List of supplementary Information

Supplementary Fig. 1: This file shows whole genome scan association results using the *HaxB4* (Black-4nSpots) assembly considering each of the four main colour pattern forms as a covariate separately. Among the 17,482,401 autosomal SNPs analysed, 388 showed a very strong association (Bayes factor > 30 db) with the Red-nSpots form proportion, the vast majority (89%) of which belonged to the extended Black4Spots allele sequence.

Supplementary Fig. 2: This file shows genome scan association results focusing on the colour pattern genomic region of the *HaxB4* (Black-4nSpots) assembly considering each of the four main colour pattern forms as a covariate separately. The 18 SNPs with the strongest association signals (Bayes factor > 100 db) delineate a 203 kb region (from position 922 to 1,123 Mb), representing the strongest candidate region for the colour pattern locus on the *HaxB4* assembly.

Supplementary Fig 3: This file shows a schematic view of the BAC contig covering the sequence of the Black-4Spots allele of the colour pattern locus.

Supplementary Fig. 4: This file shows an alignment of the sequence of the Black-4Spots allele on the *HaxR* assembly (i.e. Red-nSpots allele). Only similarity segments > 1.5 kb and with an overall identity >98% are represented. The black star indicates the position of the gene *pannier* in both sequences and the white rectangle the region showing high sequence divergence between the alleles of the two forms.

Supplementary Data 1: Consensus sequence (2.87 Mb in fasta format) including the candidate genomic region of the Black-4Spots allele of the colour pattern locus as well as adjacent regions.

Supplementary Table 1: This file contains information on statistics characterizing reads for the four different MinION sequencing flow-cells (*HaxR* assembly).

Supplementary Table 2: This file contains information on statistics characterizing the read mapping for each of the six individuals used for the sexing of *HaxR* contigs.

Supplementary Table 3: This file contains information on the ratio of female-to-male read coverage for each *HaxR* contig. For X-linked contigs, the coverage is halved in males.

Supplementary Table 4: This file contains information on the preparation kit, sequencer type and read mapping statistics for the 14 different DNA pools used in genome-wide association study.

Supplementary Table 5: This file contains information the DNA libraries used for the assembly *HaxB4*.

Supplementary Table 6: This file contains primer sequences used for qPCR.

Supplementary Information

Analysis of read mapping coverage of pool-seq NGS data on both the extended Red-nSpots (*HaxR* assembly) and Black-4Spots (*HaxB4* assembly) allele sequences reveals large scale sequence divergence among the four main color form alleles of *Harmonia axyridis*.

Motivation

Comparing *de novo* sequences surrounding *pannier* for the Red-nSpots and Black-4Spots alleles highlighted large scale divergence in an upstream region covering ca. 170 kb in the *HaxR* and *HaxB4* assemblies (see the main text and Extended Data Fig. 7). Such large scale sequence divergence explained the clustering of SNPs harbouring strong weights of evidence in favour of association signals with the proportion of Red-nSpot or Black-4Spots individuals in each sequenced pool considered as covariate in the genome-wide association analysis (Extended data Fig 2, Supplementary Fig 2). The genome-wide association design was sub-optimal for the Black-2Spots form due to allele dominance and the Black-nSpots form only represented by a single pool. We nevertheless observed a similar clustering of strongly associated SNPs in the same genomic region (Extended Figure 2, Supplementary Figure 2). This suggested extended sequence divergence in the upstream region of *pannier* for the Black-2Spots and Black-nSpots alleles too.

To further assess the level of sequence divergence between the alleles of the four main colour pattern forms, and this especially for the Black-2Spots and Black-nSpots alleles that were not *de novo* sequenced, we examined read mapping coverage of the pool-seq data representative of these forms. The rationale of this approach was that extended divergence in the sequences of an allele represented at high frequency in a given pool, relatively to the reference assembly on which reads are mapped, is expected to translate into a local decrease in read coverage.

Methods

We considered sequence data available for: (i) the four pools (CH2-R, BRG-R, WAS-R and BIO-R) including Red-nSpots individuals only (i.e., with a Red-nSpots allele frequency equal to 1); (ii) the three pools (CH2-B4, BIO-B4 and BRG-B4) consisting of Black-4Spots individuals only (i.e., with an expected Black-4Spots allele frequency ≥ 0.5 due to the possible presence of Red-nSpots alleles with an hidden expression in heterozygous Black-4Spots/Red-nSpots individuals); (iii) the CH2-B2 pool consisting of Black-2Spots individuals only (i.e., with an expected Black-2Spots allele frequency ≥ 0.5 due to the possible presence of Red-nSpots or Black-4Spots alleles with an hidden expression in heterozygous Black-2Spots/Red-nSpots or Black-2Spots/ Black-4Spots individuals); and (iv) the NOV-Bn pool consisting of Black-nSpots individuals only (i.e., with an expected Black-nSpots allele frequency close to one due to the fixation of the Black-nSpots form in the corresponding wild population). The five other remaining DNA pools that were used for the genome-wide association study were not considered here because they either consisted of mix of individuals from two melanic colour pattern forms (cf. the CH2-B2 pool) or they displayed a genome-wide median coverage ≤ 25 (Supplementary Table 4).

Based on the mpileup file combining the mapping results of the pool-seq data onto the Red-nSpots assembly *HaxR* (see Methods), we computed read coverages (using default options of the samtools 1.3.1 depth program) at each position of the utg676 contig covering the Red-nSpots allele for the above nine DNA pools of interest. After discarding positions covered by <10 reads over all the pools or with a within pool coverage >95th percentile in at least one pool, the utg676 contig was divided into consecutive windows of 10,000 positions with an overlap of 5,000 positions. Let $c_{i,p}$ represent the window coverage (computed as the average coverage over the 10,000 window positions) for window *i* in pool *p*; m_p the mean genome wide coverage (computed over all HaxR autosomal contigs excluding utg676 and over-covered position) in pool *p*; and $s_{i,p}=c_{i,p}/m_p$ the standardized window coverage for window *i* in pool *p*. To identify regions with lowered read mapping efficiency due to high sequence divergence with respect to the Red-nSpots allele (*HaxR* assembly), we computed a relative (standardized) window coverage as $rc_{i,p}=s_{ip}/s_i^R$ (for window *i* in pool *p*) where s_i^R is the standardized coverages for window *i* computed after merging reads from the four Red-nSpots representative pools. This correction allowed accounting for local variation in window coverage shared across experiments. Supplementary Information Fig. 1 plots the estimated $rc_{i,p}$ for each pools as a function of mid-window position (i.e., average of the underlying positions) on the utg676 contig of the *HaxR* assembly. Taking the *HaxB4* assembly as a reference, we similarly computed the relative (standardized) window coverage as $rc'_{i,p}=s'_{ip}/s'^{B4}_i$ (for window *i* in pool *p*) over the Black-4Spots allele sequence of the *HaxB4* assembly, where s'^{B4}_i the standardized coverages for window *i* computed after merging reads from the three Black-4Spots representative pools (see Supplementary Information Fig. 2).

Results

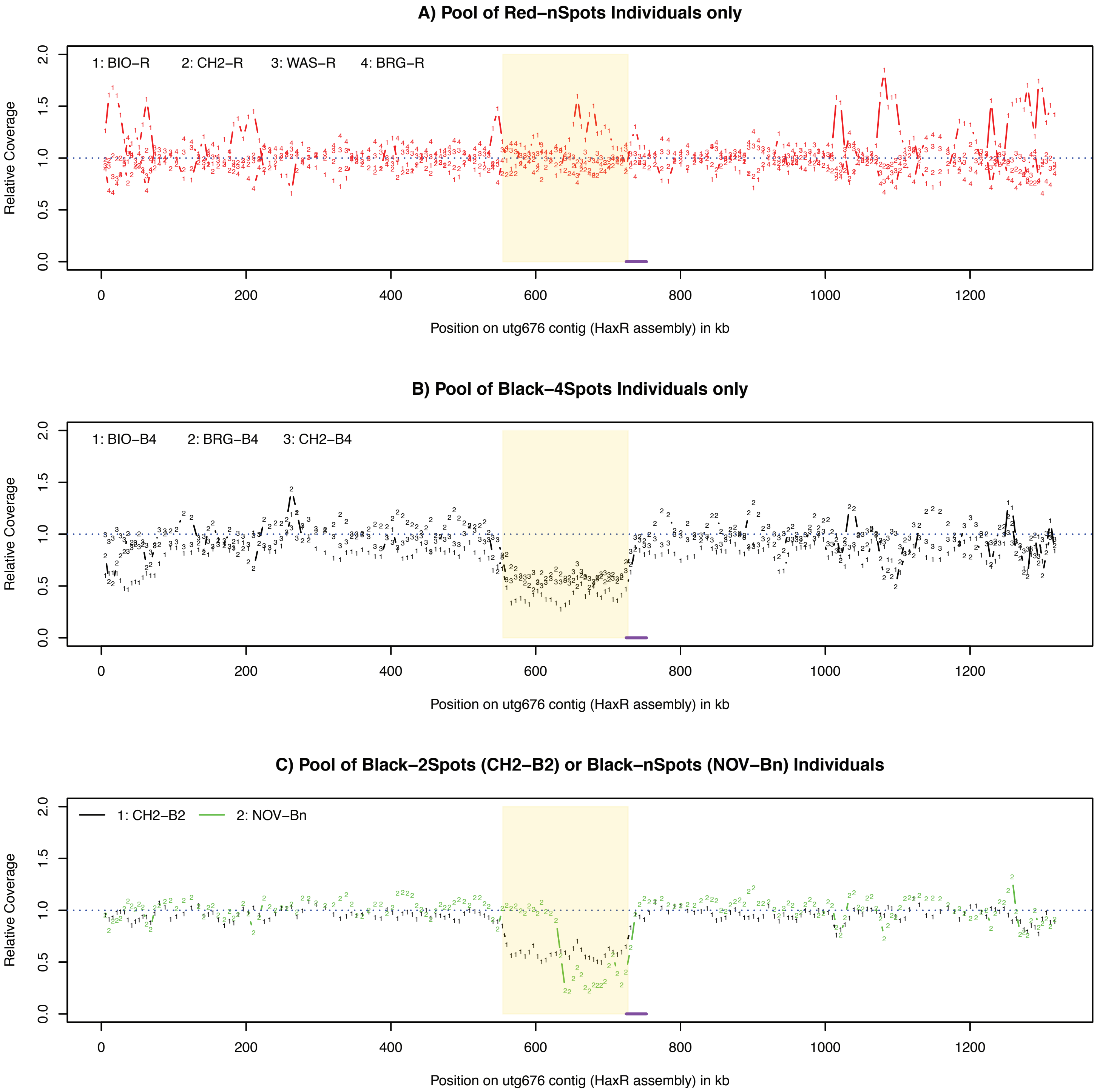
As expected, no obvious departure of the relative window coverage (from the expected value of 1) was observed over the entire Red-nSpots utg676 contig for the four pools including Red-nSpots individuals only, although the BIO-R pools displayed several regions with high relative coverage values (Supplementary Information Fig. 1A). In contrast, for the three pools including Black-4Spots individuals only, we found that the relative window coverage was almost halved in the ca. 170 kb region upstream *pannier* (Supplementary Information Fig. 1B). This pattern is in perfect agreement with the strong sequence divergence observed when comparing the *de novo* sequences of the Black-4Spots and Red-nSpots alleles. It is worth noting that the differences in sample frequency of the Red-nSpots allele in the three Black-4Spots DNA pools is likely to explain the slight pattern differences observed in their relative coverage. When considering the extended Black-4Spots allele sequence as a reference (assembly HaxB4), the pattern was reverted: no obvious departure of the relative window coverage was observed for Black-4Spots pools (Supplementary Information Fig. 2B) and the coverage was almost halved for Red-nSpots pools in the genomic region harbouring strong sequence divergence between the two colour pattern form alleles (Supplementary

Information Fig. 2A). These different analyses hence confirmed the relevance of using relative coverage to assess sequence divergence between the allele sequences of the colour pattern forms.

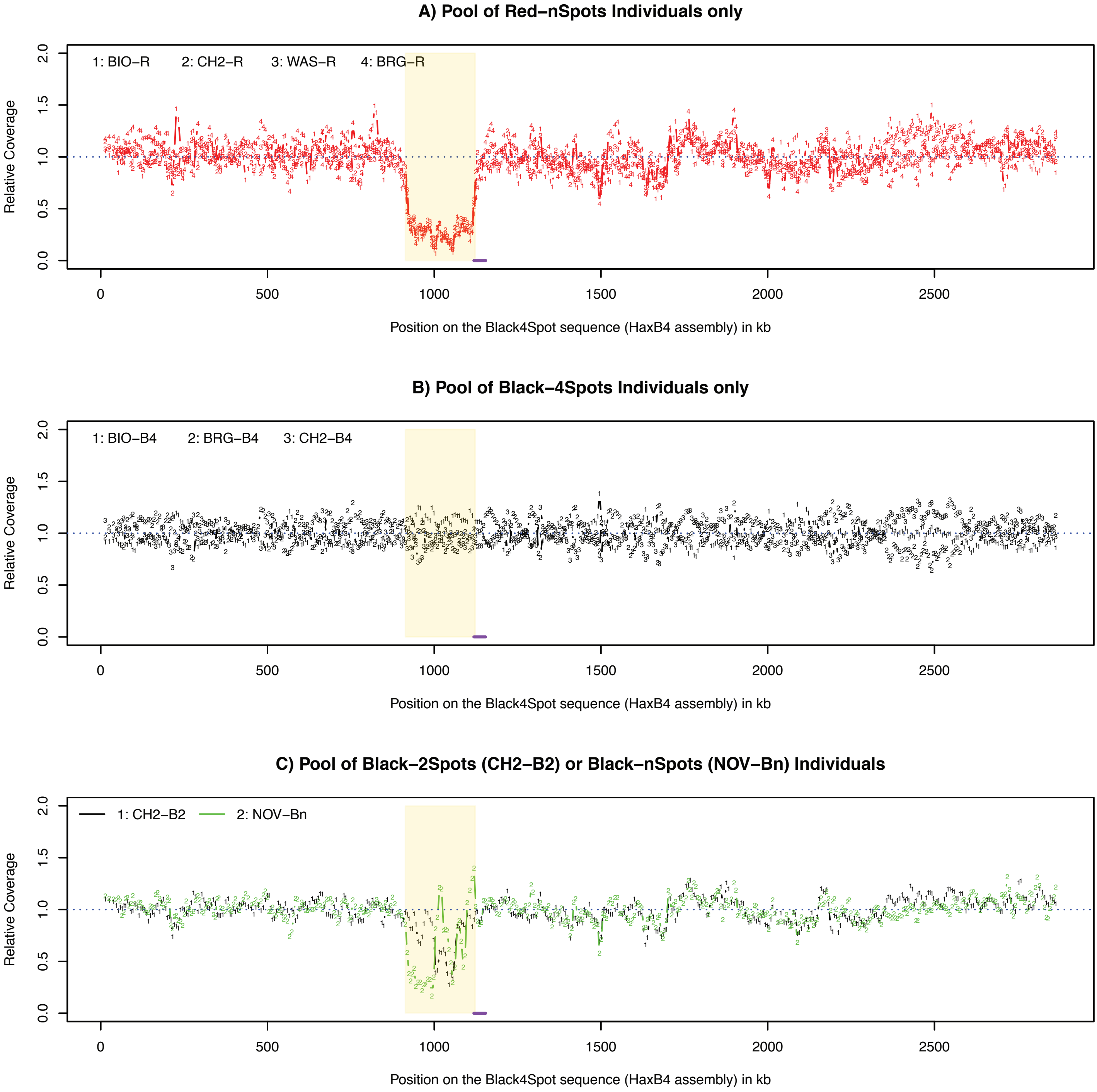
For the CH2-B2 pool including Black-2Spots individuals only, we observed a clear reduction in relative coverage over the entire genomic region where the sequences of the Red-nSpots and Black-4Spots alleles diverge, when taking as a reference the extended Red-nSpots allele sequence (Supplementary Information Fig. 1C). Interestingly, the reduction in relative coverage concerned a more restricted region (located more closely to *pannier*) when taking as a reference the extended Black-4Spots allele sequence (Supplementary Information Fig. 2C). This suggests that the Black-2Spots allele might be more similar to the Black-4Spots allele, over at least a part of the upstream region of *pannier,* although it should be kept in mind that additional large scale variation (insertion or deletion) specific to the Black-2Spots allele are not identifiable with this approach.

Finally, for the NOV-Bn pool including Black-nSpots individuals only, we observed a clear reduction in relative coverage over the second half of the region where the sequences of the Red-nSpots and Black-4Spots alleles diverge, when taking as a reference the extended Red-nSpots allele sequence (Supplementary Information Fig. 1C). This pattern suggests a close similarity in sequence of the Black-nSpots allele with the Red-nSpots allele over the first half of its sequence. When taking as a reference the extended Black-4Spots allele sequence a clear reduction of the NOV-Bn relative coverage was observed over the first half of the region where the sequences of the Red-nSpots and Black-4Spots alleles diverge, whereas coverage reduction was less obvious in the second half of the region (Supplementary Information Fig. 2C). This suggests some sequence similarity between the Black-nSpots and the Black-4Spots allele in the latter genomic region. It is worth noting, however, that for the Black-nSpots allele too, additional specific large scale variation (insertion or deletion) are not identifiable with this approach.

Overall, our analyses of read mapping coverage on both the extended Red-nSpots (*HaxR* assembly) and Black-4Spots (*HaxB4* assembly) allele sequences, strongly suggest that heterogeneous large scale sequence divergences are present among all four main color form alleles of *H. axyridis*. This promotes, for future work, the *de novo* sequencing of other colour pattern form alleles than the Red-nSpots and Black-4Spots carried out for the purpose of our study.



Supplementary Information | Figure 1; Relative read mapping coverage for window of 10,000 positions over the extended Red-nSpots allele sequence (HaxR assembly) for A) the four pools including Red-nSpots individuals only; B) the three pools including Black-4Spots individuals only; and C) the CH2-B2 and NOV-Bn pools including respectively Black-2Spots individuals only and Black-nSpots individuals only. The horizontal blue dotted line gives the value of 1 expected for the absence of departure in the relative coverage.
The genomic region where the sequences of the Red-nSpots and Black-4Spots alleles diverge is represented as a shaded orange area. The position of the gene pannier is represented by a purple line segment.

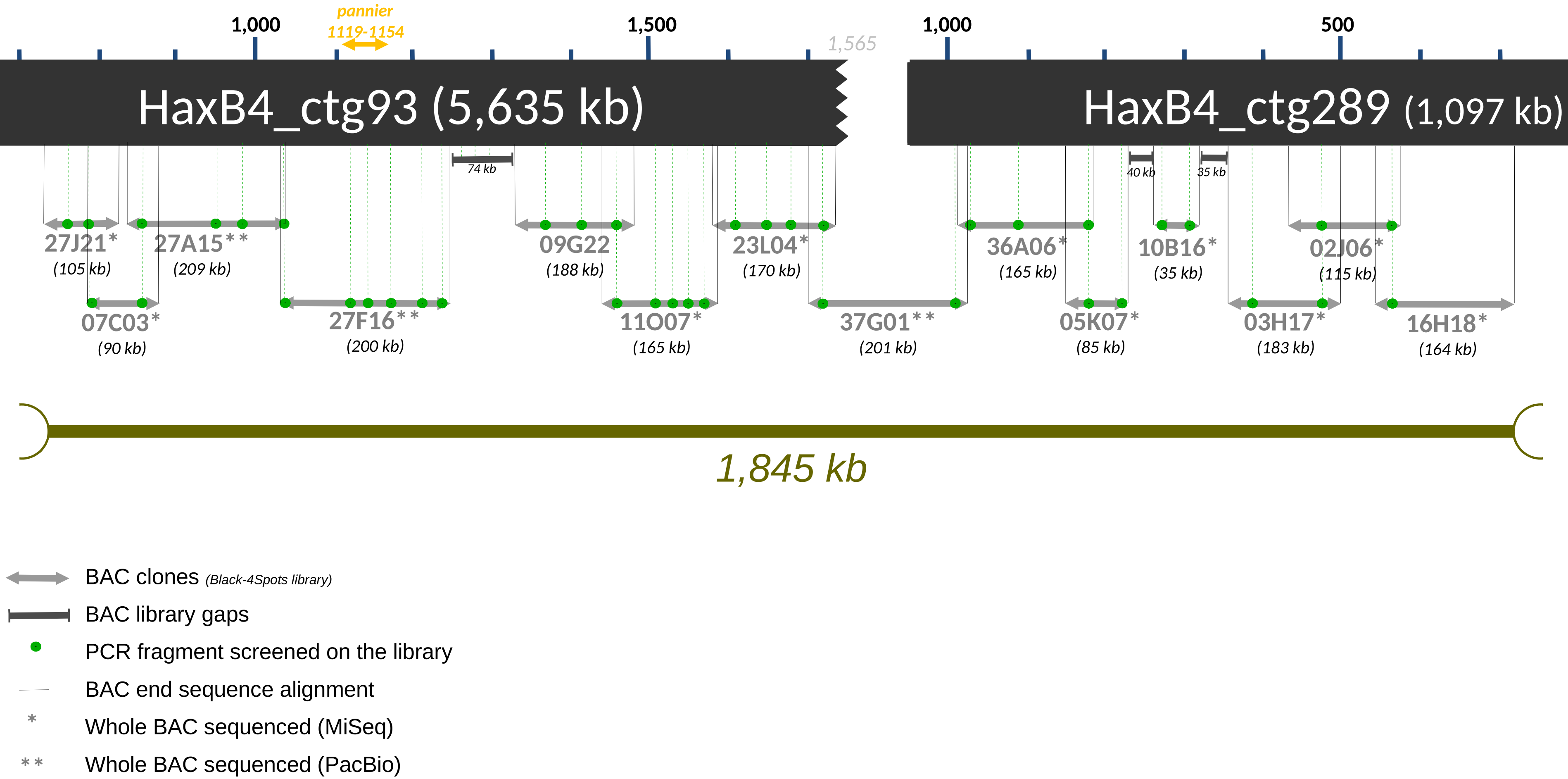


A) Pool of Red–nSpots Individuals only

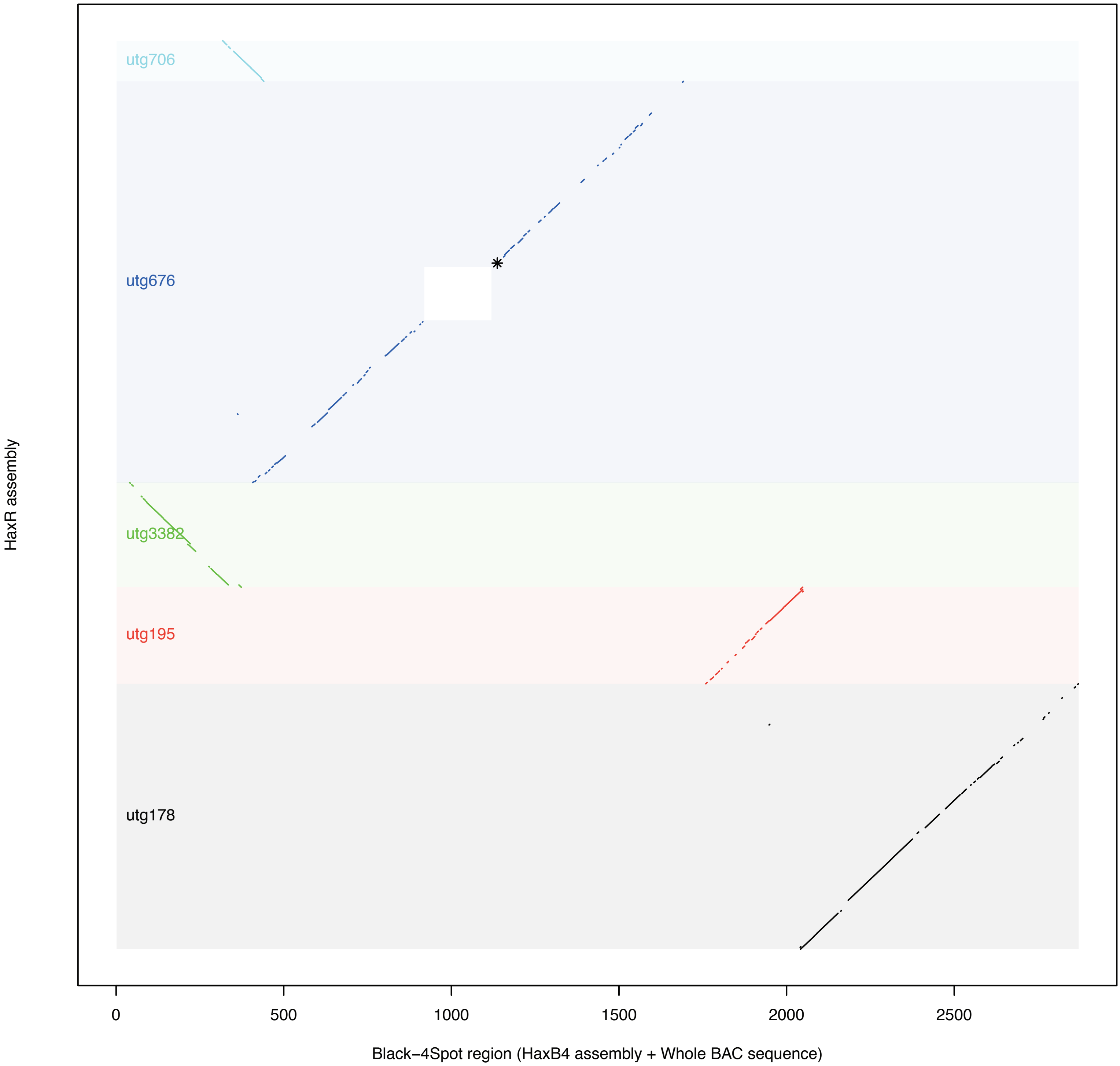
B) Pool of Black–4Spots Individuals only

C) Pool of Black–2Spots (CH2–B2) or Black–nSpots (NOV–Bn) Individuals

Supplementary Information | Figure 2: Relative read mapping coverage for window of 10,000 positions over the extended Black-4Spots allele sequence (HaxB4 assembly) for A) the four pools including Red-nSpots individuals only; B) the three pools including Black-4Spots individuals only; and C) the CH2-B2 and NOV-Bn pools including respectively Black-2Spots individuals only and Black-nSpots individuals only. The horizontal blue dotted line gives the value of 1 expected for the absence of departure in the relative coverage. The genomic region where the sequences of the Red-nSpots and Black-4Spots alleles diverge is represented as a shaded orange area. The position of the gene pannier is represented by a purple line segment.



Supplementary Fig 3: Schematic view of the BAC contig covering the sequence of the Black-4Spots allele of the colour pattern locus.



Supplementary Fig. 4: Alignment of the sequence of the Black-4Spots allele on the HaxR assembly (i.e. Red-nSpots allele). Only similarity segments > 1.5 kb and with an overall identity >98% are represented. The black star indicates the position of the gene pannier in both sequences and the white rectangle the region showing high sequence divergence between the alleles of the two forms.

Supplementary Table 1

	Flow cell 1	Flow cell 2	Flow cell 3	Flow cell 4
Targeted fragment size	25Kb	25Kb + BluePipin 10 Kb	25Kb	30Kb
Total nb. of sequences	712,089	487,071	810,601	448,579
Total sequence length (Mb)	6.9	5.8	6.4	3.8
Median sequence size (bp)	5,259	10,969	7,647	4,541
Maximum sequence size (bp)	302,741	281,965	106,757	137,546
N25 (bp)	26,36	26,482	15,545	26,335
N50 (bp)	19,436	20,236	11,962	18,603
N75 (bp)	12,244	14,634	8,787	11,065
N90 (bp)	4,743	8,417	5,559	5,401

Supplementary Table 2

Ind. sample code	Sex	Nb. of PE reads pairs (% mapped)	Nb. of filtered PE reads (% properly paired)	Median read coverage
EN-F1	F	53,813,455 (97.45%)	69,814,907 (87.13%)	21
EN-F2	F	23,650,846 (97.28%)	30,228,718 (88.49%)	8
EN-F3	F	17,467,055 (96.99%)	23,042,426 (86.55%)	6
EN-M1	M	31,627,893 (97.33%)	41,712,987 (87.18%)	12
EN-M2	M	21,047,048 (96.94%)	26,260,442 (88.83%)	7
EN-M3	M	21,073,231 (96.50%)	27,254,330 (86.10%)	7

Supplementary Table 4

Pooled-sequencing sample code	Preparation kit	Sequencer	Nb. of PE reads pairs (% mapped)	Nb. of filtered PE reads (% prop. paired)	Mean (median) coverage *	SRA accession
CH1-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	103,561,210 (96.48%)	76,554,062 (90.27%)	26.95 (18)	SRR7252403
CH1-B	Nextera DNA sample preparation	HiSeq2500 (2x125)	123,007,350 (97.06%)	105,051,083 (88.81%)	37.65 (25)	SRR7252396
CH2-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	107,545,837 (94.23%)	125,023,414 (92.77%)	44.61 (41)	SRR7252400
CH2-B4	Nextera DNA sample preparation	HiSeq2500 (2x125)	76,780,092 (97.53%)	93,590,049 (92.61%)	33.42 (31)	SRR7252397
CH2-B2	Nextera DNA sample preparation	HiSeq2500 (2x125)	72,988,079 (97.60%)	90,024,288 (92.34%)	31.60 (30)	SRR7252407
JP-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	91,767,378 (58.68%)	64,602,103 (92.54%)	22.11 (20)	SRR7252408
JP-B4	Nextera DNA sample preparation	HiSeq2500 (2x125)	98,914,000 (57.21%)	67,576,564 (92.80%)	22.68 (21)	SRR7252405
NOV-Bn	Nextera DNA sample preparation	HiSeq2500 (2x125)	98,896,363 (93.76%)	105,125,617 (83.98%)	33.04 (32)	SRR7252406
BRG-R	TruSeq DNA sample prep.	HiSeq2000 (2x100)	143,903,442 (96.38%)	174,837,995 (87.72%)	43.98 (42)	SRR7252398
BRG-B4	TruSeq DNA sample prep.	HiSeq2000 (2x100)	145,341,073 (96.48 %)	182,488,179 (88.29%)	46.55 (45)	SRR7252399
ENA-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	82,448,872 (96.80%)	91,557,435 (77.10%)	25.91 (25)	SRR7252395
WAS-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	95,858,923 (93.26%)	90,765,494 (84.35%)	28.22 (27)	SRR7252404
BIO-R	Nextera DNA sample preparation	HiSeq2500 (2x125)	61,547,164 (98.86%)	86,963,677 (96.70%)	29.55 (28)	SRR7252401
BIO-B4	Nextera DNA sample preparation	HiSeq2500 (2x125)	84,393,625 (98.88%)	120,260,171 (96.72%)	40.79 (40)	SRR7252402

Supplementary Table 5

Library name	Library type	Target insert size in bp	Realized insert size in bp	Nb. of reads (% used)	Final Coverage
PE250a	Paired End (2x150)	250	197 (+/- 70)	608,208 (40.5%)	0.1
PE250b	Paired End (2x150)	250	227 (+/- 51)	68,189,816 (80.9%)	25.7
PE400	Paired End (2x150)	400	383 (+/- 25)	63,059,050 (64.8%)	18.6
PE600	Paired End (2x150)	600	574 (+/- 29)	80,471,132 (57.7%)	20.6
LJD3	LJD* (2x100)	3000	2414 (+/- 359)	57,424,608 (61.0%)	8.9
MP3	Mate Pair (2x100)	3000	1554 (+/- 473)	78,248,120 (31.1%)	6.8
LJD8	LJD* (2x100)	8000	7626 (+/- 623)	137,521,236 (9.9%)	3.7
MP8	Mate Pair (2x100)	8000	5606 (+/- 940)	46,004,730 (35.2%)	4.2

* LJD = Long Jumping Distance library

Supplementary Table 6

Primers used for RNAi experiments

dsRNA	Name	Sequence
ds-pnr1	T7-pnr-F1	TAATACGACTCACTATAGGGAGAggacacgtttggaagaataag
	T7-pnr-R1	TAATACGACTCACTATAGGGAGAcataaggtgacgtccgttg
ds-pnr2	T7-pnr-F2	TAATACGACTCACTATAGGGAGAgttgcacggagtgaatagg
	T7-pnr-R2	TAATACGACTCACTATAGGGAGAgaaacgtcgtgggtatatg
ds-EGFP	T7-EGFP-F1	TAATACGACTCACTATAGGATGGTGAGCAAGGGCGA
	T7-EGFP-R1	TAATACGACTCACTATAGGcgttcttctgcttgtcgg
ds-GATAe	T7-elt1-F1	TAATACGACTCACTATAGGTGACGCAGACGGAGGC
	T7-elt1-R1	TAATACGACTCACTATAGGGGCGCTCGTTGGTGTGTA

Primers for qPCR comparisons

Gene	Name	Sequnece
pnr	qPnr-F1	AgAcCCgtaAgAggaAgCCGA
	qPnr-R1	TGCTTACTGGTGTCTTGCCGT
pnr	qPnr-F2	aCCaCaAgATGAATGggATGAACAgACCC
	qPnr-R2	CAGCACAGTccCAAGCGTCTGGTT
eIF4A	Ha-eIF4a-F1	caccgtggtacgtcggcg
	Ha-eIF4a-R1	ggaaactctaggcaacgtcgg
eIF5A	qHa-EIF5a-F1	TGAAATCTTGTGGTGAAGAGGTTGT
	qHa-EIF5a-R1	gtccccctatccttctttcgccaTT

Name	R^2 of standard curve
qPnr-F1	0,990
qPnr-R1	
qPnr-F2	0,996
qPnr-R2	
Ha-eIF4a-F1	0,999
Ha-eIF4a-R1	
qHa-EIF5a-F1	0,996
qHa-EIF5a-R1	