figure 3.7: 418 or 0.69% of polymorphic sites. These are the counts in the cells (0, 36) and (36, 0) of the 2D-SFS.

The net sequence divergence D_a per site between *erythropus* and *parallelus* is 0.003.

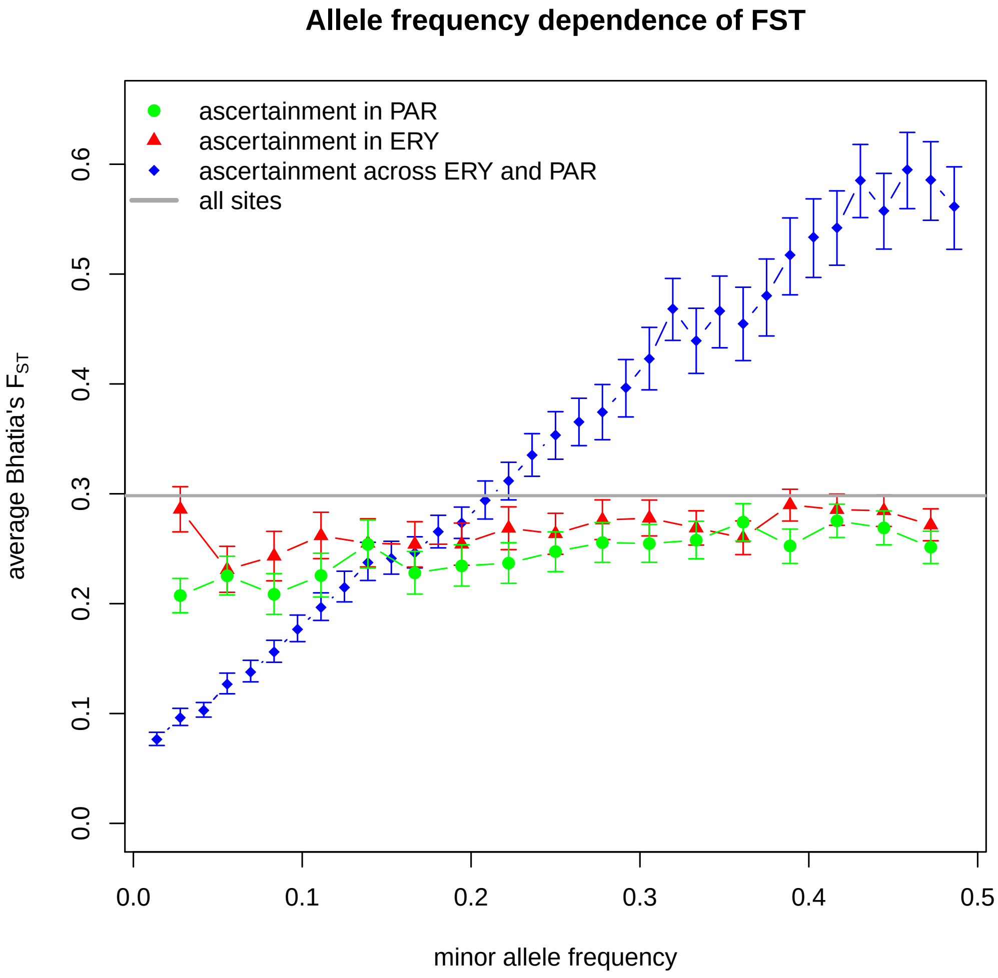


Fig. 3.10 F_{ST} by MAF class for different ascertainment schemes. The global average F_{ST} when using all sites, i.e. without SNP ascertainment, is shown as a grey line. For the subset of sites in each MAF class the median and the 95% CI of 10,000 bootstrap resamples are shown.

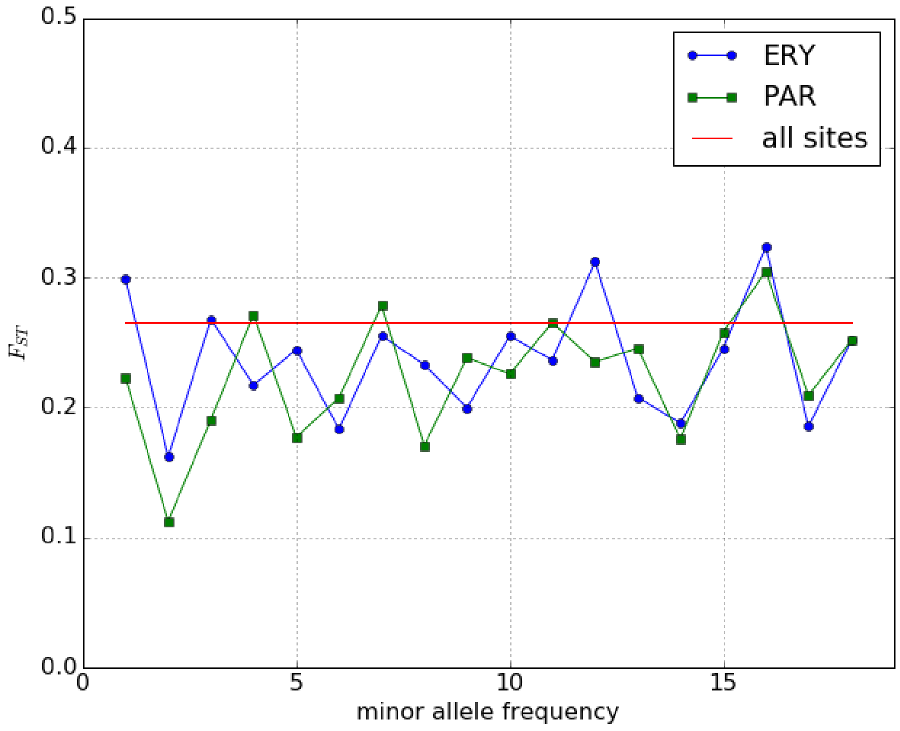


Fig. 3.11 Distributions of F_{ST} by MAF as estimated from the 2D SFS in figure 3.7 on page 78 with the function `Fst` in $\delta a \delta i$. SNP's were subset by minor allele frequency in either *erythropus* (ERY) or *parallelus* (PAR).

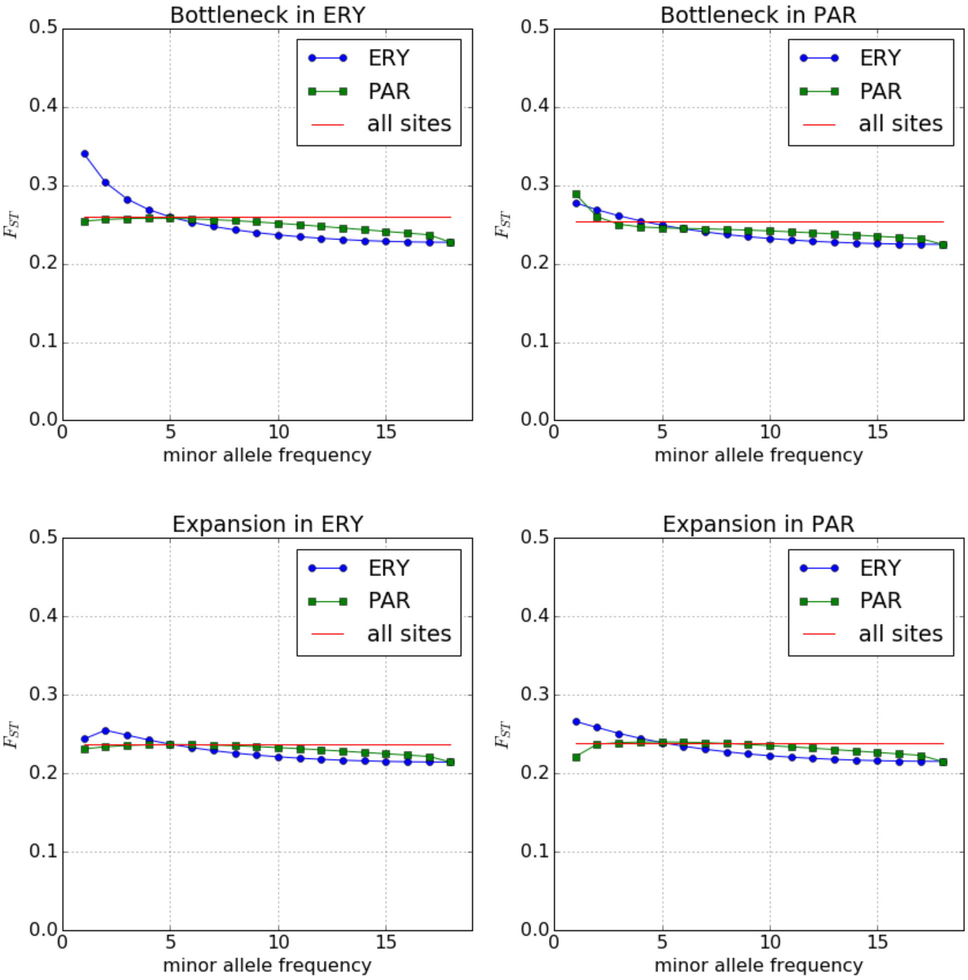


Fig. 3.12 Distributions of F_{ST} by MAF for four different demographic models. Distributions were computed from simulated model spectra in the same way as for figure 3.11.

3.3.2 Genetic diversity within *erythropus* and *parallelus*

Figure 3.13 shows the ML estimate of the global folded site frequency spectrum for *erythropus* and *parallelus*. The conspicuous spike in frequency for the minor allele count of 2 as compared to standard neutral expectations, particularly in *parallelus*, is not compatible with any demographic scenario. However, even when disregarding frequency class 2, the folded site frequency spectrum of *parallelus* seems to be more right-skewed, that is having a greater proportion of low-frequency variants, than *erythropus*. There is apparently not a big difference between spectra estimated from only overlapping sites and spectra including non-overlapping sites. Note, that this site frequency spectrum is not based on called SNPs. Instead it is based on per-site sample allele frequency likelihoods (SAF) that incorporate the genotype uncertainty of each individual due to binomial sampling of alleles from a diploid genotype – as happens during next generation sequencing – as well as sequencing and read alignment error (Li, 2011b; Li et al., 2008; Nielsen et al., 2012, 2011). It does not therefore suffer from the ascertainment bias observed in previous studies that were based on sites detected as polymorphic in a specific sample (Albrechtsen et al., 2010; Han et al., 2014; Korneliussen et al., 2013). This ascertainment bias has led to a relative excess of intermediate versus low frequency classes of SNP's. It also does not suffer (as much) from bias caused by genotype calling from low-coverage sequencing. This can either lead to an excess of low frequency variants (Nielsen et al., 2012, fig. 1), if sequencing errors are mistaken for alleles, or a deficiency of low frequency variants if genotype calling algorithms require a minimum coverage (or coverage ratio) before they call a heterozygote genotype (e. g. Liu and Fu (2015)).

The site frequency spectrum has not yet been corrected for bias due to allele-drop-out (Luca et al., 2011). Simulations by Cariou et al. 2016 indicate, however, that "this bias is of minor importance when the polymorphism is below 2 %, which is the case in most species, at least in animals". The estimated expected nucleotide heterozygosity (π_{site}) estimated here for both subspecies is well below 2% (see tab. 3.1). Allele-drop-out is a problem that has affected previous types of genetic markers like microsatellites or AFLP's, usually known under the term "null alleles". Arnold et al. (2013) and Gautier et al. (2012) have shown with simulations that loci affected by allele-drop-out show a greater proportion of intermediate allele frequencies as compared to loci not affected. My data filtering, that included filtering for minimum coverage and for minimum number of individuals with read data, should have enriched for sites less affected by allele-drop-out. Those sites show the opposite bias towards a relative excess of low-frequency variants (Arnold et al., 2013). Figure 3.13 shows site frequency spectra estimated from only overlapping sites, i.e. sites with read data for at least 9 individuals in each population sample, as well as from including non-overlapping

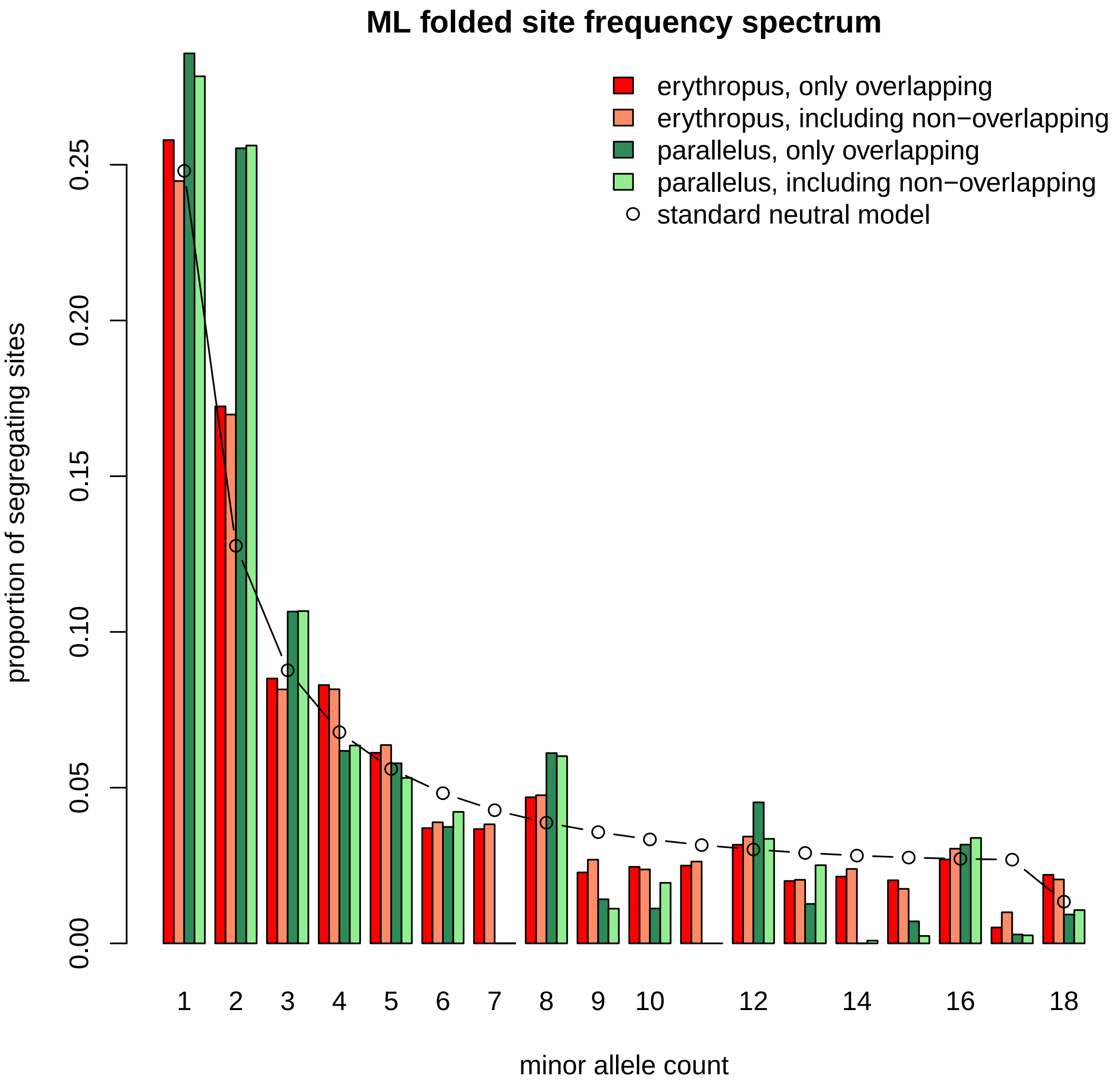


Fig. 3.13 ML estimate of the global minor allele frequency spectrum for *erythropus* (red) and *parallelus* (green). The site frequency spectra were estimated from either *only overlapping* nucleotide sites (1.13 million) or *including non-overlapping* sites (1.6 million from *erythropus* and 1.2 million from *parallelus*).

sites. Overlapping sites should be more enriched for those less affected by allele-drop-out and therefore show a greater bias towards a relative excess of low-frequency variants. This effect can be observed in the diversity statistics reported in table 3.1 where, for instance, the expected nucleotide heterozygosity (π_{site}) for *parallelus* calculated from a site frequency spectrum from only overlapping sites is lower than when calculated from a SFS including non-overlapping sites. This effect, although statistically significant, is relatively small and therefore unlikely to change any conclusions drawn from the data in a qualitative way. This makes sense since loci with greater nucleotide heterozygosity should have a higher chance of allele-drop-out due to polymorphisms in the restriction site and therefore also a lower chance of overlapping between the two populations (i.e. having 9 individuals with read data in each population).

Since the same filtering thresholds were applied to the data from *erythropus* and *parallelus* and since the *parallelus* sample has generally much lower coverage (see fig. 3.3), it would seem plausible that, when *including non-overlapping sites*, the sites used to estimate the site frequency spectrum for *parallelus* have been relatively more enriched for those less affected by allele-drop-out than the sites used to estimate the SFS of *erythropus*. However, although absolute parameter values may well be biased in both *parallelus* and *erythropus*, all analyses in this and the following section are also done from only overlapping sites. Any differences between *parallelus* and *erythropus* can therefore not be due to lower coverage in *parallelus* in conjunction with allele-drop-out. Also, the allele-drop-out problem is a missing data problem and to some extent similar to the missing data problem caused by low coverage next-generation sequencing. Both increase false homozygote calls for true heterozygotes and both therefore lead to a deviation of observed genotype frequencies from HWE. The algorithms for inferring the site frequency spectrum implemented in `ANGSD` incorporate an assumption of HWE [Nielsen et al. 2012, eq. 2 and 3 and Vieira et al. 2013]. They place a low conditional probability on combinations of genotypes in the sample that would deviate strongly from HWE. I therefore expect `ANGSD` to mitigate the effects of allele-drop-out on population genetic analyses.

Figure 3.14 shows that despite the large amount of sites included in the ML estimation, there is still considerable uncertainty in the global folded site frequency spectrum, particularly for *parallelus*. This could be explained by the lower average sequence coverage for individuals from the *parallelus* population (see fig. 3.3).

The number of segregating sites (S) and the average number of pairwise differences (π_{Tajima}) are summary statistics of the site frequency spectrum. While S weights each site equally, intermediate frequency variants contribute much more to the magnitude of π than high- and low-frequency variants (Wakeley, 2009, eq. 1.4). Both diversity estimates are significantly higher in *parallelus* than in *erythropus* (tab. 3.1).

This seems at odds with the expectation from the hitherto proposed historical biogeographic model of a postglacial expansion of *C. p. parallelus* towards central and western Europe from a glacial refuge in the Balkans (Cooper et al., 1995; Lunt et al., 1998). *C. p. erythropus* in the Pyrenees on the other hand is expected to be derived from several smaller refuges in southern Spain, i.e. its expansion after the last Ice Age would have covered a much shorter distance. According to this model one would expect the *parallelus* subspecies to have

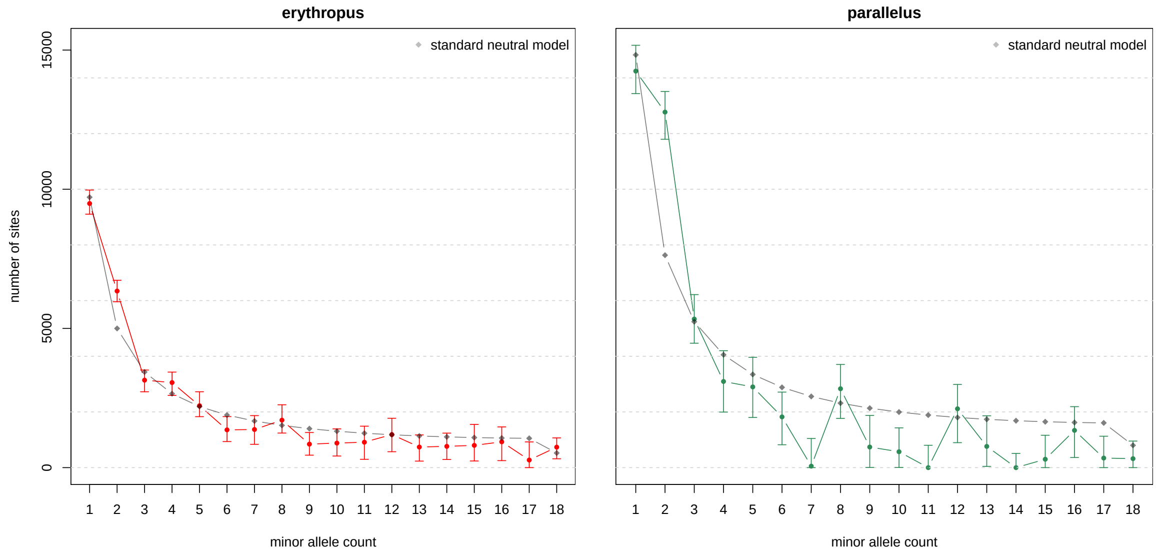


Fig. 3.14 The global folded site frequency spectrum of *erythropus* and *parallelus* with absolute counts from only overlapping sites and 95% CI limits from 200 bootstrap resamples of contigs. A standard neutral model spectrum was fit to each observed spectrum for comparison.

undergone a long series of founder events which should have reduced genetic diversity at the edge of the distribution much more so than for the *erythropus* subspecies (Luca et al., 2011).

That ancestral *parallelus* and *erythropus* expanded from their glacial refuges as a "wave", i.e. without colonisation by rare long–distance migration, is made unlikely by at least two observations:

- the colonisation of Britain by *C. p. parallelus* before the flooding of the English Channel after the last Ice Age (Cooper et al., 1995),

Table 3.1 Comparison of the proportion of segregating sites S_{prop}, the average number of pairwise differences per site π_{site} and Tajima's D for the global site frequency spectrum of *erythropus* and *parallelus*. Numbers in brackets are 95% bootstrap confidence intervals. Diversity statistics are calculated from only overlapping sites or including non-overlapping sites. There are 1,130,775 overlapping sites. When including non-overlapping sites, estimates are based on 1,214,939 sites for *parallelus* and 1,638,468 sites for *erythropus*.

	statistic		
	S_{prop}	π_{site}	Tajima's D
overlapping sites			
erythropus	0.0326 (0.0320, 0.0331)	0.00715 (0.00704, 0.00726)	-0.345 (-0.386, -0.308)
parallelus	0.0442 (0.0435, 0.0450)	0.00805 (0.00794, 0.00815)	-0.936 (-0.965, -0.908)
including non-overlapping sites			
erythropus	0.0325 (0.0321, 0.0330)	0.00735 (0.00726, 0.00745)	-0.242 (-0.282, -0.210)
parallelus	0.0451 (0.0445, 0.0457)	0.00832 (0.00820, 0.00845)	-0.897 (-0.927, -0.868)

- the clines for several markers in the Pyrenean hybrid zone between the two subspecies that are too wide to be explained by the *average* dispersal rate of this species of 30 m per generation (Nichols and Hewitt, 1994).

The higher genetic diversity in Pyrenean *parallelus* is not unique to this study though. Lunt et al. (1998) reported intraregional K_S values of a mitochondrial sequence that were 1.5 times higher in Pyrenean *parallelus* than in Pyrenean *erythropus* and Llewellyn (2008), table 3–3, found that nucleotide diversity (π_{Tajima}) was higher in *parallelus* than in *erythropus* from the Pyrenees in 5 out of 7 expressed sequences. K_S values in Cooper et al. (1995) from a single nuclear locus contradict this and show about twice as much genetic variation in Spain as in France.

Both subspecies have a significantly negative Tajima's D (tab. 3.1). This means that there is an excess of low frequency variants. Besides artefacts in the data or analysis pipeline (Shafer et al., 2016), this can only have two reasons. First, many variable sites are of recent recent origin, which would be a signal of recent population size expansion. Second, pervasive genome-wide selection, positive or negative and at the detected sites or closely linked sites. With purifying selection, low frequency variants would also be of ancient origin and be kept at low frequency by selection. The Tajima's D estimates also differ significantly between the two subspecies. The estimate for *parallelus* is far more negative than the estimate for *erythropus*. This is as expected from a greater recent population size expansion in *parallelus* than in *erythropus*. The difference in Tajima's D observed here therefore would be in line with the scenario inferred by previous phylogeographic studies. Alternatively, a greater ancient population size in *parallelus*, as indicated by the higher genetic diversity estimates, would allow for a greater strength of purifying selection in *parallelus* than in *erythropus* (Corbett-Detig et al., 2015).

3.3.3 Demographic history of the hybrid zone

If all sites were bi-allelic and if there were no linkage among sites, i.e. each site could be regarded as an independent data point, then the two-dimensional site frequency spectrum in figure 3.7 would be a complete summary of the genetic variation in the data set. Figure 3.15 shows the folded version of this. Each cell in this figure indicates the count of sites with a specific allele frequency in *erythropus* and a specific allele frequency in *parallelus* of the *joint* minor allele, i.e. the minor allele when pooling both populations.

A simple *divergence-in-isolation* model in $\delta a \delta i$ infers a similar divergence time for the two subspecies as previously estimated from mitochondrial sequence data (tab. 3.2). Assuming a mitochondrial nucleotide substitution rate of 2% per million years and no gene

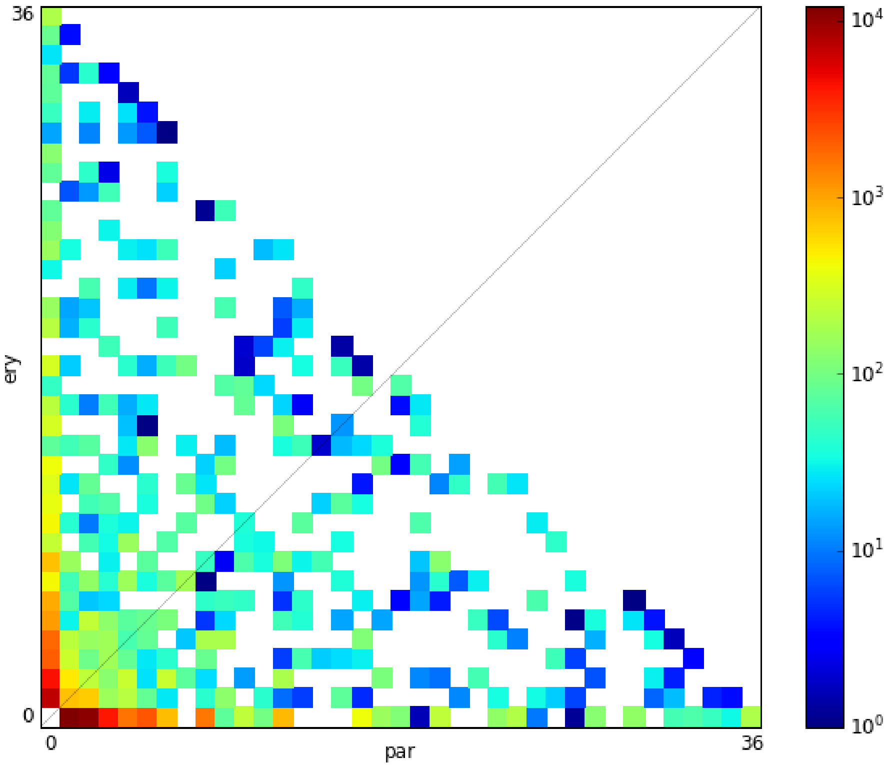


Fig. 3.15 Folded joint site frequency spectrum of *erythropus* and *parallelus* from 1,130,775 overlapping sites. The cells contain counts on a log scale of sites with a certain joint minor allele frequency. The upper right half of the matrix is empty in a folded 2D spectrum and therefore set white. Counts of zero have been set to 1 and appear white in the lower left half of the matrix. This spectrum contains 74,058 sites that are polymorphic in *erythropus* or *parallelus* or both. Monomorphic sites (with joint frequency [0,0]) have been masked for better visualisation.

flow since the split from their common ancestral population, Lunt et al. (1998) have estimated the divergence time to between 363,000 and 731,000 years. The estimate of divergence time from this model in $\delta a \delta i$ is also very similar to the estimated divergence time derived from the expectation of D_a, the net sequence divergence (see eq. 3.4): 504,166 years.

Adding gene flow to this model improves the fit of the simulated model spectrum to the observed spectrum (tab. 3.3). Allowing for gene flow also doubles the estimated divergence time. This *divergence-with-migration* model can be reduced to the *divergence-in-isolation* model (without migration) by setting the migration rate to zero. The latter model is therefore nested within the former, allowing for the application of a LRT for the statistical significance of the existence of gene flow between *erythropus* and *parallelus*. With the composite likelihood adjustment of the LRT statistic evaluated at the complex model (i.e. *divergence-with-migration*) optimal parameterisation, the addition of a migration rate parameter greater than zero is highly significant ($p \simeq 0.0$). Note, that since here a single parameter is on the boundary of the parameter space, the null distribution of the LRT statistic is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self and Liang, 1987). If the adjustment factor is computed with the optimal parameters from the nested *divergence-in-isolation* model, then a non-zero migration rate has a p-value of 0.011. The observed 2D-SFS therefore seems to provide a robust signal of low but non-zero gene flow between the two subspecies during the history of their divergence.

The *divergence-with-migration* model spectrum has markedly reduced residuals for the counts of low frequency shared polymorphisms as compared to the *divergence-in-isolation* model (fig. 3.16). An increased number of shared polymorphisms at low frequency in one or both populations is a signal of gene flow (Gutenkunst et al., 2009, suppl. mat. section 1.1). Further details can be found in section 3.4.8 on page 120 of the supplementary material.

Allowing the gene flow to be asymmetric further improves the fit to the data (tab. 3.4). Are the two migration rates, one for each direction, significantly different from each other? The model with one (symmetrical) migration rate is not just nested within a model with two different migration rates. I have therefore parameterised the *asymmetric migration* model so that the divergence with (symmetrical) migration model is nested within it by setting $m_2 = r \times m_1$. The divergence with symmetrical migration model has by definition an r of 1. I can then ask: does allowing r to be different from 1 significantly improve the fit to the observed spectrum? The p-value of the LRT statistic (D), with an adjustment factor calculated by evaluating at the complex model optimal parameterisation, is 0.0 (assuming D is χ_1^2 distributed). With an adjustment factor calculated by evaluating at the nested model optimal parameterisation, the p-value of the LRT is 0.00083. There therefore seems to be a robust signal of asymmetric migration between *erythropus* and *parallelus*.

Gene flow is estimated to have been on average ∼5 times higher in the direction *parallelus→erythropus* than vice versa (tab. 3.4). Several observations from previous studies have

Table 3.2 Parameters for a simple *divergence-in-isolation* model inferred with $\delta a \delta i$. N_{ery}: diploid population size of *erythropus*, N_{par}: diploid population size of *parallelus*, T: divergence time in years, -logL: negative log-likelihood of the model spectrum simulated with the given parameter values. The population sizes are assumed to have been constant since the split from the common ancestral population.

Parameter	ML estimate	95% CI
N_{ery}	602,490	586,288 – 618,692
N_{par}	1,281,364	1,236,331 – 1,326,398
T	486,848	476,703 – 496,994
-logL	25,375	

Table 3.3 Parameters for a *divergence-with-migration* model inferred with $\delta a\delta i$. N_{ery}: diploid population size of *erythropus*, N_{par}: diploid population size of *parallelus*, T: divergence time in years, m: proportion of new immigrant individuals each generation, -logL: negative log-likelihood of the model spectrum simulated with the given parameter values. The population sizes are assumed to have been constant since the split from the common ancestral population. Gene flow is assumed to be equal in both directions.

Parameter	ML estimate	95% CI
N_{ery}	599,837	584,697 – 614,978
N_{par}	1,167,067	1,132,862 – 1,201,271
T	1,083,296	1,062,695 – 1,103,896
m ($\times 10^{-7}$)	2.49	2.41 – 2.57
-logL	21,942	

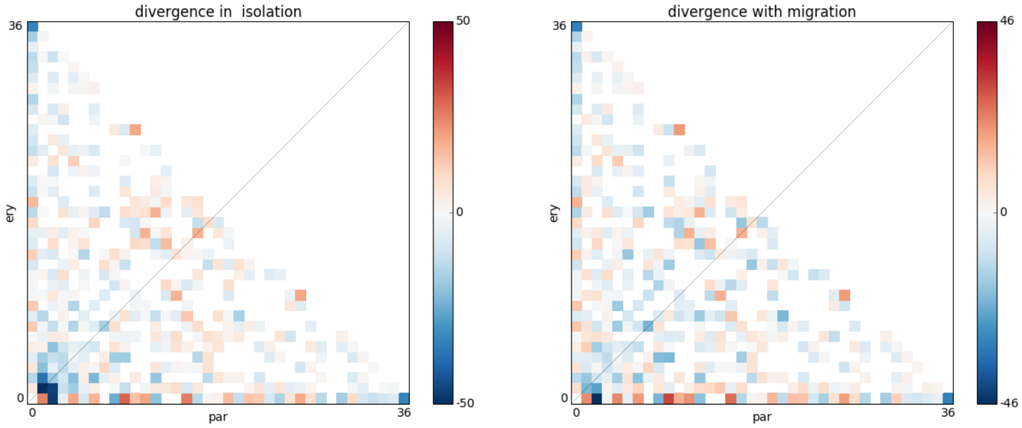


Fig. 3.16 The plots of Poisson residuals between the observed 2D-SFS (fig. 3.15) and the best-fitting model spectra from the *divergence-in-isolation* model (left) and the *divergence-with-migration* model (right). Note the markedly reduced residuals for the counts of low frequency shared polymorphisms in the residual plot of the *divergence-with-migration* model.

indicated greater introgression from *parallelus* into *erythropus* (summarised in the chapter General Introduction).

One potential reason for stronger effective gene flow from *parallelus* into *erythropus* than vice versa could be that introgressed alleles from *parallelus* had a greater chance to reach those *erythropus* populations that survived the ice ages in southern Spain than introgressed *erythropus* alleles could have reached those *parallelus* populations that survived the ice ages in the Balkans or even further southeast. This is already obvious from the much greater geographical distance of the current hybrid zone (and probably also hybrid zones of previous epochs) to the inferred location of glacial refuges of the two subspecies (Lunt et al., 1998).

Table 3.5 shows the parameters of a two-epoch model with exponential size change in the second (recent) epoch. The model allows a size change starting at time T_2 in the

Table 3.4 Parameters for an *asymmetric-migration* model inferred with $\delta a \delta i$. N_{ery}: diploid population size of *erythropus*, N_{par}: diploid population size of *parallelus*, T: divergence time in years, m: proportion of new immigrant individuals each generation, -logL: negative log-likelihood of the model spectrum simulated with the given parameter values. The population sizes are assumed to have been constant since the split from the common ancestral population.

Parameter	ML estimate	95% CI
N_{ery}	500,974	483,927 – 518,021
N_{par}	1,279,484	1,240,749 – 1,318,220
T	1,138,863	1,102,539 – 1,175,187
$m_{ery \rightarrow par}$ $(\times 10^{-7})$	1.06	0.89 – 1.23
$m_{par \rightarrow ery}$ $(\times 10^{-7})$	4.97	4.66 – 5.28
-logL	21,465	

past with the populations exponentially reaching the current population size (fig. 3.17). Each population size change is enforced to happen at the same time for both populations. Migration rates are assumed to be the same across epochs. The best-fit parameters of this model indicate a very recent population size reduction for both *erythropus* and *parallelus*, with a much more severe bottleneck in *parallelus* than *erythropus*. Does the addition of the second epoch significantly improve the fit to the observed 2D-SFS? By either setting $N^2_{ery} = N^1_{ery}$ and $N^2_{par} = N^1_{par}$ or by setting $T_2 = 0$, I can reduce this two epoch model to the one-epoch *asymmetric migration* model from above. When calculating the composite likelihood adjustment factor by evaluating at the optimal parameterisation of the complex model (treating N^2_{ery}, N^2_{par} and T_2 as nested), the p-value of the LRT is $\simeq 0.0$. Note that here one parameter (T_2) is at the boundary of the parameter space. So D is not strictly χ^2 distributed with 3 degrees of freedom (see Self and Liang 1987). However, I cannot find out the correct mixing of probability distributions, so I am using the χ^2_3 distribution in the hope that this is conservative. The adjustment factor evaluating at the optimal parameterisation of the simple model (*asymmetric-migration*) resulted in an error message and therefore could not be calculated. Still, the addition of the second epoch seems to allow for a statistically significant improvement of the fit to the data. Note, that a piecewise two-epoch model, i.e. with an instantaneous size change at T_2 years in the past, fits the data slightly better (by 21 log-likelihood units) than this model with exponential size change during the second epoch. However, the piecewise model converged on parameter values that indicated an unrealistically recent bottleneck only 51 generations ago and a current population size for *parallelus* of only 1,217.

Figure 3.18 shows the stairway plots for *erythropus* and *parallelus*. They indicate a much larger ancient population size for *parallelus* with around 4 million than for *erythropus* with around 0.7 million. In addition, the estimate of ancient population size for *erythropus* seems

Table 3.5 Parameters for a *two-epoch-exp.-size-change* model inferred with $\delta a \delta i$. N^1_{ery}: diploid population size of *erythropus* during the first epoch after the split from the common ancestral population, N^1_{par}: diploid population size of *parallelus* during the first epoch, T_1 (T_2): duration of the first (second) epoch in years, N^2_{ery} (N^2_{par}): diploid population size of *erythropus* (*parallelus*) at the end of the second epoch (i.e. at present), m: proportion of new immigrant individuals each generation, -logL: negative log-likelihood of the model spectrum simulated with the given parameter values. The population sizes are assumed to have been constant during the first epoch after the split from the common ancestral population followed by an exponential size change in the second epoch.

Parameter	ML estimate	95% CI
N^1_{ery}	477,541	457,215 – 497,867
N^1_{par}	1,742,857	1,664,480 – 1,821,233
T_1	1,054,796	899,637 – 1,209,952
N^2_{ery}	275,161	127,196 – 423,127
N^2_{par}	40,775	29,753 – 51,798
T_2	5,330	4,188 – 6,472
$T_1 + T_2$	1,060,126	1,018,721 – 1,101,530
$m_{ery \to par}$ ($\times 10^{-7}$)	0.9	0.71 – 1.08
$m_{par \to ery}$ ($\times 10^{-7}$)	5.5	5.06 – 5.95
-logL	21,009	

to have an exceptionally low uncertainty. The uncertainty for the very recent population sizes then increases sharply. This is in contrast to the stairway plot of *parallelus*, which indicates a similar uncertainty for all time intervals. There is no hint of ancient fluctuations in population size. The apparent bottlenecks at around 400 kya are likely artefacts of the estimation method (Liu and Fu, 2015). Both stairway plots indicate a very recent drastic reduction in population size. For the *erythropus* population, the stairway plot indicates a reduction to about 5% of the ancestral population size ($\sim 35,000$) within just the last 1.5 thousand years. For the *parallelus* population, `stairway-plot` infers an even more drastic reduction to about 0.5% of the ancestral population size ($\sim 20,000$). This dramatic reduction in effective population size is inferred to have happened within the last 20 thousand years.

So there is a general agreement between the stairway plots (figure 3.18), inferred from single population site frequency spectra with `stairway-plot`, and the *two-epoch-with-exp-size-change* model inferred from the two-population joint site frequency spectrum with $\delta a \delta i$ (fig. 3.17). Both infer a much larger ancient population size for *parallelus* compared to *erythropus*. However, while $\delta a \delta i$ infers only ~1.7 million as ancient population size for *parallelus*, `stairway-plot` infers ~ 4 million. Both programmes infer a very recent population size reduction for both populations. Remember, however, that the model in $\delta a \delta i$ enforced the second size change to happen over the same time period for both populations while the stairway plots have no such restriction. The inferences of both programmes agree

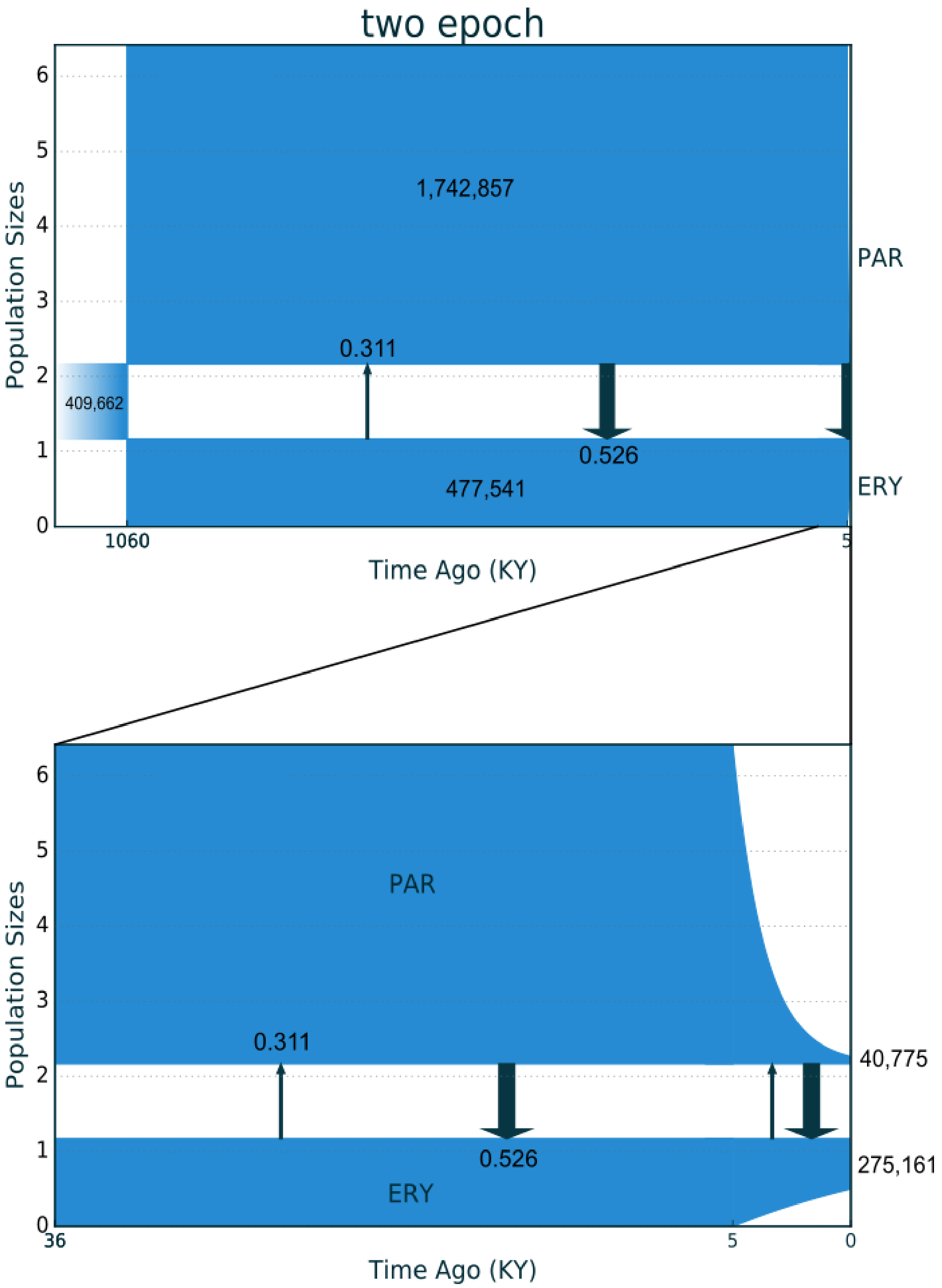


Fig. 3.17 Graphical representation of the best-fit two epoch model with exponential size change in the second epoch. Note that here numbers for migration rates (at arrows) are shown as scaled migration rates ($2Nm$), scaled by twice the population size of the receiving population.

about a very drastic population size reduction for the *parallelus* population. However, there

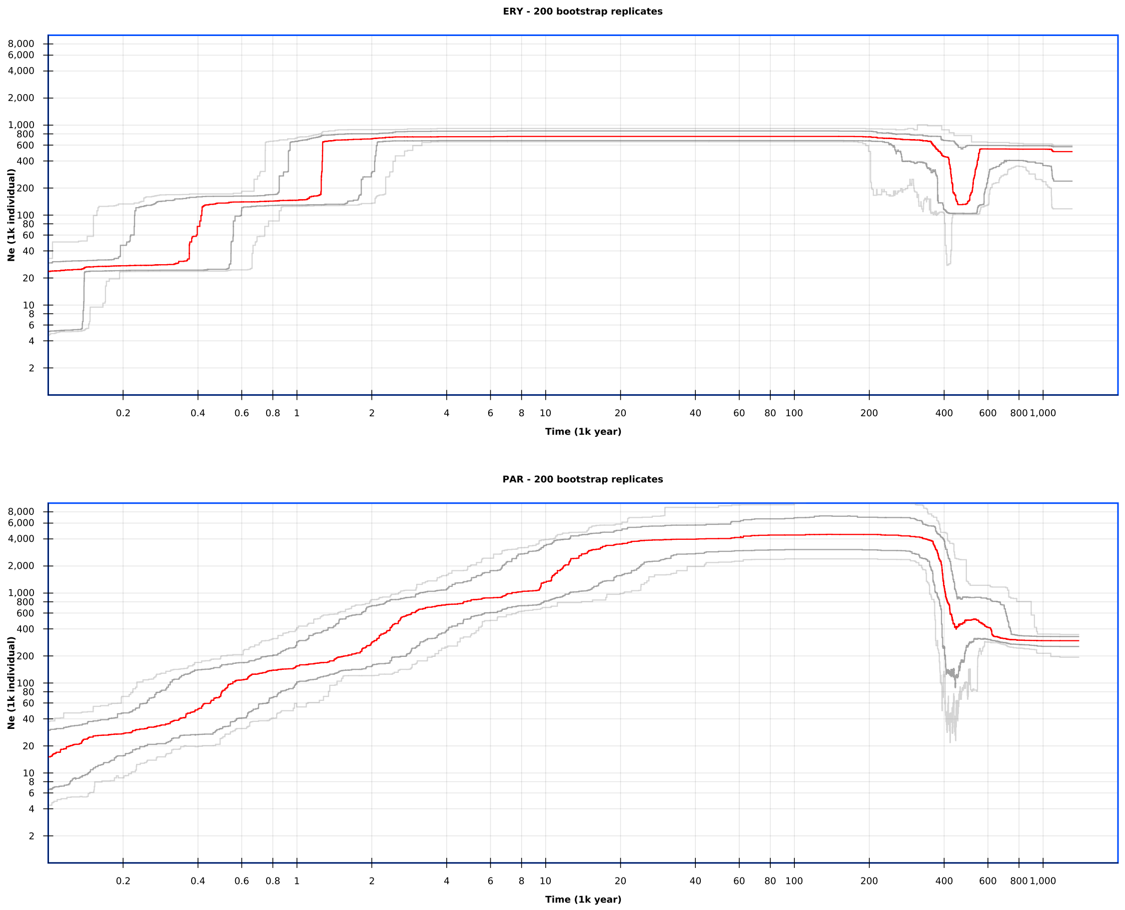


Fig. 3.18 Stairway plots for *erythropus* and *parallelus* from only overlapping sites. The red line is the median inferred population size in each time interval from 200 bootstrap replicates of the SFS. The thick grey lines define the 75% bootstrap-CI and the light grey lines define the 95% bootstrap-CI. Apparent bottlenecks in the two plots at around 400 kya are likely artefacts (see Liu and Fu (2015)).

is quite a large discrepancy about the inferred size of the recent bottleneck for the *erythropus* population. While `stairway-plot` infers a very severe bottleneck to a current population size of ∼35 thousand, $\delta a \delta i$ infers only a mild bottleneck to ∼275 thousand, which is only a reduction by less than 50% from the ancient population size of ∼477 thousand.

The recent population bottleneck shown by the stairway plots for the two subspecies agrees with expectations of range expansion from glacial refuges by long distance dispersal and serial founder events after the end of the last ice age. It also makes sense that the population size started to reduce much earlier in parallelus than erythropus as the range expansion of parallelus must have begun earlier than in erythropus for it to reach the Pyrenees at about the same time as erythropus. It is also congruent with the current biogeographic model that Pyrenean *erythropus* had experienced a much smaller bottleneck due to serial founder events during range expansion than parallelus, since the distance to the glacial refuge is much shorter for Pyrenean *erythropus* than Pyrenean *parallelus*.

In addition to allowing for two epochs with different population sizes, I have also fit two-epoch models to the data that allow for different migration rates between the epochs. The questions is: can we detect a change in the rate of migration over the history of the two subspecies? Table 3.6 shows the best-fitting parameters for an *ancient-asymmetric-migration* model in $\delta a \delta i$. This model fixes the migration rate in the second, i.e. recent, epoch (T_i) at zero. This model infers that gene flow between *erythropus* and *parallelus* has ceased ∼27 thousand years ago (fig. 3.19).

Table 3.6 Parameters for an *ancient-asymmetric-migration* model inferred with $\delta a \delta i$. N_{ery}: diploid population size of *erythropus* after the split from the common ancestral population, N_{par}: diploid population size of *parallelus*, T_c: duration of period with gene flow in years, T_i: time in years since total isolation, i.e. no gene flow, m: proportion of new immigrant individuals each generation, -logL: negative log-likelihood of the model spectrum simulated with the given parameter values. The population sizes are assumed to have been constant since the split from the common ancestral population.

Parameter	ML estimate	95% CI
N_{ery}	490,244	473,465 – 507,021
N_{par}	1,257,636	1,220,512 – 1,294,754
T_c	1,262,489	1,223,308 – 1,301,670
T_i	27,489	26,450 – 28,529
$T_c + T_i$	1,289,979	1,251,387 – 1,328,572
$m_{ery \rightarrow par}$ ($\times 10^{-7}$)	1.23	1.056 – 1.406
$m_{par \rightarrow ery}$ ($\times 10^{-7}$)	6.43	6.029 – 6.839
-logL	21,301	

Compared to the single-epoch *asymmetric-migration* model from above, this model adds a recent period of complete isolation, which improves the fit to the data by 164 log-likelihood units. The *ancient-asymmetric-migration* model can be reduced to the *asymmetric-migration* model by setting T_i to zero. The latter is therefore nested within the former and a LRT of the significance of T_i can be performed. A linkage adjusted LRT has a p-value of 0.0, when calculating the adjustment factor by evaluating at the optimal parameters of the complex model, or a p-value of 0.014, when evaluating at the optimal parameters of the simple model. On a cautionary note, the adjustment factor calculated by evaluating at the optimal parameters of the complex model is 2.22, which would indicate an *increase* in power due to linkage in the data. Obviously, this makes no sense and probably indicates that the approximations from the Godambe Information Matrix are breaking down here (Coffman et al., 2016). Despite this, adding a recent time of complete isolation between *erythropus* and *parallelus* does seem to improve the fit to the data significantly. What is more, allowing the gene flow in the second epoch to be non-zero in both directions, i.e. adding two more migration rate parameters, does not improve the fit to the data (same log-likelihood) and converges to migration rate parameter values for the second epoch that are practically zero.

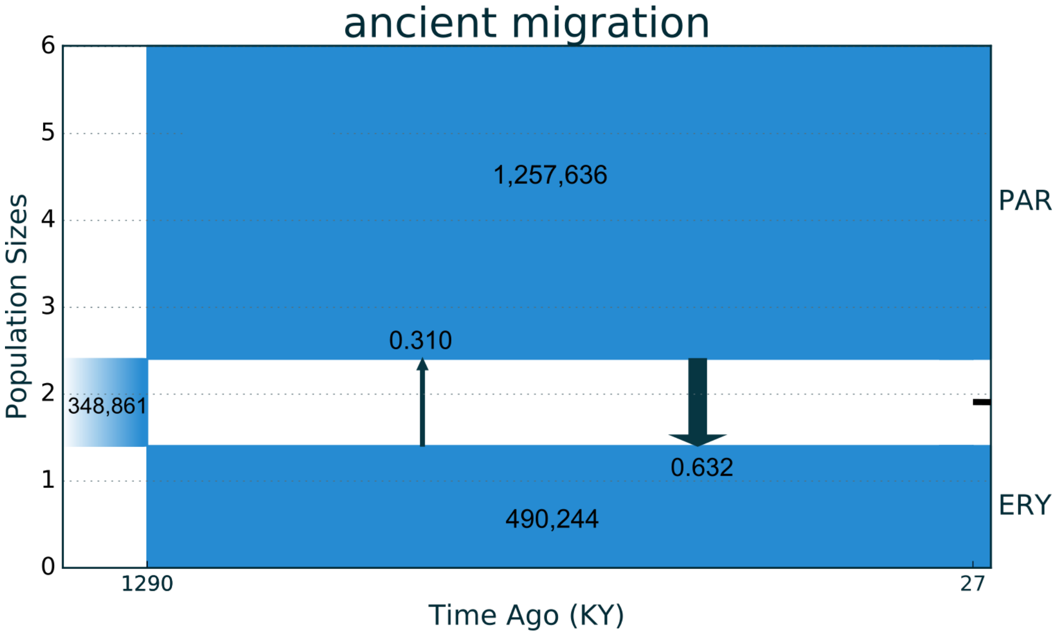


Fig. 3.19 Graphical representation of the best fit *ancient-asymmetric-migration* model. As in figure 3.17, the migration rates are shown as scaled migration rates ($2Nm$), i.e. scaled by twice the population size of the receiving population. The horizontal black bar starting at 27 thousand years in the past symbolises the cessation of gene flow.

Phylogeographic studies as well as paleoclimatic data strongly suggest that the two subspecies, *erythropus* and *parallelus*, diverged in isolation during last 2 or 3 ice ages and only recently (∼10 kya) came into secondary contact (see chapter General Introduction). Clines of several morphological, behavioural and molecular markers have been shown to be very wide (some over 40 km) (Butlin and Hewitt, 1985a,b; Butlin et al., 1991; Vazquez et al., 1994) and their centres often displaced from each other, which indicates substantial gene flow across the hybrid zone (Ferris et al., 1993). Lunt et al. (1998) have found that a population from Pyrenean Spain is genetically more similar to Pyrenean French populations at a mitochondrial sequence (COI) than to other Spanish populations. The two population samples used in this study come from populations in close proximity to the hybrid zone. They may therefore show some evidence of recent gene flow between the two subspecies. In order to detect this, I have also fit a *secondary-contact* model to the observed spectrum with $\delta a \delta i$. This model specifies divergence in isolation during an initial period followed by a period with (symmetrical) migration between the two subspecies until the present. The best fitting model parameters indicate that *erythropus* and *parallelus* diverged in isolation for 445,694 generations. The two populations then came into secondary contact 587,653 generations ago. So, the best fitting parameters clearly do not specify a recent and relatively short period of secondary contact. Also, compared to the simpler one-epoch *divergence-with-migration* model, the *secondary-contact* model improves the fit to the data by only 27 log-likelihood units. Despite this, a linkage adjusted LRT indicates that this improvement is highly significant (p-value $< 10^{-7}$).

The reverse of the *secondary-contact* model is a model that defines the first epoch as the one where gene flow is allowed, followed by a second epoch with complete isolation. It could therefore also be called a *primary-contact* model. I have also fitted such a *primary-contact* model to the data for comparison with the *secondary-contact* model. Note, that the definition of the *primary-contact* model is similar to the *ancient-asymmetric-migration* model described above with the restriction that here there is only one migration rate parameter, i.e. gene flow is assumed to be equal in both directions. Unsurprisingly, the best fit parameters of this *primary-contact* model indicate a very similar history as the slightly more complex *ancient-asymmetric-migration* model. It too infers a long initial period of contact between *erythropus* and *parallelus* ($\sim$1.2 million generations) and a very recent cessation of gene flow between the two subspecies ($\sim$28 thousand generations ago). Compared to the nested *divergence-with-migration* model, it improves the fit to the data by 91 log-likelihood units and a linkage adjusted LRT is highly significant (p < 0.004, with conservative adjustment). This confirms that the addition of an epoch without gene flow significantly improves the fit to the data. In addition, a *primary-contact* model fits the data better than a *secondary-contact* model by 66 log-likelihood units. Finally, a three-epoch model, where in the most recent epoch gene flow is allowed to recommence, converges on a time parameter for the third epoch that is practically zero. That is, the three-epoch model converges on the *primary-contact* model. This underscores that there is no significant signal of recent gene flow in the data. This may in part be caused by a lack of power to detect very low and recent gene flow with the current sample size of 18 individuals (Robinson et al., 2014). If the two populations had been affected by recent gene flow across the hybrid zone it must have been so low that the frequencies of introgressed alleles are too low to be detected with a sample size of just 18 individuals. This result confirms previous cline analyses of many other characters that indicated that these two populations could be regarded as pure representatives of their respective subspecies (Vazquez et al., 1994).

The inferred demographic models as depicted in the figures 3.17 and 3.19 should not be taken at face value. They are necessarily gross simplification of the actual demographic history. For instance, continuous migration during an epoch has been assumed in all models. The actual demographic history is certainly more complex and likely to be one of short periods of intermediate gene flow (while in secondary contact during interglacials) interrupted by long periods without gene flow (while in separate glacial refugia) (Hewitt, 1990). A model that does not contain this complexity will necessarily estimate a longterm average gene flow that is very low.

Both populations sampled are in an area that was inhospitable for these grasshoppers or even covered by glaciers during the last ice age. They must therefore have a history of range expansion from glacial refugia and colonisation of new habitat that formed between ∼20,000 and 10,000 years ago. It is not known whether a particular climatic event coincides with the estimated divergence time of about 1.0–1.2 mya. The estimated divergence time between *erythropus* and *parallelus* may simply be the time of first settlement of Iberia and the Balkans or Turkey by *C. parallelus*. The divergence time estimate is based on an estimate of per nucleotide mutation rate in other insects (Liu et al., 2017) and assumes a molecular clock, i.e. no variation in substitution rate over the time of divergence between the two subspecies.

3.4 Supplementary Material

3.4.1 How well are mapping quality scores calibrated?

The mapping quality score (mapQ) is the Phred scaled probability p that the true mapping position of the read is not the reported mapping position (rounded to the next integer):

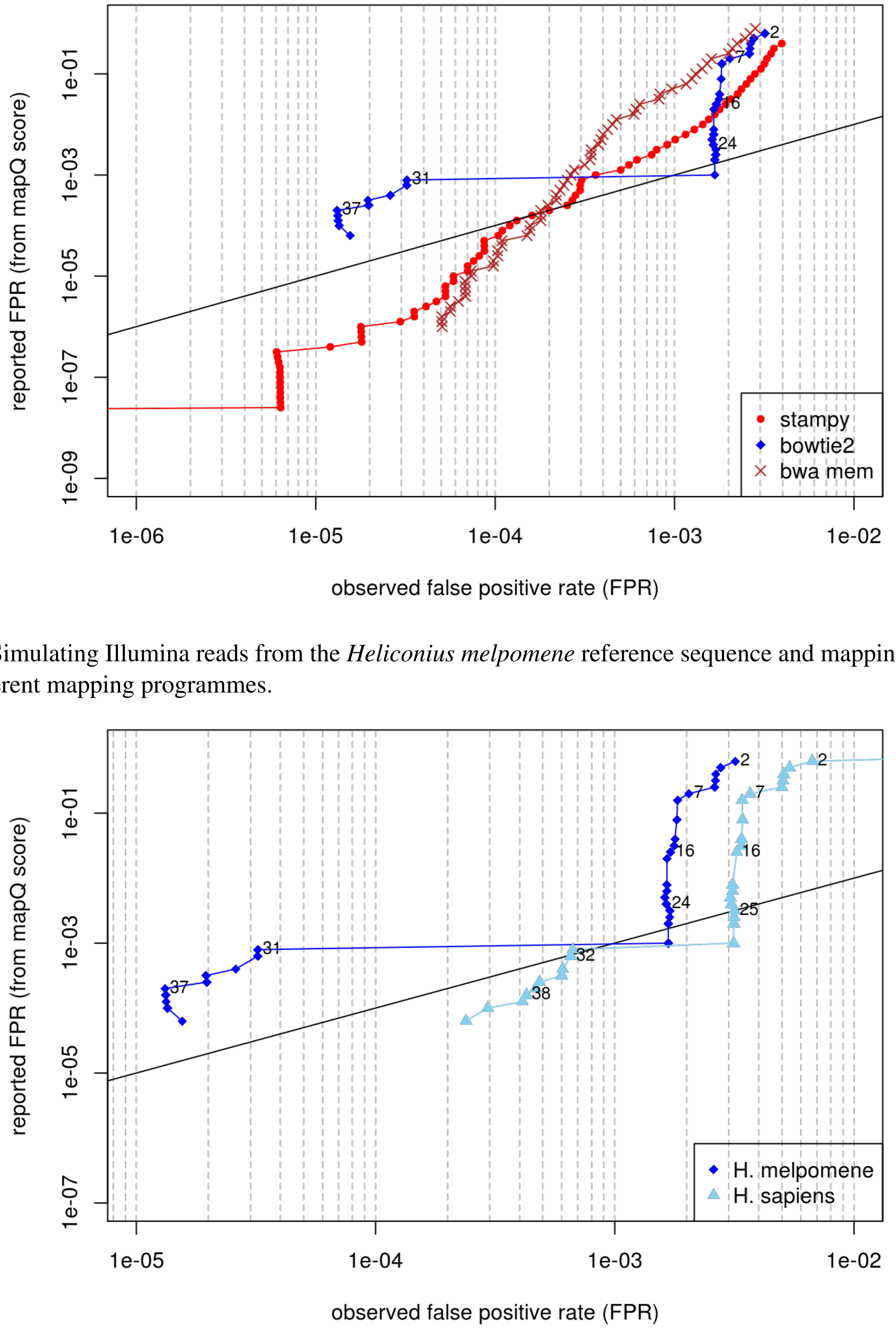
$$\text{mapQ} = -10log_{10}p$$

True mapping position is meant with respect to the provided *reference sequence*, i.e. assuming the true reference sequence is known without error. I have simulated 200,000 SE reads of 50 bp length from the *Heliconius melpomene* genome reference (Hmel1-1_Release_20120601) with the programme mason (version 0.1.2) (Holtgrewe, 2010). The *H. melpomene* reference sequence has a length of 269,658,870 bp. I have also simulated reads from the much bigger (and probably more repetitive) human genome reference sequence (version GRCh38). The human reference sequence is 11.5 times larger than the *H. melpomene* reference sequence (3,099,750,718 bp). I simulated reads without sequencing error but with a polymorphism rate of 0.01, 20% of which were indels. Indels could be up to 20 bp long and their length was drawn from a uniform distribution. bowtie2 was run in -very-sensitive and -end-to-end mode, bwa mem and stampy were run with default settings. In order to extract counts of mapped and incorrectly mapped reads from SAM files, I used the programme wgsim_eval.pl from the read simulation package wgsim of Heng Li. A read is mapped correctly if its reported mapping position is within 50 bp of the true origin. Further details can be found in Mason_sim.sh and Mapping_Tool_Test.

The observed rate of mismappings of these simulated reads should be an estimator of the true mapping quality p. I can therefore compare how well the reported mapping qualities of the different mapping programmes correspond to the observed false positive rate. Figure 3.20 shows that bowtie2, generally underestimates the true mapping quality, given the reference sequence, when simulating from the *H. melpomene* reference. When mapping against the more repetitive human reference bowtie2 underestimates mapping quality for quality scores below 25 and slightly overestimates mapping quality for reads with higher quality scores assigned.

3.4.2 Mappability across the de novo reference assembly

Mappability is a measures that can be computed for a single nucleotide position x in a reference sequence. It assumes that the reference sequence is known without error. Starting at position x a sequence of length k is extracted from the reference sequence. This kmer



(a) Simulating Illumina reads from the *Heliconius melpomene* reference sequence and mapping with different mapping programmes.

(b) Simulating Illumina reads from *Heliconius melpomene* and *Homo sapiens* reference sequence and mapping with `bowtie2`.

Fig. 3.20 Comparison of reported and observed false positive mapping rate (FPR). The reported FPR is the mapping quality score transformed back into a probability. The observed FPR is the proportion of all mapped reads with a specific mapping quality score that have an incorrect mapping position reported by the mapping programme. The diagonal line indicates 1:1 correspondence. For points above the diagonal the mapping programme underestimates the true mapping quality. For `bowtie2` several points have been annotated with the reported mapping quality score of the reads they represent.

is then searched for in the whole reference sequence allowing for up to m mismatches (or edits, i.e. including indels). The number of times the kmer is found in the reference is $F_{k,m}$. Mappability M is then defined as:

$$M_{k,m}(x) = \frac{1}{F_{k,m}(x)}$$

I have used programmes in the tool package GEM (Derrien et al., 2012) together with programmes from the tool package `BEDOPS` (Neph et al., 2012) to compute the average mappability over the positions in each contig (see `assembly.sh` for further details). I specified a kmer length of 40 and allowed up to 3 edits during the search for mapping positions in the Big Data reference assembly. The search did not include the reverse complement of the reference. Average mappability per contig was computed for all contigs in the complete unfiltered Big Data reference assembly and the subset of contigs which passed all filters, i.e. which included at least one interval after filtering that was kept for downstream analysis. The difference in location between the distributions in figure 3.5 was tested with the `R` function `wilcox.test`.

3.4.3 Filtering against deviation from HWE

I filtered out contigs that had a variable site with a *negative* inbreeding coefficient F_{IS} at the significance level of 0.05 from an LRT (Vieira et al., 2013). This was intended as a measure to reduce false-positive SNP's due to paralogous sequences mapping to the same position in the de novo reference assembly. Sites with a *positive* within population F_{IS} could be affected by allele drop out (due to polymorphisms in the restriction site), they could be affected by PCR drift or they could map to the X chromosome, since all sequenced individuals are males (and there is no homologous male sex chromosome in *Chorthippus parallelus*, Gosálvez et al. (1988)). 38% of contigs from *erythropus* and 48% of contigs from *parallelus* have a SNP with significantly positive F_{IS} at the 0.05 level. But note that Vieira et al. (2013)'s method for estimating the per-site inbreeding coefficient from genotype likelihoods shows a strong upward bias when used with very low coverage data (1×) and few individuals sequenced (10) (see their figure 1). Further analysis reveals a strong correlation of the p-value from the LRT with the MAF of the SNP (see figure 3.21). I have therefore opted not to filter contigs against *positive* F_{IS} of their SNP's. This filter would have preferentially removed SNP's (and their contigs) with high MAF and would therefore have distorted the site frequency spectrum which most analyses in this chapter are based on. Vieira et al. (2013)'s method jointly estimates genotype frequencies, minor allele frequency and per-site inbreeding coefficient. It is therefore not surprising that a site with greater MAF provides more power to detect an excess of homozygote genotypes as compared to HWE expectations. All SNP's with an excess of heterozygotes show very high MAF even for p-values as high as 0.1. This must be due to the fact that genotype frequencies and MAF are confounded parameters: a high estimate for heterozygote frequencies cannot go together with a low estimate for MAF. It is likely that an increase in sample size would increase the power to detect an excess of heterozygotes (Vieira et al., 2013).

3.4.4 Removing PCR duplicates

I have removed PCR duplicates by collapsing read pairs into unique representatives per individual with `starcode` allowing for an edit distance of 2. I have applied the same filters as before, except for the addition of positive F_{IS} filtering (which removed 225 contigs) and requiring only at least 1× coverage in at least 15 individuals for each included reference site. I only counted reads with a mapping quality score of at least 5. These filters retained 2,455,851 sites on 48,556 contigs. The average per site and per individual coverage is 1.14×. The across sample coverage per site has a modal value of 25 (see fig. 3.22). The individual coverage distributions show a median coverage of 1× for all *erythropus* individuals and a median of 0× for most *parallelus* individuals (fig. 3.23). For some more details see the notebook DEDUP.

As before, for the estimation of 1D site frequency spectra I required read data from at least 9 (of 18) individuals to include a reference position. Unfortunately, the *Big Data* set without PCR duplicates does not provide enough information for the estimation of allele frequency spectra (fig. 3.24). Different minimum coverage filtering for at least 3× in 10 or 15 individuals did not lead to improved results. It could be that the low coverage of the deduplicated read data does not allow the distinction between polymorphisms and sequencing errors.

3.4.5 PCA with SNP calling, normalisation and genotype calling

The PCA in figure 3.8 on page 86 has been done on a covariance matrix that has been estimated from genotype likelihoods without SNP and without genotype calling. Additionally, no normalisation of the genotypic covariances by the binomial variance of allele frequency has been applied. In the following I am going to show PCA's from covariance matrices that have been estimated with SNP calling, normalisation and genotype calling. SNP's were called with a LRT of the MAF being 0 at a p-value threshold of 0.001. Genotypic covariances were normalised by $p_i(1-p_i)$ as in eq. 19 in Fumagalli et al. 2013.

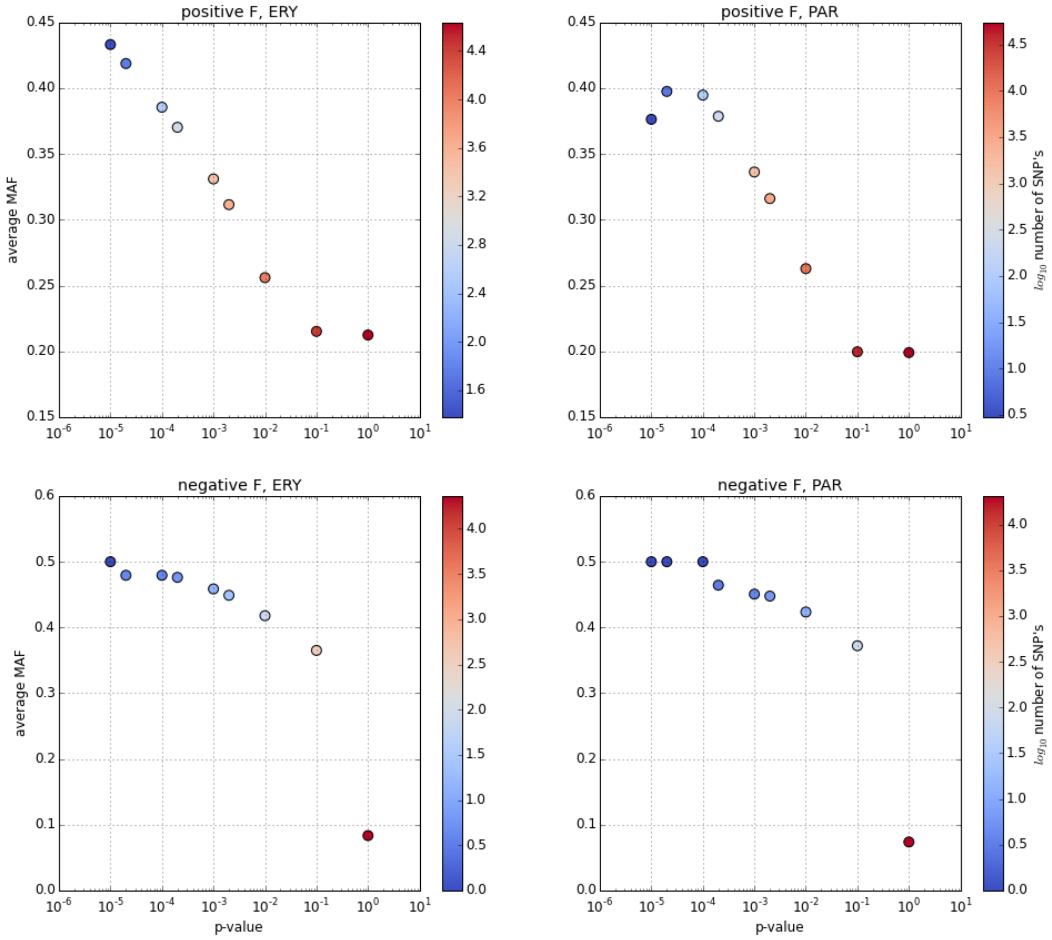


Fig. 3.21 Dependence of the p-value from an LRT test of HWE on the MAF of the SNP. The upper two panels show SNP's with $F > 0$ at the given significance level. The lower two panels show SNP's with $F < 0$. SNP's were detected with a LRT of the MAF being 0.0 with a significance level of 0.01.

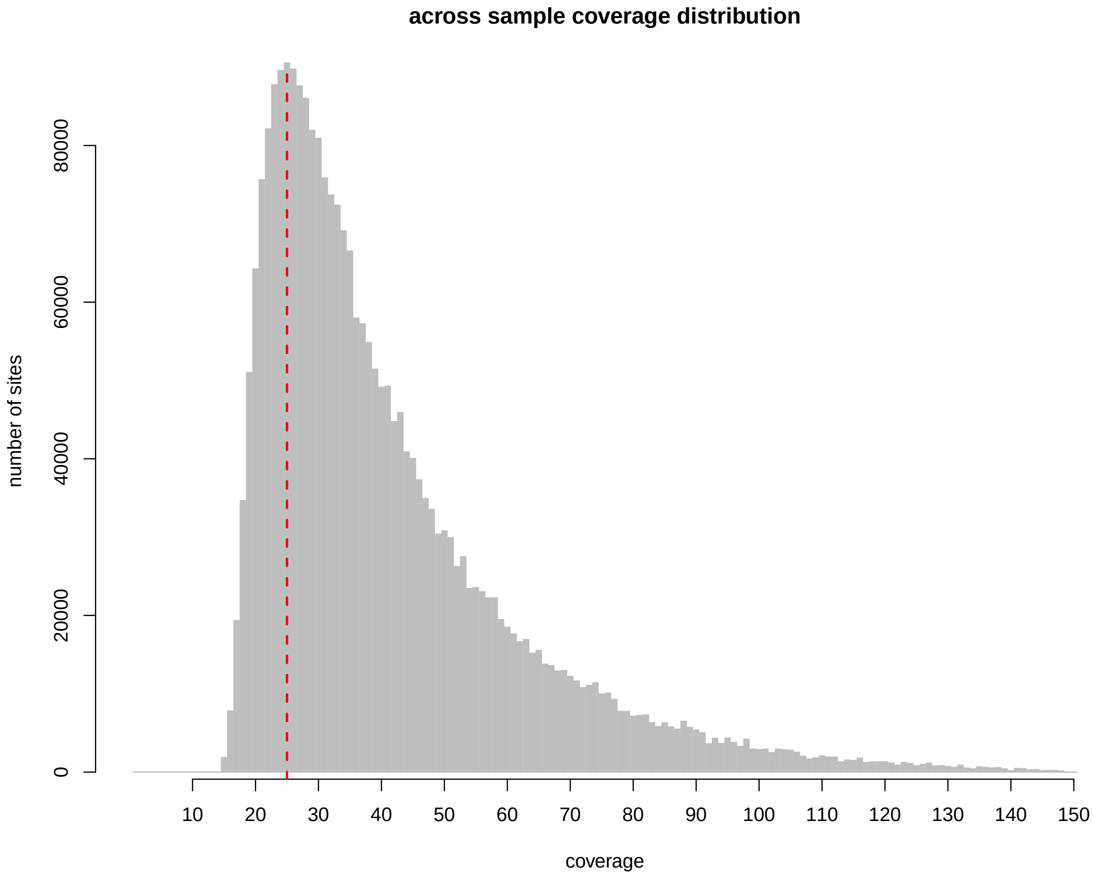


Fig. 3.22 Across sample (i.e. global) coverage distribution with de-duplicated read data after filtering (excluding HWE filtering which removed 3,486 contigs). The distribution is based 2,683,395 sites on 52,042 contigs. The red dashed line marks the modal coverage. Compare with figure 3.2 on page 71.

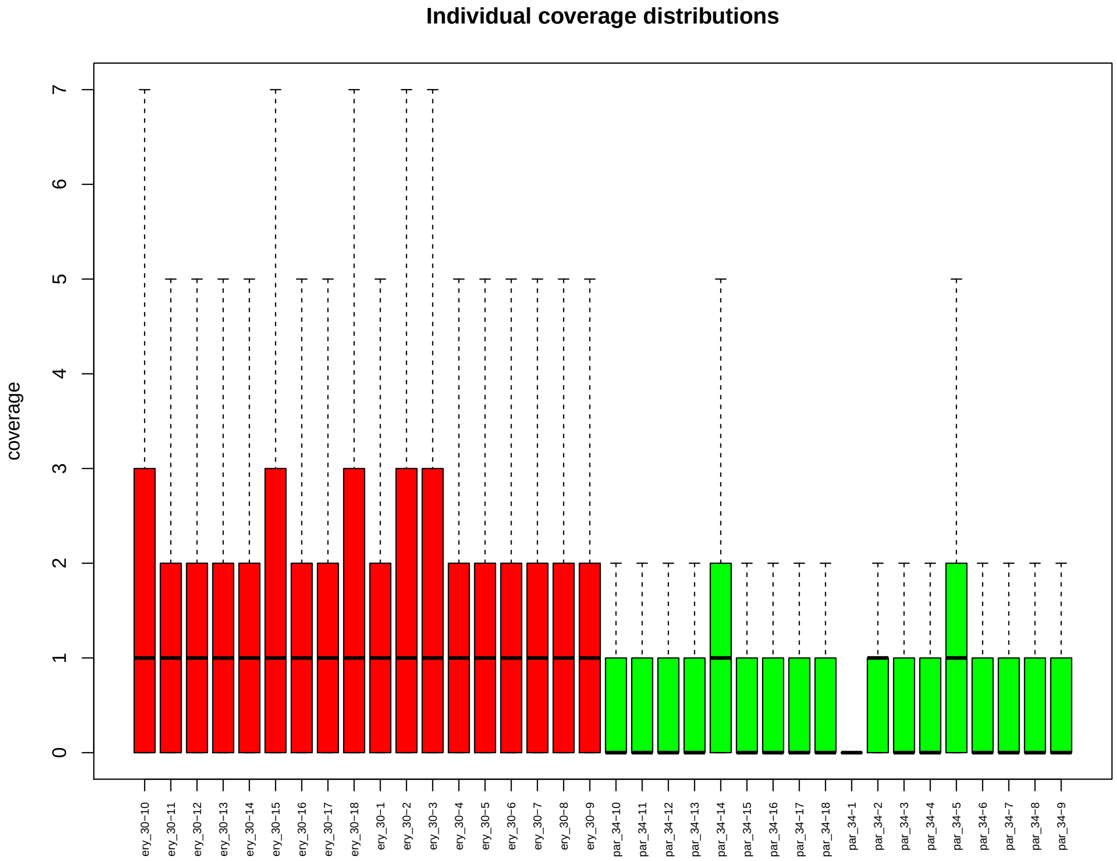


Fig. 3.23 Individual coverage distributions from de-duplicated read data after filtering (excluding HWE filtering which removed 3,486 contigs). The distribution is based 2,683,395 sites on 52,042 contigs. Compare with figure 3.3 on page 72.

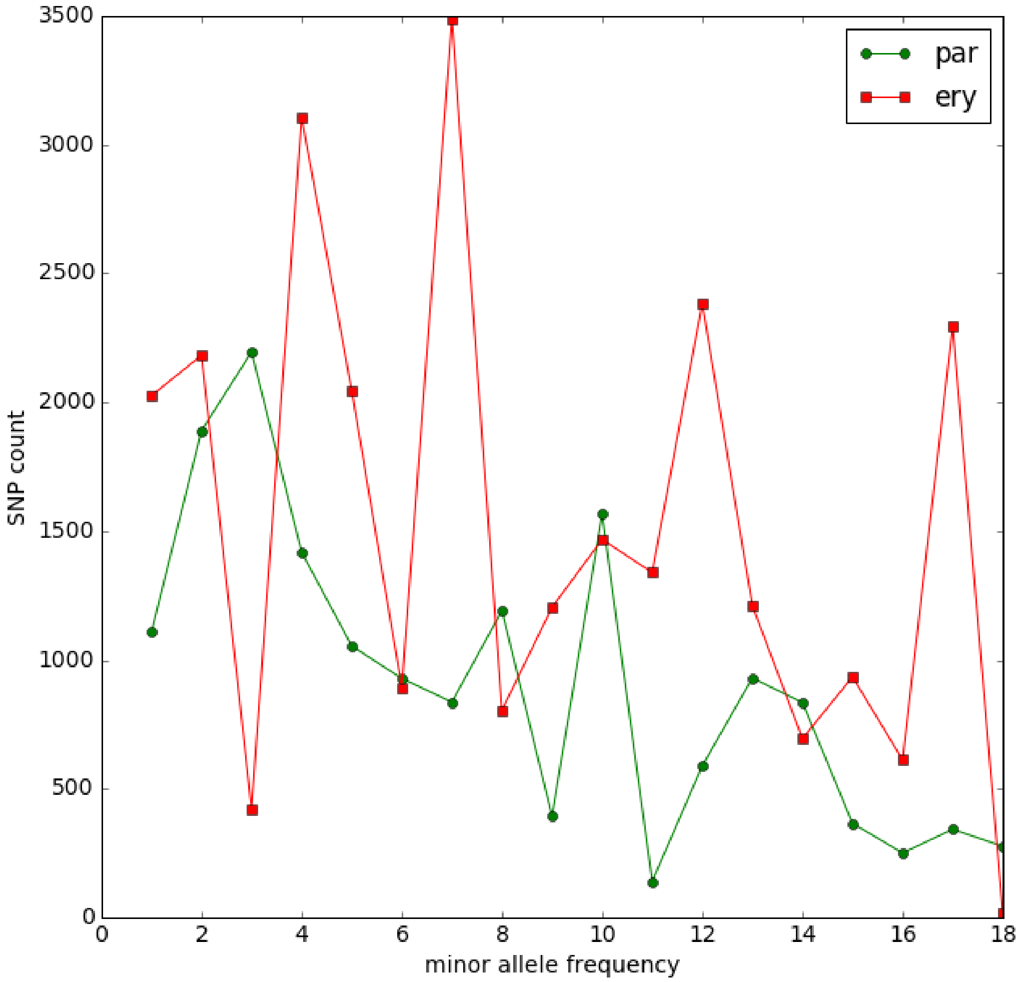


Fig. 3.24 Folded 1D site frequency spectra for *erythropus* and *parallelus* from de-duplicated read data.

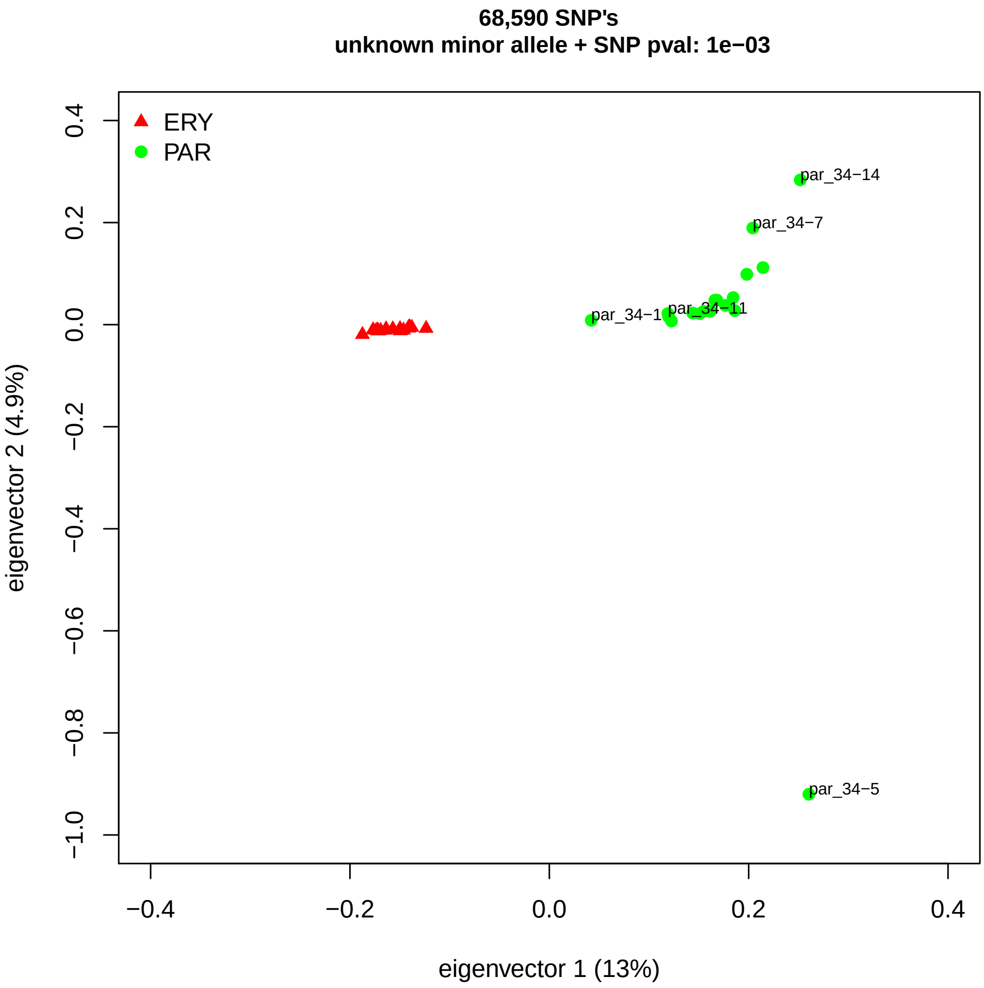


Fig. 3.25 PCA with SNP calling and normalisation.

Figure 3.25 shows the plot of the first two eigenvectors of a PCA with SNP calling and normalisation. SNP calling finds 68,590 SNP's. When comparing with the PCA from figure 3.8 on page 86 a general similarity in the relative coordinates between individuals is apparent. However, the first eigenvector in figure 3.25 explains only 13% of the total genotypic variation in the data compared to 23% in the PCA without SNP calling (fig. 3.8). Figure 3.26 shows the plot of the first two eigenvectors of a PCA with genotype calling. The genotype with the maximum posterior probability was called. Again, when compared to the PCA plot in figure 3.8 a general similarity of the relative coordinates between individuals is apparent (except for the switch in sign on the second axis, which is chosen randomly by the `prcomp` function in `R`). With genotype calling, the proportion of total variance explained by the first eigenvector is reduced even further to just 11%.

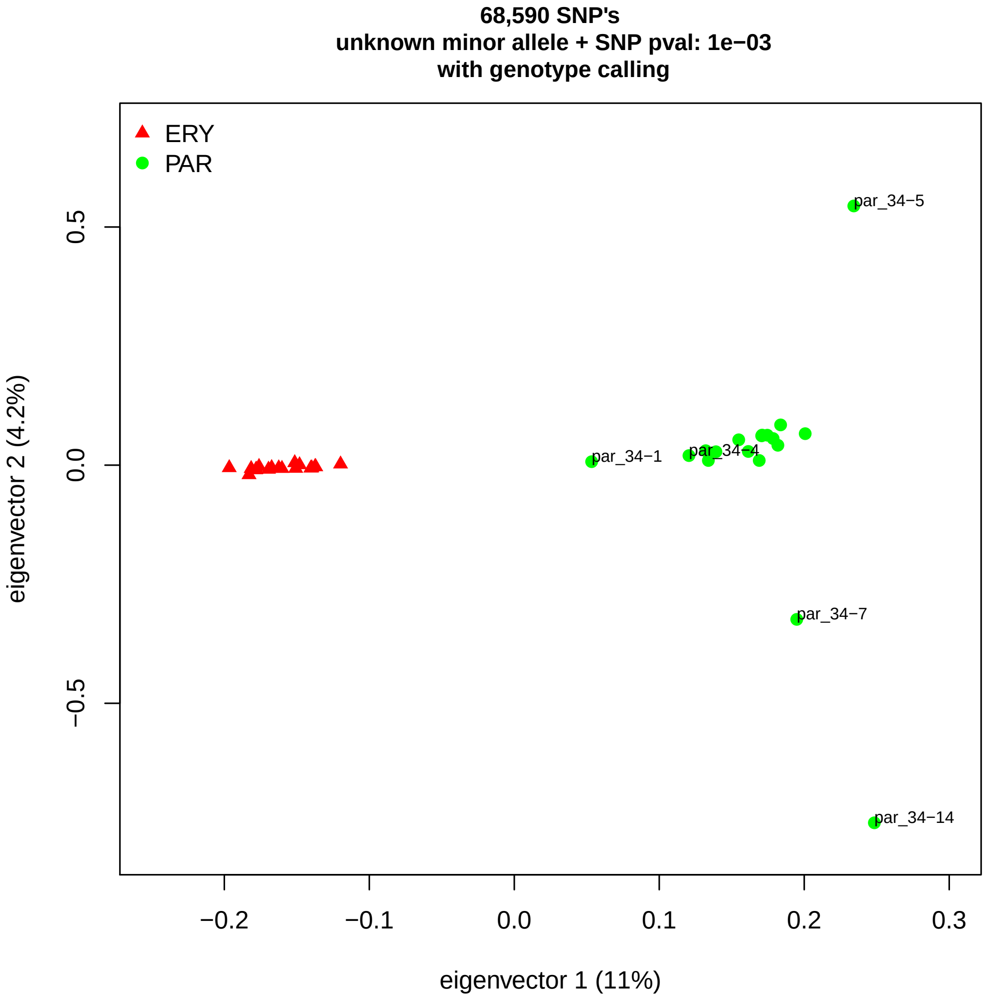


Fig. 3.26 PCA with SNP calling, normalisation and genotype calling.

These different PCA's all reveal two clusters of individuals on the first axis of variation. These clusters coincide with population label and are clearly separated on the first axis which explains from 11% to 23% of the total genotypic variation among individuals. The clear genotypic distinction of *erythropus* and *parallelus* is therefore robust to the way the genotype covariance matrix is computed.

3.4.6 F_{ST}

`realSFS` can also estimate Reynolds' F_{ST} (Fumagalli et al., 2013, eq. 1–3). Reynolds' F_{ST} includes a weighting by sample size. The genome-wide Reynolds' F_{ST} is 0.308. That is 0.01 higher than Hudson/Bhatia's version of F_{ST}.

F_{ST} is a summary statistic of the joint site frequency spectrum from two or more populations. The 2D-SFS in figure 3.7 on page 78 has been estimated from the 1,130,775 sites that are overlapping, i.e. sites with sequence reads from at least 9 individuals in each population. I have used the function `Fst` in $\delta a \delta i$ (Gutenkunst et al., 2009) to calculate Weir & Cockerham's F_{ST} (Weir and Cockerham, 1984, eq. for $\hat{\theta}$ at top of p. 1363) from this 2D-SFS. This version of F_{ST} should be very similar to Reynolds F_{ST}[3]. Both assume equal amounts of drift experienced by both populations and if this assumption is violated, they become dependent on the ratio of sample sizes. In contrast, Hudson/Bhatia's F_{ST} does not make this assumption and is independent of sample sizes (Bhatia et al., 2013). Here, however, sample sizes are the same (18). So I do not expect both versions of F_{ST} to differ because of weighting by sample size. The genome-wide F_{ST} as calculated from this 2D-SFS is 0.265 (see section "Fst" in `05_2D_models.ipynb`). This is 0.043 below Reynolds' F_{ST} as estimated from 1.6 million sites with `realSFS`.

Hudson/Bhatia's genome-wide F_{ST} has shown a bias of 0.025 as determined by the median of the empirical null distribution estimated by 100 permutations of population label. I have also estimated the empirical null distribution of Weir & Cockerham's F_{ST} estimated from the 2D site frequency spectra estimated from sample allele frequency files with permuted population labels. The median of this distribution indicates a bias in the genome-wide F_{ST} estimate of 0.037 (fig. 3.27).

Weir and Cockerham (1984), p. 1366, propose a correction of bias based on jackknife resampling of loci. I have created all delete–1 jackknife resamples over 32,706 contigs and used them for bias correction with the following formula:

$$F_{ST_{corr}} = nF_{ST} - \frac{n-1}{n}\sum_{i=1}^{n} F_{ST_i}$$

[3] this has not been fully verified due to the complexity of the formulas

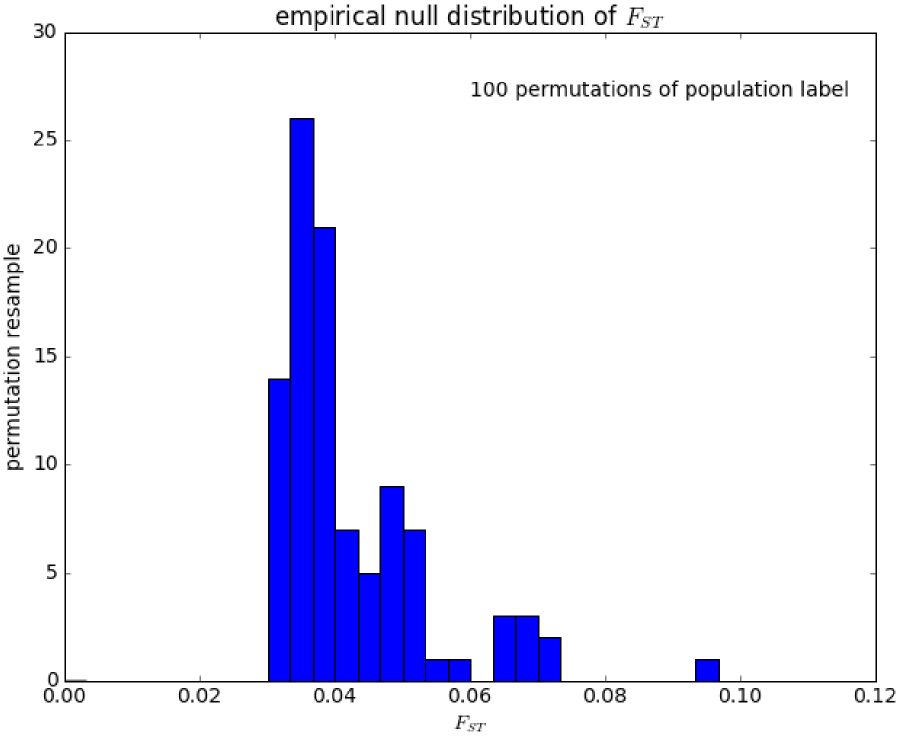


Fig. 3.27 Empirical null distribution of Weir & Cockerham's F_{ST} estimated in $\delta a \delta i$ from the 2D-SFS's of 100 permutations of population label.

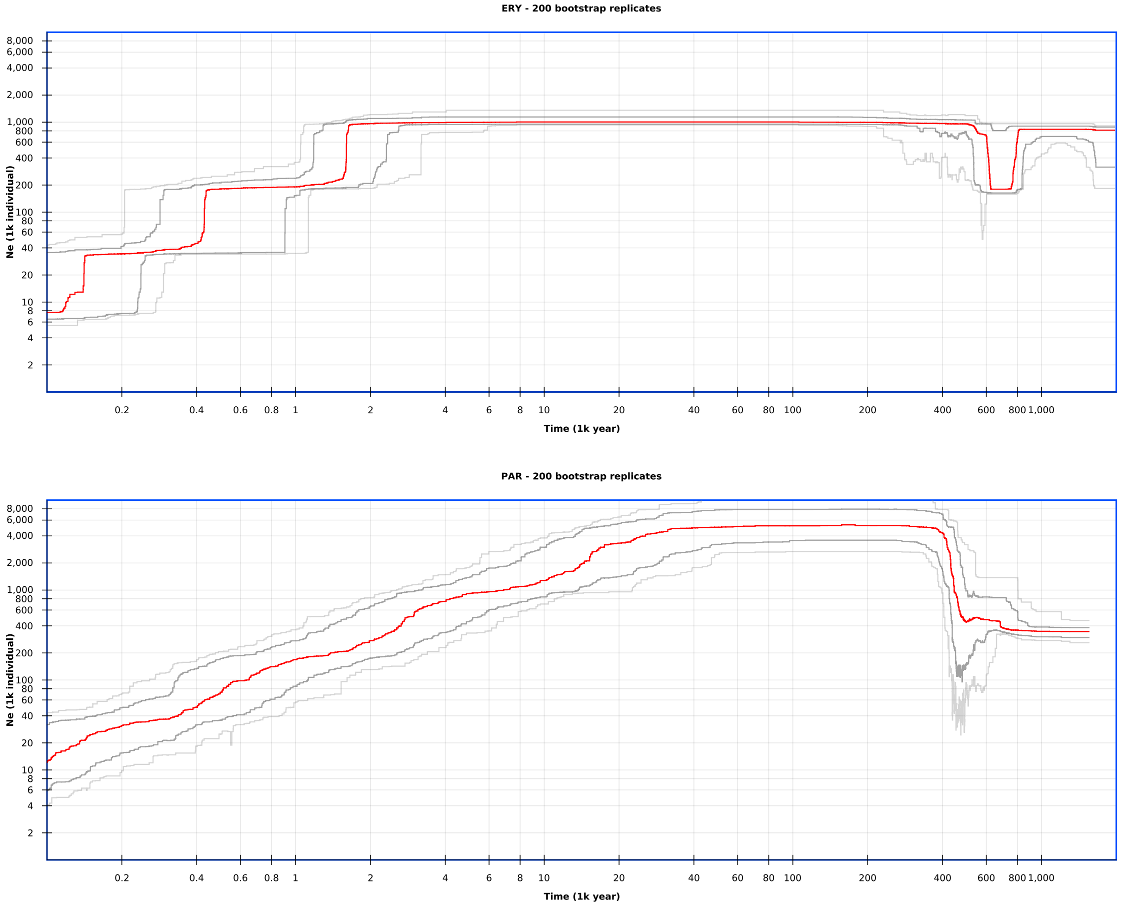


Fig. 3.28 `stairway plot` for *erythropus* and *parallelus* from including non-overlapping sites. The thick grey lines define the 75% bootstrap-CI and the light grey lines define the 95% bootstrap-CI. Compare with figure 3.18 on page 102.

F_{ST_i} is the F_{ST} from the i-th jackknife resample. This leads to only a negligible bias correction of -3.5e-06.

3.4.7 Stairway plots

The stairway plots in figure 3.18 have been created from only overlapping sites, i.e. exclusively from sites that had 9 individuals with read data in both populations. Figure 3.28 shows the corresponding stairway plots when this restriction is released. When comparing with figure 3.18, the stairway plot for *parallelus* is hardly changed. The stairway plot for *erythropus* is also very similar to the one in figure 3.18. The ancient population size is estimated to have been slightly higher, about 1.0 million, and the current population size is estimated to be about 8,000 instead of above 20,000.

3.4.8 Distinguishing reduced divergence time from increased migration

The figures 3.29 and 3.30 show a series of expected two-dimensional site frequency spectra simulated with $\delta a \delta i$ for different divergence times and migration rates. They show that qualitatively quite similar expected spectra can be generated when either reducing divergence time or increasing the migration rate. N_a stands for the inferred common ancestral population size of the two populations, which was calculated by first computing Watterson's θ from the observed 2D-SFS (see eq. 3.9 on page 81) and assuming $\theta_W \simeq 4 N_a \mu L$. With the mutation rate μ set to 3×10^{-9} and the total sequence length L set to 1,130,775, this results in an estimate of N_a of 1,126,030. A divergence time of $0.222 \times 2N_a$ therefore corresponds to 500,000 generations. Migration rates reported in figure 3.30 are scaled by $2N_a$, i.e. divide by $2N_a$ to get m, the proportion of new immigrant individuals per generation.

Figure 3.31 compares the best fitting model spectra from $\delta a \delta i$ for the *divergence-in-isolation* model and the *divergence-with-migration* model. The greatest qualitative difference between the two model spectra is the much higher number of expected low frequency shared polymorphisms in the *divergence-with-migration* model spectrum and the higher expected number of high frequency shared polymorphisms in the *divergence-in-isolation* model spectrum. The model parameters for these two spectra are given in table 3.2 on page 97 and 3.3 on page 98, respectively.

Figure 3.32 shows the plot of Poisson residuals between the best fit model spectra in figure 3.31. The large residuals for shared polymorphisms that are at low frequency in one or both populations is a signature of gene flow (Gutenkunst et al. 2009, suppl. mat. section 1.1).

3.4.9 Miscellaneous results from demographic modelling

Figure 3.33 shows the best fitting model spectra from the *asymmetric-migration* model and the *two-epoch-with-exp-size-change* model inferred with $\delta a \delta i$ together with plots of the Poisson residuals between them. The plot of Poisson residuals in the bottom left indicates that the main difference between the two models is in how they fit low frequency variants private to *parallelus*. Of these variants, the *two-epoch-with-exp-size-change* model predicts far fewer singletons and many more doublets, triplets, etc.

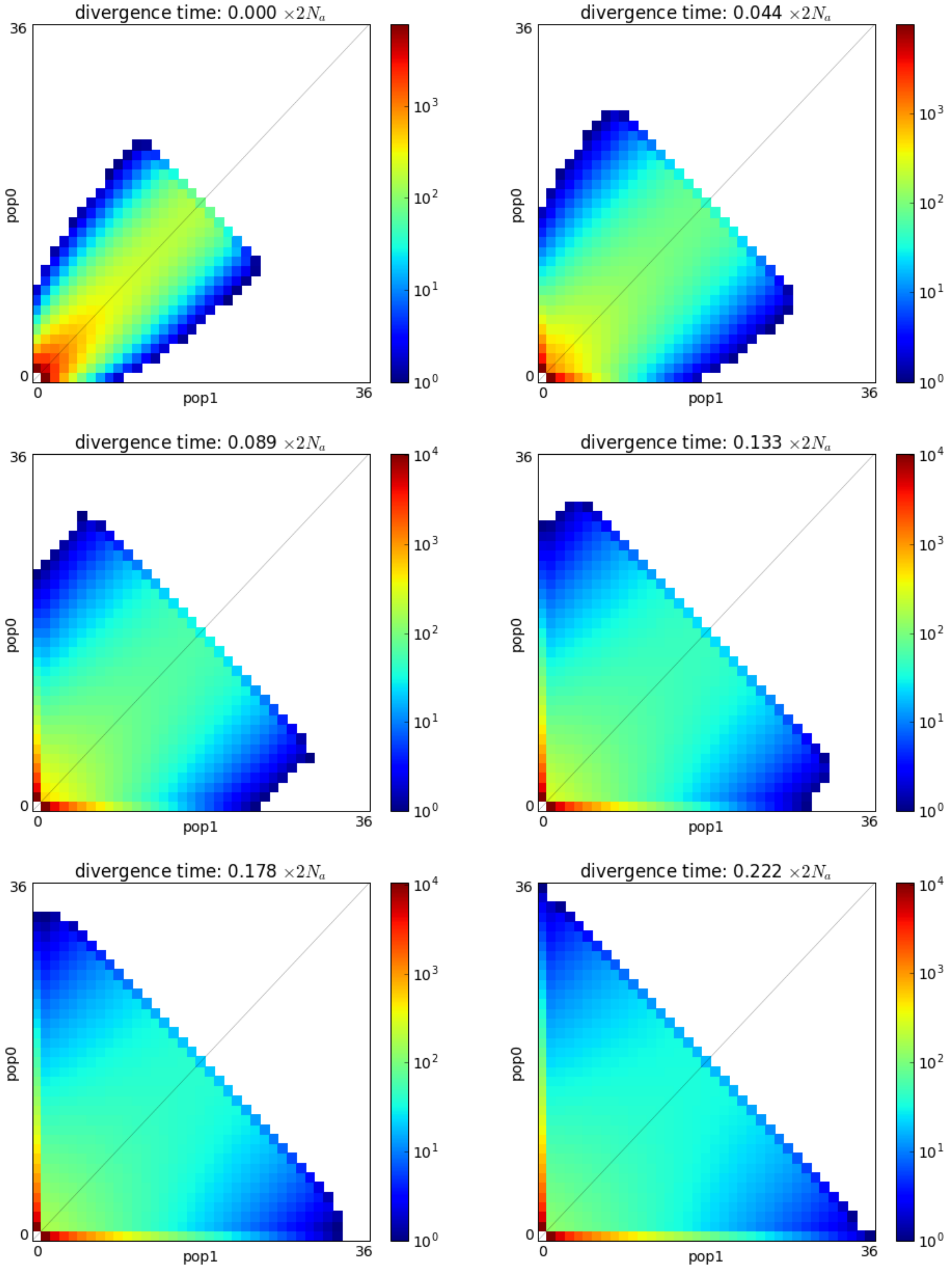


Fig. 3.29 Expected two-dimensional site frequency spectra for an increasing series of divergence times without migration.

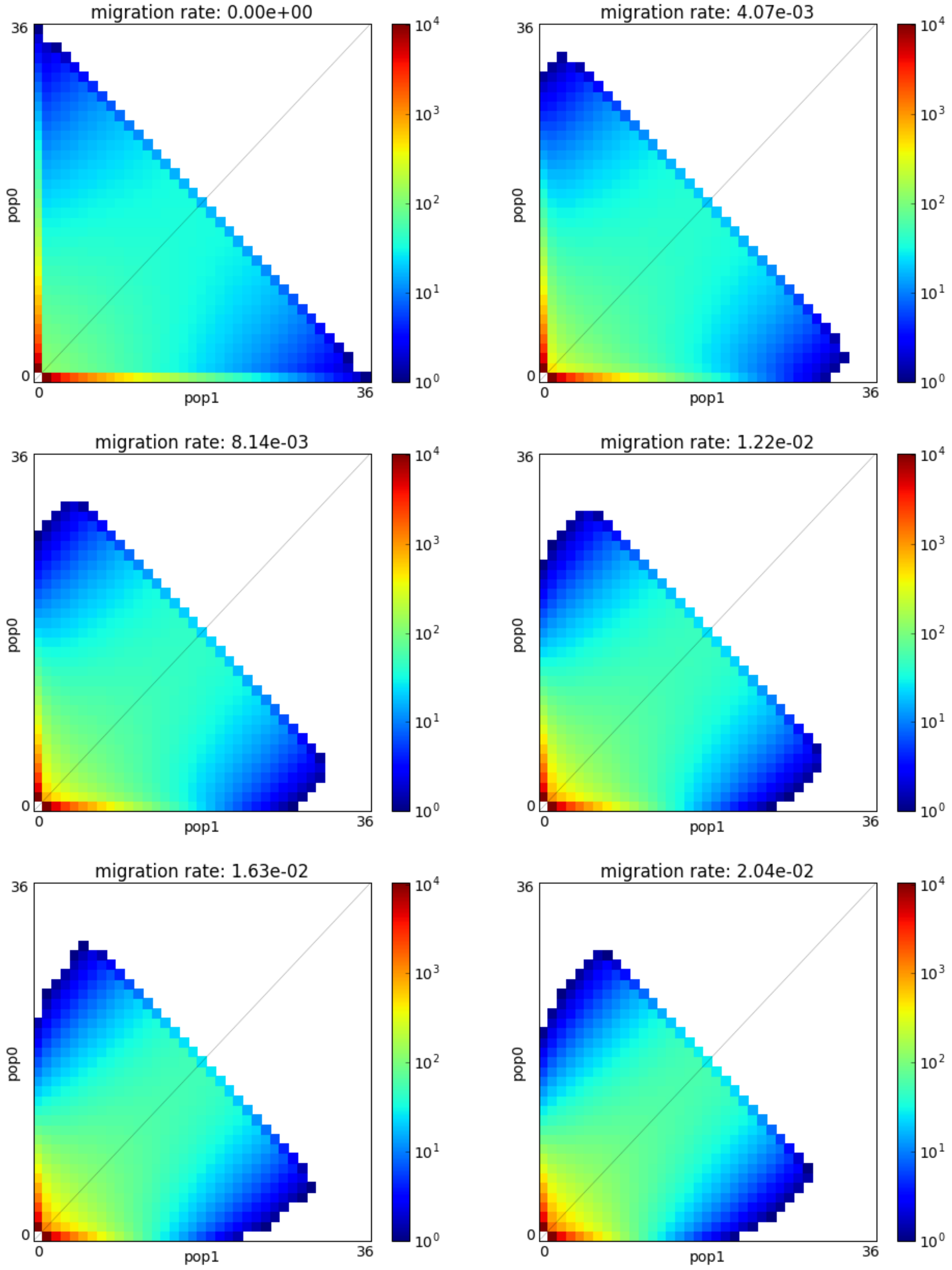


Fig. 3.30 Expected two-dimensional site frequency spectra for an increasing series of migration rates and with a fixed divergence time of 500,000 generations.

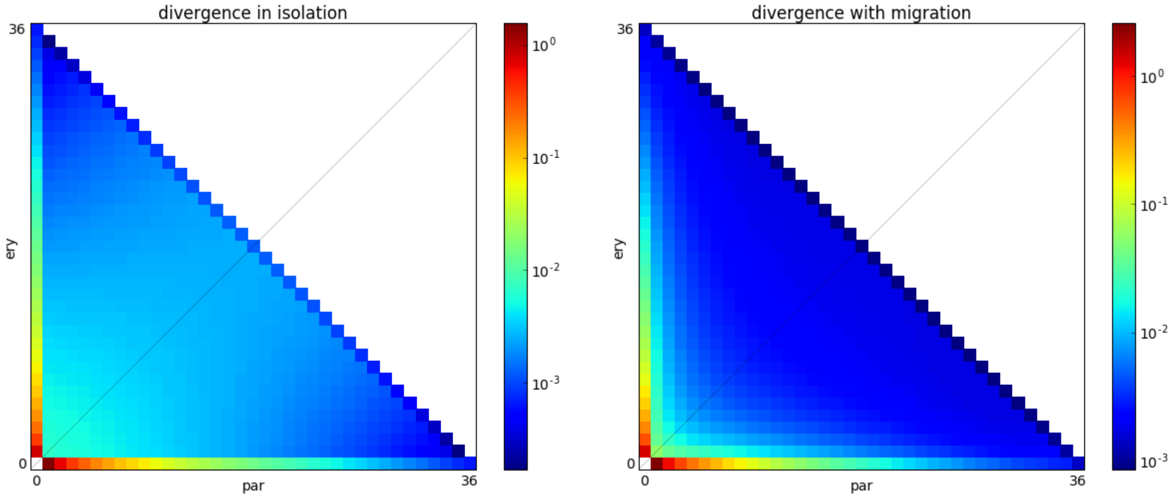


Fig. 3.31 Best fitting model spectra from $\delta a \delta i$ for the *divergence-in-isolation* model (left) and the *divergence-with-migration* model (right).

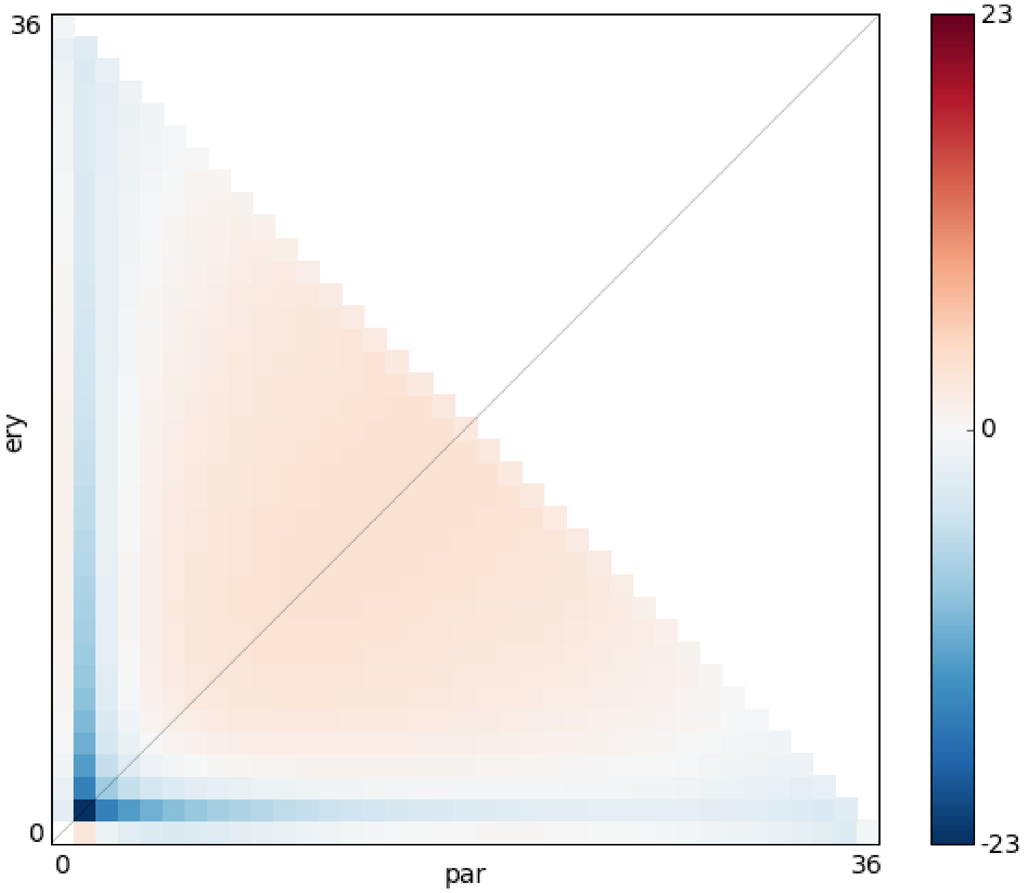


Fig. 3.32 Plot of Poisson residuals between the best fit model spectra in figure 3.31 (each scaled by the optimal θ).

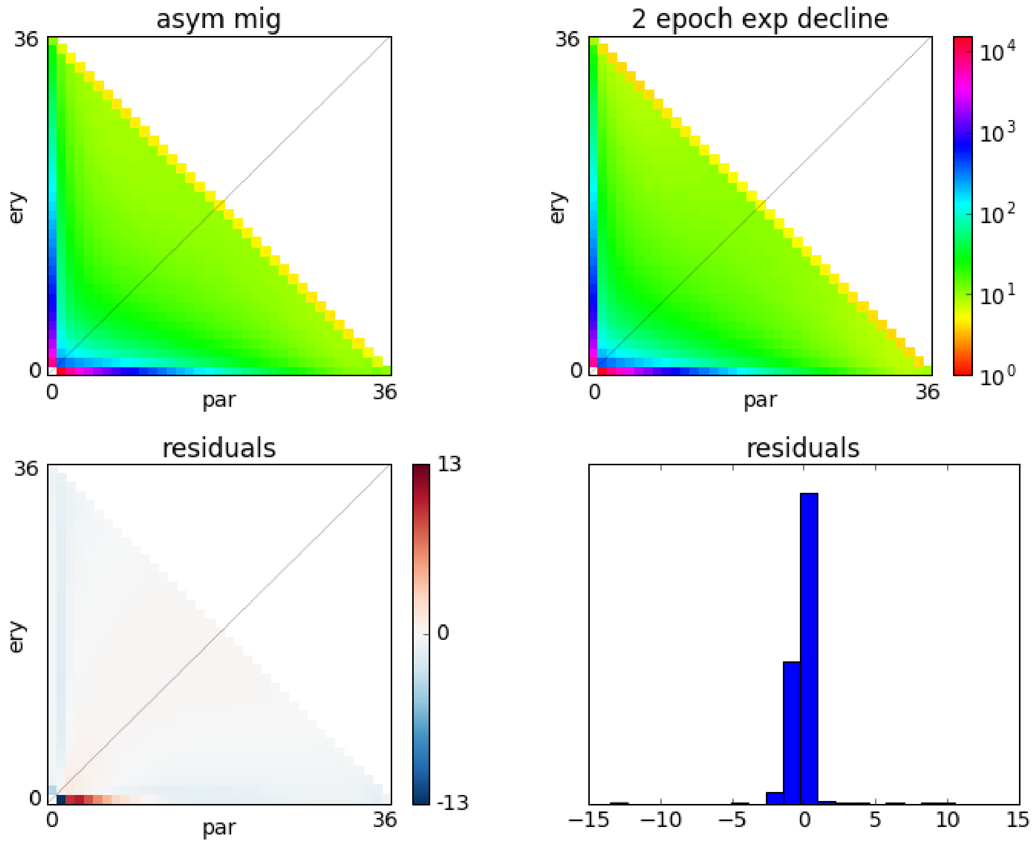


Fig. 3.33 Top: plots of the best fitting model spectra of the *asymmetric-migration* model and the *two-epoch-with-exp-size-change* model. Bottom: the Poisson residuals between the two best fitting model spectra (each scaled by the optimal θ).

Chapter 4

General Discussion

The two subspecies *erythropus* and *parallelus* have a high genome-wide average genetic differentiation (F_{ST}: 0.298, 95% bootstrap CI: 0.294 – 0.303). The net sequence divergence D_a of 0.003, however, clearly marks them as a semi-isolated species pair in the "grey zone" of the speciation continuum (Roux et al., 2016). The two subspecies have not diverged in complete allopatry. Instead there is a robust signal of ancient and asymmetric gene flow during the history of the divergence of the two subspecies (see tab. 3.4). Allowing for gene flow doubles the divergence time estimate from about 0.5 to 1.1 mya. Gene flow had been about 5 times higher in the direction from *parallelus* to *erythropus* than vice versa. This is in line with many previous studies of the Pyrenean hybrid zone that indicated an asymmetry of gene flow in the same direction (see chapter General Introduction).

Gene flow was apparently low enough and positive selection for alleles causing DMIs high enough to allow the evolution of DMIs between the two subspecies (Bank et al., 2012). The effective number of haploid migrants per generation estimated by the *two-epoch-exp.-size-change* (fig. 3.17) and *ancient-asymmetric-migration* models (fig. 3.19) are much less than 2 ($2Nm \ll 2$). A population scaled migration rate for *diploid* individuals much less than 1 allows the near fixation of alternative alleles by genetic drift (fig. 4.1) (Slatkin, 1987; Slatkin and Barton, 1989). Note that the finite island model predicts a far lower F_{ST} from the estimated migration rates in $\delta a \delta i$ than the F_{ST} calculated from the observed 2D-SFS (cf. fig. 3.9 and section 3.4.6 on page 117). Unless there are unmet assumptions of the finite island model and unknown biases in the data, which I am unaware of, this suggests that genetic drift alone cannot fully account for the genome-wide divergence of the two subspecies and divergent selection and/or selection against hybrids may need to be invoked to explain a large part of it. In addition, the clear evidence of ancient gene flow between the two subspecies also indicates that the alleles that cause hybrid male sterility cannot be of ancient origin and have spread solely via genetic drift (Bank et al., 2012; Gavrilets, 1997), as

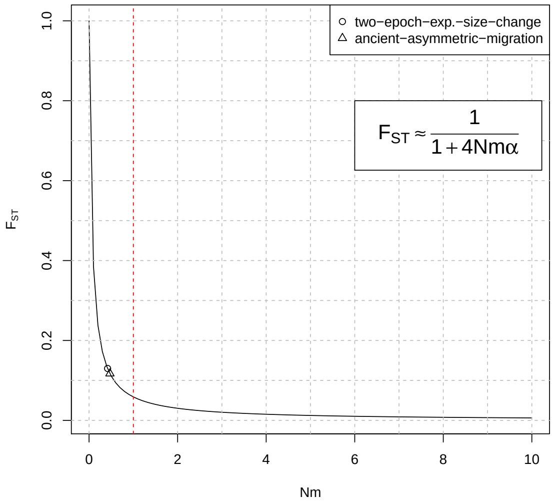


Fig. 4.1 Relationship between genetic differentiation (F_{ST}) and the effective number of diploid migrants per generation (Nm) in the finite island model (Hudson et al., 1992). $\alpha = [n/(n-1)]^2$, n is the number of subpopulations and has been set to 2. F_{ST} increases sharply for $Nm < 1$ (left of the red dashed line). The inferred population scaled migration rates (both directions combined) from the two best-fit models in $\delta a \delta i$ are plotted on the graph for comparison.

could be assumed given that they are not fixed in the two subspecies (Shuker et al., 2005). The only possibility for the evolution of a neutral DMI would be if sterility had evolved in complete allopatry during the last ice age and there had been effectively no gene flow since the secondary contact (that could have eroded that DMI) as suggested by the *ancient migration* model in $\delta a \delta i$. However, as mentioned before, the current data set probably lacks the power to detect very low and very recent gene flow due to low sample sizes (Robinson et al., 2014). So, in order to exclude the possibility of neutral DMIs underlying the sterility in hybrids of the two subspecies, the demographic modelling from chapter Investigation into the demographic history of the hybrid zone would have to be repeated with a much greater sample size. Probably more than 100 individuals per population would be required to be able to detect recent, low gene flow. Very low gene flow on the other hand would be less effective at eroding the evolved DMI.

The *parallelus* population had a much greater ancient population size than *erythropus*: at least twice as high according to δaδi and four times as high according to `stairway-plot`. There seems to be a signal of a recent (postglacial) drastic bottleneck in *parallelus* as would be expected from serial founder events during range expansion from a Balkan glacial refuge. However, it seems difficult to reconcile this with the strongly negative Tajima's *D* for *parallelus* (tab. 3.1), which would indicate a population size expansion. This discrepancy remains an open question that should be resolved. The signal for a population bottleneck in *erythropus* is more ambiguous. While `stairway-plot` infers a strong and very recent bottleneck from the 1D site frequency spectrum (fig. 3.18), δaδi infers only a mild bottleneck from the 2D site frequency spectrum (fig. 3.17). A less severe bottleneck would be consistent with the expected lower effect of serial founder events during the expansion of *erythropus* from its glacial refuge in southern Spain to the Pyrenees.

A primary contact model (i.e. with ancient gene flow) fits the data better than a secondary contact model. That means that there is no significant signal of recent gene flow in the genome-wide 2D-site frequency spectrum. This also indicates that the detected asymmetry of gene flow between *erythropus* and *parallelus* cannot just be a consequence of different distances of the sampled populations to the hybrid zone centre (fig. 2.1) or hybrid zone movement. Rather, this hints at intrinsic mechanisms for asymmetric isolation like those described in chapter General Introduction.

Underlying the analyses of demographic history is the assumption that the vast majority of RAD markers are selectively neutral. The analysis may therefore be improved by removing outlier loci that may be evolving under selection (Beaumont and Nichols, 1996; Whitlock and Lotterhos, 2015). An F_{ST} outlier scan could be attempted with this data set to detect RAD loci potentially evolving under divergent selection in the two subspecies. These may or may not be linked to loci involved in DMIs. However, given the low number of informative SNPs (less than ~74,058, Roesti et al. 2012) spread over a low number of RAD loci (34,343) in the current data set when compared with the genome size (10–14 Gbp), it is likely that only a subset of the genomic regions under selection could be detected (Lowry et al., 2016). In addition, the large CIs of the 1D site frequency spectra (see fig. 3.14) already indicate great uncertainties in allele frequency estimates of individual sites. Due to these large uncertainties in allele frequencies, a method for F_{ST} outlier detection needs to be employed that takes this uncertainty into account, like the Bayesian hierarchical F-model of Gompert and Buerkle (2011b). This detects outliers based on an empirical distribution of per-locus F_{ST}. If the data set contains many thousands of unlinked loci and the vast majority of them are neutral and only weakly affected by linked selection, then this empirical distribution should provide a good null distribution (Whitlock and Lotterhos, 2015). The simulation of a distribution of neutral F_{ST} based on the demographic models inferred in chapter 3 may therefore not be necessary.

The uncertainties in the estimates of parameters for demographic models are much larger than suggested by the reported 95% bootstrap CIs. These intervals only capture the uncertainty due to genetic sampling, i.e. sampling a finite number of independent realisations of the evolutionary process. Due to the large number of independent loci used for inference here, this uncertainty is rather small. Much greater uncertainty lies in the method of parameter inference. For instance, $\delta a \delta i$ and `stairway-plot` differ dramatically in their estimates of the ancient population size of *parallelus*: 1.26 million and 4 million, respectively (compare fig. 3.18 and 3.19). The fact that `stairway-plot` does not incorporate gene flow in its model is unlikely to be the reason for this discrepancy. If this were the case, then the *divergence-in-isolation* model in $\delta a \delta i$ should also estimate a much larger population size for *parallelus* than the models *divergence-with-migration* and *asymmetric-migration*. However, this is clearly not the case (compare N_{par} in table 3.2 with N_{par} in tables 3.3 and 3.4).

The conversion of model parameter estimates from $\delta a \delta i$ and `stairway-plot` from genetic units to absolute/physical units requires the assumption of a mutation rate. I have assumed a mutation rate of 3×10^{-9} per nucleotide site and year for *C. parallelus* throughout. This estimate was taken from the few direct estimates of mutation rates in insect species available so far, measured by detecting differences between parents and offspring (Liu et al., 2017). Those estimates are remarkably close to each other and their Poisson error 95% CIs range from 1.0×10^{-9} to 6.1×10^{-9}. Estimates of single nucleotide mutation rates are for *Heliconius melpomene*: 2.9×10^{-9} (Keightley et al., 2015), for *Drosophila melanogaster*: 2.8×10^{-9} (Keightley et al., 2015), for *Bombus terrestris*: 3.6×10^{-9} (Liu et al., 2017) and for *Apis mellifera*: 3.4×10^{-9} (Yang et al., 2015). So unless the true mutation rate in grasshoppers is an outlier, the uncertainty in demographic model parameter estimates due to uncertainty in the assumed mutation rate is likely to be small. For example the estimate of divergence time in the *divergence-in-isolation* model from $\delta a \delta i$ of 486 thousand years changes to 521 thousand years when assuming the mutation rate estimated for *D. melanogaster* (+7%) or to 405 thousand years when assuming the mutation rate estimated for *B. terrestris* (-16%).

The estimates of effective population size for the two subspecies of *C. parallelus* are most likely underestimates of the true effective population size. This is because for a wide range of sexually reproducing species across the tree of life a large reduction in genome-wide genetic diversity due to selection on linked sites has been inferred (Corbett-Detig et al., 2015). The effect of linked selection on genome-wide average genetic diversity greatly depends on the genome-wide average effective recombination rate. The recombination rate can vary greatly between fairly closely related species (e.g. 37 cM/Mb in *Apis mellifera* vs. 8.7 cM/Mb in *Bombus terrestris*, Liu et al. 2017). The *effective* recombination rate is lower than 1/2 the rate of crossing-over if there is any form of inbreeding.

Two high quality estimates of effective population size have been published recently for insect species. One for *Drosophila melanogaster*: 1.4×10^6 (Keightley et al., 2014), the other for *Heliconius melpomene*: 2×10^6 (Keightley et al., 2015). These estimates of effective population size are based on estimates of nucleotide heterozygosity (π_{Tajima}) from 4-fold degenerate sites in genome resequencing data. They are not based on demographic modelling and also likely underestimate the true effective population size due to selection at linked sites. So, the estimate of 4 million as the ancient population size of *parallelus* by `stairway-plot` can still be regarded as consistent with these estimates for the two other insect species and also when considering the generally narrow range of effective population sizes across the tree of life as compared to current census population sizes (Coop, 2016).

Only folded site frequency spectra could be analysed in this study since the sequence of close enough outgroups were not available to infer the ancestral allele at polymorphic sites. Suitable outgroups for *C. parallelus* would be the closely related *C. curtipennis* endemic to North America and the more distantly related sister species *C. montanus* (BUTLIN and HEWITT, 1987). Unfolded site frequency spectra contain more information for the inference of demographic history and selection, although the large uncertainties in the site frequency spectrum of *parallelus* would be even greater in an unfolded site frequency spectrum. A third population sample (from *curtipennis* rather than *montanus*) would also allow the application of the population branch statistic (PBS) to detect loci under recent positive selection in one of the three populations (Yi et al., 2010).

Chapter 3 is still an incomplete investigation of the demographic history of the two subspecies and important aspects of it remain unknown or uncertain. For instance, the spatial aspect of the history. Why does Pyrenean *parallelus* have a greater nucleotide diversity than Pyrenean *erythropus* if its putative glacial refuge is so much more distant? Are the glacial refugia and origins of expansions indeed in the Balkans and southern Spain as suggested by previous phylogeographic data of two loci (Cooper et al., 1995; Lunt et al., 1998)? These aspects could be further investigated with RAD data from population samples spread across the distribution range of both subspecies (He et al., 2017; Noguerales et al., 2018). Ten individuals per population should suffice, but the number of sampled populations should be maximised and their locations spread evenly across the distribution range. With such data the pairwise directionality index Ψ between populations of Peter and Slatkin (2013) could be calculated that allows the localisation of the origin of a recent range expansion. The inference of the origin of expansion could be improved by adding data from species distribution models and including distributions in the past inferred by using paleo-climatic data (Elith and Leathwick, 2009; Kozak et al., 2008). This can then be combined with spatially explicit coalescent simulations and approximate Bayesian computation (Ray et al., 2010). This whole approach has been recently documented by He et al. (2017).

References

Albrechtsen, A., Nielsen, F. C., Nielsen, R., 2010. Ascertainment biases in snp chips affect measures of population divergence. Molecular biology and evolution 27 (11), 2534–2547.

Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L., Lander, E. S., 2000. An snp map of the human genome generated by reduced representation shotgun sequencing. Nature 407 (6803), 513–516.

Andersen, E. C., Gerke, J. P., Shapiro, J. A., Crissman, J. R., Ghosh, R., Bloom, J. S., Félix, M.-A., Kruglyak, L., Mar. 2012. Chromosome-scale selective sweeps shape caenorhabditis elegans genomic diversity. Nat Genet 44 (3), 285–290.
URL http://dx.doi.org/10.1038/ng.1050

Andolfatto, P., 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the drosophila melanogaster genome. Genome Research 17 (12), 1755–1762.
URL http://genome.cshlp.org/content/17/12/1755.abstract

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., Stern, D. L., 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Research.
URL http://genome.cshlp.org/content/early/2011/02/28/gr.115402.110.abstract

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., Hohenlohe, P. A., Feb. 2016. Harnessing the power of radseq for ecological and evolutionary genomics. Nat Rev Genet 17 (2), 81–92.
URL http://dx.doi.org/10.1038/nrg.2015.28

Andrews, K. R., Hohenlohe, P. A., Miller, M. R., Hand, B., Seeb, J. E., Luikart, G., Oct. 2014. Trade-offs and utility of alternative radseq methods. Mol Ecol.
URL http://dx.doi.org/10.1111/mec.12964

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, n/a–n/a.
URL http://dx.doi.org/10.1111/mec.12276

Atwood, T. S., Gribbin, J. M., Boone, J. Q., Nipper, R. W., Lillegard, N. J., Johnson, E. A., 2011. Rad longread: a snp discovery and de novo sequence assembly strategy. In: Conference Poster.

Badisco, L., Huybrechts, J., Simonet, G., Verlinden, H., Marchal, E., Huybrechts, R., Schoofs, L., De Loof, A., Vanden Broeck, J., Mar. 2011. Transcriptome analysis of the desert locust central nervous system: Production and annotation of a *Schistocerca gregaria* EST database. PLoS ONE 6, e17274.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0017274

Bailey, R., Lineham, M., Thomas, C., Butlin, R., 2003. Measuring dispersal and detecting departures from a random walk model in a grasshopper hybrid zone. Ecological Entomology 28 (2), 129–138.
URL http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2311.2003.00504.x/full

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., Johnson, E. A., 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3 (10), e3376.
URL http://dx.doi.org/10.1371/journal.pone.0003376

Bank, C., Bürger, R., Hermisson, J., 2012. The limits to parapatric speciation: Dobzhansky–muller incompatibilities in a continent–island model. Genetics 191 (3), 845–863.
URL http://www.genetics.org/content/191/3/845.short

Barton, N. H., Hewitt, G. M., 1985. Analysis of hybrid zones. Annual Review of Ecology and Systematics 16 (1), 113–148.
URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.es.16.110185.000553

Baxter, S. W., Davey, J. W., Johnston, J. S., Shelton, A. M., Heckel, D. G., Jiggins, C. D., Blaxter, M. L., Apr. 2011. Linkage mapping and comparative genomics using next-generation rad sequencing of a non-model organism. PLoS ONE 6 (4), e19315.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0019315

Beaumont, M. A., Nichols, R. A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B 263, 1619–1626.

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., Langley, C. H., Nov. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in drosophila simulans. PLoS Biol 5 (11), e310.
URL http://dx.doi.org/10.1371/journal.pbio.0050310

Belda, J. E., Cabrero, J., Camacho, J. P., Rufas, J. S., 1991. Role of C-heterochromatin in variation of nuclear DNA amount in the genus *Chorthippus* (Orthoptera, Acrididae). Cytobios 67 (268), 13–21.

Bella, J., Butlin, R., Ferris, C., Hewitt, G., Apr. 1992. Asymmetrical homogamy and unequal sex ratio from reciprocal mating-order crosses between chorthippus parallelus subspecies. Heredity 68 (4), 345–352.
URL http://dx.doi.org/10.1038/hdy.1992.49

Bella, J., Hewitt, G., Gosalvez, J., 1990. Meiotic imbalance in laboratory-produced hybrid males of chorthippus parallelus parallelus and chorthippus parallelus erythropus. Genetical research 56 (01), 43–48.

Bella, J. L., Serrano, L., Orellana, J., Mason, P. L., 2007. The origin of the *Chorthippus parallelus* hybrid zone: chromosomal evidence of multiple refugia for iberian populations. Journal of Evolutionary Biology 20 (2), 568–576.
URL http://dx.doi.org/10.1111/j.1420-9101.2006.01254.x

Bhatia, G., Patterson, N., Sankararaman, S., Price, A. L., 2013. Estimating and interpreting fst: The impact of rare variants. Genome Research 23 (9), 1514–1521.
URL http://genome.cshlp.org/content/23/9/1514.abstract

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., Good, J. M., 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics 13 (1), 403.
URL http://dx.doi.org/10.1186/1471-2164-13-403

Bialozyt, R., Ziegenhagen, B., Petit, R. J., Jan. 2006. Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. Journal of evolutionary biology 19, 12–20.

Buckley, S. H., Tregenza, T., Butlin, R. K., 2003. Transitions in cuticular composition across a hybrid zone: historical accident or environmental adaptation? Biological Journal of the Linnean Society 78 (2), 193–201.
URL http://dx.doi.org/10.1046/j.1095-8312.2003.00147.x

Buno, I., Torroja, E., Lopez-Fernadez, C., Butlin, R. K., Hewitt, G. M., Gosalvez, J., Dec. 1994. A hybrid zone between two subspecies of the grasshopper chorthippus parallelus along the pyrenees: the west end. Heredity 73, 625.
URL http://dx.doi.org/10.1038/hdy.1994.170

Butlin, R., Hewitt, G., 1985a. A hybrid zone between chorthippus parallelus parallelus and chorthippus parallelus erythropus (orthoptera: Acrididae): behavioural characters. Biological Journal of the Linnean Society 26 (3), 287–299.

Butlin, R. K., Hewitt, G. M., 1985b. A hybrid zone between *Chorthippus parallelus parallelus* and *Chorthippus parallelus erythropus* (orthoptera: Acrididae): morphological and electrophoretic characters. Biological Journal of the Linnean Society 26 (3), 269–285.
URL http://dx.doi.org/10.1111/j.1095-8312.1985.tb01636.x

BUTLIN, R. K., HEWITT, G. M., 1987. Genetic divergence in the chorthippus parallelus species group (orthoptera: Acrididae). Biological Journal of the Linnean Society 31 (4), 301–310.
URL +http://dx.doi.org/10.1111/j.1095-8312.1987.tb01995.x

Butlin, R. K., Ritchie, M. G., Hewitt, G. M., 1991. Comparisons among morphological characters and between localities in the chorthippus parallelus hybrid zone (orthoptera: Acrididae). Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 334 (1271), 297–308.
URL http://rstb.royalsocietypublishing.org/content/334/1271/297.abstract

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T., 2009. Blast+: architecture and applications. BMC Bioinformatics 10 (1), 421.
URL http://www.biomedcentral.com/1471-2105/10/421

Cariou, M., Duret, L., Charlat, S., 2016. How and how much does rad-seq bias genetic diversity estimates? BMC Evolutionary Biology 16 (1), 240.
URL http://dx.doi.org/10.1186/s12862-016-0791-0

Casbon, J. A., Osborne, R. J., Brenner, S., Lichtenstein, C. P., Jul. 2011. A method for counting pcr template molecules with application to next-generation sequencing. Nucleic Acids Res 39 (12), e81.
URL http://dx.doi.org/10.1093/nar/gkr217

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., Postlethwait, J. H., 2011. Stacks: Building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1 (3), 171–182.
URL http://www.g3journal.org/content/1/3/171.abstract

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., Allendorf, F. W., 2017. Unbroken: Radseq remains a powerful tool for understanding the genetics of adaptation in natural populations. Molecular Ecology Resources 17 (3), 362–365.
URL http://dx.doi.org/10.1111/1755-0998.12669

Chong, Z., Ruan, J., Wu, C.-I., 2012. Rainbow: an integrated tool for efficient clustering and assembling rad-seq reads. Bioinformatics.
URL http://bioinformatics.oxfordjournals.org/content/early/2012/09/01/bioinformatics.bts482.abstract

Chutimanitsakun, Y., Nipper, R. W., Cuesta-Marcos, A., Cistué, L., Corey, A., Filichkina, T., Johnson, E. A., Hayes, P. M., 2011. Construction and application for qtl analysis of a restriction site associated dna (rad) linkage map in barley. BMC Genomics 12, 4.
URL http://dx.doi.org/10.1186/1471-2164-12-4

Coffman, A. J., Hsieh, P. H., Gravel, S., Gutenkunst, R. N., 2016. Computationally efficient composite likelihood statistics for demographic inference. Molecular Biology and Evolution 33 (2), 591.
URL +http://dx.doi.org/10.1093/molbev/msv255

Coop, G., 2016. Does linked selection explain the narrow range of genetic diversity across species? bioRxiv.
URL https://www.biorxiv.org/content/early/2016/03/07/042598

Cooper, S. J., Hewitt, G. M., 1993. Nuclear dna sequence divergence between parapatric subspecies of the grasshopper chorthippus parallelus. Insect Mol Biol 2 (3), 185–194.

Cooper, S. J., Ibrahim, K. M., Hewitt, G. M., Feb. 1995. Postglacial expansion and genome subdivision in the european grasshopper *Chorthippus parallelus*. Mol Ecol 4 (1), 49–60.

Corbett-Detig, R. B., Hartl, D. L., Sackton, T. B., Apr. 2015. Natural selection constrains neutral diversity across a wide range of species. In: Coop (2016), p. e1002112, e1002112.
URL https://www.biorxiv.org/content/early/2016/03/07/042598

Crawford, J. E., Lazzaro, B. P., 2012. Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. Front Genet 3, 66.
URL http://dx.doi.org/10.3389/fgene.2012.00066

DaCosta, J. M., Sorenson, M. D., 2014. Amplification biases and consistent recovery of loci in a double-digest rad-seq protocol. PloS one 9, e106713.

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., Blaxter, M. L., Oct. 2012. Special features of rad sequencing data: implications for genotyping. Mol Ecol.
URL http://dx.doi.org/10.1111/mec.12084

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., Blaxter, M. L., Jul. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12 (7), 499–510.
URL http://dx.doi.org/10.1038/nrg3012

DAWSON, D. A., HORSBURGH, G. J., KÜPPER, C., STEWART, I. R. K., BALL, A. D., DURRANT, K. L., HANSSON, B., BACON, I., BIRD, S., KLEIN, A., KRUPA, A. P., LEE, J.-W., MARTÍN-GÁLVEZ, D., SIMEONI, M., SMITH, G., SPURGIN, L. G., BURKE, T., 2010. New methods to identify conserved microsatellite loci and develop primer sets of high cross-species utility ‚Äì as demonstrated for birds. Molecular Ecology Resources 10 (3), 475–494.
URL http://dx.doi.org/10.1111/j.1755-0998.2009.02775.x

Derrien, T., EstellÃ©, J., Marco Sola, S., Knowles, D. G., Raineri, E., GuigÃ³, R., Ribeca, P., Jan. 2012. Fast computation and applications of genome mappability. PLOS ONE 7 (1), 1–16.
URL https://doi.org/10.1371/journal.pone.0030377

Eaton, D. A. R., 2014. Pyrad: assembly of de novo radseq loci for phylogenetic analyses. Bioinformatics 30 (13), 1844–1849.
URL http://bioinformatics.oxfordjournals.org/content/30/13/1844.abstract

Edgar, R. C., 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5), 1792–1797.
URL http://nar.oxfordjournals.org/content/32/5/1792.abstract

Elith, J., Leathwick, J. R., 2009. Species distribution models: Ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics 40 (1), 677–697.
URL http://dx.doi.org/10.1146/annurev.ecolsys.110308.120159

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., Mitchell, S. E., May 2011. A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. PLoS ONE 6 (5), e19379.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0019379

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., Sella, G., Aug. 2016. A genomic map of the effects of linked selection in drosophila. In: Coop (2016), pp. 1–24, 1–24.
URL https://doi.org/10.1371/journal.pgen.1006130

Etter, P. D., Preston, J. L., Bassham, S., Cresko, W. A., Johnson, E. A., 2011. Local de novo assembly of rad paired-end contigs using short sequencing reads. PLoS One 6 (4), e18561.
URL http://dx.doi.org/10.1371/journal.pone.0018561

Evans, B. J., Zeng, K., Esselstyn, J. A., Charlesworth, B., Melnick, D. J., Sep. 2014. Reduced representation genome sequencing suggests low diversity on the sex chromosomes of tonkean macaque monkeys. Mol Biol Evol 31 (9), 2425–2440.
URL http://dx.doi.org/10.1093/molbev/msu197

Ferris, C., Rubio, J., Serrano, L., Gosalvez, J., Hewitt, G., Aug. 1993. One way introgression of a subspecific sex chromosome marker in a hybrid zone. Heredity 71 (2), 119–129.
URL http://dx.doi.org/10.1038/hdy.1993.115

Flanagan, N., Mason, P., Gosalvez, J., Hewitt, G., 1999. Chromosomal differentiation through an alpine hybrid zone in the grasshopper chorthippus parallelus. Journal of Evolutionary Biology 12 (3), 577–585.

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., Nielsen, R., Nov. 2013. Quantifying population genetic differentiation from next-generation sequencing data. Genetics 195 (3), 979–992.
URL http://dx.doi.org/10.1534/genetics.113.154740

Fumagalli, M., Vieira, F. G., Linderoth, T., Nielsen, R., May 2014. ngstools: methods for population genetics analyses from next-generation sequencing data. Bioinformatics 30 (10), 1486–1487.
URL http://dx.doi.org/10.1093/bioinformatics/btu041

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., Estoup, A., Oct. 2012. The effect of rad allele dropout on the estimation of genetic variation within and between populations. Mol Ecol.
URL http://dx.doi.org/10.1111/mec.12089

Gavrilets, S., 1997. Hybrid zones with dobzhansky-type epistatic selection. Evolution 51 (4), 1027–1035.
URL http://www.jstor.org/stable/2411031

Gillespie, J. H., 2004. Populations Genetics - A Concise Guide.

Gompert, Z., Buerkle, C. A., May 2011a. Bayesian estimation of genomic clines. Mol Ecol 20 (10), 2111–2127.
URL http://dx.doi.org/10.1111/j.1365-294X.2011.05074.x

Gompert, Z., Buerkle, C. A., Mar. 2011b. A hierarchical bayesian model for next-generation population genomics. Genetics 187 (3), 903–917.
URL http://dx.doi.org/10.1534/genetics.110.124693

Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Dec. 2008. Widespread mito-nuclear discordance with evidence for introgressive hybridization and selective sweeps in lycaeides. Mol Ecol 17 (24), 5231–5244.
URL http://dx.doi.org/10.1111/j.1365-294X.2008.03988.x

Gompert, Z., Lucas, L. K., Nice, C. C., Fordyce, J. A., Forister, M. L., Buerkle, C. A., 2012a. Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. Evolution, no–no.
URL http://dx.doi.org/10.1111/j.1558-5646.2012.01587.x

Gompert, Z., Parchman, T. L., Buerkle, C. A., Feb. 2012b. Genomics of isolation in hybrids. Philos Trans R Soc Lond B Biol Sci 367 (1587), 439–450.
URL http://dx.doi.org/10.1098/rstb.2011.0196

Good, J. M., 2011. Reduced Representation Methods for Subgenomic Enrichment and Next-Generation Sequencing. Humana Press, Totowa, NJ, pp. 85–103.
URL http://dx.doi.org/10.1007/978-1-61779-228-1_5

Gosálvez, J., López-Fernández, C., Bella, L., Butlin, R., Hewitt, G., 1988. A hybrid zone between chorthippus parallelus parallelus and chorthippus parallelus erythropus (orthoptera: Acrididae): chromosomal differentiation. Genome 30 (5), 656–663.

Gosalvez, J., López-Fernandez, C., Esponda, P., 1980. Variability of the dna content in five orthopteran species. Caryologia 33 (2), 275–281.
URL http://dx.doi.org/10.1080/00087114.1980.10796840

Graham, C. F., Glenn, T. C., McArthur, A. G., Boreham, D. R., Kieran, T., Lance, S., Manzon, R. G., Martino, J. A., Pierson, T., Rogers, S. M., Wilson, J. Y., Somers, C. M., 2015. Impacts of degraded dna on restriction enzyme associated dna sequencing (radseq). Molecular Ecology Resources, n/a–n/a.
URL http://dx.doi.org/10.1111/1755-0998.12404

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., Bustamante, C. D., Oct. 2009. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS Genet 5 (10), e1000695.
URL http://dx.doi.org/10.1371/journal.pgen.1000695

Han, E., Sinsheimer, J. S., Novembre, J., Mar. 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. Mol Biol Evol 31 (3), 723–735.
URL http://dx.doi.org/10.1093/molbev/mst229

He, Q., Prado, J. R., Knowles, L. L., 2017. Inferring the geographic origin of a range expansion: Latitudinal and longitudinal coordinates inferred from genomic data in an abc framework with the program x-origin. Molecular Ecology 26 (24), 6908–6920.
URL http://dx.doi.org/10.1111/mec.14380

Hewitt, G. M., 1990. Divergence and speciation as viewed from an insect hybrid zone. Canadian Journal of Zoology 68 (8), 1701–1715.

Hewitt, G. M., 1993. After the ice: parallelus meets erythropus in the pyrenees. Hybrid zones and the evolutionary process, 140–164.

Hewitt, G. M., 1996. Some genetic consequences of ice ages, and their role in divergence and speciation. Biological Journal of the Linnean Society 58 (3), 247–276.
URL http://www.sciencedirect.com/science/article/pii/S0024406696900358

Hewitt, G. M., 1999. Post-glacial re-colonization of european biota. Biological Journal of the Linnean Society 68 (1-2), 87–112.
URL http://www.sciencedirect.com/science/article/B6WBR-45GW8DS-7/2/39cec29c344c530779768618d9d21787

Hewitt, G. M., Butlin, R. K., East, T. M., 1987. Testicular dysfunction in hybrids between parapatric subspecies of the grasshopper *Chorthippus parallelus*. Biol J Linn Soc 31, 25–34.

Hird, S. M., Brumfield, R. T., Carstens, B. C., Jul. 2011. Prgmatic: an efficient pipeline for collating genome-enriched second-generation sequencing data using a 'provisional-reference genome'. Mol Ecol Resour 11 (4), 743–748.
URL http://dx.doi.org/10.1111/j.1755-0998.2011.03005.x

HOHENLOHE, P. A., AMISH, S. J., CATCHEN, J. M., ALLENDORF, F. W., LUIKART, G., 2011. Next-generation rad sequencing identifies thousands of snps for assessing hybridization between rainbow and westslope cutthroat trout. Molecular Ecology Resources 11, 117–122.
URL http://dx.doi.org/10.1111/j.1755-0998.2010.02967.x

Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., Cresko, W. A., Feb. 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced rad tags. PLoS Genet 6 (2), e1000862.
URL http://dx.doi.org/10.1371%2Fjournal.pgen.1000862

Hohenlohe, P. A., Day, M. D., Amish, S. J., Miller, M. R., Kamps-Hughes, N., Boyer, M. C., Muhlfeld, C. C., Allendorf, F. W., Johnson, E. A., Luikart, G., 2013. Genomic patterns of introgression in rainbow and westslope cutthroat trout illuminated by overlapping paired-end rad sequencing. Mol Ecol 22 (11), 3002–3013.
URL http://dx.doi.org/10.1111/mec.12239

Holtgrewe, M., 2010. Mason – a read simulator for second generation sequencing data. Tech. Rep. Technical Report TR-B-10-06, Freie Universität Berlin, Institut für Mathematik und Informatik.

Hsieh, P., Hallmark, B., Watkins, J., Karafet, T. M., Osipova, L. P., Gutenkunst, R. N., Hammer, M. F., 2017. Exome sequencing provides evidence of polygenic adaptation to a fat-rich animal diet in indigenous siberian populations. Molecular Biology and Evolution 34 (11), 2913–2926.
URL +http://dx.doi.org/10.1093/molbev/msx226

Hudson, R. R., Slatkin, M., Maddison, W. P., 1992. Estimation of levels of gene flow from dna sequence data. Genetics 132 (2), 583–589.
URL http://www.genetics.org/content/132/2/583

Hurst, G. D., Jiggins, F. M., 2005. Problems with mitochondrial dna as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. Proceedings of the Royal Society B: Biological Sciences 272 (1572), 1525–1534.

Hutchison, J., 2013. Patterns of gene flow in a grasshopper hybrid zone. Ph.D. thesis, University of Sheffield - Department of Animal and Plant Sciences.

Ibrahim, K. M., Nichols, R. A., Hewitt, G. M., Sep. 1996. Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. Heredity 77 (3), 282–291.
URL http://dx.doi.org/10.1038/hdy.1996.142

Ilut, D. C., Nydam, M. L., Hare, M. P., 2014. Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. Biomed Res Int 2014, 675158.
URL http://dx.doi.org/10.1155/2014/675158

Jarvis, A., Reuter, H. I., Nelson, A., Guevara, E., 2008. Hole-filled srtm for the globe version 4, available from the cgiar-csi srtm 90m database.
URL http://www.cgiar-csi.org/data/srtm-90m-digital-elevation-database-v4-1

John, B., Hewitt, G., 1966. Karyotype stability and dna variability in the acrididae. Chromosoma 20 (2), 155–172.

Johnson, P. L. F., Slatkin, M., Jan. 2008. Accounting for bias from sequencing error in population genetic estimates. Mol Biol Evol 25 (1), 199–206.
URL http://dx.doi.org/10.1093/molbev/msm239

Jones, E., Oliphant, T., Peterson, P., et al., 2001–. SciPy: Open source scientific tools for Python. [Online; accessed <today>].
URL http://www.scipy.org/

Jones, M. R., Good, J. M., 2016. Targeted capture in evolutionary and ecological genomics. Molecular Ecology 25 (1), 185–202.
URL http://dx.doi.org/10.1111/mec.13304

Jouganous, J., Long, W., Ragsdale, A. P., Gravel, S., 2017. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics.
URL http://www.genetics.org/content/early/2017/05/08/genetics.117.200493

Kang, L., Chen, X., Zhou, Y., Liu, B., Zheng, W., Li, R., Wang, J., Yu, J., Dec. 2004. The analysis of large-scale gene expression correlated to the phase changes of the migratory locust. Proc Natl Acad Sci U S A 101 (51), 17611–17615.
URL http://dx.doi.org/10.1073/pnas.0407753101

Keightley, P. D., Ness, R. W., Halligan, D. L., Haddrill, P. R., Jan. 2014. Estimation of the spontaneous mutation rate per nucleotide site in a drosophila melanogaster full-sib family. Genetics 196, 313–320.

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., Davey, J. W., Jiggins, C. D., Jan. 2015. Estimation of the spontaneous mutation rate in heliconius melpomene. Molecular biology and evolution 32, 239–243.

Keller, I., Bensasson, D., Nichols, R. A., Feb. 2007. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3 (2), e22.
URL http://dx.doi.org/10.1371/journal.pgen.0030022

Kim, S. Y., Lohmueller, K. E., Albrechtsen, A., Li, Y., Korneliussen, T., Tian, G., Grarup, N., Jiang, T., Andersen, G., Witte, D., Jorgensen, T., Hansen, T., Pedersen, O., Wang, J., Nielsen, R., 2011. Estimation of allele frequency and association mapping using next-generation sequencing data. BMC Bioinformatics 12, 231.
URL http://dx.doi.org/10.1186/1471-2105-12-231

Kimura, M., 1986. Diffusion Models of Population Genetics in the Age of Molecular Biology. Springer New York, New York, NY, pp. 150–165.
URL http://dx.doi.org/10.1007/978-1-4613-8631-5_10

Korneliussen, T. S., Albrechtsen, A., Nielsen, R., 2014. Angsd: Analysis of next generation sequencing data. BMC Bioinformatics 15 (1), 356.
URL http://dx.doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., Nielsen, R., 2013. Calculation of tajima's d and other neutrality test statistics from low depth next-generation sequencing data. BMC Bioinformatics 14, 289.
URL http://dx.doi.org/10.1186/1471-2105-14-289

Kosugi, S., Natsume, S., Yoshida, K., MacLean, D., Cano, L., Kamoun, S., Terauchi, R., 2013. Coval: improving alignment quality and variant calling accuracy for next-generation sequencing data. PLoS One 8 (10), e75402.
URL http://dx.doi.org/10.1371/journal.pone.0075402

Kozak, K. H., Graham, C. H., Wiens, J. J., 2008. Integrating gis-based environmental data into evolutionary biology. Trends in Ecology & Evolution 23 (3), 141–148.
URL http://www.sciencedirect.com/science/article/pii/S0169534708000426

Langmead, B., Salzberg, S. L., Apr. 2012. Fast gapped-read alignment with bowtie 2. Nat Meth 9 (4), 357–359.
URL http://dx.doi.org/10.1038/nmeth.1923

Lee, H., Schatz, M. C., Aug. 2012. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics 28 (16), 2097–2105.
URL http://dx.doi.org/10.1093/bioinformatics/bts330

Li, Durbin, R., 2011. The SAM Format Specification.

Li, H., Apr. 2011a. Improving snp discovery by base alignment quality. Bioinformatics (Oxford, England) 27, 1157–1158.

Li, H., Nov. 2011b. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27 (21), 2987–2993.
URL http://dx.doi.org/10.1093/bioinformatics/btr509

Li, H., Apr. 2014. Towards Better Understanding of Artifacts in Variant Calling from High-Coverage Samples. ArXiv e-prints.

Li, H., Ruan, J., Durbin, R., Nov. 2008. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Res 18 (11), 1851–1858.
URL http://dx.doi.org/10.1101/gr.078212.108

Liu, D.-F., Dong, Z.-M., Zhang, D.-Y., ze Gu, Y., Guo, P.-J., Han, R.-H., Jiang, G.-J., 2008. Molecular phylogeny of the higher category of acrididae (orthoptera: Acridoidea). Zoological Research 29, 585–591.
URL http://www.bioline.org.br/abstract?zr08089

Liu, H., Jia, Y., Sun, X., Tian, D., Hurst, L. D., Yang, S., 2017. Direct determination of the mutation rate in the bumblebee reveals evidence for weak recombination-associated mutation and an approximate rate constancy in insects. Molecular Biology and Evolution 34 (1), 119.
URL +http://dx.doi.org/10.1093/molbev/msw226

Liu, X., Fu, Y.-X., May 2015. Exploring population size changes using snp frequency spectra. Nat Genet 47 (5), 555–559.
URL http://dx.doi.org/10.1038/ng.3254

Llewellyn, A., 2008. The genetics of male sterility in a grasshopper hybrid zone. Ph.D. thesis, Univerity of Sheffield, Department of Animal and Plant Sciences.

Lowry, D. B., Hoban, S., Kelley, J. L., Lotterhos, K. E., Reed, L. K., Antolin, M. F., Storfer, A., 2016. Breaking rad: an evaluation of the utility of restriction site-associated dna sequencing for genome scans of adaptation. Molecular Ecology Resources, n/a–n/a.
URL http://dx.doi.org/10.1111/1755-0998.12635

Lozier, J. D., 2014. Revisiting comparisons of genetic diversity in stable and declining species: assessing genome-wide polymorphism in north american bumble bees using rad sequencing. Molecular Ecology 23 (4), 788–801.
URL http://dx.doi.org/10.1111/mec.12636

Luca, F., Hudson, R. R., Witonsky, D. B., Rienzo, A. D., Jul. 2011. A reduced representation approach to population genetic analyses and applications to human evolution. Genome Res 21 (7), 1087–1098.
URL http://dx.doi.org/10.1101/gr.119792.110

Lunt, D. H., Ibrahim, K. M., Hewitt, G. M., May 1998. mtdna phylogeography and postglacial patterns of subdivision in the meadow grasshopper chorthippus parallelus. Heredity 80 (Pt 5), 633–641.

Lunter, G., Goodson, M., Jun. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. Genome Res 21 (6), 936–939.
URL http://dx.doi.org/10.1101/gr.111120.110

Mallet, J., 2008. Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1506), 2971–2986.
URL http://rstb.royalsocietypublishing.org/content/363/1506/2971.abstract

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., Turner, D. J., 2010. Target-enrichment strategies for next-generation sequencing. Nature methods 7 (2), 111–118.

Maricic, T., Whitten, M., Pääbo, S., Nov. 2010. Multiplexed dna sequence capture of mitochondrial genomes using pcr products. PLOS ONE 5 (11), 1–5.
URL https://doi.org/10.1371/journal.pone.0014004

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., Emerson, B. C., 2015. Restriction site-associated dna sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. Molecular Ecology Resources 15 (1), 28–41.
URL http://dx.doi.org/10.1111/1755-0998.12291

McKinney, G. J., Larson, W. A., Seeb, L. W., Seeb, J. E., 2017. Radseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on breaking rad by lowry etÂ al. (2016). Molecular Ecology Resources 17 (3), 356–361.
URL http://dx.doi.org/10.1111/1755-0998.12649

McVicker, G., Gordon, D., Davis, C., Green, P., 2009. Widespread genomic signatures of natural selection in hominid evolution. Plos Genetics 5 (5).

Merz, C., Catchen, J. M., Hanson-Smith, V., Emerson, K. J., Bradshaw, W. E., Holzapfel, C. M., 2013. Replicate phylogenies and post-glacial range expansion of the pitcher-plant mosquito, wyeomyia smithii, in north america. PLoS One 8 (9), e72262.
URL http://dx.doi.org/10.1371/journal.pone.0072262

Morgulis, A., Gertz, E. M., Schäffer, A. A., Agarwala, R., Jun. 2006. A fast and symmetric dust implementation to mask low-complexity dna sequences. J Comput Biol 13 (5), 1028–1040.
URL http://dx.doi.org/10.1089/cmb.2006.13.1028

Nadeau, N. J., Ruiz, M., Salazar, P., Counterman, B., Medina, J. A., Ortiz-Zuazaga, H., Morrison, A., McMillan, W. O., Jiggins, C. D., Papa, R., Aug. 2014. Population genomics of parallel hybrid zones in the mimetic butterflies, h. melpomene and h. erato. Genome research 24, 1316–1333.

Nadeau, N. J., Whibley, A., Jones, R. T., Davey, J. W., Dasmahapatra, K. K., Baxter, S. W., Quail, M. A., Joron, M., ffrench Constant, R. H., Blaxter, M. L., Mallet, J., Jiggins, C. D., 2012. Genomic islands of divergence in hybridizing heliconius butterflies identified by large-scale targeted sequencing. Philosophical Transactions of the Royal Society of London B: Biological Sciences 367 (1587), 343–353.
URL http://rstb.royalsocietypublishing.org/content/367/1587/343

Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., Stamatoyannopoulos, J. A., 2012. Bedops: high-performance genomic feature operations. Bioinformatics 28 (14), 1919–1920.
URL +http://dx.doi.org/10.1093/bioinformatics/bts277

Nichols, R. A., Hewitt, G. M., 1994. The genetic consequences of long distance dispersal during colonization. Heredity 72 (3), 312–317.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., Wang, J., 2012. Snp calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PLoS One 7 (7), e37558.
URL http://dx.doi.org/10.1371/journal.pone.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., Song, Y. S., Jun. 2011. Genotype and snp calling from next-generation sequencing data. Nat Rev Genet 12 (6), 443–451.
URL http://dx.doi.org/10.1038/nrg2986

Noguerales, V., Cordero, P. J., Ortego, J., Jan. 2018. Inferring the demographic history of an oligophagous grasshopper: Effects of climatic niche stability and host-plant distribution. Molecular phylogenetics and evolution 118, 343–356.

Nosil, P., Parchman, T. L., Feder, J. L., Gompert, Z., Aug. 2012. Do highly divergent loci reside in genomic regions affecting reproductive isolation? a test using next-generation sequence data in timema stick insects. BMC Evol Biol 12 (1), 164.
URL http://dx.doi.org/10.1186/1471-2148-12-164

Parchman, T. L., Gompert, Z., Braun, M. J., Brumfield, R. T., McDonald, D. B., Uy, J. A. C., Zhang, G., Jarvis, E. D., Schlinger, B. A., Buerkle, C. A., Jun. 2013. The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. Molecular ecology 22, 3304–3317.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., Buerkle, C. A., Jun. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol Ecol 21 (12), 2991–3005.
URL http://dx.doi.org/10.1111/j.1365-294X.2012.05513.x

Patterson, N., Price, A. L., Reich, D., Dec. 2006. Population structure and eigenanalysis. PLoS Genet 2 (12), e190.
URL http://dx.doi.org/10.1371/journal.pgen.0020190

Pérez, F., Granger, B. E., May 2007. IPython: a system for interactive scientific computing. Computing in Science and Engineering 9 (3), 21–29.
URL http://ipython.org

Peter, B. M., Slatkin, M., Nov. 2013. Detecting range expansions from genetic data. Evolution; international journal of organic evolution 67, 3274–3289.

Peter, B. M., Slatkin, M., 2015. The effective founder effect in a spatially expanding population. Evolution 69 (3), 721–734.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., Hoekstra, H. E., May 2012. Double digest radseq: An inexpensive method for *De Novo* snp discovery and genotyping in model and non-model species. PLoS ONE 7 (5), e37135.
URL http://dx.doi.org/10.1371%2Fjournal.pone.0037135

Petit, E., Excoffier, L., Mayer, F., 1999. No evidence of bottleneck in the postglacial recolonization of europe by the noctule bat (nyctalus noctula). Evolution 53 (4), 1247–1258.
URL http://www.jstor.org/stable/2640827

Pfender, W. F., Saha, M. C., Johnson, E. A., Slabaugh, M. B., May 2011. Mapping with rad (restriction-site associated dna) markers to rapidly identify qtl for stem rust resistance in lolium perenne. Theor Appl Genet 122 (8), 1467–1480.
URL http://dx.doi.org/10.1007/s00122-011-1546-3

Pool, J. E., Hellmann, I., Jensen, J. D., Nielsen, R., Mar. 2010. Population genetic inference from genomic sequence variation. Genome Res 20 (3), 291–300.
URL http://dx.doi.org/10.1101/gr.079509.108

Pujolar, J. M., Jacobsen, M. W., Als, T. D., Frydenberg, J., Munch, K., Jónsson, B., Jian, J. B., Cheng, L., Maes, G. E., Bernatchez, L., Hansen, M. M., May 2014. Genome-wide single-generation signatures of local selection in the panmictic european eel. Mol Ecol 23 (10), 2514–2528.
URL http://dx.doi.org/10.1111/mec.12753

Puritz, J. B., Hollenbeck, C. M., Gold, J. R., 2014a. ddocent: a radseq, variant-calling pipeline designed for population genomics of non-model organisms. PeerJ 2, e431.
URL http://dx.doi.org/10.7717/peerj.431

Puritz, J. B., Lotterhos, K. E., 2017. Expressed exome capture sequencing (eecseq): a method for cost-effective exome sequencing for all organisms with or without genomic resources. bioRxiv.
URL https://www.biorxiv.org/content/early/2017/12/24/223735

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., Bird, C. E., Oct. 2014b. Demystifying the rad fad. Mol Ecol.
URL http://dx.doi.org/10.1111/mec.12965

QGIS, Q. D. T., 2017. Qgis geographic information system. open source geospatial foundation project.
URL http://qgis.osgeo.org

Ray, N., Currat, M., Foll, M., Excoffier, L., 2010. Splatche2: a spatially explicit simulation framework for complex demography, genetic admixture and recombination. Bioinformatics 26 (23), 2993.
URL +http://dx.doi.org/10.1093/bioinformatics/btq579

Raychoudhury, R., Grillenberger, B. K., Gadau, J., Bijlsma, R., van de Zande, L., Werren, J. H., Beukeboom, L. W., Mar. 2010. Phylogeography of nasonia vitripennis (hymenoptera) indicates a mitochondrial-wolbachia sweep in north america. Heredity 104, 318–326.

Richards, P. M., Liu, M. M., Lowe, N., Davey, J. W., Blaxter, M. L., Davison, A., Jun. 2013. Rad-seq derived markers flank the shell colour and banding loci of the cepaea nemoralis supergene. Mol Ecol 22 (11), 3077–3089.
URL http://dx.doi.org/10.1111/mec.12262

Ritchie, M. G., 1990. Are differences in song responsible for assortative mating between subspecies of the grasshopper chorthippus parallelus (orthoptera: Acrididae)? Animal Behaviour 39 (4), 685–691.
URL http://www.sciencedirect.com/science/article/B6W9W-4JS8556-8/2/f8872e1aea24048a7f7fa1f7651cd6e0

Ritchie, M. G., Butlin, R. K., Hewitt, G. M., 1989. Assortative mating across a hybrid zone in <i>chorthippus parallelus</i> (orthoptera: Acrididae). Journal of Evolutionary Biology 2 (5), 339–352.
URL http://dx.doi.org/10.1046/j.1420-9101.1989.2050339.x

Ritchie, M. G., Butlin, R. K., Hewitt, G. M., 1992. Fitness consequences of potential assortative mating inside and outside a hybrid zone in *Chorthippus parallelus* (Orthoptera, Acrididae) - implications for reinforcement and sexual selection theory. Biological Journal Of The Linnean Society 45 (3), 219–234.

Robinson, J. D., Coffman, A. J., Hickerson, M. J., Gutenkunst, R. N., 2014. Sampling strategies for frequency spectrum-based population genomic inference. BMC Evol Biol 14, 254.
URL http://dx.doi.org/10.1186/s12862-014-0254-4

Roesti, M., Salzburger, W., Berner, D., 2012. Uninformative polymorphisms bias genome scans for signatures of selection. BMC Evol Biol 12, 94.
URL http://dx.doi.org/10.1186/1471-2148-12-94

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., Oct. 2016. Vsearch: a versatile open source tool for metagenomics. PeerJ 4, e2584.
URL https://doi.org/10.7717/peerj.2584

Roth, S., Köhler, G., Reinhardt, K., Predel, R., 2007. A discrete neuropeptide difference between two hybridizing grasshopper subspecies. Biological Journal of the Linnean Society 91 (4), 541–548.

Roux, C., Fraïsse, C., Romiguier, J., Anciaux, Y., Galtier, N., Bierne, N., Dec. 2016. Shedding light on the grey zone of speciation along a continuum of genomic divergence. PLoS biology 14, e2000234.

Rutledge, R. G., Côté, C., Aug. 2003. Mathematics of quantitative kinetic pcr and the application of standard curves. Nucleic Acids Res 31 (16), e93.

Schweyen, H., Rozenberg, A., Leese, F., 2014. Detection and removal of pcr duplicates in population genomic ddrad studies by addition of a degenerate base region (dbr) in sequencing adapters. The Biological Bulletin 227 (2), 146–160, pMID: 25411373.
URL https://doi.org/10.1086/BBLv227n2p146

Self, S. G., Liang, K.-Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82 (398), 605–610.
URL http://www.jstor.org/stable/2289471

Sella, G., Petrov, D. A., Przeworski, M., Andolfatto, P., Jun. 2009. Pervasive natural selection in the drosophila genome? PLOS Genetics 5 (6), 1–13.
URL https://doi.org/10.1371/journal.pgen.1000495

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., Wolf, J. B. W., 2016. Bioinformatic processing of rad-seq data dramatically impacts downstream population genetic inference. Methods in Ecology and Evolution, n/a–n/a.
URL http://dx.doi.org/10.1111/2041-210X.12700

Shuker, D. M., Underwood, K., King, T. M., Butlin, R. K., Dec. 2005. Patterns of male sterility in a grasshopper hybrid zone imply accumulation of hybrid incompatibilities without selection. Proc R Soc B 272 (1580), 2491–2497.
URL http://dx.doi.org/10.1098/rspb.2005.3242

Skotte, L., Korneliussen, T. S., Albrechtsen, A., Jul. 2012. Association testing for next-generation sequencing data using score statistics. Genet Epidemiol 36 (5), 430–437.
URL http://dx.doi.org/10.1002/gepi.21636

Slatkin, M., 1987. Gene flow and the geographic structure of natural populations. Science 236 (4803), 787–792.
URL http://science.sciencemag.org/content/236/4803/787

Slatkin, M., Barton, N. H., Nov. 1989. A comparison of three indirect methods for estimating average levels of gene flow. Evolution; international journal of organic evolution 43, 1349–1368.

Small, K. S., Brudno, M., Hill, M. M., Sidow, A., Mar. 2007. Extreme genomic variation in a natural population. Proc Natl Acad Sci U S A 104 (13), 5698–5703.
URL http://dx.doi.org/10.1073/pnas.0700890104

Stölting, K. N., Nipper, R., Lindtke, D., Caseys, C., Waeber, S., Castiglione, S., Lexer, C., Feb. 2013. Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. Mol Ecol 22 (3), 842–855.
URL http://dx.doi.org/10.1111/mec.12011

Sturt, F., Garrow, D., Bradley, S., 2013. New models of north west european holocene palaeogeography and inundation. Journal of Archaeological Science 40 (11), 3963–3976.
URL http://www.sciencedirect.com/science/article/pii/S0305440313001982

Taberlet, P., Fumagalli, L., Wust-Saucy, A.-G., Cosson, J.-F., 1998. Comparative phylogeography and postglacial colonization routes in europe. Molecular Ecology 7 (4), 453–464.
URL http://dx.doi.org/10.1046/j.1365-294x.1998.00289.x

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. Mega6: Molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution 30 (12), 2725–2729.
URL http://mbe.oxfordjournals.org/content/30/12/2725.abstract

Tiffin, P., Ross-Ibarra, J., 2014. Advances and limits of using population genetics to understand local adaptation. Trends in Ecology & Evolution 29 (12), 673–680.
URL http://www.sciencedirect.com/science/article/pii/S0169534714002237

Tin, M. M. Y., Rheindt, F. E., Cros, E., Mikheyev, A. S., Mar. 2015. Degenerate adaptor sequences for detecting pcr duplicates in reduced representation sequencing data improve genotype calling accuracy. Mol Ecol Resour 15 (2), 329–336.
URL http://dx.doi.org/10.1111/1755-0998.12314

Tregenza, T., Pritchard, V. L., Butlin, R. K., Dec. 2002. The origins of postmating reproductive isolation: testing hypotheses in the grasshopper *Chorthippus parallelus*. Population Ecology 44 (3), 0137–0144.
URL http://dx.doi.org/10.1007/s101440200017

Turelli, M., Orr, H. A., May 1995. The dominance theory of haldane's rule. Genetics 140 (1), 389–402.

Ulitsky, I., Oct. 2016. Evolution to the rescue: using comparative genomics to understand long non-coding rnas. Nature reviews. Genetics 17, 601–614.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., Sonstegard, T. S., Mar. 2008. Snp discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth 5 (3), 247–252.
URL http://dx.doi.org/10.1038/nmeth.1185

Vazquez, P., Cooper, S. J., Gosalvez, J., Hewitt, G. M., Oct. 1994. Nuclear dna introgression across a pyrenean hybrid zone between parapatric subspecies of the grasshopper chorthippus parallelus. Heredity 73 (Pt 4), 436–443.

Vieira, F. G., Fumagalli, M., Albrechtsen, A., Nielsen, R., Nov. 2013. Estimating inbreeding coefficients from ngs data: Impact on genotype calling and allele frequency estimation. Genome Res 23 (11), 1852–1861.
URL http://dx.doi.org/10.1101/gr.157388.113

Virdee, S., Hewitt, G., 1990. Ecological components of a hybrid zone in the grasshopper chorthippus parallelus (zetterstedt)(orthoptera, acrididae). Boletin de Sanidad Vegetal. Plagas (Spain).

Virdee, S., Hewitt, G., Dec. 1992. Postzygotic isolation and haldane's rule in a grasshopper. Heredity 69 (6), 527–538.
URL http://dx.doi.org/10.1038/hdy.1992.168

Wagner, C. E., Keller, I., Wittwer, S., Selz, O. M., Mwaiko, S., Greuter, L., Sivasundar, A., Seehausen, O., 2012. Genome-wide rad sequence data provide unprecedented resolution of species boundaries and relationships in the lake victoria cichlid adaptive radiation. Molecular Ecology, n/a–n/a.
URL http://dx.doi.org/10.1111/mec.12023

Wakeley, J., 2009. Coalescent Theory. Roberts & Company Publishers.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., Hao, S., Chen, B., Ma, Z., Yu, D., Xiong, Z., Zhu, Y., Fan, D., Han, L., Wang, B., Chen, Y., Wang, J., Yang, L., Zhao, W., Feng, Y., Chen, G., Lian, J., Li, Q., Huang, Z., Yao, X., Lv, N., Zhang, G., Li, Y., Wang, J., Wang, J., Zhu, B., Kang, L., Jan. 2014. The locust genome provides insight into swarm formation and long-distance flight. Nat Commun 5, –.
URL http://dx.doi.org/10.1038/ncomms3957

Warren, R. L., Sutton, G. G., Jones, S. J. M., Holt, R. A., 2007. Assembling millions of short dna sequences using ssake. Bioinformatics 23 (4), 500–501.
URL http://bioinformatics.oxfordjournals.org/content/23/4/500.abstract

Waters, J. M., Fraser, C. I., Hewitt, G. M., Feb. 2013. Founder takes all: density-dependent processes structure biodiversity. Trends in ecology & evolution 28, 78–85.

Weir, B. S., 1999. Genetic Data Analyses II. Sinauer Associates.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., Hill, W. G., 2005. Measures of human population structure show heterogeneity among genomic regions. Genome research 15 (11), 1468–1476.
URL http://genome.cshlp.org/content/15/11/1468.short

Weir, B. S., Cockerham, C. C., 1984. Estimating f-statistics for the analysis of population structure. Evolution, 1358–1370.

Whitlock, M. C., Lotterhos, K. E., Oct. 2015. Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of f(st). Am Nat 186 Suppl 1, S24–S36.
URL http://dx.doi.org/10.1086/682949

Wilmore, P., Brown, A., 1975. Molecular properties of orthopteran dna. Chromosoma 51 (4), 337–345.
URL http://dx.doi.org/10.1007/BF00326320

Yang, S., Wang, L., Huang, J., Zhang, X., Yuan, Y., Chen, J.-Q., Hurst, L. D., Tian, D., Jul. 2015. Parent-progeny sequencing indicates higher mutation rates in heterozygotes. Nature 523, 463.
URL http://dx.doi.org/10.1038/nature14649

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., Wang, J., Jul. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329 (5987), 75–78.
URL http://dx.doi.org/10.1126/science.1190371

Zabal-Aguirre, M., Arroyo, F., Bella, J. L., Feb. 2010. Distribution of wolbachia infection in chorthippus parallelus populations within and beyond a pyrenean hybrid zone. Heredity 104 (2), 174–184.
URL http://dx.doi.org/10.1038/hdy.2009.106

Zabal-Aguirre, M., Arroyo, F., García-Hurtado, J., de la Torre, J., Hewitt, G. M., Bella, J. L., 2014. Wolbachia effects in natural populations of chorthippus parallelus from the pyrenean hybrid zone. Journal of Evolutionary Biology 27 (6), 1136–1148.
URL http://dx.doi.org/10.1111/jeb.12389

Zerbino, D. R., 2010. Using the velvet de novo assembler for short-read sequencing technologies. In: Current Protocols in Bioinformatics. John Wiley & Sons, Inc., pp. –.
URL http://dx.doi.org/10.1002/0471250953.bi1105s31

Zerbino, D. R., Birney, E., May 2008. Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res 18 (5), 821–829.
URL http://dx.doi.org/10.1101/gr.074492.107

Zorita, E., Cuscó, P., Filion, G. J., Jun. 2015. Starcode: sequence clustering based on all-pairs search. Bioinformatics 31 (12), 1913–1919.
URL https://github.com/gui11aume/starcode

Chapter 5

Appendix

5.1 sRAD protocol

5.1.1 isolate DNA from grasshopper hindleg

- with Qiagen Dneasy Blood and Tissue Kit (spin columns)

5.1.2 check the quality of the isolations

- by running each isolation on a 0.8% gel (figure 5.1)

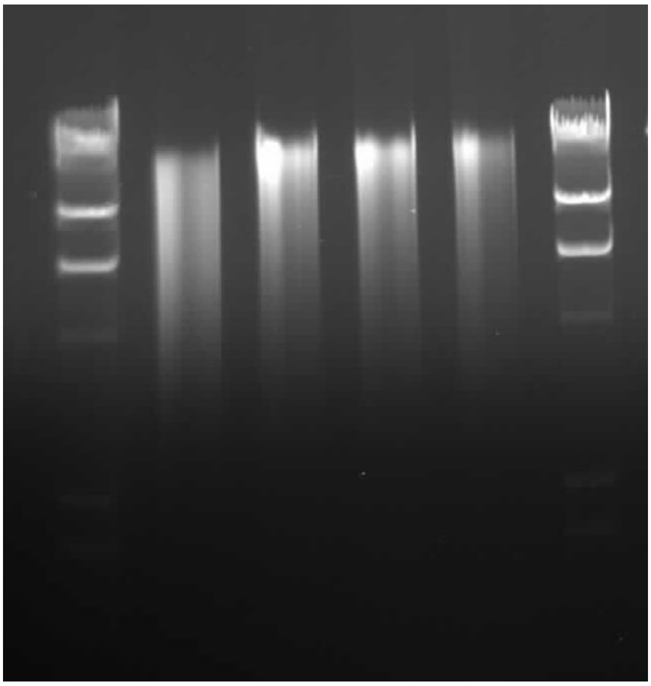


Fig. 5.1 Four spin column DNA isolation from grasshopper hindlegs next to HindIII digested lambda. The DNA is obviously already fragmented. All grasshopper DNA isolations look like that. The sRAD protocol worked anyway fine with them.

5.1.3 treat DNA isolations with RNase if necessary

- can be done on the spin column, leading to less RNase contamination

5.1.4 quantify DNA samples with fluorimeter twice

- produce at least three replicates of calf thymus standard serial dilutions for the standard curve

5.1.5 transfer 200ng of each DNA isolation into a well in a 96 well plate

- in an Excel spreadsheet, enter the DNA isolation code for each well, print it out
- beware of cross-contamination when opening the lid

5.1.6 digest 200ng of DNA from each individual with SbfI HF

- make master mix of 20×:
  - 2.0μl 10X NEBuffer 4
  - 0.25μl SbfI (20 U/μl) $\rightarrow$ 5 U/sample
  - 10μl ddH_2O
- fill with ddH_2O to 20μl endvolume
- mix by pipetting, shake the plate at the end, spin down with centrifuge
- incubate for 1.5 hours at 37°C
- heat-inactivate in thermal cycler at 65°C for 20 min [1]
- allow to cool down slowly to room temperature in thermal cycler (30 min)

5.1.7 ligate P1 adapters with different barcodes to each restriction digest

- put the sample plate on ice
- thaw P1 adapter plate on ice, shake to mix, spin down, reseal the plate after use
- first add to each heat inactivated restriction digest:
  - 1.0 μl of 100nM (=0.1pmole) barcoded P1 adapter [2]
- then make master mix of 20×:
  - 1.0 μl 10X NEB Buffer 2 [3]

[1] no “cut-ligation” when using SbfI and barcodes ending with CC

[2] 0.5pmole/μg DNA; this is probably more than necessary, however, if you size select at 350bp or above adapter dimers shouldn’t be a problem

[3] to keep the concentration of monovalent cations at 50mM

  - 0.3 μl **r**ATP (100mM → end concentration 1 mM) [4]
  - 0.2 μl concentrated T4 DNA Ligase (2,000 NEB U/μl) [5]
  - 7.5 μl ddH2O

- add 9.0 μl of master mix to each well to a 30μl endvolume
- final monovalent cation concentration should be 50 mM [6]
- incubate at room temperature (RT) for 30 min, but over night in the fridge doesn't do any harm and might give higher yield
- heat-inactivate at 65°C for 20 min in thermal cycler
- allow to cool down slowly to room temperature

5.1.8 combine samples

- pool the 18 individual ligation mixes, making up ~540μl and ~3.6 μg DNA

5.1.9 shear DNA with Covaris machine in Edinburgh

- optimization needed, test the result of each Covaris setting with Agilent Bioanalyzer 2100 DNA 100 chip or agarose gel with GeneRuler 100bp DNA ladder

Table 5.1 optimal Covaris settings

duty cycle	10%
intensity	5
cycles/burst	200
duration	100sec

5.1.10 clean up sheared library

- with one Qiagen MinElute PCR Purification column (Cat. no. 28004), capacity each 5 μg DNA
- elute with 21 μl EB

[4] rATP powder dissolved in EB (pH 8.5) is stable; reduce freeze-thawing cycles; 0.1 mM ATP is as efficient as 1mM but a 10mM ATP concentrations inhibit ligations!

[5] 400U/sample corresponding to 2000 NEB U/μg DNA

[6] NEBuffer 4 contains 50mM potassium ions

5.1.11 size selection on agarose gel

- rinse the gel tank and use fresh buffer before running the gel[7]
- run elution (20μl) in one lane of 1.25% agarose gel with 10 μl 6X Orange Dye for 1h 15 min at 6.7 V/cm [8]
- the wells should be less than half full, otherwise migration of fragments will be distorted [9]
- load 20 μl 100bp ladder (20μg) in the left lane, leave 1 lane space between standard and library
- use fresh razor blade and UV transilluminator (long wave setting to minimize mutations) to cut out a size range of ~350-750 bp [10]

5.1.12 gel extraction of DNA

- split the gel slice in half
- melt gel at RT with frequent vortexing
- use one column of Qiagen MinElute Gel Extraction kit for the gel slices [11]
- elute with 20μl

5.1.13 blunt-ending of sheared fragments

- NEB Quick Blunting Kit (Cat. no. E1201S)
- to the eluate from the previous step, add:
  - 2.5 μl 10X blunting buffer
  - 2.5 μl dNTP mix (1mM) [12]

[7] if you have run a gel from a different library before, otherwise not necessary

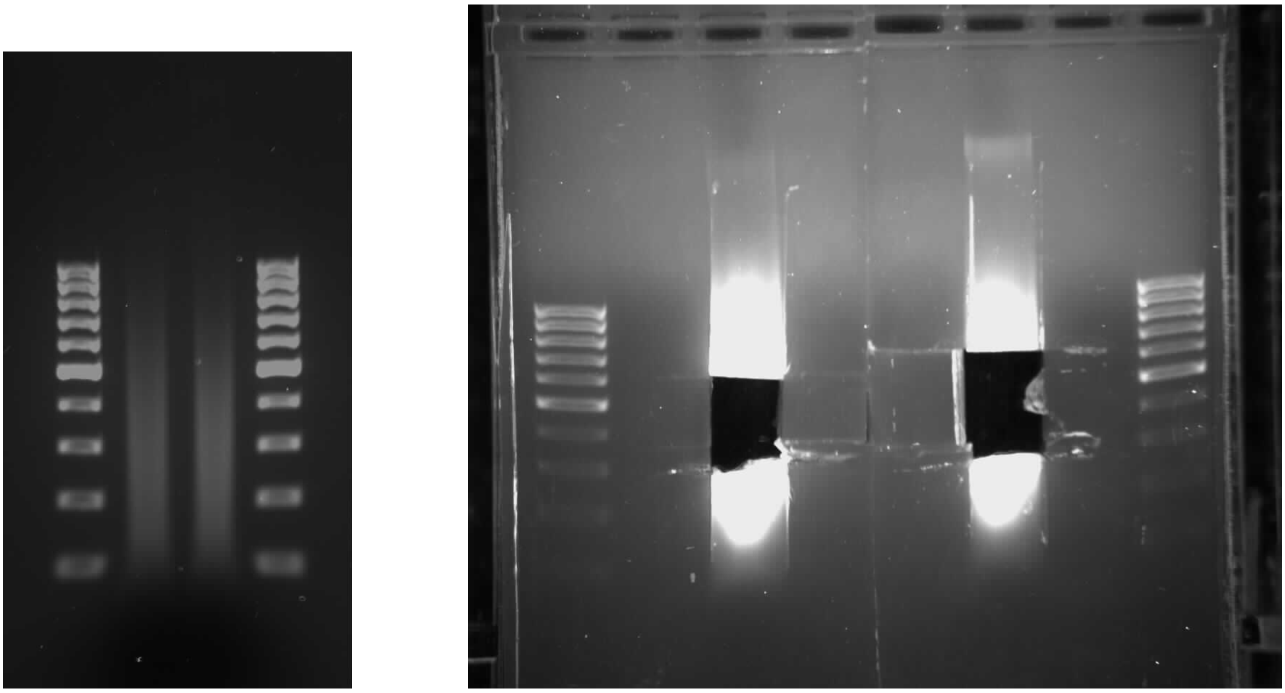
[8] ~3.6μg DNA per lane, 2X Orange Dye is necessary to make the DNA sink into the gel well, a lot of DNA could otherwise be lost at this step

[9] 5–6mm wide wells

[10] when making the vertical cuts, be sure not to go below 350bp, otherwise risk of adapter dimer contamination (Maureen Liu)

[11] even though it says otherwise in the manual of the kit, you can gel extract with just one column any amount of gel as long as it is completely melted before loading

[12] be sure to add before enzyme mix!



(a) two pools sheared with the Covaris settings from Table 5.1

(b) after size selection of the pools

Fig. 5.2 Size selection of fragments after shearing.

  - 1.0 μl Blunt Enzyme mix

- incubate at RT for 30 min

5.1.14 clean up blunt-ended library

- with Qiagen MinElute column
- elute with 42μl EB into a tube containing 5μl 10X NEBuffer 2

5.1.15 A-tailing

- to the eluate of the previous step, add:
  - 1.0 μl **d**ATP (10mM) → 0.2 mM final concentration
  - 3.0 Klenow fragment (3' – 5' exo$^-$)
  - incubate at 37°C for 30 min
- allow to cool down slowly to RT

5.1.16 clean up A-tailed library

- with Qiagen MinElute column
- elute with 43μl EB into a tube containing 5.0 μl 10X NEBuffer 2 [13]

5.1.17 P2Y-adapter ligation [14]

- to the eluate of the previous step, add:
  - 1.0 μl of 10μM P2Y adapter (=10pmole)
  - 0.5 μl **r**ATP (100mM) → 1.0 mM final concentration
  - 0.5 μl T4 DNA Ligase
- incubate at RT for 30 min, but over night in the fridge doesn't do any harm and might give higher yield

[13] final NaCl concentration 50mM, necessary to keep P2Y adapter annealed, salt concentrations of 100mM could decrease ligation efficiency (from NEB FAQ)

[14] Let the adapter oligos anneal slowly over a couple of hours in the heat block or thermal cycler. The adapter can be tested through ligation to a Taq PCR product and subsequent test PCR with the following P2Y primer: 5' – TCTCGGCATTCCTGCTGAAC – 3' (Kang-Wook Kim)

5.1.18 clean up P2 ligated library

- with MinElute column
- elute with 50 μl EB

5.1.19 RAD tag test amplification

- set up PCR:
  - 6.5 μl H2O
  - 12.5 μl 2x Phusion Mastermix [15]
  - 2.5 μl P1 primer (10μM) → 1.0μM end concentration [16]
  - 2.5 μl P2 primer (10μM)
  - 1.0 μl RAD library template
- use filter-tips or different pipettes for anything post-PCR !

Table 5.2 PCR programme

98°	30sec	
98°	10sec	×25
65°	30sec	
72°	20sec[a]	
72°	5min	
4°	∞	

[a] 30sec per kb recommended

5.1.20 check RAD amplification on a gel

- on 1% agarose gel put:
  - 5 μl of PCR product
  - 2μl GenRuler 100 bp DNA Ladder

[15] Do not use Phusion PCR kit with standard dNTP's. Phusion only works with high quality dNTP's !

[16] I found that a much higher primer concentration than usual can greatly increase yield

  - 1 μl RAD library template

- PCR product must be at least twice as bright as the template and template should be faintly visible

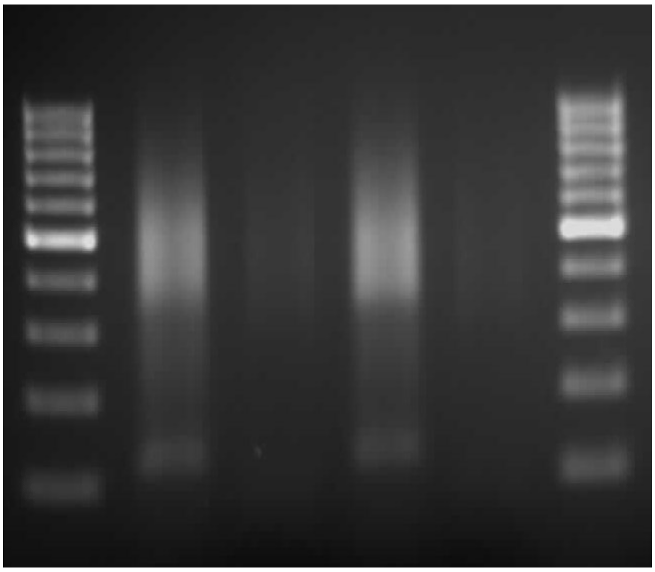


Fig. 5.3 PCR products of test amplifications of two libraries left of their respective templates. Adapter dimers are visible at ~130bp.

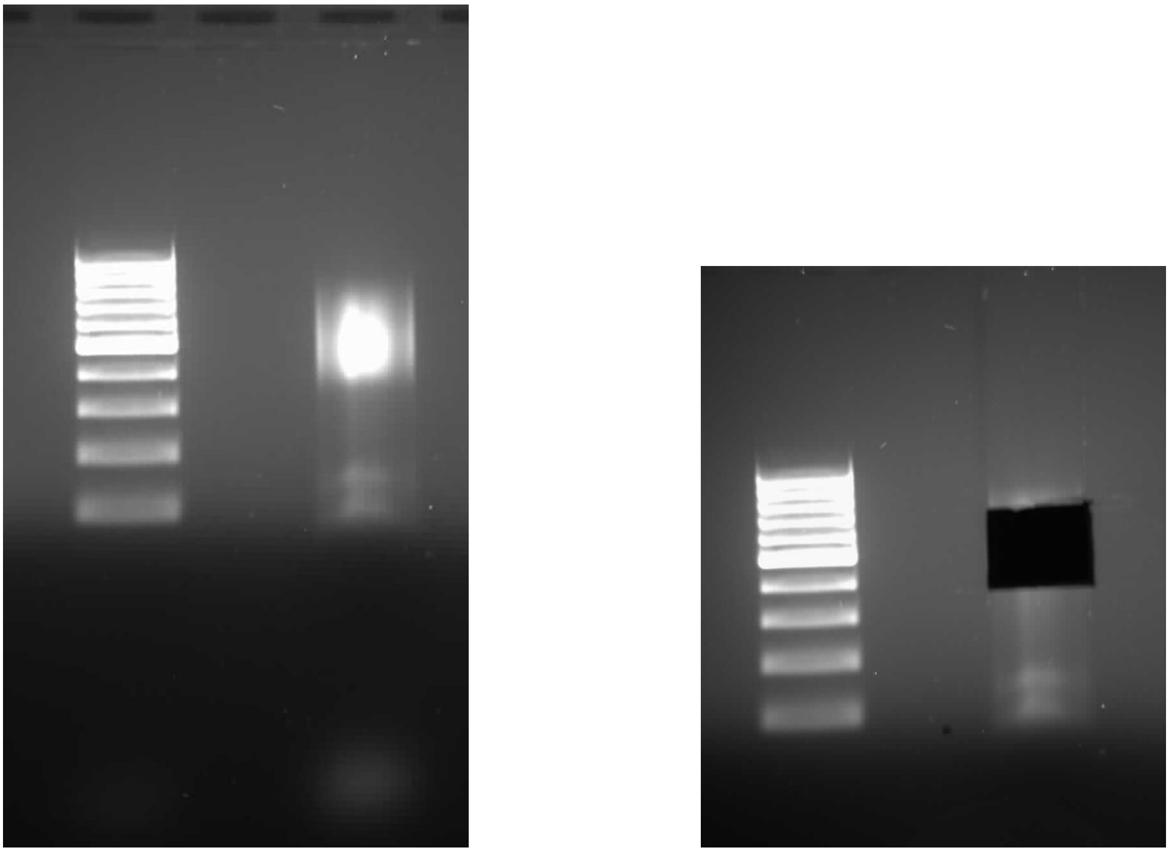
5.1.21 perform a 100μl PCR

- with 4μl template if very strong PCR product from test PCR, otherwise up to 30μl template
- 18 PCR cycles only in order to minimize PCR duplicates and bias
- purify the PCR product with Qiagen MinElute PCR purification column, elute with 20μl
- use filter-tips for anything post-PCR !

5.1.22 gel purification of amplified library

- use filter-tips for anything post-PCR !
- rinse the gel tank and use fresh buffer before running the gel if you have run a different library before

- run elution (20μl) in one lane of 1.25% agarose gel with 10 μl 6X Orange Dye for 1h 15 min at 6.7 V/cm
- the wells should be less than half full, otherwise migration of fragments will be distorted [17]
- load 20μl 100bp ladder (20μg) in the left lane, leave 1 lane space between standard and library
- use fresh razor blade and UV transilluminator (long wave setting to minimize mutations) to cut out a size range of ~350-750 bp [18]



(a) before size selection

(b) after size selection

Fig. 5.4 Size selection of the amplified RAD library. Adapter dimer bands are clearly visible. About ~0.03% of the reads from this library came from adapter dimers or sequences with very small genomic inserts.

[17] 5–6mm wide wells

[18] adapter dimers run at ~130bp, when making the vertical cuts, be sure not to go below 350bp, otherwise risk of adapter dimer contamination (Maureen Liu)

5.1.23 gel extraction

- filter-tips !
- with MinElute Gel Extraction Kit
- melt agarose slice at room temperature with frequent mixing
- elute in 20 μl EB

5.1.24 quantify molar concentration of RAD tags

- determine DNA concentration of the library with fluorimeter twice and each with at least 3 replicates of the calf thymus standard serial dilution
- determine size distribution and peak size of RAD tags with Agilent Bioanalyzer 2100 DNA chip or from agarose gel picture
- multiply peak size by 650 [g/mol] [19] to get the molecular weight of the library
- divide the DNA concentration of the library [g/μl] by it's molecular weight to get the molar concentration [nmole/L] of RAD tags in the library

5.1.25 validate library [20]

- A-tail PCR product
- T/A clone 1.0 μl of library into pGEM vector
- Sanger sequence some clones
- check for frequency of P1-P1 dimers, PCR duplicates [21] and blast the sequences

5.2 Double-Digest sRAD protocol

5.2.1 ingredients

- silica membrane genomic DNA extraction kit (e. g. Qiagen Dneasy Blood and Tissue Kit)

[19] the molecular weight of a base pair

[20] optional because of the cost and effort involved with cloning, but recommended before spending a lot of money on Solexa sequencing

[21] when P2Y adapter is at the same position in two clone sequences

- agarose
- fluorometer, Hoechst dye and standard solutions (e. g. Calf Thymus standard)
- SbfI High Fidelity from NEB
- EagI HF and AgeI HF from NEB
- thermal cycler
- 96-well PCR plates
- adhesive plate sealing film from qPCR machine
- plate centrifuge
- barcoded P1 adapters (at 100nM concentration)
- P2Y adapter with complementary sticky ends to the 6bp cutter used (and optionally containing barcodes)
- **r**ATP (100nM concentration)
- concentrated T4 DNA ligase
- Qiagen MinElute reaction cleanup kit (Cat. no. 28204)
- Glycerol
- 6x OG
- TBE
- QG buffer
- SybrSafe
- Blue-light transilluminator
- razor blades
- SpeedVac (or just table centrifuge)
- Phusion PCR mastermix
- P1 and P2 PCR primer
- filter tips
- BioAnalyzer
- ethidium bromide

5.2.2 protocol

5.2.3 isolate DNA from grasshopper hindleg

- with Qiagen Dneasy Blood and Tissue Kit (spin columns)

5.2.4 check the quality of the isolations

- by running each isolation on a 1.0% gel

5.2.5 quantify DNA samples with fluorometer twice

- produce at least three replicates of calf thymus standard serial dilutions for the standard curve
- produce at least 5 points for the standard curve spanning from ~200ng/μl to ~12.5ng/μl
- remove dodgy measurements of the standard in order to increase r^2 to at least 0.99
- get at least two concentration measurements per DNA isolation

5.2.6 digest 132 ng of DNA from each individual with SbfI HF and XhoI

- make master mix of 40×:
  - 3.0μl 10X NEBuffer 4
  - 3.0μl 10x BSA
  - 0.5μl SbfI-HF (20 U/μl) → 10 U/sample → 72U/μg DNA [22]
  - 0.5μl XhoI (20 U/μl) → 10 U/sample → 72U/μg DNA
  - 10μl ddH_2O
- based on the DNA measurements, adjust the amount of DNA isolation volume for the digestion, so that more or less equal amounts of each sample go into the library (see `Fluorimeter.ods`)
- fill with ddH_2O to 30μl endvolume

[22] This requires 20μl of the 25μl SbfI enzyme in a tube of 500 U.

- the total amount of DNA from all samples should not exceed the capacity of a MinElute spin column (5μg). Otherwise, two separate libraries have to be prepared.
- in an Excel spreadsheet, enter the code and volume of each DNA isolation in a layout that corresponds to the 96 well plate, print it out and use it as reference when pipetting
- beware of cross-contamination, particularly when opening the lid of the plate
- mix by pipetting, shake the plate at the end, spin down with centrifuge
- incubate for 3 hours at 37°C in a thermal cycler with heated lid
- heat-inactivate in thermal cycler at 65°C for 20 min, then ramp down to room temperature at no more than 2°C/min [23]

5.2.7 calculate the molar amount of sticky ends in the restriction digest

- Parameters:
  - genome size: 12×10^9bp
  - average molecular weight of a base pair: $660 \frac{g}{mol \times bp}$
  - expected number of SbfI restriction sites per genome (GC 46.5%): 135,633 [24]
  - amount of digested DNA in g per sample: 132×10^{-9}g
- SbfI sticky ends:

$$\begin{aligned}\text{molar amount of SbfI sticky ends} &= \frac{\text{amount of DNA}}{\text{MW of bp} \times \text{genome size}} \times \text{SbfI sites per genome} \times 2 \qquad (5.1)\\ &= \frac{132 \times 10^{-9} g}{660 \frac{g}{mol \times bp} \times 12 \times 10^9 bp} \times 135,633 \times 2\\ &= 4.52 \times 10^{-15}\text{mol}\\ &= \underline{\underline{\text{4.52fmol per sample}}}\end{aligned}$$

- XhoI sticky ends:

[23] adhesive films for qPCR plates only seal tight after heating via a heated lid beyond 70°C

[24] see `ComplexityReduction.xls` for the calculation of this number

- expected number of XhoI restriction sites per genome (GC 46.5%):
  2,509,115 [25] → 18.5 × SbfI sites

$$\begin{aligned}\text{molar amount XhoI sticky ends} &= 18.5 \times \text{molar amount of SbfI sticky ends} \qquad (5.2)\\ &= 18.5 \times 4.52 \times 10^{-15}\text{mol}\\ &= 83.6 \times 10^{-15}\text{mol}\\ &= \underline{\underline{83.6\text{fmol per sample}}}\end{aligned}$$

- in order to provide adapters in $\sim 10-20\times$ excess toward sticky ends, use 100fmol (=0.1pmol) P1 adapter per sample and 2pmole of P2 adapter per sample

5.2.8 set up a 10μM P2Y-XhoI adapter stock solution from oligos

- set up annealing buffer (AB) as shown in table 5.3

Table 5.3 annealing buffer set up:

NEB2 (10x)[a]	100μl
EDTA (100mM)[b]	110μl
ddH$_2$O	790μl
	1,000μl

[a] 1x NEB2 contains 10mM MgCl$_2$
[b] 0.372g EDTA dissolved in 10ml 1x NEB2

- ...split the volume into 100μl aliquots and heat them to 65° for 20 minutes [26]
- spin down lyophilised oligos in manufacturers tube for 1 min at maximum speed
- dissolve the lyophilised oligos with EB to 100μM
- then set up 10μM adapter solution with:
- ...and anneal the oligos by heating the mixture in the thermal cycler to 96°C for 2 minutes and then ramping down to RT at 2°/min

[25] see `ComplexityReduction.xls` for the calculation of this number
[26] in order to denature nuclease contamination

Table 5.4

upper oligo (100μM)	10μl
lower oligo (100μM)	10μl
AB	80μl
	100μl

5.2.9 ligate adapters to each restriction digest

- put the sample plate on ice
- thaw P1 adapter plate on ice, shake to mix, spin down, reseal the plate after use
- first add to each heat inactivated restriction digest:
  - 1.0 μl of 100nM barcoded P1 adapter → 0.1 pmol [27]
  - 2.0 μl of 1μM P2-XhoI adapter → 2.0 pmol [28]
- vortex plate and spin down
- then make master mix of 40×:
  - 0.8 μl 10X NEB Buffer 2 [29]
  - 0.4 μl **r**ATP (100mM → end concentration 1 mM) [30]
  - 0.2 μl concentrated T4 DNA Ligase (2,000 NEB U/μl) [31]
  - 5.6 μl ddH_2O
- add 7.0 μl of master mix to each well to a 40μl end volume and mix by pipetting up and down
- after carefully sealing the plate with adhesive film, vortex and spin down

[27] 0.757pmole/μg DNA; 22 fold excess of adapter to cohesive ends. If you size select at 300bp or above, adapter dimers shouldn't be a problem.

[28] 15 pmol/μg DNA; 23.9 fold excess of adapter to cohesive ends.

[29] adds 10mM NaCl to the final solution, final NaCl concentration ~50mM, which is necessary to keep the P2Y adapter double stranded; however, salt concentrations of 100mM decrease ligation efficiency (from NEB FAQ)

[30] rATP powder dissolved in EB (pH 8.5) is stable; reduce freeze-thawing cycles; 0.1 mM ATP is as efficient as 1mM but a 10mM ATP concentrations inhibit ligations!

[31] 400U/sample corresponding to 3030 NEB U/μg DNA

- final monovalent cation concentration should be ~50 mM [32]
- incubate at room temperature (RT) for 2 hours, then over night in the fridge
- heat-inactivate at 65°C for 20 min in thermal cycler, then ramp down to RT at 2°C/min

5.2.10 combine samples

- pool the 38 individual ligation mixes, making up ~1,520μl and ~5 μg DNA

5.2.11 clean up and concentrate the adapter ligated library

- with one Qiagen MinElute reaction cleanup column (Cat. no. 28204), capacity each 5μg DNA[33]
- use at least as much ERC buffer as ligation mix for the reaction cleanup kit
- elute with 15 μl EB

5.2.12 size selection on agarose gel

- rinse the gel tank and use fresh buffer before running the gel[34]
- make a 110ml 1% TBE gel with 6.3μl SybrSafe
- add 10μl 6x OG loading dye and ~5μl 100% Glycerol to the 15μl eluate of the last step [35]
- run the whole mix in one lane at 13 V/cm for ~45 min or when the orange dye just about reaches the bottom of the gel
- the wells should be less than half full, otherwise migration of fragments will be distorted → 5–6mm wide wells
- load 30μl 100bp ladder (50ng/μl) in the left lane, leave 1 lane space between standard and library

[32] NEBuffer 4 contains 50mM potassium ions

[33] ignore what the kit manual says about the maximum amount of enzymatic reaction that can be cleaned up per column

[34] if you have run a gel from a different library before, otherwise not necessary

[35] ~5μg DNA per lane, the glycerol is necessary to make the DNA sink into the gel well, a lot of DNA could otherwise be lost at this step

- use fresh razor blade and a blue light transilluminator to cut out a size range of ~300-800 bp (fig. 5.5) [36]
- cut the gel block into 4 pieces and put each into a 2ml tube
- weigh each tube and subtract the weight of an empty tube to get the gel weights in mg
- add 3× as much buffer QG to each tube as the weight of its gel piece
- rotate the tubes at RT for one hour to melt the gel pieces
- combine the dissolved gel pieces in a bigger vessel and add one gel volume (mg=ml) of isopropanol followed by mixing
- purify that solution over one MinElute spin column with a SpeedVac [37]
- elute with 30μl EB

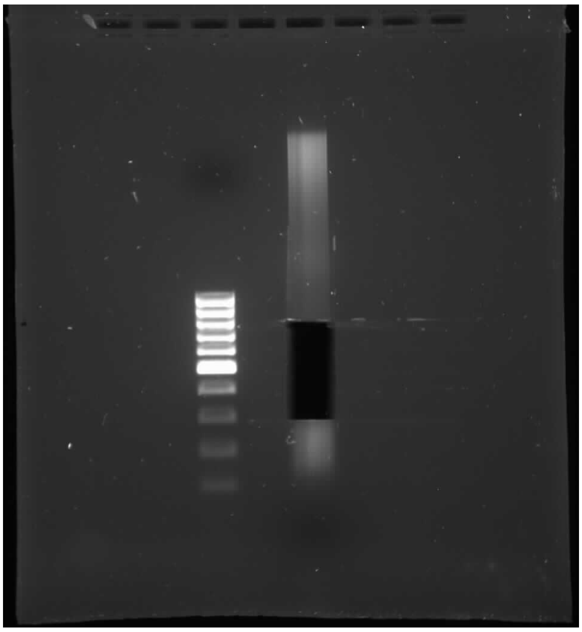


Fig. 5.5 Gel picture after size selection.

[36] be as accurate as possible with the vertical cuts, you can take your time, be sure not to go below 300bp, otherwise risk of adapter dimer contamination (Maureen Liu)

[37] even though it says otherwise in the manual of the kit, you can gel extract with just one column as long as the gel contains no more than 1% agarose and it is completely melted before loading

5.2.13 selective amplification

- save 6μl of the eluate of the last step for later qPCR
- then set up a PCR with the remaining 24μl eluate:
  - 6.0 μl H2O
  - 50.0 μl 2x Phusion Mastermix [38]
  - 10.0 μl P1-thiol primer (10μM) → 1.0μM end concentration [39]
  - 10.0 μl P2-thiol primer (10μM)
  - 24 μl RAD library template
- split the PCR mix into 5x 20μl volumes and add 10μl mineral oil to each PCR tube
- then run each tube with the PCR programme shown in table 5.5

Table 5.5 PCR programme

98°	30sec	
98°	10sec	×18
65°	30sec	
72°	30sec[a]	
72°	5min	
4°	∞	

[a] 30sec per kb recommended

5.2.14 purifcation and concentration of the selective PCR product

- use filter-tips or different pipettes for anything post-PCR !
- combine all PCR products and check 5μl of it on a 1.25% gel next to 2μl Genruler
- take 6μl of the PCR product for qPCR
- purify the rest of the PCR product over a MinElute column eluting with 10μl EB

[38] Do not use Phusion PCR kit with standard dNTP's. Phusion only works with high quality dNTP's !

[39] I found that a much higher primer concentration than usual can greatly increase yield

5.2.15 determination of template amount used for selective PCR by qPCR

- use filter-tips
- use 4μl of the 6μl of the PCR product set aside in the previous step to determine it's DNA concentration with the fluorometer and picoGreen dye and use it as a standard in the qPCR
- make 8x 1:10 serial dilutions of the PCR product in 10μl EB each to produce a standard curve
- set up 20μl qPCR reactions with SybrGreen PCR master mix and 2μl template:
  - three replicates of serial dilutions including negative control
  - three replicates of the template saved in step 5.2.13
- from the C_t values and the known amount of template molecules in the standard dilution, determine the amount of template molecules per locus and individual [40] (see `sRAD/qPCR/TRIAL_LIB_241011_data.xls`)

5.2.16 normalisation of the library

Unless the library looks like a homogeneous smear on the gel, it is advisable to normalise it.

- check activity of double strand specific nuclease (DSN) with control template from the kit
- use filter-tips
- take 6μl of the purified PCR product and add 2μl 4x Hybridization buffer [41]
- put 4μl of the mixture in each of two tubes, labeled “1/8” and “1/16”
- overlay the reaction mixtures with 10μl mineral oil and spin down for 2 min at max speed
- in a thermal cycler, heat the mixture to 98° for 2 min

[40] This can be used to predict the expected proportion of false homozygote genotype calls due to PCR drift and false heterozygote genotype calls due to high coverage PCR mutations.

[41] 500mM final NaCl concentration for annealing. EB contains 10mM Tris-HCl at pH 8.5, 1x hybridization buffer contains 50mM HEPES at pH 7.5. So the pH in the mixture should be close to 7.5.

- . . . then incubate at 68° for 5 hours
- put DSN master buffer into thermal cycler to preheat it to 68°
- make a 1/8th and 1/16 dilution of the DSN enzyme with DSN storage buffer
- keep the thermal cycler at 68°C and add 5μl preheated DSN master buffer to each tube while keeping the tube in the cycler, then flick, spin down briefly and immediately put back into the thermal cycler
- incubate for 10 minutes at 68°C
- add 1μl of 1/8th or 1/16th DSN enzyme into each tube respectively [42]
- incubate for 25 minutes at 68°C
- add 10μl DSN stop solution, mix and spin [43]

5.2.17 test PCR of normalisation

- use filter-tips
- from each vial of the last step (i. e. “1/8” and “1/16”), set up three 10μl test PCR’s with 1μl template
- run PCR for 5, 10 and 15 cycles with temperature steps as in table 5.5 and check 5μl PCR product on a 1.25% EtBr gel next to 2μl Genruler
- examine the PCR product: homogeneous smear in right size range?, 5 cycle PCR product visible? (see figure 5.6)
- decide which DSN dissolution produced the better result

5.2.18 amplification of normalised library

- use filter-tips
- set up a 50μl PCR as follow:

[42] just add the enzyme, don’t flick, don’t spin down, just be quick, i. e. no more than 10 seconds for this step!

[43] the reaction mixture contains 5mM $MgCl_2$ and the stop solution contains 5mM EDTA to neutralise it. Figure 4(B) in the TRIMMER kit manual suggests that inactivation of DSN by heating is not guaranteed to be complete.

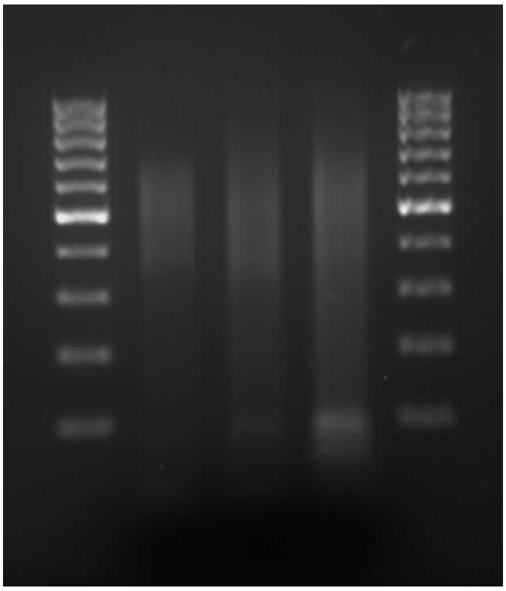


Fig. 5.6 Gel picture of 5, 10 and 15 cycle test PCR's (from left to right).

  - 25μl 2x Phusion Master mix
  - 5μl P1-thiol primer (10μM)
  - 5μl P2-thiol primer (10μM)
  - 10μl normalised library
  - 5μl ddH_2O

- run the PCR programme in table 5.5 for 5-7 cycles, depending on the outcome of the test PCR [44]

5.2.19 purification of the library

- use filter-tips
- purify the 50μl PCR product over a MinElute column
- elute with 25μl EB
- use 4μl for DNA concentration measurement with fluorometer
- calculate an estimate of the molar concentration of the library

[44] These additional PCR cycles do not further bias the library's representation of the pre-normalisation template since this PCR starts with a lot of template molecules as indicated by the few cycles necessary to create a visible PCR product.

5.2.20 validate library [45]

- A-tail PCR product
- T/A clone 1.0 μl of library into pGEM vector
- Sanger sequence a few dozen clones
- check for whether the sequences contain a P1 adapter sequence on one end and a P2 adapter sequence on the other

[45] optional because of the cost and effort involved with cloning, but recommended before spending a lot of money on Solexa sequencing