Analysis of Copy Number Variation in the *Abp* Gene Regions of Two House Mouse Subspecies Suggests Divergence during the Gene Family Expansions

Željka Pezer[1,3,†], Amanda G. Chung[2,†], Robert C. Karn[2], and Christina M. Laukaitis[2,*]

[1]Max Planck Institute for Evolutionary Biology, Plön, Germany

[2]Department of Medicine, College of Medicine, University of Arizona

[3]Present address: Ruđer Bošković Institute, Zagreb, Croatia

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: cmlaukai@email.arizona.edu.

Accepted: May 26, 2017

Abstract

The *Androgen-binding protein* (*Abp*) gene region of the mouse genome contains 64 genes, some encoding pheromones that influence assortative mating between mice from different subspecies. Using CNVnator and quantitative PCR, we explored copy number variation in this gene family in natural populations of *Mus musculus domesticus* (*Mmd*) and *Mus musculus musculus* (*Mmm*), two subspecies of house mice that form a narrow hybrid zone in Central Europe. We found that copy number variation in the center of the *Abp* gene region is very common in wild *Mmd*, primarily representing the presence/absence of the final duplications described for the mouse genome. Clustering of *Mmd* individuals based on this variation did not reflect their geographical origin, suggesting no population divergence in the *Abp* gene cluster. However, copy number variation patterns differ substantially between *Mmd* and other mouse taxa. Large blocks of *Abp* genes are absent in *Mmm*, *Mus musculus castaneus* and an outgroup, *Mus spretus*, although with differences in variation and breakpoint locations. Our analysis calls into question the reliance on a reference genome for interpreting the detailed organization of genes in taxa more distant from the *Mmd* reference genome. The polymorphic nature of the gene family expansion in all four taxa suggests that the number of *Abp* genes, especially in the central gene region, is not critical to the survival and reproduction of the mouse. However, *Abp* haplotypes of variable length may serve as a source of raw genetic material for new signals influencing reproductive communication and thus speciation of mice.

Key words: androgen-binding protein, copy number variation, gene family expansion, divergence, hybrid zone, *Mus musculus* subspecies.

Introduction

Structural variants are genomic segments >50 bp (Alkan et al. 2011) that are deleted, duplicated, inserted, inverted or translocated in the genome (Conrad and Hurles 2007). Segmental duplications (SDs), aka low copy repeats (LCRs), are large DNA sequences ranging from 1 to 400 kb in length that occur in more than one site in a genome and that characteristically share >90% sequence identity (reviewed in Eichler 2001). SDs are prone to cause copy number (CN) variation via nonallelic homologous recombination (NAHR; Shaffer and Lupski 2000; Stankiewicz and Lupski 2002a; Cooper et al. 2007), and copy number variants (CNVs) provide the raw material for gene family expansion and diversification that is an important evolutionary force (Perry 2008). Certain families of retrotransposons have been found to be enriched at the junctions of SDs in the human, bovine, mouse, rat, and grapevine genomes (Babcock et al. 2003; Bailey et al. 2003, 2004; Tuzun et al. 2004; Zhou and Mishra 2005; She et al. 2008; Liu et al. 2009; Giannuzzi et al. 2011). Alu (SINE) elements are the predominant repeat sequences at junctions in primates where they contribute to the interspersed pattern of gene duplication characterizing the human genome (Babcock et al. 2003; Bailey et al. 2003). In the cattle and rodent genomes where tandem duplication predominates, junctions are enriched for LINE and LTR elements (Bailey et al. 2003, 2004; Tuzun et al. 2004; Zhou and Mishra 2005; She et al. 2008; Liu et al. 2009).

CNVs account for as much as 13% of some mammalian genomes (Stankiewicz and Lupski 2010) and genes found in CNVs can influence human disease as well as the phenotype of nonhuman mammals (Freeman et al. 2006; Radke and Lee 2015). Two functionally separate classes of CNV genes are emerging, depending on whether or not the genes and CNVs are associated with SDs (Sjodin and Jakobsson 2012). CNVs in SD-associated genes are more likely to recur or to be maintained by NAHR in order to reach polymorphic levels (>1%) in populations. Moreover, SD-associated genes are enriched in functions involved in environmental response (e.g., immune response, chemosensation, toxin metabolism, and reproduction) and some may be under positive selection as shown for the amylase gene in human populations (Perry et al. 2007). There are numerous examples of nonpathogenic phenotypic changes produced by structural variations that may have adaptive potential in humans and other mammals (Radke and Lee 2015). Additional forces, including mutation, recombination and demographic history also influence the occurrence and impact of CNVs and structural variations have even greater potential to produce adaptive phenotypes than point mutations (Redon et al. 2006; Sjodin and Jakobsson 2012).

A recent study of genome-wide CN variation in natural populations of the house mouse revealed divergent patterns of genic CNVs and identified genes with major population-specific expansions (Pezer et al. 2015). These data show a hotspot of CNVs in the large (3 Mb) *Androgen-binding protein* (*Abp*) gene region on the proximal end of chromosome 7 in the mouse genome, consistent with extensive volatility in the mouse *Abp* gene family reported earlier (Karn and Laukaitis 2009). This gene family has served as an example of NAHR generating CNVs and SDs have been identified in the central region where significant NAHR has duplicated large blocks of genes (Karn and Laukaitis 2009; Janoušek et al. 2013).

ABPs are dimeric proteins produced primarily by glands of the face and neck of the mouse (Laukaitis et al. 2005; Karn et al. 2014) and some have the ability to bind male sex steroid hormones with a clear specificity (Dlouhy and Karn 1983; Karn 1998; Karn and Clements 1999), hence their name. ABP dimers are composed of an alpha subunit, encoded by *Abpa*, connected by disulfide bridging to a betagamma subunit, encoded by *Abpbg*. Because there are 30 *Abpa* and 34 *Abpbg* genes in the mouse genome, these have number suffixes (e.g., *Abpa27*; Laukaitis et al. 2008; reviewed in Laukaitis and Karn 2012 and Karn and Laukaitis 2014). Most of these exist in <*Abpa-Abpbg*> pairs (arrows point in the 3′ direction), called "modules" because they appear to be the original unit of duplication (Karn and Laukaitis 2009). Salivary ABPs mediate assortative mate selection based on subspecies recognition that potentially limits gene exchange between subspecies where they meet (Laukaitis et al. 1997; Talley et al. 2001; Laukaitis and Karn 2012) and there is evidence that ABP constitutes a system of incipient reinforcement across the European hybrid zone where house mouse subspecies make secondary contact (Bímová Vošlajerová et al. 2011). These findings suggest that they are "environmental genes" (Nguyen et al. 2008) because they are involved in reproduction. This category of genes constitutes ~10% of the total mammalian gene complement, which possess functions that differ from those of the conserved gene set and tend to vary widely in content between individuals, as shown for humans, inbred mouse strains and fish (Cutler et al. 2007; Schrider and Hahn 2010; Chain et al. 2014). The genes in this volatile portion of the genome are subject to frequent duplication, deletion and pseudogene formation (Waterston et al. 2002; Gibbs et al. 2004; Karn and Laukaitis 2009; Lander 2011) because of relaxed purifying selection (Nguyen et al. 2008; Pezer et al. 2015) or positive selection (Perry et al. 2007; Sjodin and Jakobsson 2012) on CNVs in SDs.

The older flanking paralogs in the recently and extensively expanded mouse *Abp* gene family are more likely to be expressed (12/15; 80%), with a clear partition between salivary gland and lacrimal expression in C57BL/6, the genome mouse (Karn et al. 2014). In contrast, the last two duplications that occurred by NAHR produced half of the 64 total *Abp* genes from the original 14 genes in the central part of the *Abp* region (Janoušek et al. 2013). Surprisingly, none of the 18 new genes created by those last two duplications are expressed in either tissue in C57BL/6 (Karn et al. 2014). One reason may be that half of the 14 genes in the progenitor segment appear to have been pseudogenes before the last two duplications occurred. The pseudogenes that are their products represent over half (56%) of the 32 total genes in the current central *Abp* region of the mouse genome. These observations raise the question: Why are there 64 *Abp* genes when only 15 are expressed in salivary and lacrimal glands? This question was one of our motivations to examine CN variation in the *Abp* gene region of wild mouse populations with CNVnator analysis of next-generation sequencing (NGS) reads.

In this study, we focused on five main questions: 1) How common is CN variation in the *Abp* gene region in wild mice? 2) Which *Abp* genes are most commonly affected and how does CN variation occur in the *Abp* region? 3) What are the organizational differences that result from CN variation of the *Abp* cluster between *Mus musculus domesticus* (*Mmd*) and *M. m. musculus* (*Mmm*), the two subspecies of house mice that make contact along a narrow hybrid zone in Central Europe? 4) How is this gene family organized in the sister subspecies *Mus musculus castaneus* (*Mmc*) and an outgroup, *Mus spretus* (*Ms*)? and 5) How does reliance on a reference genome influence the appearance of CNVs in these related taxa? We also analyzed CNVs in a number of classical inbred strains with the goal of determining the extent to which the genomes of those strains are representative of the genomes of wild mice. Finally, we analyzed sequences at breakpoints of CNVs in the central region in order to explain the possible mechanism(s) causing CN variation.

The answers to these questions were surprising. They suggest that the size of the *Abp* gene region in wild mice is

polymorphic, with wild *Mmc* mice surviving without 28% or more of the central region genes. *Mmm*, *Mmd*, and *Ms* mice may have even smaller *Abp* gene regions. Our data further suggest that the divergence of the two subspecies, *Mmd* and *Mmm*, preceded the last duplication events in the *Mmd Abp* gene regions. Consequently the architecture of the central gene region, appears to be quite different in the two, as well as in *Mmc* and *Ms*. We discuss this in light of our recent proposal that runaway gene duplication may occur during some gene family expansions (Janoušek et al. 2016) and compare our findings with the work of others on gene functionalization within short evolutionary time spans.

Materials and Methods

DNA Samples

Genomic DNA was obtained from Jackson Laboratories (Bar Harbor, ME) for C57BL/6, five wild-derived *Mmd* strains (WSB, Lewes, Pera, Tirano, Zalende) and five wild-derived *Mmm* strains (CZECH/I, CZECH/II, PWD, PWK, and Skive). DNA samples from wild mice were provided by Diethard Tautz (Max Planck Institute for Evolutionary Biology, Plön, Germany). Sampling procedure and locations are described in detail in Harr et al. (2016). In brief, wild *Mmd* mice were trapped in Cologne-Bonn (Germany), Central Massif (France), Ahvaz (Iran) and Heligoland Island (Germany, North Sea), *Mmm* mice were trapped in Studenec (Czech Republic) and Almaty (Kazakhstan), *Mmc* mice were trapped in Himachal Pradesh (India), and *Ms* mice were trapped in Madrid (Spain). The sampling scheme for trapping wild mice ensured that individuals were not related to each other and that they represent local populations (Ihle et al. 2006; Harr et al. 2016). We excluded three DNA samples of *Mmd* (JR7-F1C, TP121B, TP17-2) from experimental validation due to low DNA concentration and high level of degradation.

Calling CNVs with CNVnator

We analyzed in detail CNV calls in the *Abp* region of 60 wild mice in total: 27 *Mmd*, 16 *Mmm*, nine *Mmc*, and eight *Ms* individuals. Whole genomes of these samples were paired-end sequenced using Illumina HiSeq2000 (see Harr et al. 2016 for details on sequencing and mapping) and CNVs were predicted by CNVnator software (Abyzov et al. 2011) relative to the mm10 reference genome assembly (Pezer et al. 2015; Harr et al. 2016). We also analyzed publically available data for five wild-derived mouse strains (WSB, LEWES, PWK, CAST, and ZALENDE) and seven classical inbred mouse strains (C57BL/10J, C3H/HeJ, BALB/cJ, A/J, DBA/1J, NZB/B1NJ, and NZO/HlLtJ). We downloaded mapped sequencing reads (version 1502) for these strains from the Mouse Genome Project FTP site (Keane et al. 2011). CNVnator was used to call CNVs in each individual mouse strain relative to the C57BL/6 mm10 assembly, with bin size optimally chosen such that the ratio of the average read depth signal and its standard deviation falls between 4 and 5. Bin size ranged from 70 to 250 bp. We used Circos (Krzywinski et al. 2009) to create individual heat map tracks based on CNVnator's results. Circular images were linearized in Adobe Photoshop using the *Polar Coordinates* filter. We used the Integrative Genomics Viewer (IGV; Robinson et al. 2011) to manually inspect and visualize NGS data at the sequence level.

Abp Paralog Primer Design and Validation

We designed primer sets for the following paralogs: *bg9/14/16, a10/15/17, bg31/32* (see supplementary table S1, Supplementary Material online for primer sequences and amplicon sizes). These primer sets were designed to include two or three paralogs because of the inability to distinguish between paralogs with highly similar or identical sequences. Paralogs *bg14* and *bg16* are identical, differing by 11 bases from *bg9*; *a15* and *a17* are identical and differ from *a10* by four bases; *bg31* and *bg32* are identical (see supplementary fig. S1, Supplementary Material online). We tested the primer sets on C57BL/6 and ten wild-derived genomic DNAs using conventional PCR to amplify paralogs and we visualized amplification bands on agarose gels. Correct paralog amplification was verified by Sanger sequencing at the University of Arizona Genomics Core.

In order to derive consensus sequences for each paralog in each of the wild individuals, we first called SNPs by using *samtools mpileup* and *bcftools view* followed by *vcfutils.pl* to create consensus of read alignments (Li et al. 2009; Danecek et al. 2011). The resulting sequences were aligned with the reference genome sequence using DNASIS Max (Hitachi) to identify any SNPs in the primer regions. Several wild *Mmd* and *Mmm* samples were also tested using conventional PCR and sequenced to ensure that the primer sets correctly amplify the specific sequences.

Detection of CNs by qPCR

Validated primers were subsequently used in qPCR experiments to determine CNs of the paralog sets. qPCR was performed using the Thermo Fisher SYBR green kit with experimental conditions as previously described (Karn et al. 2014) at the Genomics Shared Resource of the University of Arizona Cancer Center. The parameters for qPCR experiments were previously optimized and standard curves were created for each primer set to determine primer efficiency (Karn et al. 2014). The samples were tested in triplicate for each primer set, and Ct values were obtained from the ABI Prism 7000 software. We used the comparative Ct method to calculate the CN for each paralog set (see supplementary file S1, Supplementary Material online for detailed calculations). The fold-change for each sample was determined by comparison to a single-copy internal control gene (*Irp3*) in the reference genome (C57BL/6). For primer sets that include multiple

Fig. 1.—Copy number variants in the *Abp* region of 27 wild *Mmd* samples relative to the mouse genome (2011, mm10) shown as a heat map. The color legend to copy numbers is on the top left (deletions relative to the reference are in red, duplications in green and no change in CN in gray). The sample designation appears to the left of each horizontal track and the country of origin appears to the right (F = France; G = Germany; H = Heligoland; I = Iran). The positions of the *Abp* genes appear at the bottom: A blue vertical line is an *Abpa* gene and a red vertical line is an *Abpbg* gene. Arrows above the heat map point to the left and right breakpoints encompassing the ultimate duplication region.

paralogs, the calculated fold-change was further multiplied by the number of genes represented in the primer set; for example, a fold-change of *bg9/14/16* was multiplied by six (three diploid genes) to account for each paralog represented in the primer set.

Comparison of CNVnator and qPCR-Derived CNs

To ascertain how well the CNs predicted by CNVnator agree with those detected by qPCR, we calculated the Pearson product-moment correlation coefficient (*r*) and Lin's concordance correlation coefficient (CCC; Lin 1989) in *R*.

Analysis of Diagnostic Nucleotides

Only a handful of sites differentiate the three paralog pairs in the ultimate duplication (14-31-15/16-32-17) from their ancestor (9-29-10), as identified previously by sequence alignment (see supplementary fig. S1, Supplementary Material online; Karn and Laukaitis 2009; Janoušek et al. 2013). Given the high repetitiveness and sequence similarity along the *Abp* cluster, we sought to identify which of these positions can unequivocally indicate presence of paralog pairs in the central region. To achieve this, we extracted a 40-nucleotide sequence containing and surrounding each diagnostic nucleotide and searched for it in the reference genome with BLAT (Kent 2002). With the exception of one sequence, all had high similarity matches to other genomic regions (see supplementary fig. S2, Supplementary Material online). We considered a position to be reliable if it contained a diagnostic nucleotide only in alignments with *a15/17*, *bg14/16*, or *bg31/32*. In every *Mmd* and *Mmm* individual, we used *bam-readcount* (D. Larson et al., https://github.com/genome/bam-readcount) to count the number of reads supporting each base at each of these reliable positions.

Analysis of Breakpoints

Consensus sequences were obtained and aligned for samples with CNV state-changes suggesting absence of the ultimate duplication. NGS reads were viewed in the IGV browser and the number of reads falling on the midpoint nucleotide of the breakpoints identified by Janoušek et al. (2013) were manually counted and compared using a two-tailed *t*-test with the assumption of unequal variance.

Results

CN Variation in the *Abp* Gene Family of Wild *Mmd* Samples

One of the main objectives of this study was to determine the impact of retrotransposon-mediated NAHR events (Janoušek et al. 2013) on the *Abp* gene region by identifying and comparing CNVs in the *Abp* gene family of wild *M. musculus* samples. Using the C57BL/6 genome (2011, mm10) as a reference, we plotted CNVnator-predicted CNV calls in the *Abp* region as a heat map which helps to visualize differences in CN. We compared the calls from 27 wild *Mmd* mice (fig. 1) from four different localities: Heligoland Island, Iran, Germany and France (see supplementary table S2, Supplementary Material online). Large segments in the central region show

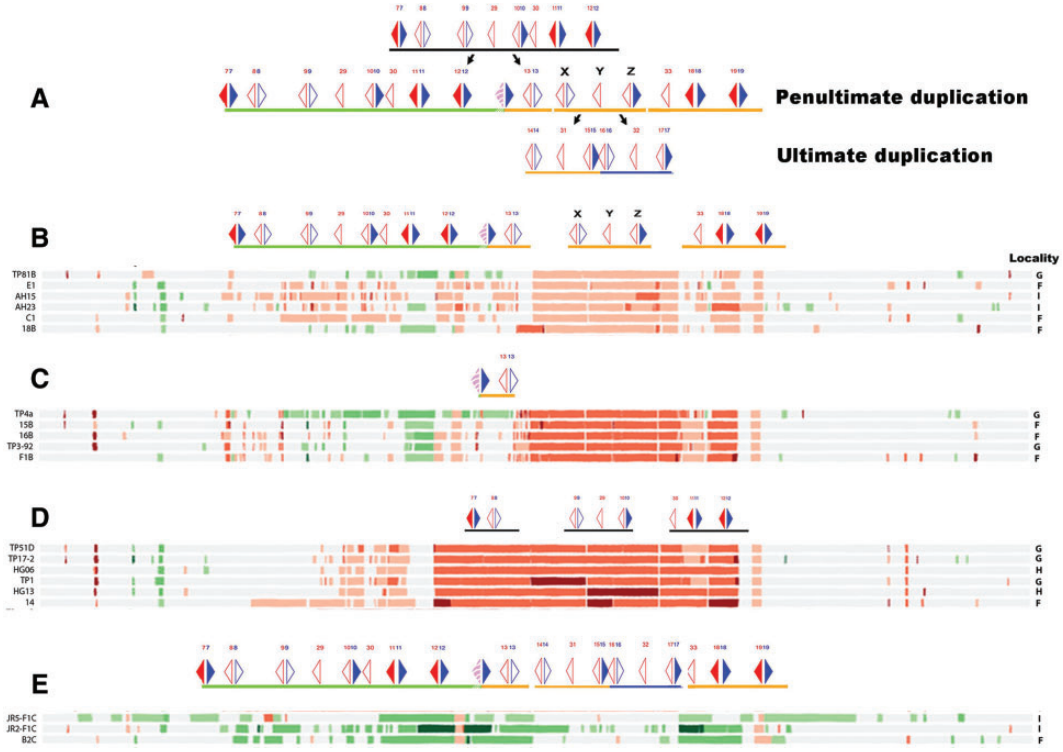

Fig. 2.—An interpretation of the CNVnator results based on the locations of the penultimate and ultimate duplications in the *Abp* gene region. (*A*) The last two duplications that led to a doubling of the size of the *Abp* gene region in the mouse genome (adapted with permission from Janoušek et al. 2013). Blue arrows are *Abpa* genes. Red arrows are *Abpbg* genes. Filled symbols indicate potentially expressed genes, whereas outlined symbols are putative pseudogenes. (*B*) Hypothetical haplotypes for six samples that lack the ultimate duplication. (*C*) Hypothetical haplotypes for five samples derived from a partial duplication involving only modules 7 and 8. (*D*) Hypothetical haplotypes for six samples that lack the penultimate duplication. (*E*) Hypothetical haplotypes for three samples with the complete set of duplications found in the mouse genome (grey) and some smaller additional duplications (green).

varying lengths of deficiencies and degrees of deletion (intensities of red color in the heatmap), indicating fewer gene copies relative to the reference genome (fig. 1). There is no clear correlation between localities of the *Mmd* mouse samples and size of CNVs and there is variability in CN as well as in the size of deficiencies within samples from the same geographical region (fig. 1). The exceptions are the three samples from Heligoland Island, all of which have the same size of gene region deficiency, but with different CN.

Figure 2*A* shows the last two duplications of the *Abp* gene region involving a larger penultimate duplication and a smaller ultimate one (Janoušek et al. 2013). Six wild *Mmd* samples (22%; TP81B, E1, AH15, AH23, C1, and 18B) have a deficiency that corresponds to the ultimate *Abp* gene duplication, involving paralogs <*bg14-a14*> *bg31* <*bg15-a15*> (hereinafter *14-31-15*) and <*bg16-a16*> *bg32* <*bg17-a17*> (hereinafter *16-32-17*; fig. 2*B*). The *Abp* regions of five samples (19%; TP4a, 15B, 16B, TP3-92, F1B) appear to have been formed by a much shorter version of the penultimate duplication, likely involving only modules 7 and 8. Alternatively this might have occurred by deletion of modules 14–19 from the penultimate duplication (fig. 2*C*). Six other samples (22%; TP51D, TP17-2, HG06, TP1, HG13, 14) are missing all genes in modules 12–19, suggesting the lack of the penultimate duplication (fig. 2*D*). Three samples (11%; JR5-F1C, JR2-F1C, B2C) have no evident absence of genetic material and may have duplicated material (fig. 2*E*), whereas seven samples (26%) have more complex deletion and duplication patterns. We focused our further analyses on three paralog sets, *bg9/14/16, a10/15/17* and *bg31/32*, due to the extent of CN variation among samples and because their involvement in the ultimate duplication caused them to be nearly or exactly identical in sequence (Laukaitis et al. 2008; Karn and Laukaitis 2009).

Comparison of *bg31/32* Paralog CN Data in Wild *Mmd* Mice

Because the members of each of these three paralog sets have >99% sequence identity (Laukaitis et al. 2008; Karn and Laukaitis 2009), it was not possible to design primer pairs to amplify individual paralogs. Instead, we used a single primer set for each of the groups *bg9*, *bg14*, and *bg16*; *a10*, *a15*, and *a17*; and *bg31* and *bg32*. Paralogs *bg31* and *bg32* are the products of duplications from *bg29* (Karn and Laukaitis 2009; Janoušek et al. 2013), but divergence at five nucleotide sites (see supplementary fig. S1, Supplementary Material online) allows production of a primer set which can

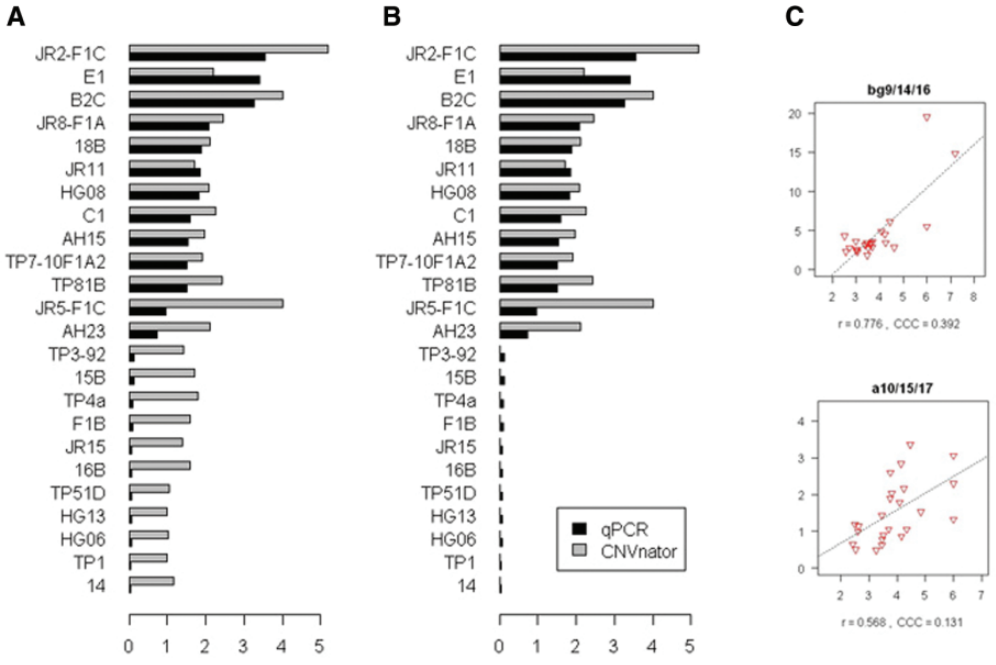

Fig. 3.—Comparison of CN values derived from CNVnator and qPCR. (*A*) and (*B*) A comparison of *bg31/32* qPCR and CNVnator data. The combined CN of the two paralogs is shown for 24 wild *Mmd* samples. Copy numbers determined bioinformatically and experimentally are juxtaposed to highlight the discrepancies between the two methods. (*A*) Values after the original calculations. (*B*) Adjusted values after correction based on the absence of a diagnostic nucleotide. (*C*) Correlation plots of CN estimates derived by CNVnator and qPCR for *bg9/14/16* and *a10/15/17*. Data for 24 samples are shown. qPCR values appear on the *y* axis and CNVnator values appear on the *x* axis. Two measures of correlation are shown below each graph (*r*, Pearson's correlation coefficient; CCC, Lin's concordant correlation coefficient). The linear model fit is indicated by a dashed line.

distinguish *bg31/32* from *bg29*. Thus bg*31/32* qPCR data are the most straightforward to interpret (fig. 3*A* and *B*), as they measure the combined CN of the two genes that fall in the centers of the *14-31-15* and *16-32-17* clusters. Red bars covering both *bg31* and *bg32* in figure 2C and *D* suggest the absence of both paralogs, and most such bars in that figure encompass genes from *bg14* to *a17*, indicating the absence of the ultimate duplication.

Figure 3*A* compares *bg31/32* CNs predicted by CNVnator to those obtained by qPCR for 24 *Mmd* samples. Eleven samples that were analyzed using qPCR showed absence of *bg31* and *bg32*, whereas CNVnator predicted presence of one to two genes (fig. 3*A*). We suspected that those samples lack the region of the ultimate *Abp* gene duplication (presented as red bars in the heatmap covering both *bg31* and *bg32* and extending at least from *bg14* to *a17* in HG13, HG06, TP1, JR15, 14, TP51D, TP3, F1_B, 15B, 16B, TP4A) and therefore we examined the read alignments that form the basis of calling CNVs. In regions of sparse read coverage, individuals with virtually no copies of *bg31* and *bg32* by qPCR also contained mapped reads with SNPs not found in these paralogs in the C57BL/6 genome sequence (see supplementary fig. S3, Supplementary Material online). Previous review of these and other "SNPs" called in the assembly process revealed that they are the diagnostic nucleotide positions differentiating paralogs *bg31/32* from *bg29* (Laukaitis et al. 2008), and upon which uniquely amplifying primers were designed (Karn et al. 2014; Chung et al. 2017). Thus, these are not truly SNPs, but reflect the presence of pseudoalleles (two or more genes that are distinct, closely linked loci). In the process of NGS read mapping, sequenced fragments originating from either *bg29* or *bg31/32* are randomly aligned to all three regions of the reference genome due to their high sequence similarity, leading to the mistaken impression that all three genes are present in the sequenced genome. To overcome this, we evaluated diagnostic nucleotides from qPCR to correct CN misassignments by CNVnator. The combined CN of *bg31/32* was set to zero in individuals not containing read evidence for diagnostic nucleotides in *bg31/32* (fig. 3*B*). After the correction, the correlation coefficient between CNs predicted by CNVnator and qPCR for *bg31/32* increased from 0.74 (*P*-value $= 3.5 \times 10^{-5}$) to 0.86 (*P*-value $= 8.2 \times 10^{-8}$) as calculated by Pearson's and from 0.53 to 0.80 as calculated by Lin's concordance.

Analysis of the Central Region of Wild *Mmd* by CNVnator and qPCR

The ultimate duplication created the *14-31-15* and *16-32-17* genomic segments in the C57BL/6 mouse strain (Karn and Laukaitis 2009; Janoušek et al. 2013). Accordingly, we would expect equal numbers of genes belonging to the same

segment, for example, if *bg31/32* genes are absent, we would also expect to find no copies of *bg14/16* and *a15/17*. Indeed, CNVnator predicts a similar CN state of genes within the same block, represented by the same heat map color of CNVs encompassing the entire *14-31-15* and *16-32-17* segments (fig. 2*B–D*). We verified this by qPCR using primers to detect the combined CN of *bg9/14/16* and the combined CN of *a10/15/17*. Comparison to the corresponding values calculated from CNVnator-predicted CNs revealed strong correlation in the case of *bg9/14/16* (Pearson's correlation coefficient; $r = 0.78$, P-value $= 8.3 \times 10^{-6}$) and moderate correlation for *a10/15/17* ($r = 0.57$, P-value $= 0.0038$; fig. 3*C*).

The relatively high correlation indicates the overall credibility of CNVnator predictions. However, similar to *bg31/32*, the sheer abundance of ambiguously mapped reads across these regions prevents the determination of true absolute CNs. Once again, we used diagnostic nucleotides to refine CN predictions where possible (i.e., in cases of complete deletions). Given that the region corresponding to the ultimate duplication in the reference genome is rich in highly similar repetitive sequences, such as SDs, we determined which diagnostic nucleotides (i.e., differentiating *bg14/16* from *bg9*, *a15/17* from *a10* and *bg31/32* from *bg29*) unambiguously indicate the presence of the *Abp* gene in question and not some other genomic region(s). There was one such reliable position for *a15/17* (position 524 in the alignment of paralogous genes; see supplementary fig. S1, Supplementary Material online), one for *bg31/32* (position 634; the same as used for qPCR primer design) and five for *bg14/16* (positions 219, 1299, 1494, 1521 and 1539). We next screened genomic read alignments of every sample for each of these positions to get the number of reads supporting the diagnostic base only. The results summarized in supplementary table S3, Supplementary Material online (see also supplementary File S2, Supplementary Material online) suggest complete absence of the genes resulting from the ultimate duplication in 12 individuals and are in agreement with qPCR results where the same individuals show zero copies of *bg31/32* (fig. 3*A*).

Other recently duplicated and thus identical *Abp* genes (e.g., modules 11 and 18) may have allowed misassignment of NGS reads. This likely explains some of the small nonuniform CN blocks, such as the light red block over *bg33* and module 18 in TP51D, TP17-2 and TP1, with corresponding light red blocks over *bg30* and module 11.

Identification of the Breakpoint in the Ultimate *Abp* Duplication in Wild *Mmd* Genomes

We identified putative breakpoints as CN state changes, that is, changes in read depth, and focused on the ultimate duplication because it is expected to contain the most complete retrotransposon sequences in the *Abp* gene region (Janoušek et al. 2013). Figure 1 shows the actual breakpoints, defined as recombination between strands, 2,770 bp downstream from

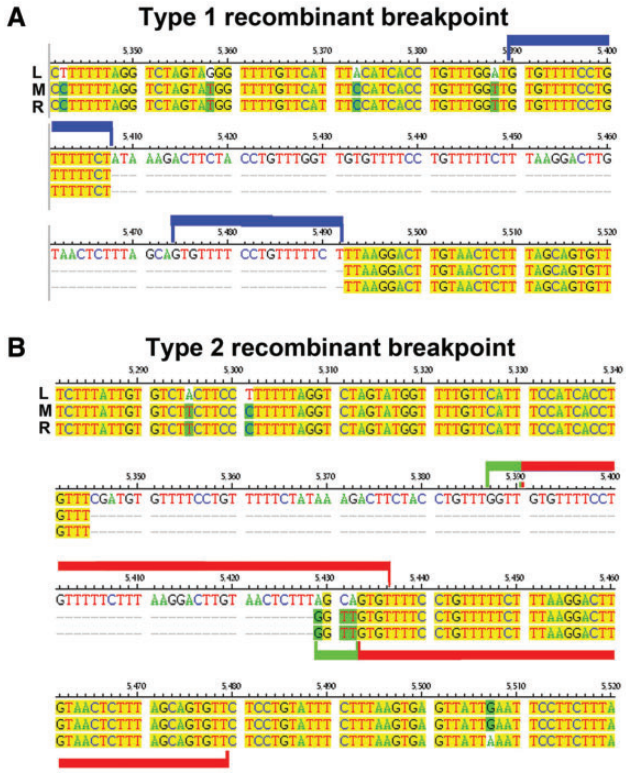

Fig. 4.—Two different types of breakpoints for the most recent *Abp* gene duplication. An alignment of the three sequences that surround the gap in the L1Md_T retrotransposon described by Janoušek et al. (2013) showing two different types of breakpoint configurations: (*A*) Type 1 is the most common type, found in seven samples, with a pair of 18 nucleotide repeats (blue brackets). (*B*) Type 2, found in three samples (red brackets with green ends), was the same type of breakpoint shown in figure 3 of Janoušek et al. (2013).

the left breakpoint (brptL) but only 50 bp downstream from the right breakpoint (brptR), consistent with the exchange between the middle (M) and right-hand (R) strands (see fig. 3*A* in Janoušek et al. 2013). In all affected *Mmd* samples, these shared the location of that breakpoint in the L1Md_T retrotransposon (Janoušek et al. 2013). Moreover, they are characterized by the same set of repeats (Janoušek et al. 2013) but varied in the organization of the residual nucleotides within the breaks (see below). This suggests recurrence of events at this location in the L1Md_T insertion.

Figure 4 shows two different types of breakpoint configuration in ten of the *Mmd* samples. We designated the more frequent (7/10; 70%) as Type 1 (samples TP81B, JR2-F1C, JR5-F1C, 18B, AH15, C1, JR8). It has a simple configuration, with an 18 bp sequence on the left flank of bg14 (strand L) that is repeated 76 bp later. Type 2 was less frequent (3/10; 30%; B2C, AH23, E1) and has the more complex arrangement described by (Janoušek et al. 2013). There was no obvious correlation of breakpoint type with heat map pattern or geography and both types of breakpoints occur in samples with different deletion patterns.

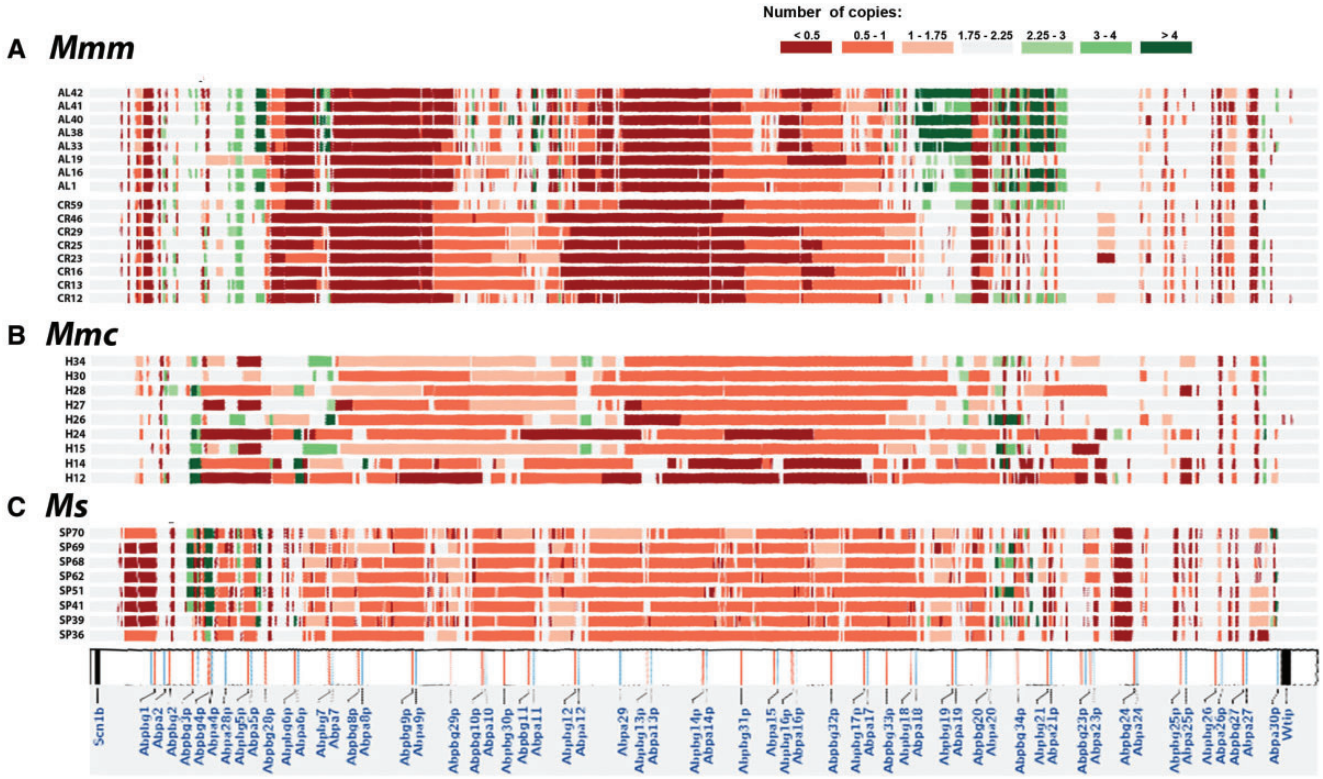

Fig. 5.—CNVs in the *Abp* region of wild *Mmm, Mmc*, and *Ms* mice. Copy numbers were calculated by CNVnator relative to the mouse genome (2011, mm10) and are shown as heat maps. A color legend to the copy numbers is on the top right (deletions relative to the reference are presented in red, duplications in green and no change in CN in gray). Sample designations are on the left of each track. (*A*) A heat map of CNVnator data in the *Abp* region for 16 *Mmm* samples. Eight individuals are from Kazahstan (AL) and eight are from the Czech Republic (CR). (*B*) A heat map of CNVnator data for CNVs in the *Abp* region for nine *Mmc* samples from India. (*C*) A heat map of CNVnator data for CNVs in the *Abp* region for eight *Ms* samples from Spain.

Six samples shaded light red over the modules 14–17 (figs. 1 and 2*B*) are missing the ultimate duplication seen in the mouse genome and therefore have only a single copy of the five-gene segment (*XYZ*) in figure 2*B*, rather than its ten gene duplication product (Janoušek et al. 2013). Samples TP51D through 14 are missing the penultimate duplication (fig. 2*D*). Therefore, we assessed read depth over the left, middle and right L1Md_T sequences we found in both groups. Supplementary table S4, Supplementary Material online (see also supplementary File S2, Supplementary Material online) shows the mean number of reads over these regions, all of which have a mapping quality of 0 to 1. There are significantly more reads in the first group (no ultimate duplication; fig. 2*B*) than in the second group (no penultimate duplication; fig. 2*D*; data in supplementary table S4, Supplementary Material online; *P*-value $= 2.3 \times 10^{-7}$) consistent with missing L1Md_T breakpoints in the genomes completely lacking the ultimate *Abp* duplication seen in the mouse genome.

Structure of the *Abp* region in *Mmm*, *Mmc*, and *Ms*

The house mouse, *M. musculus*, comprises at least three relatively distinct parapatric gene pools given subspecies status by some and full species status by others (Boursot et al. 1993; Sage et al. 1993). These three subspecies have been defined at the periphery of the distribution range of the species: *Mmd* from the near East to western Europe through the Mediterranean Basin, *Mmm* from eastern Europe to northern China, north of the Himalayas, and *Mmc* in South-East Asia, Malaysia and southern China, respectively. Most inbred laboratory strains are overwhelmingly *Mmd* in origin with very small contributions of *Mmm* and *Mmc* (<10%; Yang et al. 2011). The ancestors of the *Mmd* subspecies made secondary contact with those of the *Mmm* subspecies, creating a house mouse hybrid zone that extends across central Danish Jutland and south through Europe to the Black Sea (Boursot 1993; Sage et al. 1993). Thus, *Mmm* is the only subspecies that makes extensive contact with *Mmd* in the wild and it is therefore of special interest to compare the structure of the *Abp* cluster between these two subspecies.

The *Mmm* CNV patterns differ strikingly from *Mmd* patterns and the overall structure of the *Abp* region is more uniform between samples, suggesting deficiencies of large blocks encompassing multiple genes in all samples (areas with CN < 1 represented by red and dark red in fig. 5*A*). The combined analysis of diagnostic nucleotides and qPCR

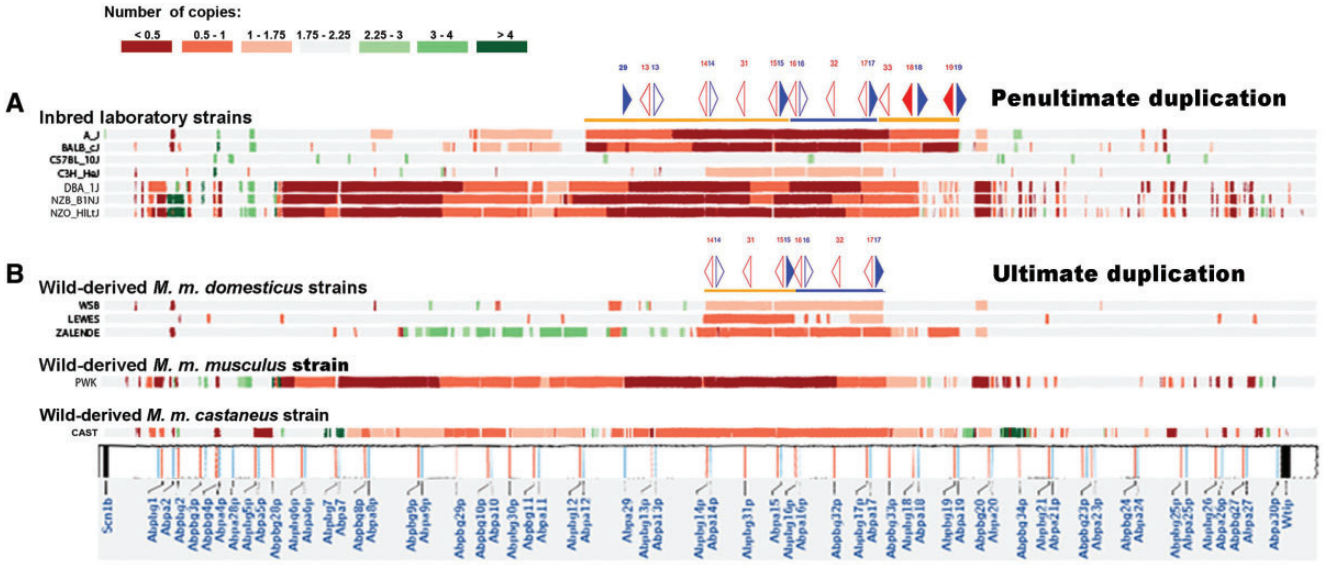

Fig. 6.—CNVs in the *Abp* region of inbred laboratory and wild-derived inbred mice. Copy numbers were calculated by CNVnator relative to the mouse genome (2011, mm10) and shown as heat map. A color legend to the copy numbers is on the top left (deletions relative to the reference are presented in red, duplications in green and no change in CN in gray). (*A*) A heat map of CNVnator data for six laboratory strains with the gene block produced by the penultimate duplication reiterated for reference. (*B*) A heat map of CNVnator data for three wild-derived *Mmd*, one wild-derived *Mmm* and one wild-derived *Mmc* strain with the gene block produced by the ultimate duplication shown for reference.

results suggest the presence of *a10* but the absence of many (possibly all) of the genes from the last two duplications of the *Abp* gene cluster in all *Mmm* samples (see supplementary fig. S4 and tables S5 and S6, Supplementary Material online; see also supplementary file S2, Supplementary Material online).

The patterns of CN variation in the *Mmc* and *Ms* samples more closely resemble those seen for *Mmm* than *Mmd*, with large areas of deficiencies in the central region (fig. 5*B* and *C*). The overall structure of the *Abp* region is more uniform between samples suggesting absence of large blocks encompassing multiple genes in all samples. A comparison of all four taxa (*Mmd*, *Mmm*, *Mmc* and *Ms*) gives the impression that more duplications occurred in *Mmd* than in any of the others.

Predictions for Inbred Strains

We predicted that mice from laboratory inbred strains and wild-derived *Mmd* strains should have similar *Abp* gene region structure to those of the wild *Mmd* samples, that is, with the majority deficient for the products of the penultimate and/or ultimate duplications proposed for the mouse genome (Janoušek et al. 2013). This is because of the disproportionate representation of the *Mmd* contribution to the genomes of classical inbred laboratory strains, and because the wild-derived inbred strains were sampled directly from a population solely of that subspecies origin.

We analyzed the *Abp* region from Sanger Mouse Genomes Project data for the inbred strains A/J, BALB/cJ, C57BL/10, C3H/HeJ, DBA/1J, NZB/BINJ, and NZO/HlLtJ. CNVnator results revealed that the *Abp* gene region of C57BL/10 is mostly diploid throughout (fig. 6*A*) with the exception of small, scattered duplications (green) but no deficiencies (red), consistent with C57BL/10 being closely related to C57BL/6 (Petkov et al. 2004). The C3H/HeJ strain lacks the ultimate duplication (fig. 2*B*), whereas both the BALB/c and A/J strains appear to lack the penultimate duplication (fig. 2*D*). In contrast, DBA/1J, NZB/BINJ, and NZO/HlLtJ have very different breakpoints and combined deficiencies in the central region, suggesting the absence of large portions of the *Abp* gene region and an appearance very similar to *Mmm*, especially to CR samples (compare figs. 6*A* and 5*A*). We had previously shown that DBA1J, DBA/2J and NZB/BINJ had the $Abpa^{b}$ allele of *Mmm* (Dlouhy and Karn 1984; Dlouhy et al. 1987) and that DBA/2J has *Abp* gene polymorphisms more similar to wild *Mmm* samples throughout the *Abp* gene region (Laukaitis et al. 2012). CNVnator analysis of three wild-derived *Mmd* inbred strains (WSB/EiJ, LEWES/EiJ and ZALENDE/EiJ) revealed typical *Mmd* patterns, whereas wild-derived *Mmm* inbred strain PWK/EiJ (fig. 6*B*) has the pattern of wild *Mmm* (fig. 5*A*) and the wild-derived inbred strain CAST has the pattern of wild *Mmc* (compare figs. 6*B* and 5*B*).

Discussion

Salivary ABPs mediate assortative mate selection based on subspecies recognition that can limit gene exchange between mouse subspecies (Laukaitis et al. 1997; Talley et al. 2001; Laukaitis and Karn 2012) and may contribute to shaping the

European hybrid zone (Bímová Vošlajerová et al. 2011). The genome mouse has 64 *Abp* paralogs, although many are pseudogenes and transcripts have not been found in inbred strains for all the potentially expressed *Abp* genes. Here, we report extensive CN variation in the *Abp* genes in wild mice from the three subspecies of *M. musculus* and the outgroup *M. spretus*, which may provide the raw genetic material for communication signals influencing reproductive behaviors and speciation of mice.

CNVs, especially events associated with SDs, characterize recently duplicated areas of a genome containing sequences that are difficult to distinguish from one another because of high sequence identity (Sjodin and Jakobsson 2012). SDs cover the 3 Mb mouse *Abp* gene region and are especially concentrated in the highly volatile center where NAHR has caused rapid gene duplication (Karn and Laukaitis 2009; Janoušek et al. 2013). The original observations of CN variation in the most recently expanded *Abp* central region genes were made with MLPA, a complex technique that allowed only a glimpse of the potential variation in that region (Karn and Laukaitis 2009). More recently, tools for genome-wide CNV detection based on read-depth analysis of NGS data have become available. In this study, we used CNVnator to analyze CN variation in the entire *Abp* gene region of individuals from natural populations of *Mmd*, *Mmm*, *Mmc*, and *Ms* and combined the results with qPCR analyses of key genes from the last two duplications of the *Abp* region in the mouse genome.

One challenge in the early stage of our study was reconciling the discrepancies between bioinformatic and experimental results due to technical difficulties in analysis of highly similar repetitive sequences. CNVnator and qPCR struggle with different aspects of this defining characteristic of CN variable regions. While CNVnator tends to distribute reads between nearly identical regions, qPCR is limited by the lack of unique sequences for primer placement and poor resolution when amplifying large numbers of genes in the same assay. Our experience with the two methods suggests that employing them in concert can lead to a more reliable interpretation of CNV data when sequence identity is as high as 98–100%, as in the case of the rapidly evolving *Abp* gene region (Laukaitis et al. 2008; Karn and Laukaitis 2009; Janoušek et al. 2013). An added advantage of using both techniques together is that qPCR does not rely on a reference genome to make its gene count estimates, whereas CNVnator does. This raises the issue of when it is appropriate to base genome annotation in a new species on a reference genome. Using a reference genome helped to clarify the fulfillment of hypotheses about the duplication history in the genome mouse in wild samples of *Mmd* (Janoušek et al. 2013); however, it proved to be a liability when trying to determine the organization of the same gene region in the genomes of more distantly related mouse taxa, *Mmm, Mmc*, and *Ms*.

The *Abp* Gene Cluster in Most Wild and Inbred *Mmd* Mice Differs from the C57BL/6 Reference Genome

One of the most interesting findings of our study is the unexpected and very frequent absence of many *Abp* genes in this collection of wild *Mmd* samples compared with the C57BL/6 genome mouse. It is important to note that genes suggested to be deficient by CNVnator analysis do not necessarily represent evolutionary deletions, but could indicate the absence of the most recent duplication event(s) seen in the reference genome. The ultimate and penultimate *Abp* gene duplications (Janoušek et al. 2013) are absent in most wild and wild-derived *Mmd* samples and also in most inbred strains. This causes the lack of ten to18 genes compared with the reference genome, possibly including all *Abp* genes between *bg29* and *a19*.

Since recurrent deletions and duplications can be caused by recombination or misalignment of highly similar sequences, as well as by segregation of a haplotype allele in the population, understanding the breakpoints found in wild populations can give a sense of the sources of variation within the populations. Within this collection of *Mmd* subspecies samples, geography does not predict CNV breakpoints or CN, with wide geographical distribution of polymorphism of simpler genomes within the *Mmd* range. Samples from Germany, France and Iran all have examples with the same variations in breakpoints and CN, which suggests that the last two duplication events seen in the reference mouse genome were present in an ancestor and continue to segregate in wild *Mmd* populations, although recurrent duplication cannot be entirely excluded. The exceptions to this are the samples from Heligoland Island, all of which have the same size of gene region deficiency, although with different CN. This kind of reduced polymorphism is expected for Heligoland mice, given the high estimated inbreeding coefficient, most likely due to small population size and a strong founder effect (Harr et al. 2016).

What are the genetic consequences of the high frequency of "incomplete" *Abp* haplotypes for *Mmd* populations? Given the frequency of genotypes without the ultimate and/or penultimate duplications in this collection of *Mmd* mice from natural populations, it appears that lack of these genes does not significantly compromise fitness in the wild. This is consistent with the less stringent evolutionary constraints that are proposed to act on environmental genes (Nguyen et al. 2008; Pezer et al. 2015) and higher tolerance for null alleles of genes that are members of large gene families (Feuk et al. 2006).

Alternatively, *Abp* haplotypes of variable length may serve as a source of raw genetic material for new signals influencing reproductive communication and thus speciation of mice. The segregation of *Abp* haplotypes of variable lengths allows generation of new diploid combinations in the *Mmd* population. Moreover, the mix of alleles with *Mmd* central-region

haplotypes that range from simpler to more complex suggests that additional duplications are possible and may be an ongoing phenomenon in *Mmd* populations. Indeed, other areas of the *Abp* gene region have smaller recurrent CNVs that do not necessarily involve genes and/or occur in just a few samples. These are expected in a region of such high repeat structure, since factors including the length, sequence identity, orientation, and distance between duplications influence the probability of meiotic misalignment (Stankiewicz and Lupski 2002b). To the extent that additional genetic material may lead to new adaptations, animals that lack some *Abp* genes may also lack potential for adapting to changing fitness requirements. This makes the population ripe for selective sweeps by haplotypes containing genes with more adaptive potential, as has been suggested for the *Abpa27* and *Abpbg27* genes expressed in mouse saliva (Karn and Nachman 1999; Karn et al. 2002).

There is reason to suspect that reproductive genes could be subjected to a variety of evolutionary forces promoting this type of variation. Numerous previous studies support selection on some ABP proteins including the Hudson, Kreitman, and Aguade (HKA) test (Karn and Nachman 1999), the CODEML analysis for site-specific selection (Laukaitis et al. 2012) and other methods (Laukaitis et al. 2003). Janoušek et al. (2013) identified an association of LINE1 (L1) elements with the acceleration of *Abp* gene duplication by NAHR and suggested that selection for increased gene CN resulting from densely packed repeat elements is a cause of the association. Subsequent divergence of these initially identical duplication products may have resulted in subfunctionalization, possibly producing different communication functions in paralogs with expressions partitioned between the lacrimal and salivary glands (Karn et al. 2014).

Architectural Differences in the *Abp* Gene Regions of *Mmd*, *Mmm*, *Mmc* Subspecies and *M. spretus*

The *Mmm* population as a whole generally shows fewer CNs in the *Abp* region, whereas the *Mmd* population shows more variation in CN both between geographical locations (Germany, France, Iran and Heligoland), and within the same location. On the other hand, *Mmm* mice show a different pattern of genes that are missing, extending beyond blocks that were a part of the last two duplication events in the reference genome. *Mmm* mice show relatively more similar gene deficiency patterns among individuals, with concordance of patterns within each of the two populations. Interpretation of this apparently clear-cut difference in the structure of the *Abp* region between the *Mmm* and *Mmd* samples is challenging because NGS data rely on comparison with the reference genome, which is primarily derived from the *Mmd* genetic background. Perhaps the most parsimonious interpretation relies on the idea that the "missing" segments may not reflect deletions from an ancestral state similar to the reference mouse genome, but may instead represent lack of duplications, that is, a simpler ancestral genome. We propose that the last duplication(s) observed in *Mmd* and represented in the reference genome occurred after divergence of the two *M. musculus* subspecies, and that independent duplications may have occurred in *Mmm*, which would explain differences in breakpoints.

The CNV patterns for *Mmc* and *Ms* reinforce the impression that *Abp* gene duplication patterns in *Mmd* diverged from those of the other two *M. musculus* subspecies and the outgroup *Ms*. This is consistent with an estimate that the three subspecies diverged 0.5 Myr (Chevret et al. 2005) whereas the ultimate duplication in the progenitor of the mouse genome occurred only 0.2–0.4 Myr (Laukaitis et al. 2008). Furthermore, the CNV patterns suggest that the *Abp* gene regions of *Mmm, Mmc*, and *Ms* may be simpler than that of at least some members of the *Mmd* population. It is even possible that some of those gene regions lack duplication by NAHR and thus may lack most or all of the central gene region seen in the reference genome.

A Number of Inbred Strains Have the *Mmm Abp* Gene Region

DBA1/J, NZB, and NZO strains (fig. 6*A*) have an *Abp* gene region closely resembling that of PWK (fig. 6*B*) and the wild *Mmm* samples (fig. 5*A*). Fixation of this region is surprising in two independently derived mouse lines that are >90% *Mmd* in content (DBA and NZ; Petkov et al. 2004; Yang et al. 2011). Whether this results from chance or from a biologically important role of genes in this region in mate choice influencing strain formation is unclear, however, it seems clear that having an *Mmm Abp* gene region is not incompatible with having an *Mmd* genetic background, at least in laboratory strains (Laukaitis et al. 2012). A major survey of inbred laboratory strains will be required to determine how widespread this assimilation of an *Mmm Abp* genotype has been.

The Potential Role of Runaway Gene Duplication

We recently proposed that runaway duplication resulting from NAHR acting on LINE1 sequences inserted at the edges of the SDs has created nearly half of the *Abp* genes in the genome mouse (Janoušek et al. 2016). This rapidly evolving central region of the *Abp* gene cluster contains volatile genetic material and is the same region that is extensively polymorphic in several forms in the *Mmd* populations. We propose that the CNV sequences in this region can potentially become subject to evolutionary functionalization within a relatively short evolutionary time span via de novo gene evolution (Neme and Tautz 2016). In other words, polymorphisms in the form of CNVs in the *Abp* gene region may provide an evolutionary potential similar to the polymorphisms implicated in continuous generation of new genes in *Drosophila* (Palmieri et al. 2014; Zhao et al. 2014) and mice (Neme and Tautz

2013, 2014, 2016), and runaway duplication may provide raw genetic material for this functionalization process.

In the Introduction we asked why there are 64 *Abp* genes in the mouse genome when only 15 are expressed in salivary and lacrimal glands? The short answer is that very few wild *Mmd* mice have 64 genes. Some may have as few as 46 genes represented essentially by the flanking genes that have the lion's share of expression, and there appear to be even fewer *Abp* genes in the other taxa we studied (*Mmm*, *Mmc*, and *Ms*). Probably the ancestor of the genus *Mus* had a much simpler *Abp* gene region and proportionally more of its genes were expressed. In the case of *Mmd*, the extensive accumulation of LINEs may have set the stage for a runaway style of duplication not seen in the other taxa and the 64-gene *Abp* region is the exception and not the rule. This raises a new question: What is unique about the *Mmd* subspecies that gives it an *Abp*-region architecture different from those we found in the other taxa we studied? The answer may lie in studies showing significantly different behavior patterns between *Mmd* and *Mmm* mice, such as aggression (Ganem 2012; Hiadlovska et al. 2015). However, this will require additional work to correlate the *Abp* genome with specific behaviors using tools such as congenic and transgenic mouse lines (Laukaitis et al. 1997; Chung et al. 2017).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Diethard Tautz for providing the genomic DNA and the sequencing data before the original manuscript was published. We also thank Nicole Thomsen for technical support in DNA preparation and shipment. The bioinformatic analyzes for this study were performed on the computer cluster of the Max Planck Institute for Evolutionary Biology in Plön, Germany. This work was supported by the National Cancer Institute at the National Institutes of Health (grant numbers U54 CA143924 supporting CML and P30 CA023074 supporting CML and the Genomics Shared Resource of the University of Arizona Cancer Center).

Literature Cited

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 21:974–984.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet. 12:363–376.

Babcock M, et al. 2003. Shuffling of genes within low-copy repeats on 22q11 (LCR22) by Alu-mediated recombination events during evolution. Genome Res. 13:2519–2532.

Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE. 2004. Hotspots of mammalian chromosomal evolution. Genome Biol. 5:R23.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 73:823–834.

Bímová Vošlajerová B, et al. 2011. Reinforcement selection acting on the European house mouse hybrid zone. Mol Ecol. 20:2403–2424.

Boursot P, Auffray J-C, Britton-Davidian J, Bonhomme F. 1993. The evolution of house mice. Annu Rev Ecol Syst. 24:119–152.

Chain FJ, et al. 2014. Extensive copy-number variation of young genes across stickleback populations. PLoS Genet. 10:e1004830.

Chevret P, Veyrunes F, Britton-Davidian J. 2005. Molecular phylogeny of the genus *Mus* (Rodentia:Murinae) based on mitochondrial and nuclear data. Biol J Linn Soc. 84:417–427.

Chung AG, Belone PM, Bimova BV, Karn RC, Laukaitis CM. 2017. Studies of an Androgen-binding protein knockout corroborate a role for salivary ABP in mouse communication. Genetics 205:1517–1527.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. Nat Genet. 39:S30–S36.

Cooper GM, Nickerson DA, Eichler EE. 2007. Mutational and selective effects on copy-number variants in the human genome. Nat Genet. 39:S22–S29.

Cutler G, Marshall LA, Chin N, Baribault H, Kassner PD. 2007. Significant gene content variation characterizes the genomes of inbred mouse strains. Genome Res. 17:1743–1754.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dlouhy SR, Karn RC. 1983. The tissue source and cellular control of the apparent size of *Androgen binding protein (Abp)*, a mouse salivary protein whose electrophoretic mobility is under the control of sex-limited saliva pattern (Ssp). Biochem Genet. 21:1057–1070.

Dlouhy SR, Karn RC. 1984. Multiple gene action determining a mouse salivary protein phenotype: identification of the structural gene for *Androgen-binding protein (Abp)*. Biochem Genet. 22:657–667.

Dlouhy SR, Taylor BA, Karn RC. 1987. The genes for mouse salivary androgen-binding protein (ABP) subunits alpha and gamma are located on chromosome 7. Genetics 115:535–543.

Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. Trends Genet. 17:661–669.

Feuk L, Marshall CR, Wintle RF, Scherer SW. 2006. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet. 15(Spec No (1)):R57–R66.

Freeman JL, et al. 2006. Copy number variation: new insights in genome diversity. Genome Res. 16:949–961.

Ganem G. 2012. Behavior, ecology and speciation in the house mouse. In: Macholan M, Munclinger P, Baird S, Pialek J, editors. Evolution of the house mouse. Cambridge: Cambridge University Press.

Giannuzzi G, et al. 2011. Analysis of high-identity segmental duplications in the grapevine genome. BMC Genomics. 12:436.

Gibbs RA, et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521.

Harr B, et al. 2016. Genomic resources for wild populations of the house mouse, *Mus musculus* and its close relative *Mus spretus*. Sci Data. 3:160075.

Hiadlovska Z, et al. 2015. Shaking the myth: body mass, aggression, steroid hormones, and social dominance in wild house mouse. Gen Comp Endocrinol. 223:16–26.

Ihle S, Ravaoarimanana I, Thomas M, Tautz D. 2006. An analysis of signatures of selective sweeps in natural populations of the house mouse. Mol Biol Evol. 23:790–797.

Janoušek V, Karn RC, Laukaitis CM. 2013. The role of retrotransposons in gene family expansions: insights from the mouse *Abp* gene family. BMC Evol Biol. 13:107.

Janoušek V, Laukaitis CM, Yanchukov A, Karn RC. 2016. The role of retrotransposons in gene family expansions in the human and mouse genomes. Genome Biol Evol. 8:2632–2650.
Karn RC. 1998. Steroid binding by mouse salivary proteins. Biochem Genet. 36:105–117.
Karn RC, Chung AG, Laukaitis CM. 2014. Did androgen-binding protein paralogs undergo neo- and/or subfunctionalization as the *Abp* gene region expanded in the mouse genome? PLoS ONE. 9:e115454.
Karn RC, Clements MA. 1999. A comparison of the structures of the alpha:beta and alpha:gamma dimers of mouse salivary androgen-binding protein (ABP) and their differential steroid binding. Biochem Genet. 37:187–199.
Karn RC, Laukaitis CM. 2009. The mechanism of expansion and the volatility it created in three pheromone gene clusters in the mouse (*Mus musculus*) genome. Genome Biol Evol. 1:494–503.
Karn RC, Laukaitis CM. 2014. Selection shaped the evolution of mouse androgen-binding protein (ABP) function and promoted the duplication of Abp genes. Biochem Soc Trans. 42:851–860.
Karn RC, Nachman MW. 1999. Reduced nucleotide variability at an androgen-binding protein locus (Abpa) in house mice: evidence for positive natural selection. Mol Biol Evol. 16:1192–1197.
Karn RC, Orth A, Bonhomme F, Boursot P. 2002. The complex history of a gene proposed to participate in a sexual isolation mechanism in house mice. Mol Biol Evol. 19:462–471.
Keane TM, et al. 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477:289–294.
Kent WJ. 2002. BLAT – the BLAST-like alignment tool. Genome Res. 12:656–664.
Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.
Lander ES. 2011. Initial impact of the sequencing of the human genome. Nature 470:187–197.
Laukaitis C, Karn RC. 2012. Recognition of subspecies status mediated by androgen-binding protein (ABP) in the evolution of incipient reinforcement on the European house mouse hybrid zone. In: Macholan M, Munclinger P, Baird SJ, Pialek J, editors. Evolution of the house mouse. Cambridge: Cambridge University Press.
Laukaitis CM, Critser ES, Karn RC. 1997. Salivary androgen-binding protein (ABP) mediates sexual isolation in *Mus musculus*. Evolution 51:2000–2005.
Laukaitis CM, Dlouhy SR, Emes RD, Ponting CP, Karn RC. 2005. Diverse spatial, temporal, and sexual expression of recently duplicated androgen-binding protein genes in *Mus musculus*. BMC Evol Biol. 5:40.
Laukaitis CM, Dlouhy SR, Karn RC. 2003. The mouse salivary *Androgen-binding protein (Abp)* gene cluster on Chromosomes 7: characterization and evolutionary relationships. Mamm Genome. 14:679–691.
Laukaitis CM, et al. 2008. Rapid bursts of *Androgen-binding protein (Abp)* gene duplication occurred independently in diverse mammals. BMC Evol Biol. 8:46.
Laukaitis CM, Mauss C, Karn RC. 2012. Congenic strain analysis reveals genes that are rapidly evolving components of a prezygotic isolation mechanism mediating incipient reinforcement. PLoS ONE. 7:e35898.
Li H, et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.
Lin L. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268.
Liu Y, et al. 2009. Bos *taurus* genome assembly. BMC Genomics. 10:180.
Neme R, Tautz D. 2014. Evolution: dynamics of de novo gene emergence. Curr Biol. 24:R238–R240.
Neme R, Tautz D. 2016. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. Elife 5:e09977.
Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. BMC Genomics. 14:117.
Nguyen DQ, et al. 2008. Reduced purifying selection prevails over positive selection in human copy number variant evolution. Genome Res. 18:1711–1723.
Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of Drosophila orphan genes. Elife 3:e01311.
Perry GH. 2008. The evolutionary significance of copy number variation in the human genome. Cytogenet Genome Res. 123:283–287.
Perry GH, et al. 2007. Diet and the evolution of human amylase gene copy number variation. Nat Genet. 39:1256–1260.
Petkov PM, et al. 2004. An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Res. 14:1806–1811.
Pezer Z, Harr B, Teschke M, Babiker H, Tautz D. 2015. Divergence patterns of genic copy number variation in natural populations of the house mouse (*Mus musculus* domesticus) reveal three conserved genes with major population-specific expansions. Genome Res. 25:1114–1124.
Radke DW, Lee C. 2015. Adaptive potential of genomic structural variation in human and mammalian evolution. Brief Funct Genomics. 14:358–368.
Redon R, et al. 2006. Global variation in copy number in the human genome. Nature 444:444–454.
Robinson JT, et al. 2011. Integrative genomics viewer. Nat Biotechnol. 29:24–26.
Sage RD, Atchley WR, Capanna E. 1993. House mice as models in systematic biology. Syst Biol. 42:523–561.
Schrider DR, Hahn MW. 2010. Gene copy-number polymorphism in nature. Proc Biol Sci. 277:3213–3221.
Shaffer LG, Lupski JR. 2000. Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annu Rev Genet. 34:297–329.
She X, Cheng Z, Zöllner S, Church DM, Eichler EE. 2008. Mouse segmental duplication and copy number variation. Nat Genet. 40:909–914.
Sjodin P, Jakobsson M. 2012. Population genetic nature of copy number variation. Methods Mol Biol. 838:209–223.
Stankiewicz P, Lupski JR. 2002a. Genome architecture, rearrangements and genomic disorders. Trends Genet. 18:74–82.
Stankiewicz P, Lupski JR. 2002b. Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev. 12:312–319.
Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. Annu Rev Med. 61:437–455.
Talley HM, Laukaitis CM, Karn RC. 2001. Female preference for male saliva: implications for sexual isolation of *Mus musculus* subspecies. Evolution 55:631–634.
Tuzun E, Bailey JA, Eichler EE. 2004. Recent segmental duplications in the working draft assembly of the brown Norway rat. Genome Res. 14:493–506.
Waterston RH, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562.
Yang H, et al. 2011. Subspecific origin and haplotype diversity in the laboratory mouse. Nat Genet. 43:648–655.
Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. Science 343:769–772.
Zhou Y, Mishra B. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. Proc Natl Acad Sci U S A. 102:4051–4056.

Associate editor: Bill Martin