SUPPLEMENTARY DATA

This supplement contains additional information on the results and the datasets as well as the description of our novel in-house SOLiD RAD tag sequencing protocol.

5 ADDITIONAL INFORMATION

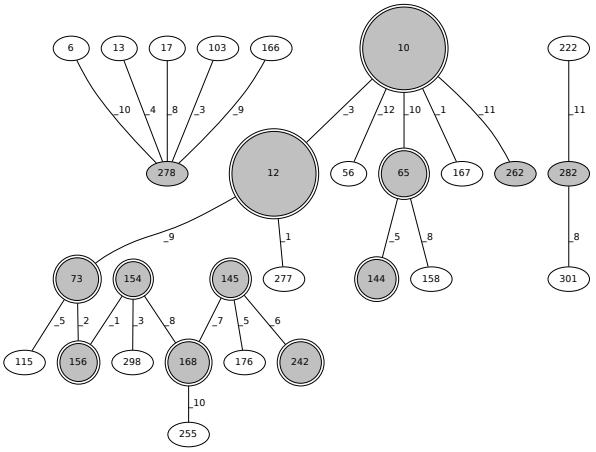

Fig. 4. ScaffoldHMM analysis assigns grey paternal linkage groups (= bins) to chromosome 2. By manual inspection, nodes with double lines are decided to be correct bins for this chromosome. An edge is drawn between linkage groups if their segregation patterns have a Hamming distance one, thus a single recombination in one individual explains the difference in the patterns. The number on the edge indicates which individual (1..12) has the Hamming error. The size of the grey nodes is proportional to the number of markers in the corresponding group. In this case, the order of the bins is found to be 144, 65, 10, 12, 73, 156, 154, 168, 145 and 242.

Chromo-some	Supporting markers	Scaffolds	Length (Mbases)	Length (cM)	Bins	NimbleGen length (cM)	Bins covered
1 (Z)	1566	202	14.0	50.0	7	33.0 / 32.7	6
2	1419	130	11.6	75.0	10	37.7 / 31.3	6
3	1237	124	11.5	66.7	9	45.4 / 40.6	8
4	1203	125	11.4	50.0	7*	66.2 / 61.2	7
5	1355	114	10.9	33.3	5	64.3 / 62.1	4
6	1212	121	10.6	58.3	8	66.3 / 39.7	4
7	1276	127	10.4	58.3	8	32.6 / 32.6	7
8	1094	108	10.4	33.3	5	67.8 / 55.4	5
9	1204	107	10.4	66.7	9	52.8 / 52.6	8
10	1039	127	10.1	25.0	4	56.4 / 38.9	4
11	1063	120	10.1	58.3	8	44.9 / 40.7	4
12	1104	112	9.8	66.7	9	58.8 / 51.9	5
13	958	109	9.6	50.0	7	60.5 / 59.7	5
14	998	112	9.6	50.0	7	61.8 / 61.0	6
15	1059	108	9.2	58.3	8	90.0 / 68.6	6
16	939	100	9.0	58.3	8	48.1 / 41.9	8
17	1000	93	9.0	58.3	8	78.4 / 71.5	4
18	951	98	8.9	58.3	8	64.5 / 39.0	5
19	868	100	8.4	41.7	6	55.4 / 47.5	3
20	951	87	8.2	66.7	9	55.3 / 41.1	9
21	892	100	8.0	66.7	9*	61.8 / 53.6	8
22	642	91	7.8	66.7	9	46.8 / 46.4	5
23	736	105	7.8	58.3	8	84.9 / 65.4	6
24	673	74	6.4	66.7	9*	36.0 / 35.7	3
25	635	87	6.3	50.0	7	47.6 / 32.5	3
26	679	73	6.2	58.3	8	43.1 / 42.2	8
27	455	63	5.4	25.0	4	65.0 / 64.9	3
28	408	70	3.9	50.0	7	42.0 / 36.2	3
29	434	81	3.2	50.0	7	37.4 / 37.4	6
30	346	59	3.0	41.7	6**	33.5 / 25.1	4
31	327	78	2.3	25.0	4	65.6 / 57.3	3
total	28723	3205	263.4	1641.6	228	1703.5 / 1466.3	

Table 2. The number of supporting markers, scaffolds, bins, and the total (non-chimeric) scaffold length of each chromosome of the Glanville fritillary linkage map based on RAD tag data (left). The length in cM is computed as $\frac{100(x-1)}{12}$ cM for a chromosome with x bins. Asterisk (*) after number of bins indicates missing bins in the order. For comparison, the right-hand side gives the length of each chromosome given by the NimbleGen data and the number of bins that are covered by the data. Two lengths are given for the NimbleGen data, the first one is obtained without modeling genotyping errors and the second one with an error model.

5.1 The Glanville fritillary butterfly data

Of the 4,989 SNPs in the NimbleGen dataset, 4,633 SNPs (probe sequences) map uniquely to the partial reference genome of the Glanville fritillary, which has been recently sequenced and will be published in the near future (Lehtonen *et al.*, in prep.). 14 individuals have been genotyped twice and two individuals three times (duplicated individuals) and about 100 SNPs are included twice in the datasets (duplicated SNPs). Given the genotype calls provided by Roche, the error rate based on duplicated individuals is 2.1% and 747 SNPs have genotype errors. Moreover, there are 840 SNPs with at least one Mendelian error taking into account only the 74 male offspring (errors in females could be due to ZW sex chromosome). The Mendelian error rate is 5.4%, but if we discard 747 SNPs with errors on duplicates, the error rate decreases to 3.8%. The final dataset was filtered by setting all genotype calls with a posterior probability (provided by Roche) lower than 0.99 as missing.

For the RAD tag data, the genotypes were called from mapped read counts with module Counts2Genotypes of Lep-MAP. The counts for the homozygous genotypes were modeled as following binomial distributions $\mathrm{Bin}(a+b,\epsilon)$ and $\mathrm{Bin}(a+b,1-\epsilon)$, where a and b are the counts for the two SNP alleles and ϵ is the read error probability. Furthermore, the counts for the heterozygous genotype were modeled with distribution $\mathrm{Bin}(a+b,0.5)$. The parameter ϵ was chosen to maximize the joint likelihood of the data. The modeling assumed that the genotypes obey Mendelian inheritance, and hence no Mendelian errors were present in the final genotype calls. The error rate in the genotype calls was estimated to be less than 0.5%, based on the number of small maternal LGs whose segregation pattern was not any of the 31 patterns corresponding to the chromosomes.

The number of offspring (12) in the RAD tag data was close to the minimum for linkage map construction. The probability of being able to separate the 31 chromosomes with these data is 80% (probability that 31 maternal inheritance patterns out of $2048 = 2^{12-1}$ are all different is $\prod_{i=1}^{31}\frac{2048-i+1}{2048}$).

Figure 4 shows how chromosome 2 was divided into 10 "bins". Mapped markers inside the scaffolds containing markers from the bins' LG were inspected to rule out false bins caused by genotyping errors. All bins are illustrated in Figure 5. Table 2 compares the two linkage maps, constructed from NimbleGen and RAD tag data, for each chromosome.

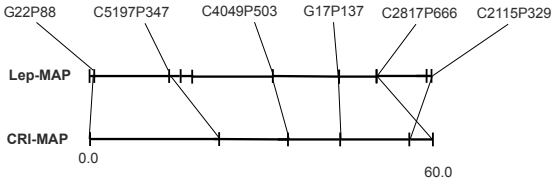

Fig. 6. The comparison of marker orders for one chromosome of the squinting bush brown butterfly. Above is the result obtained with Lep-MAP and below is the result reported in (Beldade *et al.*, 2009) (CRI-MAP).

5.2 The squinting bush brown butterfly data

Figure 6 shows one example of different marker orders on the public squinting bush brown butterfly data (Beldade *et al.*, 2009), obtained by Lep-MAP and reported in (Beldade *et al.*, 2009), where the difference is in the position of SNP C2817P666.

6 SOLID RAD TAG SEQUENCING PROTOCOL

The Restriction site Associated DNA (RAD) tag sequencing (also called RAD-Seq) methods have gained popularity among methods for studying polymorphisms and genotyping in various genomes (Miller *et al.*, 2007; Baird *et al.*, 2008). The advent of NGS (Next Generation Sequencing) methods has opened the possibility of performing population wide studies in species without a reference genome. We modified the original RAD tag sequencing protocol for use with the SOLiD sequencer.

6.1 Genomic DNA digestion

The samples (1 µg gDNA) were digested using 2 µl BspHI (PagI) (Thermo Scientific) and 2.5 µl 10x buffer O (rounded to 25 µl) at +37 °C for 1 hour, and deactivated at +80 °C for 20 minutes.

6.2 Annealing of BC and Multiplex P1 Adapters

The oligos were annealed according to Solid 4 Fragment library preparation guide to produce adaptors for multiplexing samples. Briefly, 4 µl of adaptors A and C (125 µM) were separately mixed with 4 µl of B and D (125 µM), respectively, before 1 µl 10x T4 DNA ligase buffer and 1 µl water were added, totaling 10 µl in volume. The oligos were annealed gradually from +95 °C in a thermocycler and the sequences for each are listed below.

BC and Multiplex P1 Adapters:

A: BC_adapter_Amino 5' Amino-C6-CTG CTG TAC GGC CAA GGC G
B: P2_BC_a_NcoI 5' **CAT G**CG CCT TGG CCG TAC AGC AG
C: Multiplex P1_a 5' CGC TTT CCT CTC TAT GGG CAG TCG GTG *A*T
D: Multiplex P1_b 5' Phos TCA CCG ACT G*T*T*T*T

In the adapter sequences, asterisk (*) marks a phosphorothioate bond, Phos marks a 5' phosphorylation and the restriction site overhang is bolded.

6.3 Ligation of the BC Adapter

The BC adaptor was ligated to digested gDNA, using 1 µl of 50 µM BC adapter, 2 µl of 10 mM rATP, 4 µl 10x T4 DNA ligase buffer, 1 µl T4 DNA ligase (30U/ µl), made up to 40 µl with water. The samples were incubated at +25 °C for 2 hours before heat inactivation at +65 °C for 20 minutes. DNA was purified by DNA Clean & Concentrator-5 kit (Zymo Research) and eluted in 20 µl.

6.4 Fill-in reaction

Next, the ligated complex (20 µl) went through a fill-in reaction with 10x Biotools buffer, 10 µl, dNTPmix (25mM) 0.7 µl, Biotools Polymerase (5U/ µl) 1.0 µl, water 68.3 µl with total total 100 µl. This mixture was incubated at +68 °C for 20 minutes in a PCR block.

6.5 Shearing

After the fill-in reaction, the mixture was sonicated according to Solid 4 protocol using a Covaris sonicator with set up as following; water bath temperature below +5 °C, tubes: Microtube (6x16mm) 100 µl/tube, program: microtube 165bp. After sonication, the mixture was purified with Qiagen QIAquick PCR Purification kit according to the manufacturer's instructions with elution volume 2x25 µl using EB-buffer from the purification kit. Concentrations after sonication were measured by NanoDrop.

6.6 End repair

The DNA fragments were blunt-end repaired using a mixture of enzymes 5 µl 10x T4 PNK buffer, 10 µl 5x T4 DNA Polymerase Buffer, 1.6 µl dNTP mix (25 mM), 1 µl T4 PNK (10 U/ µl), 1 µl T4 DNA polymerase (5 U/ µl), 0.4 µl DreamTag polymerase (5 U/ µl), 0.5 µl BSA (10 mg/ µl) and 2.5 µl ATP (10 mM) were mixed with 50 µl sheared DNA before being made up to 100 µl with 28 µl water. The solution was incubated for 1 hour at +25°C and then at +65 °C for 20 minutes to deactivate the enzyme mix and add the A-tail. DNA was purified with Qiagen QIAquick PCR purification kit and eluted in 50 µl EB.

6.7 Multiplex P1 Adapter ligation

The DNA fragments were ligated with the Multiplex P1 Adapter using 10x T4 DNA ligase buffer, 10 µl, Multiplex P1-Adapt.(50 µM), 1 µl. T4 DNA ligase (HC;30U/ µl) 1 µl, DNA 50 µl, and water 38 µl with a total volume of 100 µl. The mixture was then incubated for 1 hour at +25 °C, then at +65°C for 15 minutes in a PCR block.

6.8 Purification and size selection

The samples were purified using Agencourt AMPure XP magnetic beads (Beckman Coulter) according to the manufacturers instructions to remove unligated P1 adaptors and eluted in 20 µl EB.

6.9 PCR

The purified library was amplified using Multiplex PCR primer 1 and Multiplex P2_BCXX (Barcode primer). The reaction mixture consisted of 10x Maxima Hotstart buffer 5 µl, dNTP mix (25mM) 0,4 µl, Multiplex PCR primer 1 (10 µM), 1 µl, BCXX primer (10 µM) 1 µl, Maxima Hotstart polymerase 0.5 µl, $MgCl_2$ (25mM) 5.0 µl, water 34.1 µl, adding 3 µl of template with total volume of 50 µl. The mixture was amplified by PCR (1: 95 °C 4min, 2: 95 °C 30s, 3:

55 °C 30s, 4: 72 °C 1min, go to 2. x34 5: 72 °C 5min 6: 4 °C ∞). The PCR products were checked in 2% agarose gel.

Primer sequences:

Multiplex PCR primer 1
5' CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCGGTGAT
Multiplex P2_ BCXX
5' CTGCCCCGGGTTCCTCATTCTCTXXXXXXXXXXCTGCTGT ACGGCCAAGGCG
X= barcode 10 bp

6.10 Purification, concentration measurement and pooling

The PCR reaction was purified using AMPure XP and concentration measurement of the libraries were measured by Qubit HS and size selection and quality checked with Bioanalyzer DNA High Sensitive DNA chip. Equal amounts of the barcoded samples were pooled for run on the SOLiD sequencer following the SOLiD protocol.

REFERENCES

Baird, N. *et al.* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE*, **3**(10), e3376.

Beldade, P. *et al.* (2009). A gene-based linkage map for *bicyclus anynana* butterflies allows for a comprehensive analysis of synteny with the lepidopteran reference genome. *PLoS Genet*, **5**(2), e1000366.

Miller, M. *et al.* (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res.*, **17**(2), 240–248.

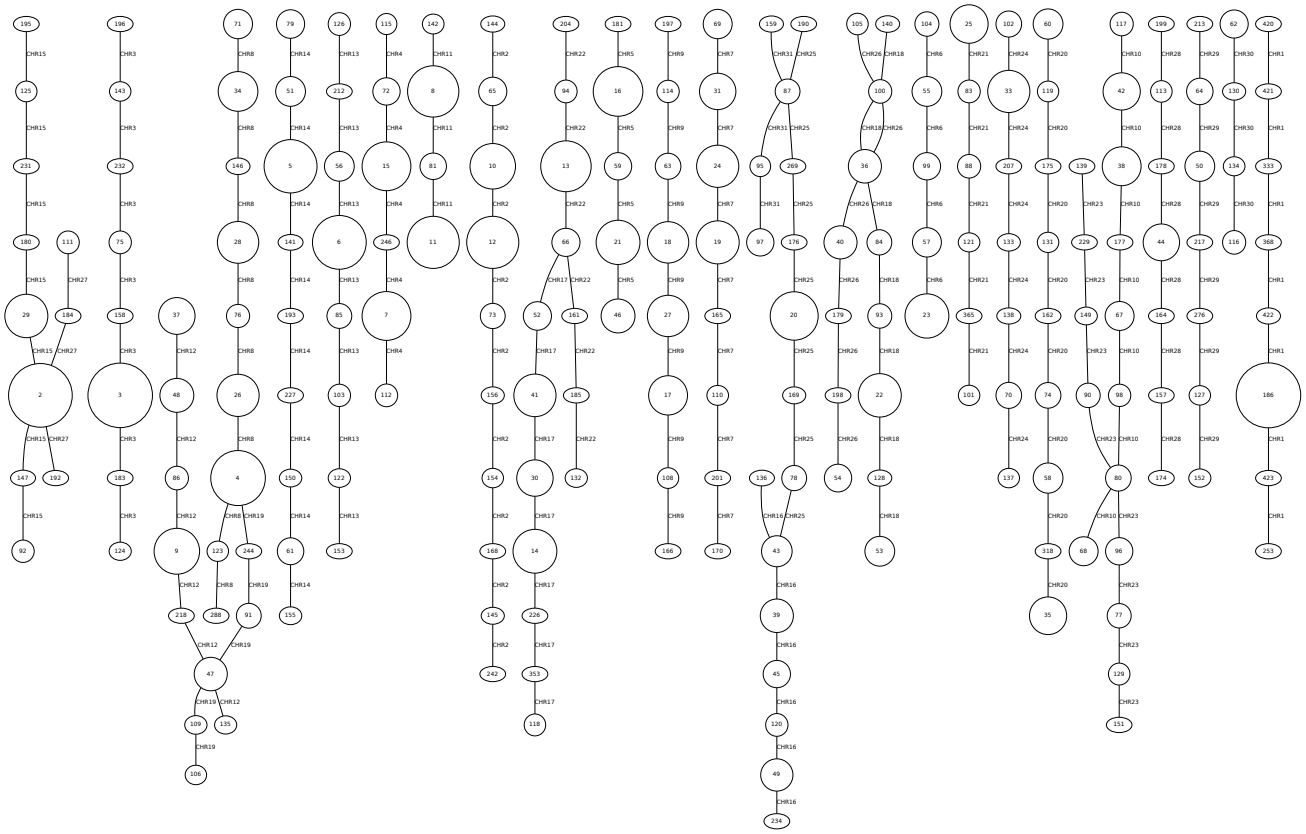

Fig. 5. Visualization of all the bins. Edges are drawn between adjacent bins and each edge is labeled with the corresponding chromosome. Nine bins are shared by two or more chromosomes and five bins are missing, hence five adjacent bins have a Hamming distance of two.