## File S1

A dense linkage map for a large repetitive genome: discovery of the sex-determining region in hybridizing fire-bellied toads (*Bombina bombina* and *B. variegata*)

Beate Nürnberger, Stuart J.E. Baird, Dagmar Čížková, Anna Bryjová, Austin B. Mudd, Mark L. Blaxter, Jacek M. Szymura

## Supplemental Methods

 page

*Bombina* husbandry and cross design 2

Selection of candidate sequences for bait design 3

Filtering of high-coverage regions in Platanus contigs 4

ARC assemblies and mapping reference 5

Processing of read files from enriched libraries 6

FastVec Mathematica Script 8

Lep-MAP3 commands 14

Reduced linkage map with one locus per cM position 15

### *Bombina* husbandry and cross design

The F0 parents, a female *B. bombina* and a male *B. variegata*, were collected in late May 2014. The pair was placed in a glass aquarium (20 x 20 x 30 cm) filled with aged tap water and some water plants. Spontaneous amplexus followed soon, as the male was in breeding condition (dark nuptial pads present).

Attempts to obtain F1 hybrids in the reverse combination (*i.e.*, *B. bombina* male and *B. variegata* female) failed. This direction of cross is more difficult to perform, requires hormonal stimulation and produces fewer offspring due to lower *B. variegata* fecundity (Szymura and Farana 1978; Rafińska 1991; Hofman and Szymura 2000).

After egg laying, the parents were removed and kept separately. Tadpoles were fed boiled dandelion (*Taraxacum officinale*) leaves and nettle (*Urtica dioica*), both homogenised in a blender. Water was changed twice a week. As the animals grew larger, they were distributed among more aquaria.

Once their front limbs emerged in early August, metamorphosing tadpoles were transferred to new aquaria with some water and gravel. Metamorphosis can be associated with an increase in mortality, but this was not observed here. After tail resorption, metamorphs were fed *Drosophila* flies dusted with calcium and vitamins, small Tubifex worms at first, then small crickets. As the toadlets grew, larger crickets and mealworms (*Tenebrio molitor*) were gradually introduced as food.

No hibernation was attempted. Toadlets overwintered at 21˚C and were fed *ad libitum*. In the summer of 2016, some F1 males began calling. Several F1 x F1 crosses were set up and produced variable numbers of eggs and tadpoles. No inspection of early embryo mortality was made to avoid any disturbance that might result in additional deaths. Throughout the experiment, no hormone stimulation was used to exclude interference with gonad development, meiosis or gamete release during the mating ritual.

We report here on two F2 crosses (ENA sample accessions in parentheses):

Family 6: F1 male 2016\_6\_M\_07B2 (#ERS5083645) x F1 female 2016\_6\_F\_12A3

(#ERS5083565)

Family 7: F1 male 2016\_6\_M\_07B2 (#ERS5083645) x F1 female 2016\_7\_F\_13A5 (#ERS5083646)

## Selection of candidate sequences for bait design

For each CLC contig that was part of a SCUBAT2 path, those sections were identified that were deemed ‘unique’ based on a CLC assembly self-blast or had exact BLAST+ matches in the SGA or the Platanus assemblies, as described in the Materials and Methods. The following schematic shows an example. The candidate bait region (grey shading) is the sequence section where all three criteria applied. Note that candidate regions were chosen irrespective of exon locations.





For 2/3 of CLC contigs in SCUBAT2 paths, each of the three criteria was met in some contig subset (with or without overlap among subsets). For this contig group, the following plot shows the distribution of contig length proportions matching each criterion to illustrate their relative selectiveness. While just over 20% of contigs had a full-length match in the SGA assembly, only about 12% were deemed entirely ‘unique’. Exact matches between the CLC and Platanus assemblies were the shortest on average.

Candidate regions were further filtered for particularly variable regions (see the following section of this file) and for matches to the REPdenovo assembly, as described in the Materials and Methods. The vast majority of candidate regions had no overlap with any exon.

## Filtering of high-coverage regions in Platanus regions

Platanus bubbles (variant sequences extracted from the de Bruijn graph) were aligned to the Platanus assembly with BLAST+ and assigned to the contig with the highest scoring match. Per contig, a variant count was computed as the sum of BLAST+ mismatch counts across aligned bubbles. The mean read coverage of Platanus contigs with inferred variants was 67.3, and the distribution was skewed to the right (left plot below). Highly-covered contigs (mean coverage > 85 [= vertical line in the plots]) had more variants on average (3.2 versus 1.4) and a larger proportion of contigs with excessive levels of variation (right plot below). They were presumably overmerged. These contigs were flagged and any matching CLC sequence was barred from bait design.



## ARC assemblies and mapping reference

From the enriched read set of the F0 *B. variegata* male, ARC produced 4,850 assemblies. For 77% of these, the entire sequence was an exact subset of the targeted CLC contig. The remainder included chimeric assemblies with BLAST+ matches to two or, in a few cases, three different enrichment targets. In particular, there were 76 assemblies for which the highest scoring match was to a target other than the one that seeded them, as



depicted in the figure above. For assembly 3, BLAST+ score of match *a* is higher than that of *b*. This precluded the use of the ‘reciprocal best BLAST hit’ criterion to select the most appropriate assembly, because none would have been found for target 3. For those targets with BLAST+ matches to more than one assembly (*n* = 208), we inspected their length and coverage of the bait region. In all but four cases, targets were best covered by the assembly that they had seeded. Overall, BLAST+ matches with a minimum length of 300 bp and full coverage of the bait region were obtained for 4,763 targets. For each of these, the match coordinates were used to extract the reference sequence from the CLC contig. For the remainder, the full CLC contig was used.

A self-blast was carried out on this reference, and all positions in matches longer than 50 bases and with sequence identity greater than 92.5% were hard masked. This was done, because some sequence sections were outside the range that had been curated for bait design.

**Processing of read files from enriched libraries**

A given <*readset*> was processed as follows:

(a) mapping to the REPdenovo assembly

bombina\_repeats is the Bowtie2 index for File S10.

bowtie2 \

 -x bombina\_repeats \

 --phred33 \

 -X 600 \

 --sensitive-local \

 -t \

 --rg-id ID:<*ID*> \

 --rg PL:Illumina \

 --rg LB:<*library*> \

 --rg SM:<*sample*> \

 -p 5 \

 -q \

 -1 <*readset*>\_1\_trimmed.fastq.gz \

 -2 <*readset*>\_2\_trimmed.fastq.gz \

 -S <*readset*>.sam

(b) extracting unmapped reads

samtools fastq -n -f 0x1 -f 0x4 -f 0x8 -1 <*readset*>\_norep\_1.fastq -2 <*readset*>\_norep\_2.fastq <*readset*>.sam

rm <*readset*>.sam

(c) mapping the repeat-subtracted reads to the *B. variegata* reference

(mixed\_masked\_bt2 is the Bowtie2 index of File S11 [in fasta format, referred to in the following as <*reference*>]). Note that .fai and .dict indices are needed below for this reference.

bowtie2 \

 -x mixed\_masked\_bt2 \

 --phred33 \

 -X 1200 \

 --sensitive-local \

 -t \

 --rg-id ID:<*ID*> \

 --rg PL:Illumina \

 --rg LB:<*library*> \

 --rg SM:<*sample*> \

 -p 5 \

 -q \

 -1 <*readset*>\_norep\_1.fastq \

 -2 <*readset*>\_norep\_2.fastq \

 -S <*readset*>.sam

(d) Sorting, reordering and deduplicating

picard SortSam I=<*readset*>.sam O=<*readset*>\_sorted.bam SO=coordinate

picard ReorderSam I=<*readset*>\_sorted.bam O=<*readset*>\_reordered.bam R=<*reference*>

picard BuildBamIndex I=<*readset*>\_reordered.bam

If there were two readsets per sample, these were processed separately up to here and then combined in the following step.

picard MarkDuplicates \

 <*readset*>\_reordered.bam [<*readset2*>\_reordered.bam] \

 OUTPUT=<*sample*>\_dedupped.bam \

 METRICS\_FILE=<sample>\_dedupping\_metrics.txt \

 TMP\_DIR=tmp

(d) Indel realignment with GATK v.3.7

GATK -T RealignerTargetCreator \

 -nt 4 \

 -I <*sample*>\_dedupped.bam \

 -R <*reference*> \

 -o <interval.list>

GATK -T IndelRealigner \

 -I <*sample*>\_dedupped.bam \

 -targetIntervals <interval.list> \

 -R <*reference*> \

 -o <sample>\_realigned.bam \

 -compress 0 \

 --maxReadsInMemory 200000 \

 --filter\_bases\_not\_stored

(e) Samtools mpileup and PoPoolation2 mpileup2sync

samtools mpileup \

 -a \

 --no-BAQ \

 --fasta-ref <*reference*> \

 --output <*sample*>.mpileup \

 <sample>\_realigned.bam

popoolation2 mpileup2sync \

 --input <*sample*>.mpileup \

 --output <*sample*>.sync \

 --threads "$PBS\_NUM\_PPN"

The \*.sync files contain the counts of reads supporting each sequence state (A:T:C:G:N:D – in this order), where N is IUPAC code (‘any base’) and D is a deletion. Rows have the following format:

|  |  |  |  |
| --- | --- | --- | --- |
| Locus | position | reference state | *<sample>* |
| Bv\_contig17\_asmbl\_rev | 192 | T | 0:37:0:0:0:0 |

## FastVec Mathematica script

### Design overview

The script is best understood with respect to the goal of inference: on multilocus input data from F2 crosses, output diplotype estimates suited to map estimation software such as Lep-MAP3. For an individual, at a locus, Lep-MAP3 takes as input diplotype estimates encoded by letters (*e.g.*, {AA, AC, CC}) and support for those estimates. If the F2 cross progeny were assayed at SNPs of state known *a priori* (*e.g.*, {A,C}), then the inference problem would reduce to, for each locus, for each interval, measure the likelihood of the data given three diplotype hypotheses {AA, AC, CC}.

Here, instead, we treat the case where the F2 progeny are assayed at bait intervals of unknown state. We treat each bait interval as a locus (*i.e.*, with no *B. bombina/B. variegata* source-recombinant alleles in the dataset). This is reasonable. There are no *B.b./B.v.* recombinants in the F1s. Each F2 chromosome likely has 1-2 recombination events per meiosis. As the total baited interval is a tiny fraction of the genome size, the probability of one of these recombination events falling in or near any bait is small. The inference problem is then a co-estimation of, for each interval (locus) two haplotype states {*B. bombina, B. variegata*}, and for each individual, for that interval, the likelihood of the data given three (diplotype) hypotheses {BbHOM, HET, BvHOM}.

Co-estimations can often be decomposed by noting what would be possible if we knew one part of the co-estimation with certainty (for example the Expectation Maximisation algorithm). If we knew a haplotype state pair {*B. bombina, B. variegata*} with certainty, then it is straightforward to construct a diplotype hypothesis space for reads resulting from some mixture of these haplotypes. We could then estimate the mixture coefficient for each individual’s read data. In the ideal case, an individuals’s estimates would fall into 3 categories: {Low, Intermediate, High}, corresponding to reads arising from {BbHOM, HET, BvHOM} diplotypes, allowing for some reference (mixture) bias in the bait design. Further, if we had some reasonably small set of C candidate haplotypes {cand1, cand2, …, candC-1, candC}, we could compare the total likelihood of all individual’s data for each of all possible C(C-1)/2 pairwise mixtures of candidate haplotypes, co-estimating a ML haplotype pair (with support) and individual ML diplotypes (with support), as required by our goal of inference.

From this point of view, the inference problem appears to reduce to the construction of a reasonably small set of haplotypes that are good candidates for those which make up all the diplotypes in the total dataset. The grandparents are assumed BbHOM, BvHOM, the F1s HET, the F2s sampled from {BbHOM, HET, BvHOM} in approximately Mendelian proportions. A co-ordinate space that separated the total data over all individuals into diplotypically similar clusters would greatly aid inference for large crosses: we expect two clusters to correspond to BbHOM and BvHOM diplotypes. Estimating {*B. bombina, B. variegata*} haplotypes by strict consensus over the total read information of each of these homozygous-by-source-grandparent clusters will give (very) high read depth estimates that are robust to cluster mis-assignments of a minority of individuals and overmerging signal.

By initial consideration of the goal of inference, we have therefore outlined an inference approach for WGS data that involves delayed calling (*cf.* Nielsen et al. 2012): there is little logic in starting analyses by estimating haplotypes for each of hundreds of individuals when a maximum of four inherited haplotypes (excepting rare recombination) underlies all their data. Instead, haplotype calling is delayed to a post diplotype-clustering stage. There, calling is simplified by large combined coverage and the assumption that source-homozygous diplotypes from divergent sources will form clusters separate from each other and their heterozygotes. The FastVec script is fast because it avoids the computational load of individual haplotype calling by (1) placing diplotypes in a vector co-ordinate space where such homozygous clusters will occur, allowing straightforward haplotype calling over a small number of diplotype clusters, producing a small number of haplotype estimates. From that point forward, (2) maximum likelihood inference, as decomposed above, is straightforward to the goal of inference.

### Implementation details

(1) Data input format

(2) A co-ordinate system for diplotype clustering

(3) Maximum likelihood given haplotype estimates

(4) Data output format

### (1) Data input format

For each bait interval (locus), the mapped read data (BAM format) was summarised using Samtools (v. 1.4) (Li et al. 2009) mpileup and PoPoolation2 (Kofler et al. 2011) mpileup2sync. The resulting summary files contain, for each individual’s sample *i* and locus *l*, a matrix M*i,l* of *n* columns (*n* = number of reference positions) and six rows (sequence states of A, C, G, T, DEL, and N; Main text, Figure 2A, B) of the counts of reads supporting each sequence state at each position.

Note:

a/ Insertions cannot be represented in this matrix of *reference* coverage.

b/ Within-read phase information is discarded in the mpileup2sync transition.

c/ Matrix addition gives combined sequence state support over sets of individuals. For example M*{i,j},l* = M*i,l* + M*j,l*

The delayed-calling inference approach implies decisions about the nature of original sequence should be based on combined sequence state support over the largest relevant set of individuals.

### (2) A co-ordinate system for diplotype clustering

**2.1 Discarding rare states**

For each locus, presence/absence of sequence states is calculated over all individuals. For any state present with frequency less than *RareStateThreshold*, that state count is set to zero for all individuals’ data. This is the only thresholding decision in the pipeline. Its rationale is simple: as the goal involves clustering, states so rare they cannot be shared even across the smallest expected cluster can contribute little to inference. For *Bombina,* *RareStateThreshold* was set to 1/32.

**2.2 Discarding fixed positions**

After rare states are discarded, any position with only one state in M*ALL,l* is discarded.

**2.3 Calculating a taxon vector**

Individuals believed to have genetic material sourced from each taxon are used to construct a taxon vector **M**p. Currently, two taxa are assumed. At each locus, the taxon vector is designed to represent the difference between M*TAXON1,l* and M*TAXON2,l*. Signal of simple differences in the sizes and/or coverages of the sets TAXON1 and TAXON2 is removed by taking state frequencies f over each matrix, leaving the difference matrix:

Δfl = f(M*TAXON1,l*) − f(M*TAXON2,l*)

This difference matrix has the undesirable property that frequency differences supported by very little data have equal weight to difference supported by large N data. One way to introduce more appropriate weightings is to measure the significance of taxon count differences at each position *x*. As these counts are multinomial, a likelihood ratio test (LRT) contrasting one pooled *versus* two separate multinomial hypotheses seems appropriate. 2ΔLL is distributed as ChiSquared(df) where the degrees of freedom is the total number of states-1.

Sig*l,x* = LRT(M*TAXON1,l,x*, M*TAXON2,l,x*)

and

**M**p,l = Δfl (1-Sigl)

Thus (perhaps poorly supported) differences in estimates of taxon frequencies at positions with poor evidence of taxon difference are downweighted. The data for all individuals at a locus is placed in the diplotype co-ordinate system by matrix multiplication Mp,l Ml. The positive (TAXON1-like) and negative (TAXON2-like) components of each individual’s matrix are then separated and totalled to give (x,y) in the diplotype co-ordinate system.

In the *Bombina* analysis, **M**p is calculated twice. Initially, the sets assumed to represent different taxa have one member each: TAXON1 = {F0 *B. variegata*}, TAXON2 = {F0 *B. bombina*}. Once all individuals are placed in the co-ordinate space, these sets can be updated to TAXON1 = {y-axis cluster}, TAXON2 = {x-axis cluster}. More generally (*e.g.,* for data from a natural hybrid zone), initial TAXON sets could be constructed from reference individuals or sets of individuals from reference localities.

**2.4 Diplotype clustering**

We use the expectation maximisation (EM) algorithm to cluster individuals by their (x,y) co-ordinates. Diplotypes are expected to form lines emanating from the origin of the space: individuals with low coverage/noisy data close to the origin *versus* strongly supported individuals far from the origin. The action of the EM algorithm can then be simplified to clustering of individuals to angle hypotheses, each angle corresponding to a line through the origin. EM convergence is improved by working with discrete angles ({0->89 degrees} + 1/2) and rescaling the axes such that Mean(x) = Mean(y). The EM algorithm implementation is guaranteed to stop. Halting states include a failure to distinguish clusters. The initial state allows for seven clusters and this cannot be exceeded.

### (3) Maximum likelihood given haplotype estimates

**3.1 Haplotype estimates from diplotype clusters**

It is assumed two of any diplotype clusters estimated at step 2.4 are majority homozygous by source, one of haplotype(s) inherited from TAXON1, another of haplotype(s) inherited from TAXON2. Majority consensus calling of combined sequence state support M*CLUSTER,l* over homozygous clusters of individuals will give highly robust haplotype estimates, given calling coverage = mean individual coverage x cluster size. In the current analyses calling coverage is nearly always >1000. Where three clusters form at step (2.4), those nearest the axes are clear homozygous cluster candidates. Where more than three clusters form, we do not presume to know which are the ‘best’ candidates. Instead, as a general approach for all cases, we define the ‘best’ candidate pair as that pair which maximises the log likelihood of the data. The maximum number of diplotype clusters is seven. The maximum number of pairs to be analysed in the likelihood framework is therefore (7 × 6) / 2 = 21. The log likelihood of the data for a haplotype estimate pair is the total of the log likelihood of the data of each individual, given that pair.

**3.2 The likelihood of an individual’s data**

An individual’s data is assumed to comprise reads from the two haplotypes in ratio A. The ratio A hypothesis space is discretised to {0, 0.01, 0.02, …, 0.99, 1}. We expect this ratio to be low for one homozygous diplotype, intermediate for heterozygotes, and high for the other homozygous case.

**3.3 Reference bias and contamination**

Given that we wish our approach to be robust to reference bias and contamination, we do not presume to know where the boundaries lie between these {low, intermediate, high} read mixture regimes (3.2). Instead, we construct the histogram of mixture MLEs across all individuals. If the data does indeed arise from (unknown ratio) read mixtures of two-taxon homozygotes and heterozygotes, then this histogram should be trimodal and can be divided into categories at two minima between the three maxima. If the histogram cannot be divided in this way, the algorithm halts. If it can, then the output is a likelihood-based assignment of individuals to modelled-diplotypes. This is in contrast to the stage (2.4) heuristic-based assignment of individuals to (unmodelled) diplotype clusters.

**3.4 MLEs and supports**

In particular, for the ML candidate haplotype pair all individuals are now forced by the model into one of exactly three categories. The maximum likelihood of an individual’s data falling in an alternate category is computed as the likelihood of its data given the closest mixing parameter to its ML value that can be found in the alternate category. Individuals with mixing parameter MLEs close to category boundaries will therefore have poorly supported categorisation.

**(4) Data output format**

The script is designed to always halt: all analysis results up to the halting point are output, plus an English language message describing the halting state. Thus, for example, any run halting after stage 2.4 will output the diplotype clustering results. For the *Bombina* analysis (~5,000 bait loci), the script runs within 48 hours on eight cores of a laptop. Returning to the goal of inference “On multilocus input data from F2 crosses, output diplotype estimates suited to map estimation software such as Lep-MAP3. For an individual, at a locus, Lep-MAP3 takes as input diplotype estimates encoded by letters (*e.g.,* {AA, AC, CC}) and support for those estimates. If the F2 cross progeny were assayed at SNPs of state known *a priori* (*e.g.,* {A,C}), then the inference problem would reduce, for each locus, to measuring the likelihood of the data given three diplotype hypotheses {AA, AC, CC}”. We see that stage 2.4 satisfies our goal, but we also see that numerous auxiliary parameters necessary to compute likelihoods have also been estimated. These include estimates of the taxon haplotypes for each locus, and estimates of reference bias and/or contamination for each locus and each individual. All these data are output in a single text file for each locus analysed. Subtables within that file are of TAB separated values (TSV format). We do not expect users to directly inspect all these per-locus files for analyses, which in the case of *Bombina* involve thousands of loci. Instead the locus files are set out such that scripts to summarise over loci are easy to design. We use one such script to extract only diplotype estimates and their supports over all loci as the input to Lep-MAP3.

## Lep-MAP3 commands

(executed as part of a shell script)

LodLimit=19

tolerance=0.001

**ParentCall2**

lep-map3 ParentCall2 data=lepmap\_input.txt removeNonInformative=1 halfSibs=1 > data.call

**Filtering2**

lep-map3 Filtering2 data=data.call dataTolerance="$tolerance" noSexFiltering=1 outputHWE=1 > data"$LodLimit"\_f.call

**SeparateChromosomes2**

lep-map3 SeparateChromosomes2 data=data"$LodLimit"\_f.call lodLimit="$LodLimit" numThreads=2 > map"$LodLimit"\_all.txt

MAXCHROM=$(cat map"$LodLimit"\_all.txt | sort -rn | head -n 1)

**JoinSingles2All**

lep-map3 JoinSingles2All map=map"$LodLimit"\_all.txt data=data"$LodLimit"\_f.call lodLimit=4 > map"$LodLimit"\_js.txt

**OrderMakers2**

for (( i=1; i<="$MAXCHROM"; i++ ))

do

 lep-map3 OrderMarkers2 map=map"$LodLimit"\_js.txt data=data"$LodLimit"\_f.call useKosambi=1 grandparentPhase=1 numThreads=2 chromosome="$i" sexAveraged=1 > order\_"$i".txt

done

The OrderMarkers2 command was rerun for each LG to produce a total of 40 (LG1,LG2) or 20 (all other LGs) replicates.

## Reduced linkage map with one locus per cM position

An initial outlier screen excluded from the dataset all loci with larger than expected segregation distortion in at least one family. Against the backdrop of variation in χ2 estimates along the map (Figure 5), loci above and below the significance thresholds were deemed to be outliers if their χ2 estimates clearly deviated from the family-specific pattern. This was done liberally with the aim to remove any potentially distorting effect of diplotyping errors and removed 69 loci (Figure S7).

For the remainder of the dataset (2,062 distinct map positions), there were 1,107 bins (= map positions to which more than one locus had been assigned). Within each bin, we counted for each locus pair the differences in inferred diplotypes across all 162 F2 individuals. For 922 bins, the maximum observed pairwise difference count was no greater than three and we picked one locus at random for the reduced map. For each of the remaining 185 bins, we sorted for each of its *n* loci the array of its associated *n*-1 difference counts and computed the average of the lower *n*-2 of these. The figure on the left illustrates this for a bin with four loci (a – d). Vertices of the graph are labelled with pairwise difference counts. Locus d is clearly the outlier with an average over the lower two difference counts of 6.5, whereas that average for the other three loci is smaller than 3. For essentially all bins with three or more loci, removal of a single outlier locus produced a set of loci with no more than three differences in all pairwise comparisons. As before, one of these was picked at random. For 33 bins with *n* = 2 and a pairwise difference greater than three, we picked the locus with the fewest differences to the immediately preceding and following loci on the map.