bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Transition from environmental to partial genetic sex determination in *Daphnia* through the evolution of a female-determining incipient W-chromosome.

Céline M.O. Reisser[1,2,3,*], Dominique Fasel[2], Evelin Hürlimann[2], Marinela Dukic[4], Cathy Haag-Liautard[2], Virginie Thuillier[2], Yann Galimov[5], and Christoph Haag[1,2].

July 26, 2016

[1]Centre d'Ecologie Fonctionnelle et Evolutive CEFE UMR 5175, CNRS Université de Montpellier Université Paul-Valéry Montpellier EPHE, campus CNRS, 1919, route de Mende, 34293 Montpellier Cedex 5, France.
[2]Université de Fribourg, Ecology and Evolution, Ch. du Musée 10, 1700 Fribourg, Switzerland.
[3]IFREMER Centre du Pacifique, UMR 241 EIO, Labex CORAIL, BP 49, 98719 Taravao, Tahiti, Polynésie Française.
[4]Universität Basel, Zoology Institute, Evolutionary Biology, Vesalgasse 1, 4051 Basel, Switzerland.
[5]Koltsov Institute of Developmental Biology RAS ul. Vavilova 26, 119334 Moscow, Russia.

∗**Corresponding author**: Celine.Reisser@ifremer.fr

Running title: Partial genetic sex determination in *D. magna*.

Keywords: sex determination, sex chromosome, turnovers, gynodioecy, male sterility mutation, *Daphnia magna*.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ABSTRACT

Sex chromosomes can appear through the evolution of genetic sex determination (GSD) from hermaphroditism or environmental sex determination (ESD). However, despite their extensive theoretical description, the early mechanisms involved in the transition from ESD to GSD have yet to be observed in nature, as no mixed ESD-GSD species has been reported in the literature and studied on the molecular level. Here, we focused on *Daphnia magna*, a small freshwater crustacean in which sex is determined solely by the environment, but for which a dominant female sex-determining locus segregates in multiple populations. We found that the sex determining genomic region shares a common location in all populations studied, in the peri-centromeric region of linkage group 3, in a region with reduced but non-zero recombination. The region also harbors numerous genes known to be involved in female and male sex determination/differentiation in other taxa, including *transformer 2* and *sox9*, as well as genes involved in chromatin remodeling. Overall, our results suggest that *D. magna* has evolved an incipient W chromosome. In addition, the occurrence of the sex-determining mutation in an area of pre-existing low recombination contributes to the debate on the degree of involvement of sexually antagonistic selection in early stages of recombination suppression in sex chromosomes. As such, *D. magna* represents the first animal species for which transition from ESD to GSD is evidenced at the genetic level in multiple populations, and could serve as a model to empirically study the role of selective forces in the early stages of sex chromosome evolution.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

Sex chromosomes have evolved multiple times in many taxa, and comparative genomics have recently highlighted that sex chromosomes of distant species may share a common evolutionary origin, while isolated populations of the same species may have evolved unique sex chromosome systems independently (Miura et al. 2008; Pokorná and Kratochvíl 2009; Stöck et al 2011; Tree of Sex Consortium 2014). Two evolutionary routes are thought to lead to the birth of sex chromosomes. The first one consists in evolving separate sexes from hermaphroditism, which is believed to be the major route in plants (Charlesworth and Mank 2010). In theory, this transition occurs through the emergence of a sex-sterility mutation on an autosome which prevents a hermaphrodite individual from reproducing via this sex (usually, the male). These individuals thus become genetically determined females, and they co-occur with fully functional hermaphrodites to form a mixed breeding system called gynodioecy, found rather frequently in plants (Charlesworth and Charlesworth 1978; Charlesworth and Mank 2010). The mutation will be favoured if it has an intrinsic selective advantage, for example due to obligate outcrossing (whether obligate outcrossing is advantageous will depend on the selfing rate and inbreeding depression), if the sex-ratio is biased toward the making of one sex, or if there is a fitness disadvantage to investing resources in both male and female functions compared to investing in only one sex (Charlesworth and Charlesworth 1978; Innes and Dunbrack 1993). Additional mutations occurring around the sex determining locus may arise and be favored if their effects are sex-antagonistic (i.e., if they have a negative effect on performance of the non genetically determined sex, but a positive effect on the genetically determined sex; Ellegren 2011). Similarly, mutated alleles benefiting males (and deleterious in females) will be selected for if they occur on the homologous sex-determining genomic region (the region corresponding to the sex determining region but on the homologous chromosome). The extreme case of such a mutation is a female-sterility mutation occurring on the homologous chromosome, which may for instance be favoured due to sex-ratio selection. Since recombination would shuffle these mutations into the opposite background, closer linkage will be favored in this region (Bull, 2006) which may lead to a strong positive selection for complete linkage through suppression of recombination. Eventually, this may result in a system, in which both sexes are determined genetically, and in which each homologous chromosome (now called proto-sex chromosomes) contain linked genes that determine

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the two sexes.

The second route towards evolving sex chromosomes is thought to occur through evolution of genetic sex determination (GSD) from environmental sex determination (ESD). This is the hypothesized prominent evolutionary route in animals, as hermaphrodites are rarer than in plants (Jarne and Auld 2006; Charlesworth and Mank 2010) and because ESD may be the ancestral state in several major animal groups (Ohno 1967; Pokorná and Kratochvíl 2016). The evolutionary transition from ESD to GSD may be gradual, for instance due to a shift in genotype-specific thresholds for male vs. female development under fluctuating environmental conditions (Van Dooren and Leimar 2003). However, the classic theory of these transitions posits a scenario resembling that of the transition from hermaphroditism to separate sexes: If a female-determining mutation occurs in a population with pure ESD, this leads to a mixed sex determination system (partial GSD), in which some individuals (the carriers of the female -determining mutation) have a genetically determined sex, while others have ESD (the route through an initial male-determining mutation is also possible). The mutation will be favoured if it optimizes the sex ratio of the population (for example if the mutation occurs after a shift in environmental conditions, favouring the differentiation of one sex over the other (Edwards 1998) or if it has some other intrinsic advantage. The evolution toward the occurrence of a pair of proto-sex chromosome is then identical to that described above.

Whereas the general theory on the evolution of sex chromosomes is widely accepted, the early stages of the transition, especially the sequence of steps and the evolutionary forces involved in the accumulation of further sex-specific mutations and/or mutations determining the other sex as well as the evolution of reduction/suppression of recombination remain poorly understood (Wright et al. 2016). Insight into these questions can be obtained from studying intermediate systems such as gynodioecy or partial GSD. Prior research has mainly focused on gynodioecy. However, in almost all extant gynodioecious species, gynodioecy is controlled by cyto-nuclear interactions (e.g., an interaction between mitochondrial male sterility mutations and nuclear restorer genes (McCauley and Bailey 2009; Beukeboom and Perrin 2014) rather than being under full nuclear control, as described in the theory of sex chromosome evolution. Only a few species appear to have pure nuclear control (Kohn 1988; Connor and Charlesworth 1989; Weller and Sakai 1991; Spigler et al. 2008). To date, no animal species in an intermediate state of transition from ESD to GSD has been reported and studied at the molecular level. Such a species would provide an extremely

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

useful model to study sex chromosome evolution in animals, especially if partial GSD was under full nuclear control. However, the theoretically described rapid nature of the transition from ESD to GSD (Pokorná and Kratochvíl 2009) may explain why no such system has yet been studied in detail.

Here, we describe the genomic basis of partial GSD and initial stages of an incipient sex chromosome evolution in a small crustacean species, *Daphnia magna*, in which pure ESD individuals are co-occurring with genetically determined females. Most *Daphnia* species are cyclical parthenogens (phases of clonal reproduction with live born offspring, interspersed with by sexual production of diapause stages). Sex determination is usually environmental: Depending on environmental triggers (which vary among populations (Roulin et al. 2013), the mother may emit a juvenile hormone, which induces male development of the clonal offspring present in the ovaries (Olmstead and Leblanc 2002). Male production can also be artificially induced by adding hormone analogs to the culture medium (Olmstead and Leblanc 2002). However, certain strains of *Daphnia* species never produce males, neither under natural conditions nor when artificially exposed to hormone (Innes and Dunbrack 1993; Tessier and Caceres 2004; Galimov et al. 2011). Hence, these strains contribute to the production of diapause stages only via their female function. The trait segregates in a manner corresponding to single locus (or single region) Mendelian inheritance with a dominant female-determining allele W. Heterozygous individuals (ZW genotypes) are genetically determined females (called non male producers NMP phenotypes), and homozygous individuals (ZZ genotypes) are normal cyclical parthenogens with ESD (i.e., participating in the production of diapause stages either through male or female function, called male producers, MP phenotypes; Galimov et al. 2011).

We first performed classical linkage mapping using three different test crosses involving individuals from populations with divergent mitochondrial lineages, to locate and characterize the female sex-determining region, and to test if the same region(s) is/are responsible for the NMP phenotype in these different lineages. Using different divergent mitochondrial lineages was relevant because the W allele and the mitochondria are transmitted maternally, and hence the presence of NMP in divergent mitochondrial lineages could potentially be due to parallel evolution (Galimov et al. 2011). Second, we used RAD-sequencing to carry out an association analysis in a random sample of NMP and MP individuals from a single natural population. To corroborate the role of identified

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

regions(s) in sex determination/differentiation, we searched for the presence of potential candidate genes involved in sex determination/differentiation using resources of the *Daphnia* Genomics Consortium (DGC; http://daphnia.cgb.indiana.edu/). To identify signatures of early sex chromosome evolution, we investigated levels of differentiation between NMP and MP individuals in the sex determining region(s), and looked for evidence for recombination reduction using linkage mapping and linkage disequilibrium analyses.

Results

Linkage mapping of the NMP-determining region with microsatellites markers

We found thirteen markers in the cross NMPxMP_1 and thirteen markers in the cross NMPxMP_2 (partly overlapping) that were significantly linked with the NMP phenotype or showed significant linkage to one of the other linked markers. Eleven of these markers within each cross were successfully genotyped in <70% of offspring and were thus used for mapping (S1 Table). These markers were found to span regions of a total map length of 68 cM and 87 cM in NMPxMP_1 and NMPxMP_2, respectively, corresponding to a part of the linkage group 3 (LG3) of the MPxMP mapping cross (Fig. 1). In the NMPxMP_1 cross, three markers were in full linkage with the NMP phenotype, and in the NMPxMP_2 cross, five markers were in full linkage with the NMP phenotype. Only two of these markers were genotyped also in the NMPxMP_3 cross, and they equally showed full linkage to the NMP phenotype (the third tested marker in that cross was not polymorphic and could thus not be assessed). In all three crosses, these fully linked markers map to the same region between cM-positions 87.8 and 94.0 on the genetic map, which we called NMP genomic region (Fig. 1), and which also contains the centromere at 90.8cM.

Microsatellite loci based rates of recombination (as assessed by genetic distance) between adjacent, but not fully linked markers tended to be somewhat lower in the NMPxMP crosses compared to the MPxMP cross of the genetic map (Fig. 1, Table 1). These reductions in genetic map distance around the NMP region were significant for two intervals in each of the two crosses. When mapped on a Marray map of the LG3 chromosome (from a SNP based genetic map of a MPxMP cross), we see that the NMP genomic region itself contains a large non-recombining region around the

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

centromere, and extends also to the peri-centromeric regions (Fig. 2). The NMP region is further characterized by relatively short scaffolds in the 2.4 assembly of the draft genome of *D. magna*. This is likely indicative of a high abundance of repetitive elements (complicating the assembly). This is in stark contrast to the long scaffolds in the second non-recombining region on this linkage group (Fig. 2). This second non-recombining region remains unexplained, although may likely an artifact from the crosses used for the mapping.

Association mapping of SNPs in the NMP region, differentiation, and heterozygosity levels

The genome-wide association analysis using RAD-sequencing revealed 43 SNPs that were significantly (FDR $<10^{-5}$) associated with the NMP phenotype (Fig. 3a, S3 Table). Of these, 36 mapped to LG3 between cM-positions 72.3 and 95.7 of the genetic map, which extends a bit outside of the previously defined NMP genomic region (Fig. 3b). Additionally, five SNPs mapped to two scaffolds on LG1, and one SNP to each of LG2 and LG4 (Table 2; Fig. 3). The number of associated SNPs per scaffold was not correlated to scaffold size (Pearson's correlation coefficient r^2= - 0.117). When plotted on a physical scale, the main association on LG3 is distributed between 5.4Mb and 9Mb of the physical map inferred by LD-mapping (Fig. 4). The 36 significantly associated SNPs of LG3 were distributed across 15 scaffolds, with a combined length of 2.42 Mb. Thus, even if we do not have the correct physical map, there is strong evidence that significantly associated SNPs are distributed across a large proportion of LG3 (22% of the total size of LG3 in bp).

Strong differentiation between NMP and MP individuals, as assessed by F_{ST}, occurred in the NMP genomic region, with values as high as F_{ST}=0.7 (Fig. 4). Levels of heterozygosity in the NMP region were also higher for NMP individuals than for the MP individuals (ţ= 0.50 and ţ= 0.33 respectively, P<0.0001; Figure 4.b), while in the rest of the genome, there was no difference in heterozygosity (ţ= 0.34 and ţ= 0.32 respectively, P= 0.149) indicating that the heterozygosity difference in the NMP region is not an artifact of a general difference at the genome scale. However, four MP individuals carried only half of the average population heterozygosity (S9 Figure).

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Linkage disequilibrium and recombination in the NMP region

The pattern of linkage disequilibrium among the 36 associated SNPs on LG3 is very strong (Fig. 5a showing only the LD r^2 among the 36 SNPs). However, these associated SNPs are separated by regions were other SNPs show low or no association with the NMP phenotype (Fig. 5b). In agreement with results from the LD analysis, the four-gamete tests showed the presence of four gametes (i.e. recombination) in 37 pairs of adjacent polymorphic sites, implying the occurrence of recombination among the NMP and the MP haplotypes (Fig. 4d).

SNP effect and identification of candidate genes involved in sex determination

Of the 283 protein sequences BLASTed, 184 returned a BLAST result. Of these, 39 were described as hypothetical proteins in the *D. pulex* draft transcriptome DAPPUDRAFT but did not reliably match any proteins in the databases. The remaining 145 sequences returned a BLAST result, among which 121 (82%) had a top hit on *Daphnia pulex* or *Daphnia magna* sequences. The remaining sequences had top hits on various arthropods (15 sequences), and invertebrates (ten sequences). Among the genes annotated, the three most represented molecular functions were ion binding, hydrolase activity and transferase activity (Fig. 6). The top three cellular components were cell, cell part, and organelle. The top three biological processes were cellular process, metabolic process, and single-organism process (Fig 6.).

When cross referencing the 145 genes annotated with the NCBI list of 601 genes known to be involved in sex determination or sex differentiation in other invertebrates, we identified 14 candidate genes (Table 3). These genes include splicing factors (*transformer 2*, *serine-arginine rich splicing factor 7*, *Half Pint*), transcription factors (*sox9*-like), and genes involved in hormonal pathways (*zip9*, *zip11*, *Broad complex*, aldo-keto reductase). In addition, a few genes involved in methylation/demethylation activity (lysine-specific histone demethylase, histone deacetylase) are also present in the region. The 14 genes are located on 6 different scaffolds, with a single scaffold (scf02569) harboring eight of these genes (scf00848 and scf02003 harboring two genes each, and scf00027, scf02723 and scf03156 harboring one gene each).

Within the NMP region, nine of the 176 SNPs identified were located in genes within which the

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

exon-intron model could not be resolved (likely errors of assembly or gene structure definition), 69 were located in intergenic regions, 17 in 5'UTR regions, 8 in 3' UTR regions, 19 in introns, and 53 within a gene (S2 Table). Of those 53 SNPs, 23 corresponded to synonymous mutations, while 30 corresponded to non-synonymous mutations. One SNP located on scaffold02003 in position 98755 introduced a stop codon in the gene (the gene was annotated as hypothetical protein from the *D. pulex* draft genome). In total, 32 of the 53 coding SNPs fell in non-annotated genes labelled Hypothetical protein of the *D. pulex* transcriptome, and only one significantly associated SNP induced a non-synonymous mutation in a candidate gene identified above: the SNP at position 4433 on scf03156, inducing a change from Valine to Leucine in the lysine specific histone demethylase. Note, however, that RAD-sequencing covers only a small fraction of the regions (here 154745 loci were mapped, with reads of 95bp, representing an estimated 6.12% of the genome). Hence, additional non-synonymous SNPs may be present in non-sequenced parts of the candidate genes or other genes in the region.

Discussion

The results of our study indicate that *D. magna* has an incipient sex chromosome showing a high resemblance to an incipient ZW system (Beukeboom and Perrin 2014; Bachtrog et al. 2014). Hence, *D. magna* represents an ideal model of sex chromosome evolution through the transition from ESD to GSD. The classical genetic mapping as well as the GWAS results strongly suggests that the NMP phenotype is determined by a single, large genomic region located on LG3. The NMP region is highly heterozygous in NMP females, and is highly differentiated from its homologous region in the MP individuals. The contribution of the few significantly linked SNPs on other linkage groups is unclear. It is possible that these regions are in linkage disequilibrium with the major region on LG3 due to pleiotropic effects. However, it is also possible that they are explained by errors in the genetic map or in the genome assembly (i.e., they may in fact be within the NMP-region on LG3). The same region on LG3 was consistently associated with NMP in all three crosses with divergent mitochondrial lineages (Galimov et al. 2011). This indicates either a single evolutionary origin of NMP in *D. magna* or parallel evolution involving repeatedly the same genomic region. A single evolutionary origin of NMP in *D. magna* would indicate that the female-determining mutation is

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

old. The mitochondrial haplotypes of the females used in the three crosses span almost the entire known divergence present in the species today, and, due to exclusive maternal transmission, both mitochondria and the female-determining mutation are co-inherited (Galimov et al. 2011). The alternative explanation of rare paternal transmission of mitochondria or of transmission of the female-determining mutation through rare males cannot be ruled out, but is unlikely (Galimov et al. 2011; Svendsen et al 2015). Parallel (convergent) evolution remains a distinct possibility. Indeed, the NMP region contains multiple genes involved in sex determination and sex differentiation. It is possible that several of these genes represent mutational targets for NMP-inducing mutation, and hence the exact mutation may not be the same from one population to the next. Moreover, the NMP phenotype has also been described in *D. pulex* (Innes and Dunbrack 1993) and in *D. pulicaria* (Tessier and Caceres 2004). Considering that an estimated 150 MYA separates *D. magna* and *D. pulex* (Kotov and Taylor 2011), parallel evolution of the NMP phenotype appears to be the most parsimonious explanation, at least at the among-species level.

The NMP region mapped to the peri-centromeric region of LG3, which has a low recombination rate not only in the heterogametic sex, but also in the homogametic sex (Duki et al. unpublished), as is typical for peri-centromeric regions. This location implies that the recombination suppression around the sex determining locus cannot entirely be attributed to the events that took place after the occurrence of the sex determining mutation (e.g., progressive restriction of recombination following accumulation of further mutations with sex-specific effects, through sex antagonistic selection). As noted by Ironside (2010), sex-antagonistic selection is not the only evolutionary process that can lead to a reduction in recombination. Indeed low-recombination regions also occur on autosomes, for instance due to inversions, proximity to the centromere or the presence of supergenes (Hoffman and Riesberg 2008; Ironside 2010). If a new sex-determining mutation occurs in such a pre-existing, low-recombination region, this could favor the evolution of proto-sex chromosomes, by rendering it more likely that, from the outset, several potential target loci for additional mutations with sex-antagonistic effects are linked to the locus at which the initial sex-determining mutation occurs To our knowledge, the only other species in which a sex-determining mutation has been found in the peri-centromeric region of a chromosome is papaya (Yu et al. 2007). In *D. magna*, our work shows that this chromosomal location contributes to the low levels of recombination of the sex-determining region, suggesting that at least a part of the reduced recombination level (com-

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

pared to other parts of the genome) is not attributable to sex-antagonistic selection occurring after the establishment of the sex-determining mutation. However, sex-antagonistic selection might still have contributed to further reducing the recombination rate around the region as our results suggest that recombination (i.e., genetic map length of the region) was further reduced in the MP x NMP crosses (i.e., specifically in ZW meiosis compared to ZZ meiosis; Charlesworth 1991). Under this scenario, multiple loci would be expected to contribute to the extended NMP phenotype (i.e. not only determine the female sex, but be involved in the expression/fitness of the female phenotype or enhance maleness of the ZZ individuals) in the NMP region. We indeed have identified many NMP-associated SNPs continuously distributed across 3Mb, in strong linkage within NMP, and separated by genomic areas in which we found evidence for recombination between the Z and W chromosome (either due to low historical crossover recombination, or gene conversion, leading to reduced levels of LD in-between the strongly associated regions). We also found many genes known to be involved in sex determination/sex differentiation in other taxa. Overall, this suggests that the sex-determining mutation may be surrounded by other loci with sex-specific beneficial alleles that are positively selected when occurring with it (maybe negatively selected in a MP genetic background). However, we cannot exclude other possibilities for the additional reduction in recombination in NMP individuals, such as localized chromosomal inversions.

The NMP region on LG3 contains multiple genes that are known to be involved in sex determination and differentiation in other taxa. Here we identified some of the usual key players involved in the regulation of the sex-determining cascade, such as transcription factors, post-transcriptional regulators and genes controlling the activation/inactivation of sex hormones/pheromones. However, we did not find a gene that shows an exclusive role in sex determination, such as *doublesex* (*dsx*), the terminal effector of the male sex determining cascade in all insects investigated to date (Salz 2011; Beukeboom and Perrin 2013) as well as in *D. magna* (Kato et al. 2011). This indicates that the NMP mutation does not impact the terminal effector of the sex determining/differentiating cascade in *D. magna*. Interestingly, the splicing regulator involved in the control of the sex-specific splicing of *dsx* was found on scf02569: *transformer 2* (*tra2*). *Tra2* is part of the spliceosome that is required to regulate female-specific splicing and polyadenylation of *dsx* pre-mRNA. The absence of *tra2* during the splicing of the *D. magna*'s *dsx* might then drive the embryo to develop as female. Scf02569 also harbors the transcription factor *sox9*, which acts to inactivate the female differen-

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

tiation pathway and promote spermatogenesis in males in mammals, hence it is conceivable that a mutation in this gene might lead to a loss of male function. Other genes on the same scaffold are involved in sex-specific endocrine signaling pathways (*zip9* membrane androgen receptor, the *Broa- complex*; Karim et al. 1993), as well as a member of the aldo-keto reductase family (Penning et al. 2000). Although significantly associated SNPs were found across a large (3Mb) region, the 400 kb long scaffold02569 contains about 25% of the significant SNPs (among which some of most strongly associated ones), as well as eight of the 14 candidate genes identified in this study, hence making it a strong candidate for a central role in the NMP phenotype.

The reasons underlying the maintenance of the NMP mutation through time remain unclear. A model that was specifically developed for *Daphnia* suggests that the mutation could be maintained if there was a fitness cost of within-clone mating (Innes and Dunbrack 1993). Inbreeding depression in Daphnia is known to be strong (Lohr and Haag 2015). In our sample set, we identified four individuals that were likely first generation offspring of within-clone mating of an MP clone (genetically identical to self fertilization). This represented 5.7% of our sampled population, and highlights the strong occurrence of inbreeding in *Daphnia*, which could explain an intrinsic advantage of NMP due to obligate outcrossing. Despite the mutation being maintained and despite this mutation likely inducing selection for a more even sex ratio (Innes and Dunbrack, 1993), an evolution towards full GSD might be hindered by the very nature of the reproductive mode of *D. magna*. Under cyclical parthenogenesis, individuals hatching from resting eggs are all females, and undergo several cycles of parthenogenetic reproduction before switching to sexual reproduction. For a complete GSD system to evolve in *Daphnia*, a mutation preventing MP individuals to produce haploid sexual egg should occur, followed by all the molecular changes necessary reshaping of the developmental processes to obtain both male and females hatching from the resting stages. Once males and female hatch from resting eggs, additional mutation(s) would be necessary to obtain the production of haploid eggs that do not depend on environmental cues. This chain of events seems highly unlikely, and, therefore, partial GSD is may be evolutionary stable in *Daphnia*. However, even with stable partial GSD, MP clones in mixed populations may still evolve male specialization to a certain degree compared to pure ESD population, due to sex-ratio selection during the sexual reproduction phase. This could favor the accumulation of male-positive recessive alleles on the Z-homolog of the NMP region, to limit genomic conflict between the genetic females and the

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

specialized male producers.

Conclusion

Here, we described for the first time the genetic evidence for an animal system representing an intermediate step in the evolution from ESD to GSD. The specific chromosomal location of the female sex-determining region calls suggests a possible role of pre-existing low levels of recombination in the early evolution of sex chromosomes. Many genes acting as key players in sex determination and differentiation in other taxa have been found in the sex determining region, and the genes on the 400kb scaffold02569 are a particularly likely candidates for genes containing the sex determining mutation. Although the system may never evolve into a full GSD system, further steps of sex chromosome evolution may occur due to male due to sex-ratio selection favoring male specialization of the remaining ESD individuals. Together with the amenability of the life history for experimental approaches (e.g., short generation time; production of high numbers of individuals, standardized breeding conditions for experimental testing) this makes *Daphnia* an excellent experimental model for research on the early evolutionary events shaping animal sex chromosomes.

Material and Methods

Linkage mapping of the NMP-determining region

In order to map the genomic region responsible for the NMP phenotype, we performed experimental crosses between known NMP and MP genotypes. Microsatellite distributed across the genome, were used to investigate the parental lines. Markers that were heterozygous in the NMP mother and for which, at the same time, the father genotype differed from that of the mother were genotyped in the offspring. For all these markers, it could unambiguously be determined which of two maternal alleles was inherited by a given offspring. Hence, linkage to the NMP-determining region could be assessed, by assaying co-transmission of maternal alleles with the phenotype, which is determined by a heterozygous locus or region (ZW genotype). Before genotyping, the offspring were phenotyped using the juvenile hormone Methyl Farnesoate, which triggers the production of males in MP strains but not in NMP strains of Daphnia (Galimov et al. 2011).

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Three crosses were investigated, involving NMP females with strongly divergent mitochondrial haplotypes. All three crosses involved outcrossing between two populations to ensure that a sufficient number of markers had different genotypes between fathers and mothers. One of the crosses also involved a NMP mother that was already a hybrid between two populations (in order to maximize heterozygosity). Specifically, the first cross called NMPxMP_1 involved a NMP female from Volvograd, Russia (N48ř31'48.00", E44ř29'13.00") and a male from Orog-Nur, Mongolia (N45ř1'57.75", E100ř39'37.73") as well as 66 of their F1 offspring. The second cross (NMPxMP_2) used a hybrid NMP female (produced by crossing a NMP female from Moscow, Russia, N55ř45'48.65", E37ř34'54.00", with a male from Orog-Nur) and a male from Vääränmaanruskia, Finland (N60ř16'17.82", E21ř53'46.74"), as well as 54 of their offspring. The third cross (NMPxMP_3) involved an NMP female from Yakutsk, Russia (N61ř57'50.57", E129ř37'51.44") and a male from Rybnoye, Russia (N56ř25'30.01", E37ř36'9.62") and 22 of their offspring.

Phenotyping methods followed the hormone test as described in Galimov et al. (2011). Microsatellite loci were amplified using the M13-protocol (Schuelke 2000): For each locus, unlabeled forward and reverse primers were used together with fluorescently labelled, universal M13 primer. The forward primer consisted of a locus-specific part as well as an overhang complementary to M13. PCR reactions were carried out using the Type-it Microsatellite PCR Kit (Qiagen) according to the manufacturer's protocol with an annealing temperature of 60řC. After 22 cycles, the annealing temperature was lowered to 53řC for another 20 cycles in order to allow for proper M13 annealing. The resulting PCR product was diluted four times and mixed with a LIZ5000 size ladder (Applied Biosystems). Samples were genotyped using ABI 3730 capillary sequencer and GENEMAPPER software v. 3.0 (Applied Biosystems). A total of 81 microsatellite loci (S1 Table) were tested in the parents. Of these, 60 were polymorphic in one or both parents and thus genotyped in the offspring (47 in NMPxMP_1 and 21 in NMPxMP_2, partially overlapping). Linkage to the NMP phenotype was assessed with a Fisher's Exact tests (two-tailed). Some of the markers were specifically designed in regions for which linkage to the NMP phenotype was suspected based on information from an earlier version of the genetic map (Routtu et al. 2010; Routtu et al. 2014) and the initial finding of weak but significant linkage of one marker (dm_scf00243_208642) in the NMPxMP_1 cross. Therefore the markers do not represent a random sample throughout the genome. The NMPxMP_3 cross was done at a later stage, and thus was only used as a validation for the results

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

obtained with the previous two crosses. As such, only three loci closely linked to NMP in the first two crosses were also genotyped in the offspring from this cross.

Linkage mapping of the NMP region was carried out in R/qtl (Broman et al. 2003). It became evident that NMP mapped to a region of linkage group 3 (LG3) of the *D. magna* genetic map v.4.0.1 (Svendsen et al. 2015; Duki et al. unpublished data). Hence, map construction was done using markers that either showed significant linkage with the NMP phenotype (P <0.01 in pairwise Fisher's exact tests) or were found on scaffolds of the *D. magna* genome v2.4 that had been mapped to LG3 (Svendsen et al. 2015; Duki et al. unpublished data) . Markers that had more than one third of missing genotypes (amplification failures, etc.) were discarded.

To construct the genetic map of the NMP genomic region, we first ordered those markers that corresponded to scaffolds on the genetic map v4.0.1 according to the cM position of the nearest SNP in v4.0.1, and then mapped the additional markers (found on scaffolds that had not been mapped in v4.0.1) using R/qtl (based on pairwise map distances). The only exception to this procedure was done for microsatellite marker scf02066_483524, which is located on a mis-assembled part of scf02066, closely linked to the end of scf00494 (Duki et al. unpublished data), and thus was ordered according to this position. Once ordered, Kosambi-corrected genetic map distances among all markers were inferred from the offspring genotypes using R/qtl (with the option sliding-window = 8 markers).

To test for reduced recombination around the NMP genomic region in the three MPxNMP crosses compared to the MPxMP mapping cross, we compared the genetic distances for intervals between adjacent markers between the crosses. Specifically, for each interval, we assessed the number of recombinant vs. non-recombinant individuals in each cross and tested for significant differences using Fisher's Exact tests (two-tailed) implemented in R core package stats (R Development Core Team, 2008).

Association mapping of SNPs in the NMP region, differentiation, and heterozygosity levels

Linkage mapping using microsatellites allowed us to map the NMP genomic region to a low-recombining region of LG3 and confirmed the mode of transmission of NMP, with NMP individuals

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

being heterozygous (ZW) and MP individuals being homozygous (ZZ) for the NMP-determining part of that region. In an attempt to delimit the NMP-determining part of this genomic region more precisely, we used SNP data obtained by RAD-sequencing of a random sample of 72 individuals (17 NMP and 55 MP; demultiplexed FASTQ files are available on the SRA database: reference XXX) from the Moscow population using BWA (Li and Durbin 2009) and the Stacks software (Catchen et al. 2013) (S2 Text, for details of the RAD-sequencing protocols and SNP calling, and S3 Table for all the SNP obtained and analysed subsequently). We first performed a genome-wide association study, using the expectation that any bi-allelic SNP functionally related to NMP or tightly linked to it should be heterozygous in all NMP individuals and homozygous for the more frequent of the two alleles in MP individuals (corresponding to the ZW and ZZ genotypes, respectively). To test for an association with NMP we grouped individuals into four categories for each bi-allelic site (only sites with a minor allele frequency over 0.1 and less than one third of the individuals with missing genotypes): heterozygous NMP individuals, non-heterozygous NMP individuals, MP individuals homozygous for the major allele, and MP individuals non-homozygous for the major allele. For each site, we counted the number of individuals in each of the four categories and calculated the expected number of individuals (under the null-hypothesis of no association) using standard Hardy-Weinberg proportions with allele frequencies estimated across all individuals. We then used Pearson's Chi-square tests with two degrees of freedom to evaluate the genotype-phenotype association at each site. However, in order to only test the hypothesis specified above, any excess that went in the direction opposite the hypothesis (for instance an excess of non-heterozygous individuals among NMP) was discarded (i.e., was not taken into account for the overall Chi-square value; S4 Script for the R script of the association analysis). Significance of association was assessed by correcting the P-value of the Chi-square test according to an overall false discovery rate (FDR) of 10^{-5} using the p.adjust function of the R core stats package. As an alternative test of differentiation between MP and NMP individuals, we also used classical F_{ST} for each SNP, estimated with the R package PopGenome (Pfeiffer et al. 2014). We also investigated levels of relative heterogysosity in the NMP region, as well as at the genome scale for MP and NMP individuals, to test if a particular difference in heterozygosity levels in the NMP region is or not correlated with a general difference in the rest of the genome.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Linkage disequilibrium estimation and detection of recombination events

While the genome-wide approach tested for the presence of associated SNPs within and outside the NMP genomic region on LG3,it does not indicate whether the association of all the SNP in the NMP region is due to physical linkage disequilibrium (association because of a lack recombination) rather than statistical linkage disequilibrium (association because positive selection on the genetic background in the NMP region). We restricted the analysis in a second step to just the part of the NMP genomic region (corresponding to LG3 centromeric linked region at 90 cM (Svendsen et al. 2015) and flanking peri-centromeric region, corresponding to positions between 85 and 95cM on the genetic map). However, due to the dearth of recombination in this region, the relative position and orientation of many of the scaffolds is unknown (several entire scaffolds having the exact same cM position). We hence first inferred the likely relative position and orientation of these scaffolds by linkage disequilibrium (LD) mapping in MP individuals, assuming no structural rearrangement of those scaffold between MP and NMP individuals. Using only MP individuals (in order to avoid circularity in later testing for differentiation between MP and NMP individuals), we estimated LD between terminal regions of different scaffolds (averaged across the three terminal SNPs) and ordered and oriented scaffolds within each cM position in a way that minimizes LD (S5 Text, for the methodology). To obtain a physical map of this region, we then used the inferred ordering and orientation of these scaffolds, and distance in base pairs estimated from the position of the SNPs within the scaffolds as well as the cumulative length of intervening scaffolds. Note that this most likely underestimates the true length in bp of the NMP region because it may also contain unassembled sequences and unmapped scaffolds between the mapped scaffolds.

Once the inferred physical map was established, we used it to map again the genotype-phenotype association in the region. We estimated LD by calculating pairwise r2 values on all individuals for each pair of SNPs across the region (S6 Table). Pairwise r2 values were estimated with MCLD (Zaikin et al. 2008), which uses genotypic data without the need to infer the (unknown) haplotypic phase. Significance was tested using 9999 permutations in MCLD, and the extent of LD was visualized using a heatmap constructed in R using the LDheatmap package (Shin et al. 2006). To assess the minimum number of historical recombination events between MP and NMP haplotypes in the region, we phased the data using the GERBIL program implemented in the package GEVALT

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

V2.0 (Davidovich et al. 2007), which results in two MP haplotypes for each MP individual and in one NMP and one MP haplotype for each NMP individual (NMP haplotypes were identified according to the presence/absence of the two most strongly associated SNPs named scf2723_2194 and scf2723_13482). We used a filtered dataset composed of 140 polymorphic sites in the region, retaining just one site per read (a maximum of two polymorphic sites on the same read were present in the whole data set, but SNPs on the same read were always in full linkage). We did not allow the program to infer missing genotypes, because this would have resulted in a data set biased towards the more common MP alleles (only 17 out of 144 haplotypes are NMP haplotypes). We then used Hudson's four-gamete test to infer the minimum number of historical recombination events needed to explain the data. Because we were interested in estimating the minimum number of recombination events, and because we could not exclude genotyping nor phenotyping error (the latter only in the direction of falsely identifying an individual as NMP), we used conservative criteria for the test: We first removed the NMP individual (RM1-01) that resulted the highest evidence for recombination (assuming that it may be the result of a phenotyping error). Furthermore, before carrying out the test we corrected singleton variants within each haplotype group: if a variant was present in only one haplotype in the group, we reverted its state to the majority allele in this group (overall, 27 loci and 6 loci out of 140 loci were reverted in NMP and MP respectively). This conservatively assumes that all these singleton variants were due to genotyping error (note that loci with a minor allele frequency of <0.1 across both groups had already been excluded during the initial filtering; S7 Table for the list of haplotypes). Finally, to test only for recombination between NMP and MP haplotypes (as opposed to recombination within MP), we inspected all instances where recombination was detected by the four-gamete test and retained only those where the inferred recombination had occurred between the two classes of haplotypes.

SNP effect and identification of candidate genes involved in sex determination

To assess whether the NMP region contains any candidate genes with already known functions related to sex differentiation or sex determination, we extracted all 1306 protein sequences corresponding to transcript sequences mapping to the scaffolds in the NMP region, and reduced isoform redundancy using BlastClust (available at http://toolkit.tuebingen.mpg.de/blastclust) with the following parameters: minimum length coverage of 60%, minimum identity of 90%, minimum tran-

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

script size of 100 amino acids. This resulted in a set of 361 protein sequences, which we trimmed by hand to remove redundancy (we only kept one transcript for each gene). The retained 283 protein sequences (S8 Text for the complete list) were blasted against the Blast2GO database (Conesa & Götz 2008), using blastp and a maximum e-value of 10^{-10}. Annotated genes were then compared with a list of 601 genes obtained from the NCBI gene data base using the keywords sex determination and sex differentiation. We also used the GFF file (available at *Daphnia* Genomic Consortium, WFleaBase), which contains gene features of *D. magna*, to classify each SNP in the NMP region according to whether it induces a synonymous or a non-synonymous change. This analysis was done using the software tool IGV (Robinson et al. 2011).

Acknowledgements

We thank the Zoo of Moscow and N. I. Skuratov for sampling permits, and David Frey for help with culture maintenance indoors. We thank the Department of Biosystem Science and Engineering of the ETH Zurich, in particular C. Beisel and I. Nissen for Illumina sequencing, and we gratefully acknowledge support by M.-P. Dubois, the platform Service des Marqueurs Génétiques en Ecologie at CEFE, and the genotyping and sequencing facilities of the Institut des Sciences de l'Evolution-Montpellier and the Labex Centre Méditerranéen Environnement Biodiversité (CeMEB). We thank M. Rösti for the modified RAD-seq protocol and for the discussions and scripts on the linkage disequilibrium analysis. We thank the University of Fribourg and the Montpellier Bioinformatic Biodiversity plateform and the Labex CeMEB for access to high-performance computing clusters. The sequence data for the *D. magna* genome project V2.4 were produced by The Center for Genomics and Bioinformatics at Indiana University and distributed via wFleaBase in collaboration with the *Daphnia* Genomics Consortium (project supported in part by NIH award 5R24GM078274-02 *Daphnia* Functional Genomics Resources). We also thank Peter Fields for their constructive comments on earlier versions of the paper. This work was supported by the Swiss National Science Foundation (Grant no. 31003A_138203), the Russian Foundation of Basic Research, and by the European Union (Marie Curie Career Integration Grant PCIG13-GA-2013-618961, DamaNMP).

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman TL et al. 2014. Sex Determination: Why So Many Ways of Doing It? PLoS Biology. 12:e1001899.

Beukeboom L, Perrin N. 2014. The evolution of sex determination. Oxford University Press, ISBN: 9780199657148.

Broman KW, Wu H, Sen S, Churchill GA. 2003. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 19:889–890.

Bull JJ. Evolution of Sex Determining Mechanisms. 1983. Benjamin/Cummings Pub. Co., Advanced Book Program, Menlo Park, CA.

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology. 22:3124-3140.

Charlesworth B. 1991. The evolution of sex chromosomes. Science. 251:1030–1033.

Charlesworth B, Charlesworth D. 1978. A model for the evolution of dioecy and gynodioecy. The American Naturalist. 112:975–997.

Charlesworth D, Mank JE. 2010. The birds and the bees and the flowers and the trees: lessons from genetic mapping of sex determination in plants and animals. Genetics. 186:9–31.

Conesa A, Götz S. 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics. Article ID 619832.

Connor HE, Charlesworth D. 1989. Genetics of male sterility in gynodioecious *Cortaderia* (Gramineae). Heredity. 63:373–382.

Davidovich O, Kimmel G, Shamir R. 2007. GEVALT: An integrated software tool for genotype analysis. BMC Bioinformatics. 8:36–43.

Duki M, Berner D, Roesti M, Haag CR, Ebert D. A high-density genetic map reveals variation in recombination rate across the genome of *Daphnia magna*. BMC Genetics, in review.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Edwards AWF. 1998. Selection and the sex ratio: Fisher's sources. American Naturalist. 151:564-569.

Ellegren H. 2011. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nature Reviews Genetics. 12:157–166.

Galimov Y, Walser B, Haag CR. 2011. Frequency and inheritance of non-male producing clones in *Daphnia magna*: evolution towards sex specialization in a cyclical parthenogen? Journal of Evolutionary Biology. 24:1572–1583.

Hoffmann AA, Rieseberg LH. 2008. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptative shifts and speciation? Annual Review of Ecology Evolution and Systematics. 39:21–42.

Innes DG, Dunbrack RL. 1993. Sex allocation variation in *Daphnia pulex*. Journal of Evolutionary Biology. 6:559–575.

Ironside JE. 2010. No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution. BioEssays. 32:718–726.

Jarne P, Auld JR. 2006. Animal mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution. 60:1816–1824.

Joron M, Frezal L, Jones RT, Chamberlain N, Lee SF, Haag CR et al. 2011. Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. Nature. 477:203–206.

Karim FD, Guild GM, Thummel CS. 1993 The *Drosophila* Broad-Complex plays a key role in controlling ecdysone-regulated gene expression at the onset of metamorphosis. Development. 118:977–988.

Kato Y, Kobayashi K, Watanabe H, Iguchi T. 2011. Environmental Sex Determination in the Branchiopod Crustacean *Daphnia magna*: Deep Conservation of a Doublesex Gene in the Sex-Determining Pathway. Plos Genetics. 7:e1001345.

Kohn J. 1988. Why be female? Nature. 335:431–433.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Kotov A, Taylor DJ. 2011. Mesozoic fossils (>145MYA) suggest the antiquity of the subgenera of *Daphnia* and their coevolution with chaoborid predators. BMC Evolutionary Biology. 11:129–138.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754–60.

Lohr JN, Haag CR. 2015. Genetic load, inbreeding depression, and hybrid vigor covary with population size: An empirical evaluation of theoretical predictions. Evolution. 69:3109-3122.

McCauley DE, Bailey MF. 2009 Recent advances in the study of gynodioecy: the interface of theory and empiricism. Annals of Botany. 104:611-620.

Miura I. 2008 An evolutionary witness: the frog *Rana rugosa* underwent change of heterogametic sex from XY male to ZW female. Sexual Development. 1:23–331.

Ohno S. 1967. Sex chromosomes and sex-linked genes. Springer-Verlag, Berlin, Heidelberg, New York.

Olmstead AW, Leblanc GA. 2002. Juvenoid hormone methyl farnesoate is a sex determinant in the crustacean *Daphnia magna*. Journal of Experimental Zoology. 293:736–739.

Penning TM, Burczynski ME, Jez JM, Hung CF, Lin HK, Ma H et al. 2000. Human 3alpha-hydroxysteroid dehydrogenase isoforms (AKR1C1-AKR1C4) of the aldo-keto reductase superfamily: functional plasticity and tissue distribution reveals roles in the inactivation and formation of male and female sex hormones. Biochemistry Journal.351:67–77.

Pfeifer B, Wittelsbuerger U, Ramos Onsins SE, Lercher MJ. 2014. PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses. Molecular Biology and Evolution. 31:1929–1936.

Pokorná M, Kratochvíl L. 2009. Phylogeny of sex-determining mechanisms in squamate reptiles: Are sex chromosomes an evolutionary trap? Zoological Journal of the Linnean Society. 156:168–183.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Pokorná M, Kratochvíl L. 2016. What was the ancestral sex-determining mechanism in amniote vertebrates? Biological Reviews. 91:1–12.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robinson JT, Thorvaldsdóttir, Winkler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative Genomics Viewer. Nature Biotechnology. 29:24–26.

Roulin AC, Routtu J, Hall MD, Janicke T, Colson I, Haag CR, Ebert D. 2013. Local adaptation of sex induction in facultative sexual crustacean: insights from QTL mapping in natural populations of *Daphnia magna*. Molecular Ecology. 22:3567–3579.

Routtu J, Jansen B, Colson I, De Meester L, Ebert D. 2010. The first-generation *Daphnia magna* linkage map. BMC Genomics. 11:508–514.

Routtu J, Hall MD, Albere B, Beisel C, Bergeron RD, Chaturvedi A, et al. 2014. An SNP-based second-generation genetic map of *Daphnia magna* and its application to QTL analysis of phenotypic traits. BMC Genomics. 15:1033.

Salz HK. 2011. Determination in Insects: a binary decision based on alternative splicing. Current Opinion in Genetics and Development. 21:395-400.

Shin JH, Blay S, McNeney B, Graham J. 2006 LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. Journal of Statistical Software. 16:Code Snippet 3.

Shuelke M. 2000. An economic method for the fluorescent labeling of PCR fragments. Nature Biotechnology. 18:233–234.

Spigler RB, Lewers KS, Main DS, Ashman TL. 2008. Genetic mapping of sex determination in a wild strawberry, *Fragaria virginiana*, reveals earliest form of sex chromosome. Heredity. 101:507-517.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Stöck M, Horn A, Grossen C, Lindtke D, Sermier R, Betto-Colliard C, et al. 2011. Ever-Young Sex Chromosomes in European Tree Frogs. PLoS Biology. 9.

Svendsen N, Reisser CMO, Dukiç M, Thuillier V, Ségard A, Liautard-Haag C et al. 2015. Identification of cryptic asexuality in *Daphnia magna* by RAD-sequencing. Genetics. 201:1143:1155.

Tessier AJ, Caceres CE. 2004. Differentiation in sex investment by clones and populations of *Daphnia*. Ecology Letters. 7:695-703.

The Tree of Sex Consortium. 2014. Tree of Sex: A database of sexual systems. Scientific Data. 1:140015.

Van Dooren TJM, Leimar O. 2003. The evolution of environmental and genetic sex determination in fluctuating environments. Evolution. 57:2667-2677.

Weller SG, Sakai AK. 1991. The genetic basis of male sterility in *Schiedea* (Caryophyllaceae), an endemic Hawaiian genus. Heredity. 67:265–273.

Wright AE, Dean R, Zimmer F, Mank JE. 2016. Nature Communications 7, Published 04 July 2016 How to make a sex chromosome. Nature Communications. 7:12087.

Yu Q, Hou S, Hobza R, Feltus FA, Wang X, Jin W. 2007. Chromosomal location and gene paucity of the male specific region on papaya Y chromosome. Molecular Genetics and Genomics. 278:177–185.

Zaykin, DV, Pudovkin AI, Weir BS. 2008. Correlation-based inference for linkage disequilibrium with multiple alleles. Genetics. 180:533–545.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

List of Figures

Figure 1. Genetic map of the two NMPxMP crosses (microsatellite markers) and of a MPxMP cross (SNP markers), showing Linkage Group 3. Map distances are in centiMorgans, calculated with the Kosambi mapping function in R/qtl. Areas in light blue / light red show a non significant reduction / expansion of recombination by comparison to the MP cross. while areas in bright blue indicates a significant reduction of recombination. For the two NMPxMP crosses, one marker per position is represented. In NMPxMP_1, dm_scf02569_310402, dm_scf00933_2550 and dm_scf00700_81490 were in full linkage with the NMP locus. In NMPxMP_2, dm_scf00532_1398 was fully linked with dm_scf02121_20555; also, dm_scf02569_317703, dm_scf01492_1407, dm_scf00933_2550, dm_scf03156_57375 and dm_scf00966_75426 were fully linked with the NMP locus.

Figure 2. Marray map of LG3 in the MPxMP cross. Pointed lines delimitate the NMP linked region according to the microsatellite mapping. The centromeric region (90.8cM) is highlighted in red. The X axis show the physical color-coded distribution of the scaffolds.

Figure 3. Genome-Wide association results. Association of SNP loci with the NMP polymorphism in a sample of 53 MP and 17 NMP individuals (a) across the entire genome and (b) centered on LG3. On LG3, markers between 72.3cM and 95.7cM show significant association with the NMP phenotype (the red line shows significance, with FDR-corrected P-values $<10^{-3}$).

Figure 4. Differentiation, heterozygosity and association levels between NMP and MP individuals along LG3. Evolution of (a) the F_{ST} values, (b) the heterozygosity difference between NMP and MP, (c) the log transformed P-values for the association analysis. (d) shows the association in the NMP region and the minimum number of recombinant haplotypes found at particular positions. In addition, centiMorgan position from the genetic map and inferred physical position (linkage disequilibrium mapping) and length of scaffolds are represented.

Figure 5. Linkage Disequilibrium Heatmap (r2 coefficients) in the NMP region. Re-

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sults are shown for (a) the NMP associated SNPs , (b) all SNPs mapping within the genomic region corresponding to the NMP region using the 72 individuals (55 MP, 17 NMP). Black triangles represent scaffolds, and the bi-colored band represents the centiMorgan values at each SNP position. + and represent the orientation of the scaffold.

Figure 6. Gene content of the NMP region in Daphnia magna using the Gene Ontology (GO) annotations. Results are shown for (a) biological processes, (b) molecular functions and (c) cellular components.

List of Tables

Table 1. Fisher's Exact test P-values for all pairs of markers considered in (a) the NMPxMP_1 cross and (b) the NMPxMP_2 cross.

Table 2. List of scaffolds containing SNPs significantly associated to NMP (Chi square test; P<10^{-5}). LG: linkage group; Size: total size of the scaffold (in basepair); Nb. SNPs: number of associated SNPs on the scaffold; SNPs position: position of the SNP on the scaffold; P value: resulting Chi square P value.

Table 3. Results of the Blast run showing the 14 candidate genes and their location on the *D. magna* genome. The table reports the start and end position of the gene on the scaffold it maps to, the size of the expected protein (number of amino-acid), the NCBI attributed gene name, the taxon with the best blast hit, the corresponding e-value, and the percentage of sequence similarity.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

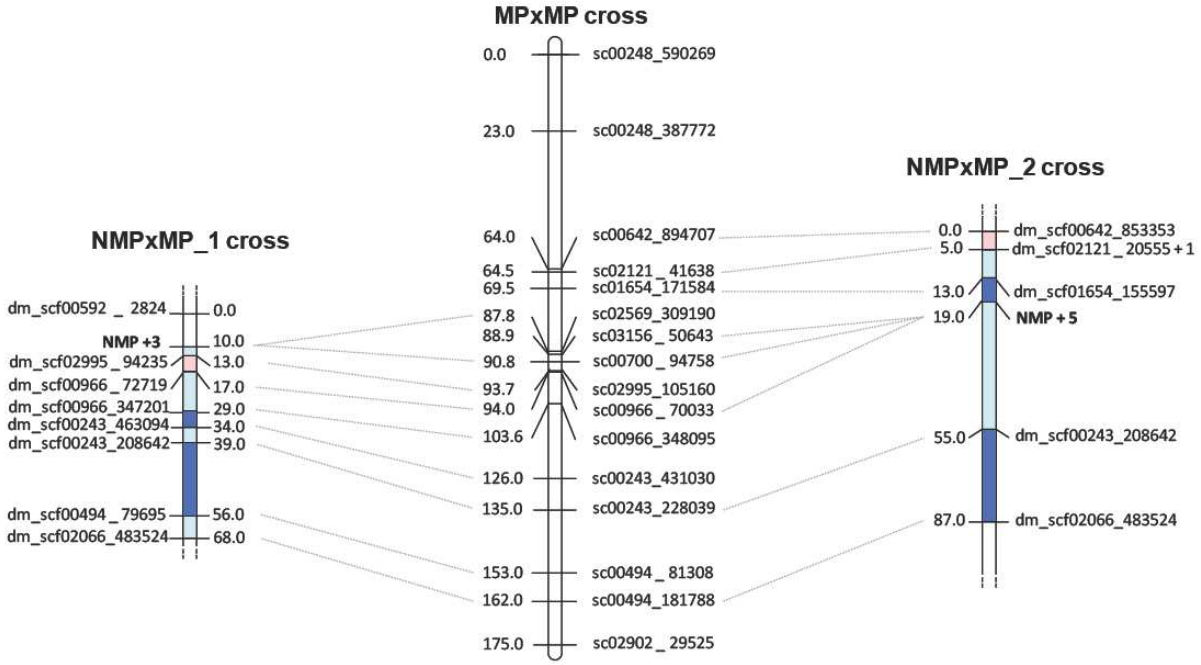

Figure 1: Genetic map of the two NMPxMP crosses (microsatellite markers) and of a MPxMP cross (SNP markers), showing Linkage Group 3. Map distances are in centiMorgans, calculated with the Kosambi mapping function in R/qtl. Areas in light blue / light red show a non significant reduction / expansion of recombination by comparison to the MP cross. while areas in bright blue indicates a significant reduction of recombination. For the two NMPxMP crosses, one marker per position is represented. In NMPxMP_1, dm_scf02569_310402, dm_scf00933_2550 and dm_scf00700_81490 were in full linkage with the NMP locus. In NMPxMP_2, dm_scf00532_1398 was fully linked with dm_scf02121_20555; also, dm_scf02569_317703, dm_scf01492_1407, dm_scf00933_2550, dm_scf03156_57375 and dm_scf00966_75426 were fully linked with the NMP locus.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

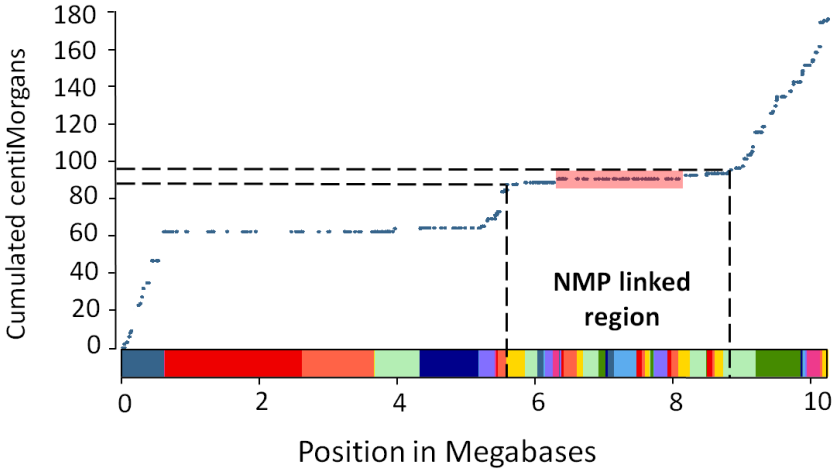

Figure 2: Marray map of LG3 in the MPxMP cross. Pointed lines delimitate the NMP linked region according to the microsatellite mapping. The centromeric region (90.8cM) is highlighted in red. The X axis show the physical color-coded distribution of the scaffolds.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

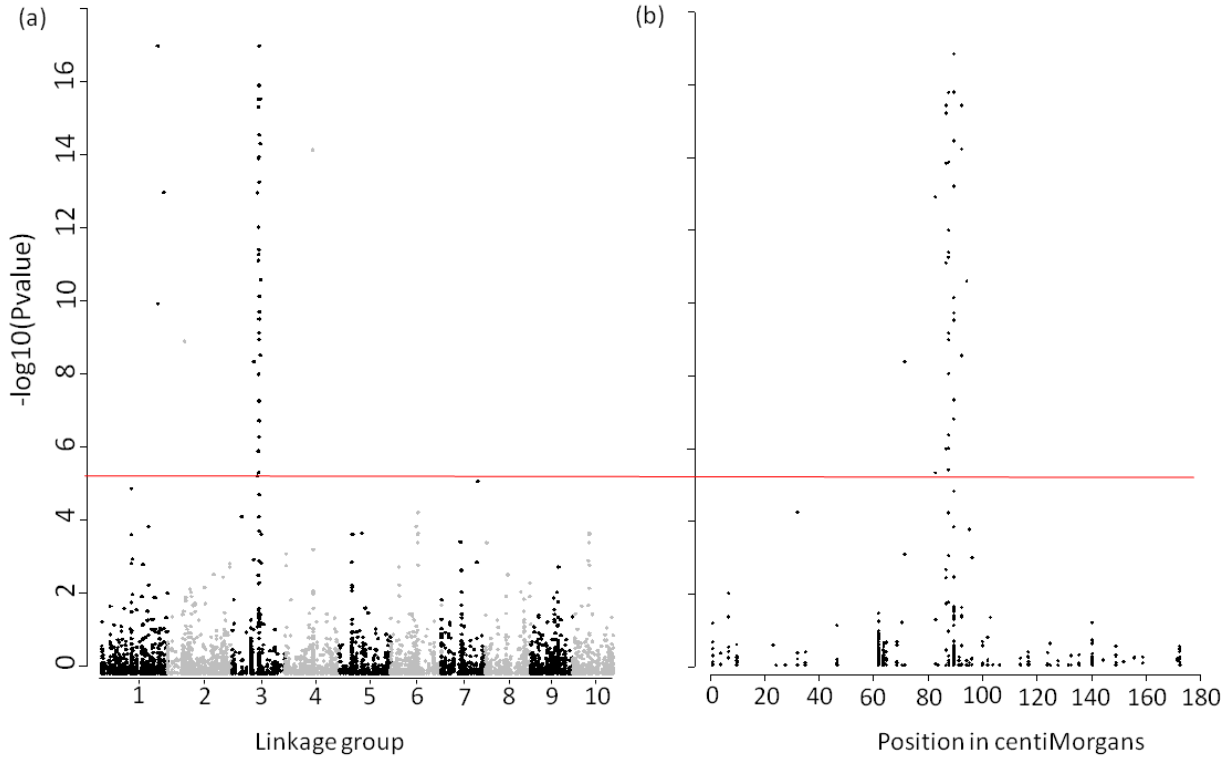

Figure 3: Genome-Wide association results. Association of SNP loci with the NMP polymorphism in a sample of 53 MP and 17 NMP individuals (a) across the entire genome and (b) centered on LG3. On LG3, markers between 72.3cM and 95.7cM show significant association with the NMP phenotype (the red line shows significance, with FDR-corrected P-values $<10^{-3}$).

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

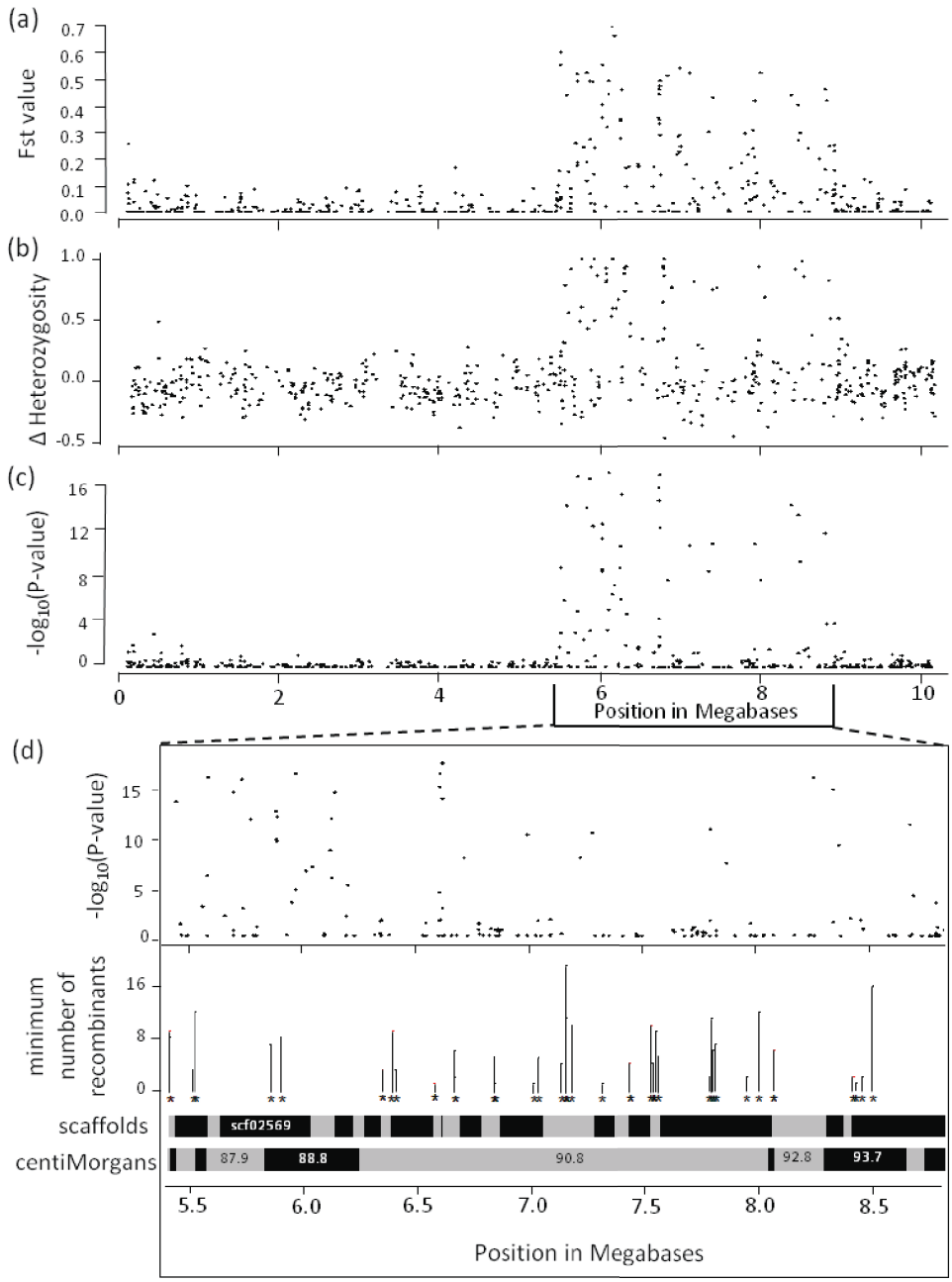

Figure 4: Differentiation, heterozygosity and association levels between NMP and MP individuals along LG3. Evolution of (a) the F_{ST} values, (b) the heterozygosity difference between NMP and MP, (c) the log transformed P-values for the association analysis. (d) shows the association in the NMP region and the minimum number of recombinant haplotypes found at particular positions. In addition, centiMorgan position from the genetic map and inferred physical position (linkage disequilibrium mapping) and length of scaffolds are represented.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

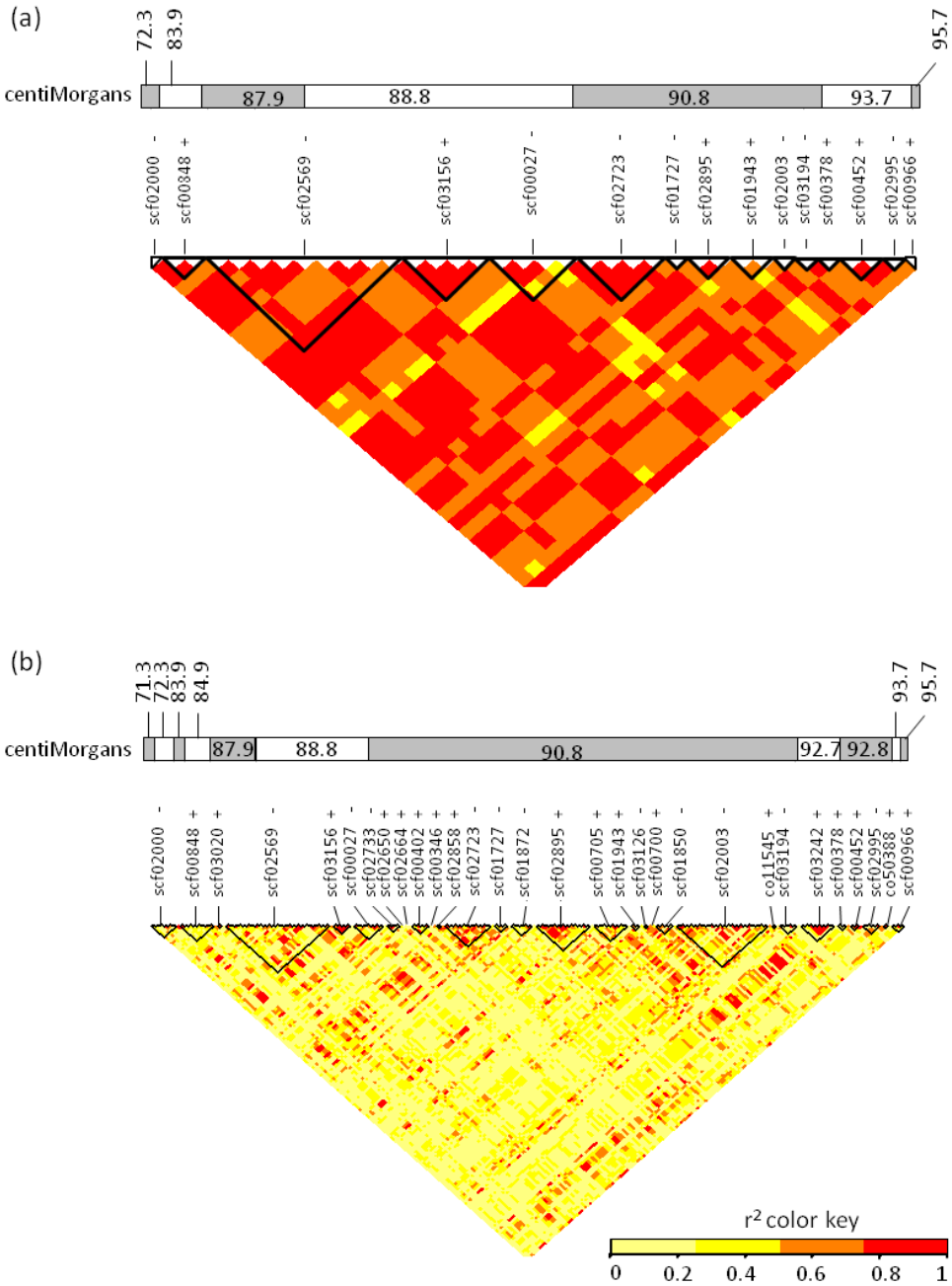

Figure 5: . Linkage Disequilibrium Heatmap (r2 coefficients) in the NMP region. Results are shown for (a) the NMP associated SNPs , (b) all SNPs mapping within the genomic region corresponding to the NMP region using the 72 individuals (55 MP, 17 NMP). Black triangles represent scaffolds, and the bi-colored band represents the centiMorgan values at each SNP position. + and represent the orientation of the scaffold.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

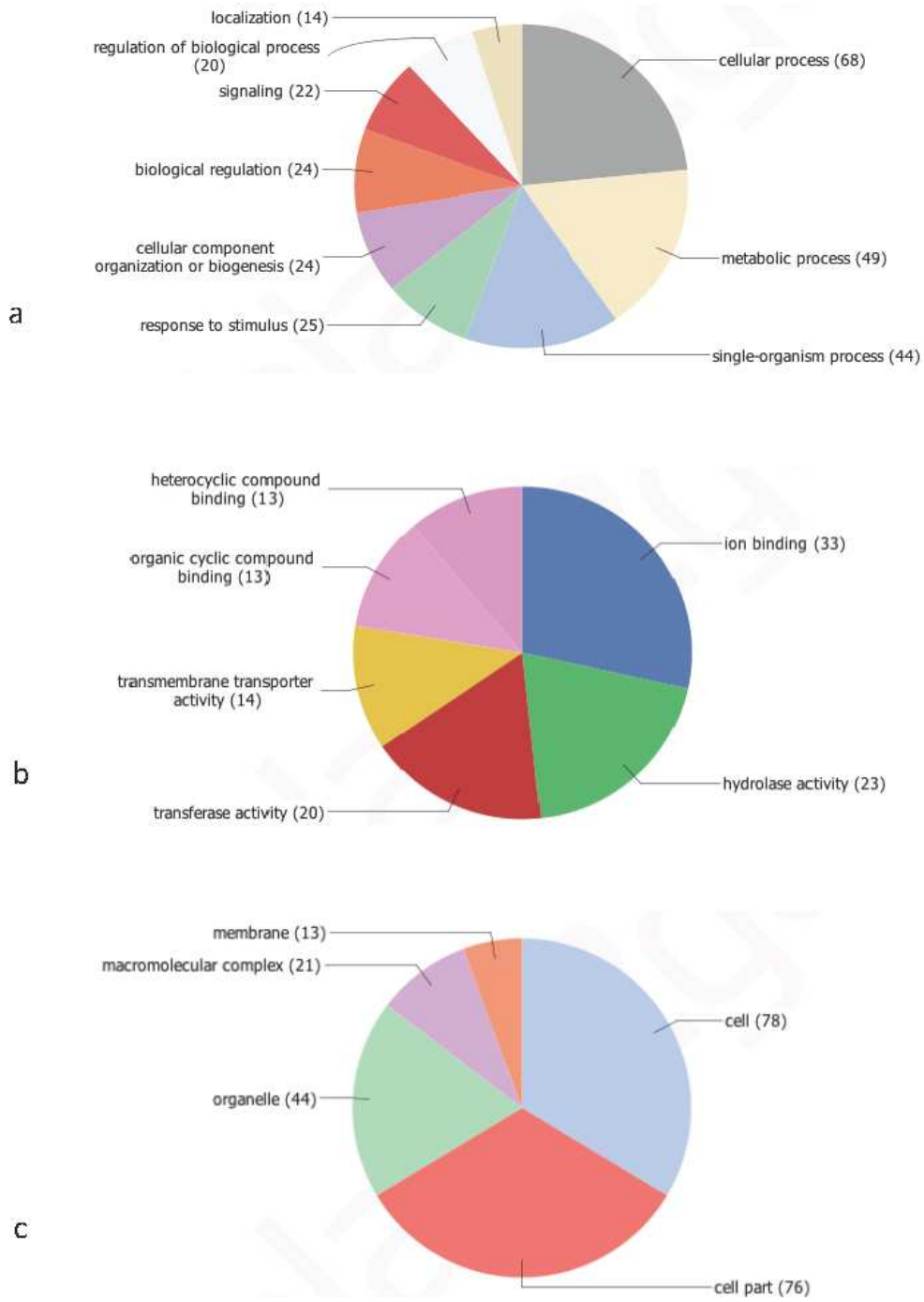

Figure 6: Gene content of the NMP region in Daphnia magna using the Gene Ontology (GO) annotations. Results are shown for (a) biological processes, (b) molecular functions and (c) cellular components.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 1: Fisher's Exact test P-values for all pairs of markers considered in (a) the NMPxMP_1 cross and (b) the NMPxMP_2 cross.

(a)	**NMPxMP_1 versus MPxMP**	**P value**	**% recombinant NMPXMP_1**	**% recombinant MPxMP**
	nmp - dm_scf02995_94235	1.000	1.515	1.905
	dm_scf02995_94235 - dm_scf00966_72719	0.300	4.545	0.952
	dm_scf00966_72719 - dm_scf00966_347201	0.655	12.121	15.238
	dm_scf00966_347201 - dm_scf00243_463094	<0.001	4.545	30.476
	dm_scf00243_463094 - dm_scf00243_208642	0.070	4.545	13.333
	dm_scf00243_208642 - dm_scf00494_79695	0.016	13.636	30.476
	dm_scf00494_7969 - dm_scf02066_483524	1.000	12.121	13.333
(b)	**NMPxMP_2 versus MPxMP**	**P value**	**% recombinant NMPxMP_2**	**% recombinant MPxMP**
	dm_scf00642_853353 - dm_scf02121_20555	0.115	3.704	0
	dm_scf02121_20555 - dm_scf01654_155597	1.000	7.404	7.692
	dm_scf01654_155597 - **nmp**	0.015	7.404	23.077
	nmp- dm_scf00243_208642	0.474	27.778	34.615
	dm_scf00243_208642 - dm_scf02066_483524	0.016	25.926	46.154

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 2: List of scaffolds containing SNPs significantly associated to NMP (Chi square test; P<10^{-5}). LG: linkage group; Size: total size of the scaffold (in basepair); Nb. SNPs: number of associated SNPs on the scaffold; SNPs position: position of the SNP on the scaffold; P value: resulting Chi square P value.

LG	Scaffold	Size (bp)	Position (cM)	Nb.SNPs	SNPs position (bp)	P values
1	scf00512	3718170	188.7	2	1898250; 1898257	2.04E-17; 1.48E-10
1	scf00205	87516	210.8	3	7803; 7820; 7827	1.63E-13; 1.63E-13; 1.63E-13
2	scf02190	2111488	58.2	1	49642	1.46E-09
3	scf02000	31222	72.3	1	4181	5.10E-09
3	scf00848	142519	83.9	2	34647; 68264	5.62E-06; 1.67E-13
3	scf02569	397658	87.9 - 88.8	9	384651; 381715; 268729; 232532; 193880; 80640; 80555; 77438; 77432	1.25E-06; 5.24E-16; 1.99E-14; 8.42E-16; 1.04E-11; 1.35E-12; 8.62E-10; 1.29E-09; 5.30E-12
3	scf03156	106208	88.8	4	4433; 50649; 50655; 79120	2.30E-16; 1.20E-06; 1.20E-06; 5.14E-07
3	scf00027	84679	88.8	4	34526; 29953; 26878; 12813	1.11E-08; 7.27E-12; 4.61E-06; 1.84E-14
3	scf02723	46460	90.8	5	2032; 2194; 13416; 13482	8.44E-14; 2.04E-17;2.30E-16; 4.78E-15
3	scf01727	97879	90.8	1	6345	5.62E-08
3	scf02895	190495	90.8	2	190140; 190147	3.76E-10; 3.76E-10
3	scf01943	93640	90.8	2	13984; 67370	5.79E-08; 2.42E-10
3	scf02003	324342	90.8	1	27223	9.30E-11
3	scf03194	167565	90.8	1	124008	1.91E-07
3	scf00378	73438	93.7	1	22714	5.24E-16
3	scf00452	41028	93.7	2	34895; 34916	8.23E-15; 8.23E-15
3	scf02995	126150	93.7	1	108378	3.47E-09
3	scf00966	460511	95.7	1	174527	3.33E-11
4	scf00311	941766	95.0	1	669922	1.18E-14

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Table 3: Results of the Blast run showing the 14 candidate genes and their location on the *D. magna* genome. The table reports the start and end position of the gene on the scaffold it maps to, the size of the expected protein (number of amino-acid), the NCBI attributed gene name, the taxon with the best blast hit, the corresponding e-value, and the percentage of sequence similarity.

scaffold	Start position	End position	Size (aa)	NCBI name	Taxon	e value	% similarity
scf00027	2877	6078	316	Serine arginine-rich splicing factor 7	*Harpegnatos saltator*	4.0E-50	82.8
scf00848	96321	97283	136	Aldo-keto reductase family 1, member C4	*Riptortus pedestris*	4.1E-59	72.5
scf02003	35289	35935	136	Poly-U-binding splicing factor Half Pint	*Acyrthosiphon pisum*	5.9E-67	95.1
scf02003	213333	214454	115	Cytochrome P450 314 family	*Daphnia magna*	3.8E-46	88.3
scf02569	3227	4315	108	Zinc transporter zip11	*Tribolium castaneum*	3.9E-29	80.2
scf02569	9179	10907	300	Zinc transporter zip9	*Poecilia formosa*	2.1E-75	74.8
scf02569	35151	44725	606	SOX-9-like transcription factor	*Acromyrmex echinatior*	5.0E-48	89.8
scf02569	218892	220701	292	DnaJ homolog dnaj-5	*Acromyrmex echinatior*	4.4E-109	72.1
scf02569	334258	337000	462	Broad-complex	*Oncopeltus fasciatus*	8.3E-50	85.2
scf02569	340469	342584	281	Transformer 2	*Daphnia pulex*	2.9E-119	88.7
scf02569	76814	79370	558	Protein SPT2 homolog	*Acyrthosiphon pisum*	2,00E-33	63.9
scf02569	228772	229714	158	Histone deacetylase complex subunit sap18	*Metaseiulus occidentalis*	1.1E-55	82.1
scf02723	1124	6033	287	Epidermal growth factor receptor kinase	*Zootermopsis nevadensis*	1.7E-30	71.2
scf03156	4200	8559	794	Lysine-specific histone demethylase 1A	*Stegodyphus mimosarum*	0.0	85.3

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Material

S1 Table: microsatellites.xls

Excel file giving the list of the 81 microsatellite markers tested in this study.

S2 Text. RAD-sequencing and SNP calling protocol.

We used the RAD-sequencing protocol developed by Etter et al. (2011) with a few modifications. The 72 individuals were divided in 2 libraries. Prior to DNA extraction, individuals were treated for 72 hours with three antibiotics (Streptomycin, Tetracyclin, Ampicilin) at a concentration of 50 mg/L of each antibiotic and fed with microscopic glass beads (Sephadex Small by Sigma Aldrich: 50 ţm diameter) at a concentration to 0.5g/100 mL. The aim of this treatment was to minimize contaminant DNA (i.e., bacterial DNA or algal DNA) in in the gut and on the surface of the carapace. Genomic DNA was extracted using the Qiagen Blood and Tissue kit following manufacturer's instructions and digested with PstI (New England Biolabs). Digested DNA was barcoded with individual-specific P1 adapters and pooled to create a library containing 2100ng DNA. The pooled library was sheared on a Bioruptor using 2 times 3 cycles (1 cycle 30 seconds ON, 1 minute OFF), and fragments between 300 and 500bp were selected through agarose gel electrophoresis. DNA fragments were blunted and a P2 adapter was ligated. The library was amplified through PCR (30 seconds at 98řC, followed by 18 cycles of 10 sec. at 98řC, 30 sec. at 65řC and 30 sec. at 72řC; a final elongation step was performed at 72řC for 5 min.). A final electrophoresis was performed to select and purify fragments between 350 and 600bp. Each library were sequenced on a single lane of an Illumina HiSeq 2000, using single-end 100 cycle sequencing by the Quantitative Genomics Facility service of the Department of Biosystem Science and Engineering (D-BSSE, ETH), Basel, Switzerland. The quality of the raw sequencing reads (library-wide and per-base) was assessed with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and reads were checked for barcode integrity, absence of adapter sequences within the reads, and integrity of the PstI cut site. The reads were sorted individually by barcode and filtered to remove reads with uncalled bases and an overall base quality score of less than 24. Reads were subsequently aligned to the *Daphnia magna* genome (V2.4; *Daphnia* Genomic Consortium, WFleaBase) using BWA v.0.7.10 (Li and Durbin 2009). Reads that did not map to the reference genome or that mapped to more than one

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

place were discarded. The sucessfully mapped reads were filtered according to mapping quality (end-to-end mapping with a mapping quality score of at least 25, no more than eight high quality substitutions).

Assignement of reads to RAD loci (defined by unique 95 bp locations on the reference genome) and genotype calling was performed in Stacks V1.19 with a bounded SNP model in pstacks (–bound_high of 0.04, according to the base call error rate provided by the sequencing facility) and allowing a maximum of two high frequency haplotypes (i.e. alleles) per locus per individual. Loci with more than two high frequency alleles were excluded because of a too high risk of falsely mapping paralogous reads to a single locus. Cstacks and sstacks were operated with default settings and with the -g option to use genomic location as method to group reads. The distribution of the minor allele frequency indicated that heterozygous loci usually had a minor allele frequency ranging between 0.2 and 0.5 within an individual. We thus fixed the max_het_seq parameter to 0.2 in the program genotypes. As such, potentially heterozygous genotypes with a minor allele frequency of between 0.05 (default homozygote cut-off) and 0.2 were considered ambiguous and were scored as missing in the results. Loci were also filtered according to sequencing depth: Loci with less than 20 reads were discarded (to reduce uncertainty in genotype calls) as were reads with a more than five times higher depth than the average depth across individuals (to reduce the risk of including repetitive elements).

After final genotype calling, loci were mapped to the Daphnia magna genetic map v.3.0 (Duki et al, submitted). This was done by extracting for each RAD locus the linkage group and cM position of the nearest map-markers on the same scaffold and, if needed, by extrapolating the cM position of the RAD locus by linear extrapolation between the two nearest map-markers.

References

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA. 2011. Local de novo assembly of RAD paired-end contigs using short sequencing reads. PLoS ONE,6:755,e18561.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S3 Table: snp_data.xls

Excel file containing the list of SNPs obtained from the RAD-sequencing pannel for all individuals. Information listed: Linkage Group; order of SNP on the genetic map; CentiMorgan position on the genetic map; basepair position on the physical map; MegaBase position; scaffold where the snp maps to; position of the SNP on the scaffold (in bp); major allele; minor allele. Additionnal information: Chi square value; associated P-value; mutation location and type (intergenomic, intron, 5' / 3' UTR; aminoacid substitution); synonymous or non synonymous mutation; gene impacted.

S4 Script: association.R

R script for performing the association analysis at the genome wide level.

S5 Text. Protocol for the physical ordering of scaffold in the NMP associated non-recombining region of LG3.

The region controlling the NMP phenotype maps to a low recombination. Because of absence of recombination in the genetic map data, the relative position and orientation of the scaffolds that have been mapped to this region are not resolved in the genetic map (i.e., several entire scaffolds mapped to the exact same cM position across all markers on these scaffolds). Hence no physical order of the scaffolds in the region can be obtained from the genetic map. This is problematic for genome-wide association studies and fine mapping of the NMP locus, especially for determining whether the NMP phenotype maps to one or multiple specific sub-regions. To obtain a potential physical order of scaffold in the region, we performed linkage disequilibrium (LD) mapping, which uses data on LD from a single population and therefore can make use of historical recombination events present in the data. LD can be measured by estimating the correlation of the allelic composition between two loci (r^2). If little historical recombination occurred between two loci, alleles at one locus should tend to co-occur with specific alleles at the other locus (i.e., the correlation should be high).

Loci that show high divergence between MP and NMP phenotypes are expected to have high LD,

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

if LD estimates are based on a mix of MP and NMP individuals. Hence, these loci would also tend to group closely together in LD mapping. Thus, in order to avoid a circular argument we based the physical ordering using LD mapping only on the 54 MP individuals sampled from the MOS population. We performed LD mapping on a somewhat larger region of LG3, between 85cM and 95cM, in order to also include SNPs just outside the NMP linked region. The MOS dataset contains SNPs on 30 mapped scaffolds in this region. Among these, there are three groups of scaffolds for which the relative position and orientation could not be resolved with the genetic map: two scaffolds at position 88.8 cM, 17 scaffolds at 90.8 cM and 4 scaffolds at 93.7 cM. For physical ordering of the scaffolds within each of these groups, we used only biallelic SNPs with a minor allele frequency over 0.1 and less than 33% missing genotypes among the MP individuals. We first calculated pairwise r^2 values for each pair of SNPs with MCLD (Zaykin 2008), which uses genotypic data irrespective of the haplotypic phase. To position the scaffolds relative to each other and to orient them, we averaged the r^2 values of the three terminal SNPs on either side of each scaffold and created a matrix of pairwise average r^2 values between each pair of scaffold extremities for each of the groups separately, also including the adjacent extremities of the two scaffolds that mapped immediately outside that cM position. When more than two scaffolds had to be ordered in a group, we perform a hierarchical clustering analysis to identify starting clusters (highly linked scaffold extremities). The hierarchical clustering was performed in R using the hclust function of the R core package *stats*. Scaffolds were then added one by one to the starting clusters, as indicated by the dendogram, and positioned next to the scaffold and oriented in a way that maximized the average r^2 values between adjacent scaffold extremities.

S6 Table: LD_nmp_region.xls

Excel document containing calculated r^2 values for the SNPs present in the NMP non recombining region.

S7 Table: phased_haplotypes.xls

Excel document containing the list of the raw and the corrected haplotypes phased in this study, along with the SNP coordinates for each position.

bioRxiv preprint doi: https://doi.org/10.1101/064311; this version posted July 27, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

S8 Text: nmp_region_gene_content.fasta

FASTA formatted document listing the 283 genes used in the analysis.

S9 Figure:

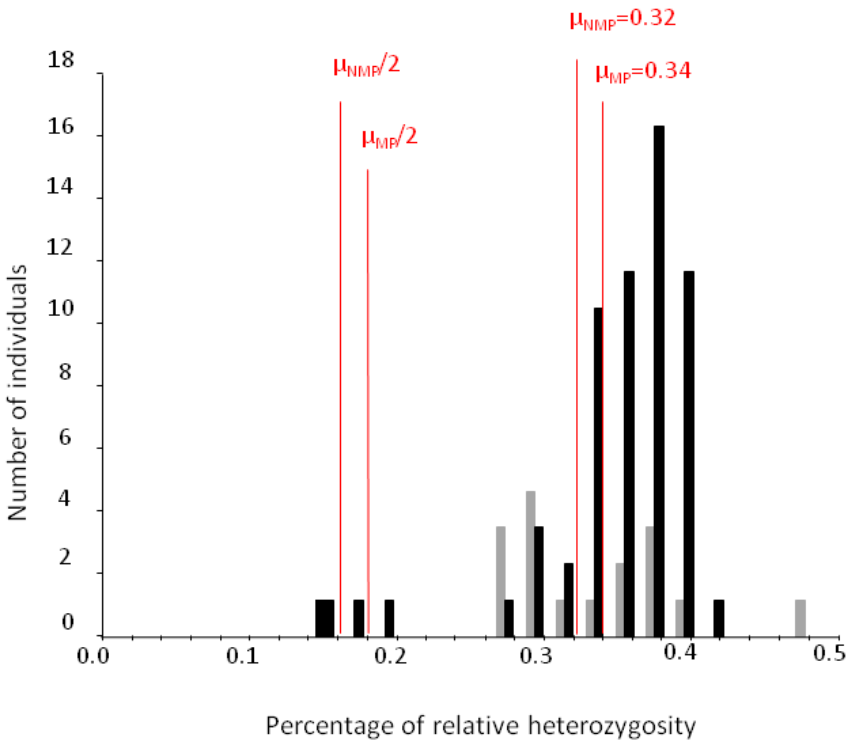

Figure S9: Distribution of the percentage of relative heterozygosity in MP (black) and NMP (grey) individuals. Calculations were performed without LG3, as this chromosome shows a higher heterozygosity in NMP individuals.

Raw genomic data:

The FASTQ files of all the individuals used in this study will be available on the SRA database upon acceptance of the manuscript.