Introgression patterns between house mouse subspecies and species reveal genomic windows of frequent exchange

Kristian Karsten Ullrich[1], **Miriam Linnenbrink**[1], **Diethard Tautz**[1*]

[1]Max-Planck Institute for Evolutionary Biology, 24306 Plön, Germany

***For correspondence:**
tautz@evolbio.mpg.de (DT)

Abstract Based on whole genome sequencing data, we have studied the patterns of introgression in a phylogenetically well defined set of populations, sub-species and species of mice (*Mus m. domesticus*, *Mus m. musculus*, *Mus m. castaneus* and *Mus spretus*). We find that many discrete genomic regions are subject to repeated and mutual introgression and exchange. The majority of these regions code for genes that are involved in parasite defense or genomic conflict. They include genes involved in adaptive immunity, such as the MHC region or antibody coding regions, but also genes involved in innate immune reactions of the epidermis. We find also clusters of KRAB zinc finger proteins that control the spread of transposable elements and genes that are involved in meiotic drive. These findings suggest that even well separated populations and species maintain the capacity to exchange genetic material in a special set of evolutionary active genes.

Introduction

Patterns of introgression are frequently observed between taxa that can still hybridize (***Green et al., 2010***; ***Salazar et al., 2010***; ***Song et al., 2011***; ***Staubach et al., 2012***; ***Hedrick, 2013***; ***Liu et al., 2015***; ***Zhang et al., 2016***; ***Stukenbrock, 2016***; ***Kumar et al., 2017***, and refs therein). The house mouse (*Mus musculus*) forms a species complex with several described and not yet fully described sub-species (***Guénet and Bonhomme, 2003***; ***Phifer-Rixey and Nachman, 2015***; ***Hardouin et al., 2015***) that are distributed in allopatric patterns across the whole world but can still hybridize, in particular at secondary contact zones (***Sage et al., 1986***; ***Teeter et al., 2008***; ***Janoušek et al., 2012***; ***Turner and Harr, 2014***). The sister species *Mus spretus* lives in sympatry with *M. m. domesticus* in Western Europe and can form hybrids with this subspecies, but maintains its own species status. Patterns of introgression between the Western and Eastern house mouse (*M. m. domesticus* and *M. m. musculus*) have been intensively studied along the hybrid zone in Europe (***Teeter et al., 2008***; ***Wang et al., 2011***; ***Janoušek et al., 2015***). Genome-wide comparisons based on SNP array data have further revealed that introgression of haplotypes can also occur across large distances, possibly mediated through human transport of mice (***Staubach et al., 2012***). Since these introgressed haplotypes are much larger than the average linkage disequilibrium (LD) blocks in wild populations (***Laurie et al., 2007***), their spread in distant populations is apparently driven by positive selection, since they would rapidly break up under drift conditions (***Staubach et al., 2012***). A prominent example of an adaptive introgression of a *M. spretus* related haplotype into *M. m. musculus* populations is the locus that confers resistance to the rodenticide Warfarin, *Vkorc1*(***Song et al., 2011***; ***Liu et al., 2015***). But haplotypes may also introgress between separated populations of the same subspecies, as it

has been shown for the MLV virus receptor *Xpr1*(**Hasenkamp et al., 2015**).

A variety of tests have been developed to trace introgression patterns and to distinguish them from incomplete lineage sorting (**Durand et al., 2011**; **Yu et al., 2012**; **Pease and Hahn, 2015**; **Martin et al., 2014**). While these have been assessed for their power to identify introgression, they have the disadvantage that they require relatively rigid assumptions about the assumed history of introgression between them. These model assumptions are violated when introgression occurs from non-sampled sources, or repeatedly between the taxa. While such more complex scenarios could potentially be integrated into more formalized models, this would increase vastly the number of test comparisons in the statistics. In fact, there is no simple solution for this problem, implying that the procedure used to trace introgression need to be somewhat adapted to the taxa and the question that one wants to ask.

We have used here a framework of mouse populations and outgroups for which the phylogenetic histories are well known, based on fossil and phylogeographic evidence (**Guénet and Bonhomme, 2003**; **Phifer-Rixey and Nachman, 2015**; **Hardouin et al., 2015**). Figure 1 depicts these populations and their relationships.

Our focal *M. m. domesticus* populations come from France (Fra) and Germany (Ger). These are derived from a population from Iran (Ira), which invaded Western Europe about 3,000 years ago (**Cucchi et al., 2005**; **Hardouin et al., 2015**). The Fra and Ger populations have split shortly after arrival in Southern France and have since developed largely independently (**Ihle et al., 2006**; **Teschke et al., 2008**; **Staubach et al., 2012**). Hence, if a secondary introgression has occurred after the split, it should become visible in only one of the populations. While any population could serve as a source for secondary introgression, including the one from Ira (**Hasenkamp et al., 2015**), our focus in this study are the two subspecies *M. m. musculus* (MUS - represented by a population from Afghanistan) and *M. m. castaneus* (CAS - represented by a population from India), as well as the species *M. spretus* (SPRE - represented by a population from Spain) as possible source populations.

To study introgression, we use full genome sequence information and a branch-length comparison approach (Figure 1). This approach allows to detect the most prominent regions of unequivocal introgression. We first applied it to a previously identified introgression event at the amylase gene cluster and provide evidence that the introgression is likely related to a rescue of a pseudogenized *Amy2b* allele. However, we noticed also that the introgression region as a whole has a much more complex history. By assessing the whole genome for similar patterns, we found that regions exist that are subject to apparent mutual introgression and haplotype exchange between the hybridizing taxa. Most notably, many of these regions code for loci involved in adaptive and innate immune defense, in the defense against transposable elements and some appear to be involved in meiotic drive. Several affect parts of the olfactory and vomeronasal receptor clusters. Many, but not all are in regions of gene clusters with copy number variation. These findings suggest that mechanisms exist that allow the frequent exchange of genes involved in frequent adaptive processes between the taxa, even though most of them are regionally separated and/or hybrids are sub-fertile.

Results

We have previously generated genomic re-sequencing data for multiple individuals for each of our focal populations (**Harr et al., 2016**). We have used these reads and created a consensus sequence representative for each population. This was done to ensure that only haplotypes with a frequency higher than 0.5 are represented. Hence, we are tracing introgression patterns that are either on the way to fixation or fixed. Our previous study (**Staubach et al., 2012**) had shown that the majority of recently introgressed haplotypes segregate only at low frequency, hence by focusing on the high frequency variants in the present study, we are tracing genomic introgression regions that have most likely been subject to recent positive selection in the respective populations (see modeling in **Staubach et al. (2012)**).

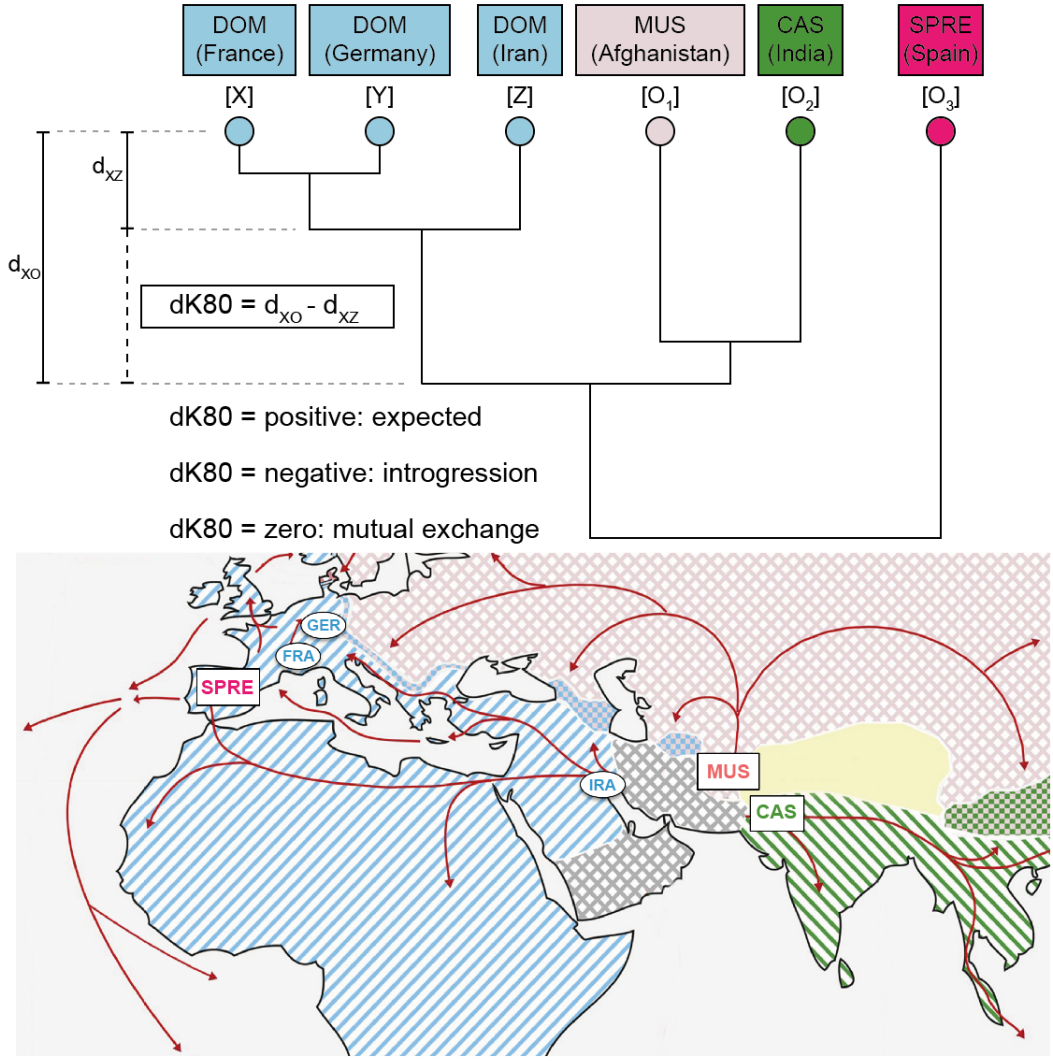

Figure 1. Relationships and origins of the mouse populations in the study. Three populations represent *M . m. domesticus* (DOM) (from France = Fra, Germany = Ger and Iran = Ira), the population from Afghanistan represents *M. m. musculus* (MUS), the one from India *M. m. castaneus* (CAS) and the one from Spain *M. spretus* (SPRE). The map shows the approximate locations, as well as the known dispersal routes (picture modified from ***Harr et al.*** (***2016***)). The tree represents the relationships. The principle of calculating dK80 is depicted to the left for one particular example, namely the comparison of the distance between Fra-Ira with the one between Fra-MUS, whereby Ira is the in-group [Z] and MUS the out-group [O_1]. The two alternative outgroups used in this study are CAS [O_2] and SPRE [O_3].

Figure 1–Figure supplement 1. dK80 q-q plots for different population combinations. (A) [X]: Fra [Z]: IRA [O_1]: AFG, (B) [Y]: Ger [Z]: IRA [O_1]: AFG, (C) [X]: Fra [Z]: IRA [O_2]: CAS, (D) [Y]: Ger [Z]: IRA [O_2]: CAS, (E) [X]: Fra [Z]: IRA [O_3]: SPRE, (F) [Y]: Ger [Z]: IRA [O_3]: SPRE. Mean of dK80 distribution is highlighted as solid lines (black: simulated data; grey: real data), 3.89 standard deviations are highlighted by dashed lines (red: simulated data; orange: real data).

Figure 1–Figure supplement 2. Preview of supplementary Table 1. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: AFG.

Figure 1–Figure supplement 3. Preview of supplementary Table 2. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: CAS.

Figure 1–Figure supplement 4. Preview of supplementary Table 3. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: SPRE.

Figure 1–Figure supplement 5. Preview of supplementary Table 4. Lists of outlier windows for the three quartet comparisons.

Figure 1–Figure supplement 6. Scheme of the simulation approach. Taking the mm10 reference sequence (yellow) as a start point, genomes were constructed in a phylogenetic context mimicking the real data including the construction of a 'new' reference (orange). Nucleotides were randomly altered given a percentage divergence value including ancestral states (grey). The resulting distances represents the phylogenetic context obtained as described in the method section.

Population-specific introgression patterns

To identify genomic regions of introgression, we used a sliding window approach (25kb per window) and generated a phylogenetic tree for each window. Linkage disequilibrium drops fast within 20kb in wild populations (***Laurie et al., 2007***; ***Staubach et al., 2012***), i.e. by focusing on 25kb window sizes, we trace mostly relatively recent events that have not been subject to much recombination. We noticed that tree lengths can vary considerably along the chromosomes, which makes a simple dxy analysis for tracing tree incongruence as indicators of introgression less suitable. To compensate for this, we use the subtraction measure deltaK80 (dK80) depicted in Figure 1. The trees serve to calculate dK80 by subtracting the distance of the focal groups (either Fra or Ger) to the founder population (Ira) from the respective distances to the tested outgroups (MUS, CAS or SPRE). dK80 is expected to be positive when no introgression has occurred and negative when introgression from the tested outgroup has occurred. Based on simulations of our tree configuration without introgression, we find that dK80 values have a normal distribution (see q-q plots in Figure 1 - figure supplement 1).

The data for the dK80 values are provided in suppl. Table 1 (comparisons with MUS as outgroup [O_1]), suppl. Table 2 (comparisons with CAS as outgroup [O_2]) and suppl. Table 3 (comparisons with SPRE as outgroup [O_3]). We have also plotted these actual data distributions in q-q plots and find that they are less dispersed and more skewed than the ones of the simulations (suppl. Figure 1). Interestingly, the skews occur not only in the negative direction, but also in the positive one. In the following we focus the analysis on the negative side and come back to the skew on the positive side in the discussion.

dK80 values can be easily visualized as genome browser tracks allowing to recognize even complex patterns of introgression (see below). We have generated a custom track set on the UCSC genome browser named "wildmouse-introgression" that includes all data discussed here. Examples of these tracks are in the figures below.

We surveyed all negative outlier windows beyond a cutoff of 0.01% of the distribution of the real data (3.89 SD). Since about 100,000 windows were surveyed, one would expect around 5 by chance in a normal distribution, but we find between 99 to 352 in the different comparisons per population (Table 1; genome positions in suppl. Table 4).

Table 1. dK80 averages and outliers in the populations.

	tested outgroup population		
	MUS	CAS	SPRE
windows (N)	102,281	102,299	102,009
dK80 average (x103)	6.08	4.15	14.94
dK80 StDEV (x103)	3.54	2.97	4.22
0.01% cutoff (x103)	-7.69	-7.40	-1.50
smallest dK80 (x103)	-55.07	-56.17	-45.53
windows (N) smaller than cutoff at 0.01%			
total in Fra	114	115	282
total in Ger	99	134	352
total in Fra and Ger	55	65	178
only in Fra	59	50	104
only in Ger	44	69	174

Among these we find between 55 - 178 in both, the Fra and the Ger population. This can be interpreted either that both populations were introgressed, possibly already in their direct colonizing ancestor, or that the Ira population has itself become introgressed from an unknown source after splitting from the ancestor of the Western European populations. Note that there

are several further mouse lineages as potential donors in Iran that have not been characterized in detail yet (***Hardouin et al., 2015***). We have inspected the regions including outlier windows in the genome browser tracks and found that they represent often complex patterns of introgression. We first focus on a region coding for amylase genes.

Amylase 2b introgression

We had previously identified the region around the pancreas Amylase 2b (*Amy2b*) as having been subject to adaptive introgression in the Fra population (***Staubach et al., 2012***). The dK80 approach and the outlier windows identify the same region, but the better resolution available through the full genome sequences allows now a much more detailed picture (Figure 2). The overall introgression (i.e. negative dK80 values) from *M. m. musculus* into Fra is as broad as originally found (approx. 0.5Mb), but the corresponding track lines for introgression from *M. m. castaneus* and *M. spretus* are more complex. A particularly strong introgression signal for all three outgroups, identified by outlier windows (Figure 2, bottom tracks), covers part of the non-coding region between *Amy2b* and its duplicated paralogs (Figure 2). Hence, the historical introgression dynamics at this locus appears to have been much more complex than originally anticipated. Note that there are also complex introgression patterns in the neighboring *Amy1a* gene (see data tracks "wildmouse-introgression") that are not further discussed here.

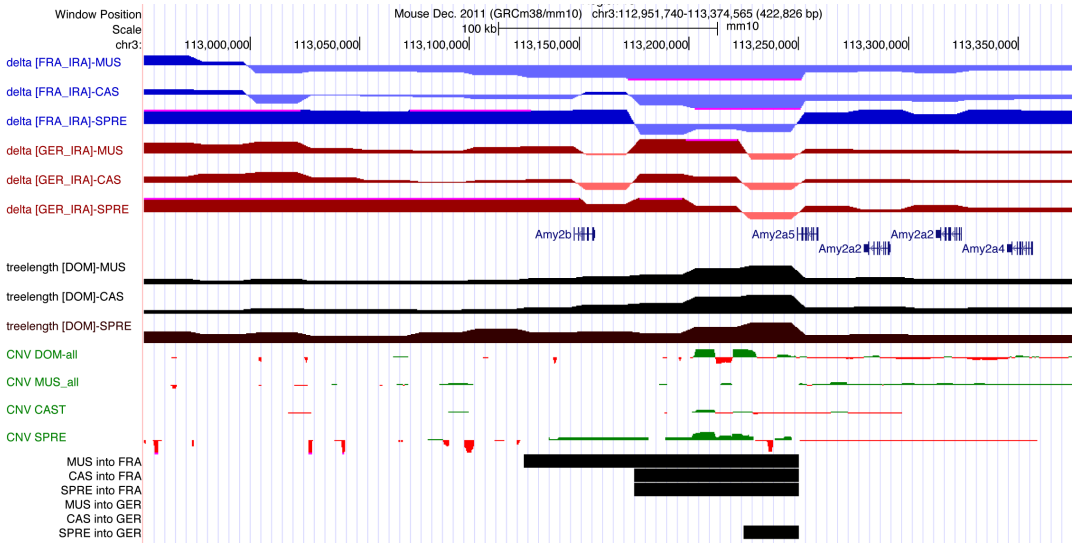

Figure 2. Introgression patterns around the amylase 2 gene cluster. Depiction of UCSC browser tracks for a region of chromosome 3 (positions in header). The rows show from top to bottom: the tracks for the dK80 values for the FRA comparisons (blue/magenta) and the GER comparisons (brown/red); gene annotations from the "UCSC Genes" track implemented in the UCSC Mouse Genome Browser (GRCm38/mm10); tree length tracks for the three outgroup comparisons (black); copy number variation tracks summed across all individuals of the respective populations, taken form ***Pezer et al.*** (***2015***) (green is more copies than reference, red is fewer copies than reference); windows with significant introgression in the indicated directions.

Interestingly, the sequence analysis provides a hint towards the possible reason for the apparently very recent introgression of a *M. m. musculus* haplotype into Fra. All of the Ger individuals sequenced harbor a mutation in the first exon that leads to a premature stop codon (Figure 3A). Hence, the Ger individuals carry a pseudogene for *Amy2b*. We have typed this variant for an extended sample of animals and populations and find the pseudogene variant to be prevalent in Germany, but rare in France (Figure 3b). The *Amy2b* sequences of the eight fully sequenced animals from Fra are all very similar and cluster closely with the *M. m. musculus* consensus sequence (Figure 3c). A native gel electrophoresis from pancreas tissue shows that there are still amylase variants in the Ger individuals, but the band patterns shows a composite between those found in the Fra animals and in *M. m. musculus / M. m. castaneus* animals. These are are most likely derived from

the paralogs in the region, but a full resolution of the CNV haplotypes would be required to verify this. A quantitative activity assay shows that the overall amylase activity is lower in Ger than in Fra or *M. m. musculus* animals (Figure 3d). Hence, this pattern is compatible with the assumption that the mice colonizing Western Europe had lost the original *Amy2b* gene, but partially replaced this by activity from the duplicated copies in the vicinity. The Fra population has apparently captured a haplotype from *M. m. musculus* representing the originally active *Amy2b* gene. This haplotype has then spread rapidly in the Fra populations and is apparently now also invading Ger populations. Because an apparent rescue of an enzyme function is involved, this would have occurred through an adaptive spread, which explains also the selective sweep signature that we had detected in this region (***Staubach et al., 2012***).

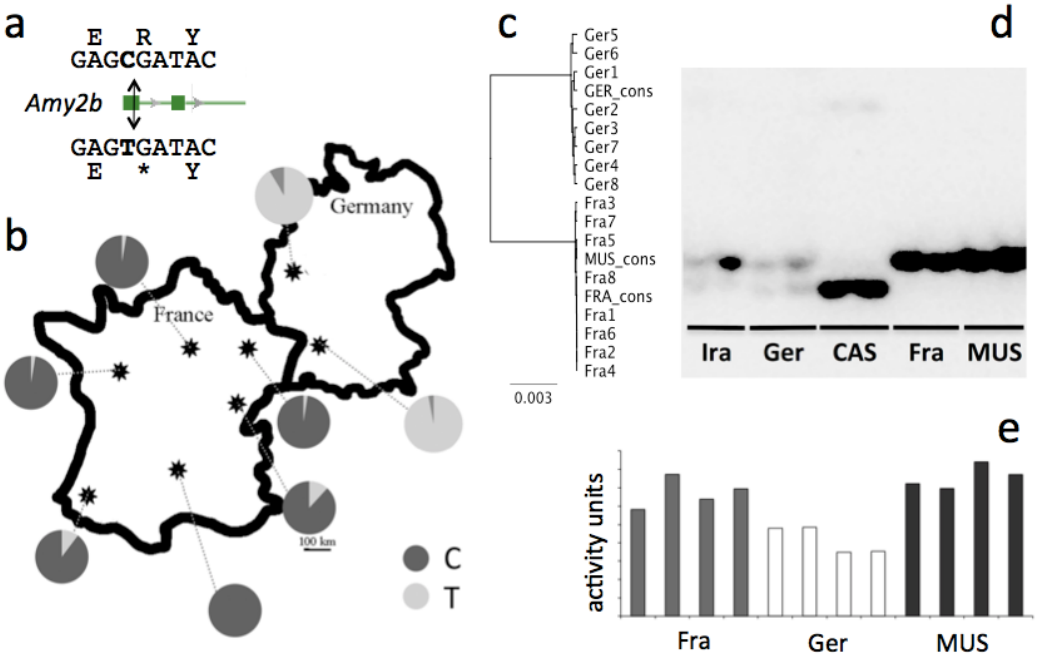

Figure 3. *Amy2b* introgression patterns. (a) Depiction of of the disabling mutation in exon1 of *Amy2b*. (b) Frequencies of the disabling (T-allele) versus enabling (C-allele) in different populations in France and Germany. (c) Phylogeny of *Amy2b* alleles of individuals and consensus sequences of the Fra and Ger population, plus consensus sequence of the MUS population. (d) Westernblot from a native gel of pancreas samples of two animals each of five populations, stained with an *Amy2b* antibody. Note that a denaturing gel does not resolve these variants (data not shown). (e) Amylase activity assays of four animals each from the three indicated populations.

The case of the *Amy2b* introgression presents thus a model of how one can envisage adaptive introgression taking place. An allele that has an advantage, either because it has itself acquired a new adaptation in a source population, or because it replaces a slightly deleterious mutation in the receiving population can apparently be specifically acquired, even from distant populations and can then reach high frequencies in the receiving population.

Genomic introgression windows

The complexity of the introgression pattern shown in Figure 2 suggests that a region subject to introgression may actually be invaded repeatedly from different sources. If this would happen frequently, one would expect that dK80 would assume values around zero because the mutual invasions would equilibrate the distances between the respective groups.

We have therefore sought to systematically identify regions of such repeated introgression. However, because of the complexity of the possible patterns, they can not be easily captured by general statistical criteria. But the regions can readily be identified when one inspects the dK80 browser tracks. Since most of the genome follows the expected distribution of dK80 values, individ-

ual introgression regions become visible as noticeable dips in this background. Figure 4 shows two examples of 10Mb windows from chromosome 12 and chromosome 17. The chromosome 12 region includes an approximately 2Mb window with a complex mixture of strong introgression (identified by outlier windows) and flat distributions. The region codes for the mouse immunoglobulin heavy gene genes (*Igh*) encoding the exons that make up the variable antibody parts. In fact, another such region on chromosome 6 shows a similar pattern (see below). The chromosome 17 region includes an approximately 5Mb region that is relatively flat for comparisons between the sub-species *M. m. musculus* and *M. m. castaneus*, plus two shorter dips where also the comparison with M. spretus shows introgression, including significant outlier windows. These are enlarged at the bottom of Figure 4 and show that they code for key genes of the major histocompatibility complex (MHC), namely the antigen fragment binding receptors *H2-A*, *H2-B*, *H2-O* and *H2-Q*. Transspecies alleles have been described for *H2-A* and *H2-B* genes before, supporting the notion that our procedure picks out such introgression regions faithfully.

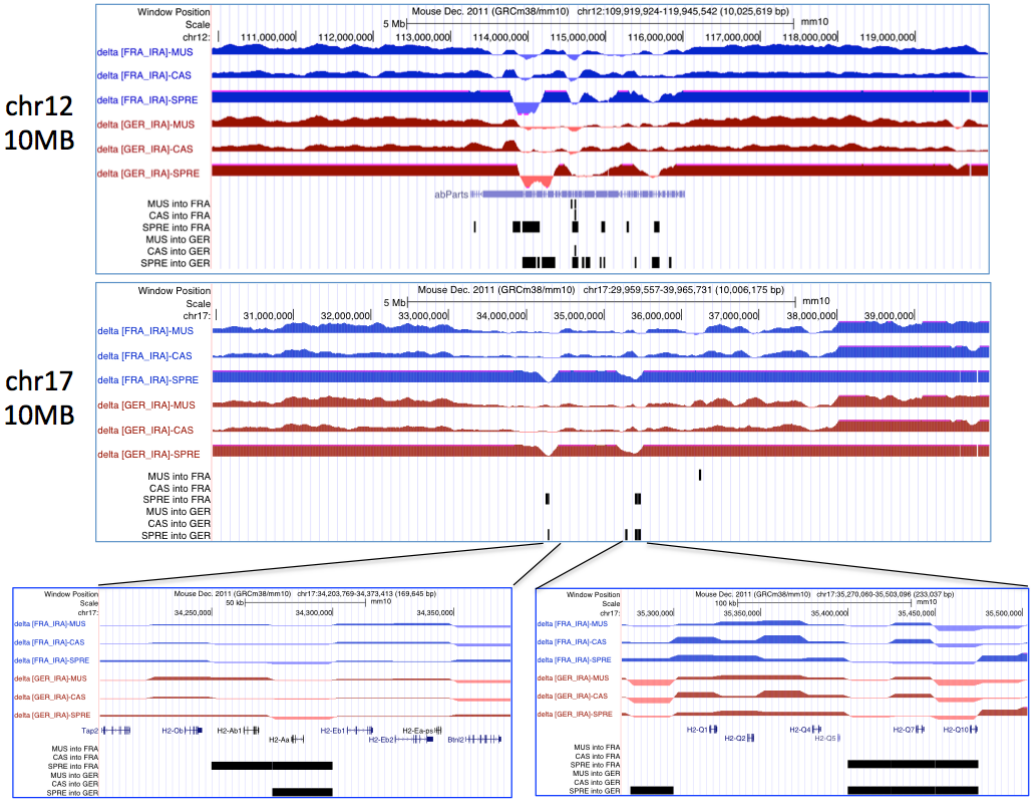

Figure 4. Two examples of mutual introgression windows. Depiction of UCSC browser tracks for 10MB regions of chromosome 12 and chromosome 17 (positions in headers). The rows show from top to bottom: the tracks for the dK80 values for the FRA comparisons (blue/magenta) and the GER comparisons (brown/red); gene annotations from the "UCSC Genes" track implemented in the UCSC Mouse Genome Browser (GRCm38/mm10); windows with significant introgression in the indicated directions.

Figure 4–Figure supplement 1. Collection of screen shots of browser windows for all mutual introgression regions listed in Table 2. For full track description see legend of Figure 2.

Figure 4–Figure supplement 2. Neighbor-joining (NJ) tree comparisons with consensus sequences in extreme introgression regions. (A) Region 11 from Table 2 (tree constructed for chr11:71,150,213-71,301,792) including *Nlrp1b*. (B) Region 7 from Table 2 (tree constructed for chr4:41,807,517-42,760,683) including a chemokine ligand cluster. (C) Standard tree structure represented by a tree of whole chr19. NJ (***Saitou and Nei, 1987***) trees based on pair-wise K80 distances (***Kimura, 1980***) with the R ape package (***Paradis et al., 2004***). Prior the NJ tree calculation all masked sites were removed from all included sequences.

Figure 4–Figure supplement 3. Preview of suppl. Table 5. This table represents an extended version of Table 2 including all identified introgression regions between subspecies and species, as well as the respective genomic locations.

We have inspected all 10MB windows across the whole genome (excluding the Y-chromosome

because of too many missing data) to identify a set of windows that show unusual patterns across all three outgroup comparisons. Table 2 lists all of these regions with a length of at least 100kb (i.e. 4 consecutive windows). The majority (12 out of 18) have well defined functions in adaptive immunity (5), innate immunity (7) and three represent KRAB Zn-finger genes which function in the repression of transposable elements (***Jacobs et al., 2014***; ***Imbeault et al., 2017***; ***Kauzlaric et al., 2017***). Two cover parts of olfactory receptor gene clusters whereby region 9 includes besides an olfactory receptor cluster also the hemoglobin beta-chain genes, which are known to be subject to complex selection and gene replacement in mouse (***Storz et al., 2007***). The track patterns for the 18 regions listed in Table 2 are compiled in Figure 4-Figure supplement 1.

Table 2. Introgression between all in-groups and out-groups. #chromosomal genome positions are provided in Figure 4-Figure supplement 3; dK80 and tree length represent averages x1000; "immunity" is classified into three categories: "adaptive" = adaptive immunity, "innate" = innate immunity and "transposons" = transposon repression

region	chr	size (Mb)	dK80	tree length	annotated coding genes	general classification	comments
1	chr1	0.7	2.2	4.0	Csprs, Sp110, Sp140	immunity (innate)	SP110 is involved in inflammation and resistance to tuberculosis
2	chr1	0.3	-1.8	19.6	Ifi204, Mndal, Ifi203, Ifi202b, Ifi205	immunity (innate)	interferon inducible genes
3	chr1	0.17	-4.0	26.1	Olfr gene cluster	sensory	olfactory receptor genes
4	chr2	0.11	-0.2	19.4	Ttpal, Serinc3	immunity (innate)	Serinc3 involved in defense response against viruses
5	chr2	2.08	-0.4	3.0	KRAB zinc finger protein cluster	immunity (transposons)	repression of transposable elements
6	chr4	2.28	-2.1	27.0	Skint gene cluster	immunity (adaptive)	required for intraepithelial T cell development
7	chr4	0.95	0.1	1.2	Ccl - chemokine ligand gene cluster	immunity (adaptive)	chemotactic factors attracting skin-associated memory T-lymphocytes
8	chr6	2.47	2.5	17.8	abParts	immunity (adaptive)	parts of antibodies, mostly variable regions
9	chr7	0.49	-4.4	19.3	Olfr receptor cluster, Hbb genes	sensory	olfactory receptor genes; hemoglobin beta genes
10	chr8	0.12	-1.7	27.7	Defb8	immunity (innate)	innate immune response
11	chr11	0.15	-36.4	59.5	Nlrp1b	immunity (innate)	sensor component of the NLRP1 inflammasome
12	chr12	1.84	1.2	20.5	abParts	immunity (adaptive)	parts of antibodies, mostly variable regions
13	chr12	1.2	0.7	3.3	KRAB zinc finger protein	immunity (transposons)	repression of transposable elements
14	chr12	2.73	1.2	4.9	9030624G23Rik	unknown	KRAB-A box protein; unknown function
15	chr13	1.66	1.4	9.6	KRAB zinc finger protein	immunity (transposons)	repression of transposable elements
16	chr14	0.42	-0.7	21.9	Ear gene family	immunity (innate)	involved in virus defense
17	chr16	0.12	-5.2	26.2	Cd200r genes	immunity (innate)	regulation of inflammation response
18	chr17	0.17	0.4	11.6	Tap2, H2-Ob, H2-Ab,H2-Aa,H2-Eb,H2-Ea, Btnl2	immunity (adaptive)	MHC core region coding for antigen fragment binding receptors

In addition to the regions involving mutual introgression with *M. spretus*, we have also systematically surveyed all regions >100kb that involve mostly exchange between the three sub-species (*M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*) by the same criteria. These include a total of 67 regions (suppl. Table 5), including the overlaps with the ones listed in Table 2. 29 of the regions have clear immune functions, eight represent olfactory and vomeronasal receptor clusters and another eight represent testis specific genes, or genes highly expressed in testis. Six regions each cover non-coding parts of the genome (i.e. no annotations in the respective window) and genes with unknown functions (suppl. Table 5).

Inspection of the windows shows that tree lengths differ much in the different regions. Some introgression regions have very long trees, others very short ones (Table 2 and suppl. Table 5). On average, there is a good correlation between tree length and dK80 score (r2 = 0.69 for the data in Table 2 and 0.45 for the data in suppl. Table 5), with more negative scores showing the longer trees. The most negative score combined with the longest average tree is found for *Nlrp1b*, which codes for the sensor component of the inflammasome (region 11 in Table 2). This pattern is apparently caused by an introgression of a highly unrelated haplotype into the Ira population (tree in Figure 4-supplement 2), which renders all other comparisons negative. The shortest average tree length is found for the window coding for a chemokine ligand cluster (region 7 in Table 2), rendering this 0.92Mb region highly similar between all analyzed taxa (tree in Figure 4-supplement 2). This suggests that a given haplotype variant has introgressed into all of them, possibly located on an inversion, since the size appears to be the same in all taxa.

The gene clusters identified within the windows show often copy number variation (see suppl. Figure 4-Figure supplement 1). Of the three single genes with major copy number variation in natural populations identified in ***Pezer et al.*** (***2015***), two show patterns of introgression. One is *Cwc22*, which encodes a splicing factor. Its introgression pattern is apparent in the comparison with CAS and SPRE (Figure 5a). The other is *Hjurp*, which codes for a holiday junction recognition protein. For this gene, introgression is mostly evident for the Ger population with respect to MUS and CAS

introgression (Figure 5b).

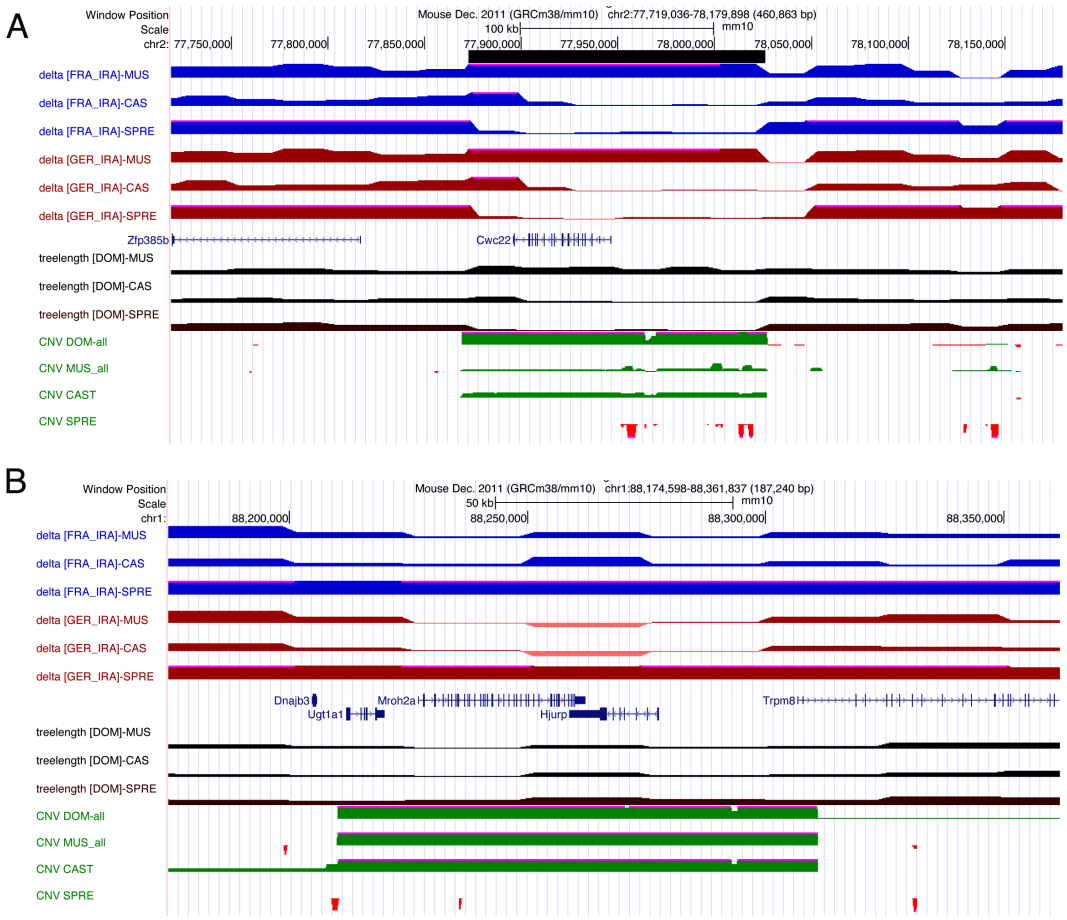

Figure 5. Introgression patterns in highly copy number variable regions. Depiction of UCSC browser tracks for chromosomal regions of *Cwc22* (A) and *Hjurp* (B) (positions in headers). Rows from top to bottom as in Figure 2. *Cwc22* and Hjurp were previously found to be among the most copy number variable genes between mouse populations (Pezer et al. 2015). *Cwc22* shows introgression in comparison to *M. m. castaneus* and *M. spretus*, for *Hjurp*, the pattern is most evident for the comparison to *M. m. musculus* and *M. m. castaneus* in the Ger population.

Discussion

Our initial data analysis was geared towards identifying specific regions of introgression in populations that have established only a few thousand years ago (Fra and Ger). One such region that we had identified before is a region on chromosome 3 coding for an amylase gene cluster. Here we confirmed this region, based on a procedure specifically developed to identify regions of potentially adaptive introgression. In the case of the amylase region, we show that a secondary replacement of pseudogenized version of *Amy2b* was apparently the reason for the introgression in one of the *M. m. domesticus* sub-populations. However, our analysis revealed that the patterns are much more complex. There are apparently many regions in the genome that are subject to repeated introgression in different directions, with apparently different histories, and including haplotypes from sources that have not been sampled. Given the abundant evidence for introgression in many studies, this should not be so surprising. In fact, this complexity makes average genome-wide statistical analyses complicated, since these can only work within the framework of given scenarios and do not account for repeated or mutual introgression. Our study has initially also focused onto a particular scenario, namely introgression into newly established populations from other sub-species and species. But the fact that many of the outlier windows identified through this

procedure overlap and cluster in specific regions revealed an underlying complexity that could eventually only be resolved through direct inspection of introgression patterns.

By using consensus sequences from the populations, our approach is conservative, since we miss out cases where an introgressing haplotype is still at a frequency below 50%. This is for example the case for the otherwise well supported *Vkorc1* region that was suggested to be in a phase of adaptive introgression in some Western European populations (***Song et al., 2011***; ***Liu et al., 2014***). In our data, we find only a subset of animals in the Fra and Ger populations carrying this haplotype, i.e. it does not show up among the consensus sequences.

Our focus was on low or negative dK80 values, since these would indicate introgression within our general tree topology. However, the q-q plots in Figure 1-Figure supplement 1 show that deviations from the expected distributions occur also at high dK80 values. We have inspected these outliers and find that they represent another level of complexity, such as introgression into the outgroup population from unknown sources, making the distances much larger than average. It is known that there are further sub-species or species in Asia (***Hardouin et al., 2015***; ***Hamid et al., 2017***) that could be the source for such introgression. Given our general finding of mutual introgression among our sampled populations, it is not unexpected that this should also occur with non-sampled populations.

Introgression in evolutionary active regions

The inspection of the dK80 tracks across the whole genome and across the different populations allowed to pinpoint regions of particularly active introgression. Intriguingly, the majority of these regions is related to adaptive and innate immune functions. This includes the MHC, where trans-species alleles are often found (***Parham et al., 1996***). These have usually been ascribed to be the result of balancing selection and/or incomplete lineage sorting after the splitting of the species. However, for species that have remained at least partially inter-fertile, there is increasing evidence that introgression must play a role as well (***Abi-Rached et al., 2011***; ***Grossen et al., 2014***). This is now fully confirmed in our study. Key genes of the core MHC region, including H2-Aa, H2-Ab, H2-Ea and H2-Eb show significant introgression windows for the comparisons with *M. spretus*. Further, an extended 2.9Mb region shows a flat dK80 pattern for the comparisons with the subspecies (Figure 2). These regions are too long to be considered as remnants from the progenitor populations maintained by balancing selection, since they would have been broken up by recombination. Interestingly, in spite of the apparent frequent introgression, at least the loci coding for the H2-Aa, H2-Ab receptors show long trees, which is compatible with balancing selection in this region to maintain a high diversity, which may even be fueled by capturing alleles from other sub-species and species.

While the introgression pattern within the MHC is not unexpected, it had remained so far unnoticed that many other immune genes are also subject to introgression. Most notably, these encode the antibody coding regions on chromosomes 6 and 12, which are also part of the adaptive immune system. Since the antibody diversity is generated through splicing among variable exons, it seems that this diversity is also enhanced by taking up alleles from other populations.

Another immune region with strong introgression and long trees is the *Skint* gene family cluster on chromosome 4 (region 6 in Table 2). *Skint*s code for proteins containing trans membrane spanning domains and extracellular IgV and IgC domains that are specifically expressed in dendritic epidermal T-cells which play a crucial role in immune defense during wound healing (***Keyes et al., 2016***). Interestingly, a second introgression region on chromosome 4 (region 7 in Table 2) codes for a cluster of cytokine genes including *CCL27*, which has been shown to selectively recruit cutaneous memory T lymphocytes into the skin (***Morales et al., 1999***) and is generally implicated in regulating wound repair (***Hocking, 2015***). This region shows, however, a very flat signature, both for dK80, as well as for tree length, indicating mutual introgression or a recent sweep of a whole haplotype across the tested species and sub-species (Figure 4-Figure supplement 2. A third introgression region associated with epidermis is the one that encodes the *Defb8* gene (region 10 in Table 2).

ß-defensins code for antimicrobial and chemo-attractant peptides, especially attracting CD4+ T-cells (**Taylor et al., 2009**) and high expression of *Defb8* occurs in epidermal tissues.

Three introgression regions (region 1, 11 and 17 in Table 2) are involved in inflammation regulation. This includes the *SP110* gene that is involved in resistance to tuberculosis (**Wu et al., 2016**), *Nlrp1b*, a sensor component of the inflammasome (**Chavarría-Smith and Vance, 2015**) and *CD200r* genes regulate macrophage function in inflammation reactions (**Snelgrove et al., 2008**; **Fraser et al., 2016**).

Another three regions code for genes that are interferon inducible and/or involved in virus defense. These include a cluster of interferon inducible genes on chromosome 1 that function in virus recognition (*Ifi203* - region 2) (**Stavrou et al., 2015**), activity against retroviruses (*Serinc3* - region 4) (**Usami et al., 2015**) and influenza virus protection in epithelia (*Ear1* - region 16) (**O'Reilly et al., 2012**). The fact that all of these regions are similarly, or even more strongly affected by introgression than the MHC genes suggests that they play a similar major role in counteracting fast evolving pathogens. Hence, the current focus on MHC genes as the major driver of evolutionary responses to pathogens may be too limited. It seems warranted to pay additional attention to immune processes in the epithelial cells, as well as innate mechanisms of virus defense. Differences in innate immune responses in human populations have also been ascribed to adaptive introgression of Neandertal alleles (**Quach et al., 2016**).

Immunity is not only relevant against pathogens, but also against transposable elements. KRAB zinc finger proteins have been implicated in this function, whereby there is an evolutionary arms race between adaptation of the zinc fingers to the recognitions sites in the transposable elements and the acquisition of new mutations in the recognition sites (**Jacobs et al., 2014**). We find a total of four clusters of such genes plus a single gene one among the introgression regions (suppl. Table 4). One of them (region 13 - Table 2) covers a 2Mb region with almost identical sequences between all taxa (i.e. very short trees). In fact, one should expect that active transposable elements could introgress as well between the taxa, i.e. it makes sense when they share the same set of defense genes.

The extended table with introgression regions that show a signal with only two of the three outgroups (suppl. Table 5) provides some further interesting insights into the type of processes affected most by introgression in mice. There are six regions that code for olfactory receptor clusters, some as specific parts of larger such clusters and two of the several vomeronasal receptor clusters in the genome show an introgression signal. This would suggest that they code for sub-types that may be particularly relevant for an evolutionary active process.

There are four regions that code for SPEER/Takusan domain genes. These were originally identified as testis-specific genes (**Spiess et al., 2003**) and they occur in several clusters in the genome. The genes have specifically evolved in the rodent lineage, possibly through some protein domain fusions and fast subsequent evolution. The variants occurring on chromosome 14 were also implicated in the regulation of synaptic activity (**Tu et al., 2007**) but these are also more expressed in testis than in brain (see also expression data provided in **Harr et al.** (**2016**)). In their study on postmeiotic transcription in sperm cells, **Moretti et al.** (**2016**) speculate that these genes may act in a meiotic drive dynamics, possibly as suppressors of the intragenomic conflict between *Slx* and *Sly* gene families (**Helleu et al., 2015**). Interestingly, the *Slx* gene cluster on the X chromosome shows also corresponding introgression patterns (suppl. Table 5) and, the *Sly* gene cluster on the Y-chromosome shows similar patterns (although we have not included these in our analysis because of too many missing data, we provide the partial data in the track patterns of "wildmouse-introgression").

Another region that was suggested to be involved in transmission ratio distortion is *Cwc22* or *R2D2* on chromosome 2 (**Didion et al., 2016**). Population genetic studies of this region have shown that it shows major copy number variation changes between populations (**Pezer et al., 2015**; **Didion et al., 2016**) and that it was involved in recent selective sweeps (**Didion et al., 2016**). Hence, this is also a region with clear introgression patterns that has a proven high evolutionary dynamics.

Conclusions

Although the populations we have included in our analysis are all very distinct, either due to allopatry or long evolutionary separation, they exchange nonetheless genes. Most of the genes that are exchanged have evident adaptive significance, including the case of the *Amy2b* gene, where a pseudogene allele became replaced by an active allele from another subspecies. Given that our study has focused on high frequency variants and that most of them are associated to various forms of immune defense, we can assume that most of the introgression events that we are tracing here have been adaptive. Hence, by identifying these introgression regions, we find at the same time candidate genes for frequent adaptations in mice.

Methods and Materials

Ethic statement

Mice were caught as described in ***Harr et al.*** (***2016***). Transportation of live mice to the animal facility, maintenance and handling were conducted in accordance with German animal welfare law (Tierschutzgesetz) and FELASA guidelines. Permits for keeping mice were obtained from the local veterinary office 'Veterinäramt Kreis Plön' (permit number: 1401-144/PLÖ-004697).

Data and mouse samples

Genome data used in this study (*M. m. domesticus* GER - 8 individuals, *M. m. domesticus* FRA - 8 individuals, *M. m. domesticus* IRA - 8 individuals, *M. m. musculus* AFG - 6 individuals, *M. m. castaneus* CAS - 10 individuals, and *M. spretus* SPRE - 8 indiciduals) were taken from Harr et al. (2016). Mouse samples were taken from the collection at our institute - sources are described in ***Harr et al.*** (***2016***) and ***Linnenbrink et al.*** (***2013***).

dK80 calculation

Quartets were used for all calculations of dK80. They included always the three DOM populations (Ger [X], Fra [Y] and Ira [Z]), as well as one of the three outgroups each, MUS [O_1], CAS[O_2] or SPRE [O_3] (compare Figure 1). Within the quartets, dK80 was calculated for trios (either with Ger or Fra) on non-overlapping sequence windows (w) throughout the genome between this population triplet on a window (w_i) as:

$$dK80_{XZOw_i} = d_{XOw_i} - d_{XZw_i} \quad (1)$$

where d_{XOw_i} and d_{XZw_i} are defined as the average Kimura's 2-parameter sequence distance (***Kimura, 1980***) between the corresponding two populations calculated with the function 'dist.dna' of the R package 'ape' (***Paradis et al., 2004***) using the model 'K80'. Prior the calculation of dK80 all sites with missing data (see below) within the specified window (w_i) and the specified populations ([X], [Z], [O]) were removed across the whole quartet with the 'Biostrings' R package (***Pages et al., 2009***) and only those windows retained with a missing rate lower than 50%.

Population specific masking

Sequences with low coverage (missing data) were excluded from the analysis by masking them for each quartet under analysis. Masking files were generated by processing the BAM files with 'genomeCoverageBed' and 'mergeBed' from the bedtools2 software suite (***Quinlan and Hall, 2010***, v2.26.0) to obtain site specific genome coverage per individual. Sites with a coverage smaller than 5 were used as masking regions for each individual. Further, to obtain population specific masking regions, site specific genome coverage files were united with 'unionBedGraphs' and sites with a united coverage smaller than 5 were used as masking regions for the populations. Suppl. Tables 1-3 provide for each window the numbers of masked ("missing") sites. Windows that had more than 50% sites missing were not included in the overall analysis.

Consensus sequences

To construct individual and population specific consensus sequences, we conducted first a SNP calling on the mapped BAM files taken from ***Harr et al.*** (***2016***). Briefly, BAM files for individuals grouped by population were processed with 'samtools mpileup' and 'bcftools call' (***Li, 2011***, v1.3.1) with relaxed mapping quality options (samtools mpileup: -q 0 -Q 10 -A -d 99999 -t DP,AD,ADF,ADR -uf mm10.fasta; bcftools call: -f GQ –m -v) to also retain information in CNV regions. We corrected also for individual sex and ploidy (chrY F 0; chrX M 1; chrY M 1; chrM F 1; chrM M 1). We note that many of the here described introgression regions occur in copy number variable regions, which are often filtered out in standard analyses that constrain the data to high quality mapped reads. In our simulations we paid particular attention to such regions to assess whether the inclusion of low quality mapped reads would lead to artifacts with respect to introgression signals, but we did not find this to be the case. If anything, the inclusion of these reads would only make distance larger rather than smaller.

The resulting population specific VCF files were recoded with 'vcftools' (***Danecek et al., 2011***, v0.1.15) to remove indels and to only retain variant sites (vcftools: –recode –remove-indels -recode-INFO-all -non-ref-ac-any 1) either for the complete population (-keep pop) or for the individuals (-keep ind). Further, the recoded population specific VCF file containing multiple individuals was parsed with 'vcfparser.py mvcf2consensus' (https://gitlab.gwdg.de/evolgen/introgression; -cdp 11) to obtain a CONSENSUS VCF file for each population. In brief, the allelic depth information for the reference and alternative allele was summed over all individuals per population by simultaneously removing all sites which had a total depth (reference plus alternative allelic depth) of smaller than 11 in the population (DP<11).

The CONSENSUS VCF was used in combination with the population specific masking region to construct chromosome alignments with the reference (GRCm38/mm10) and the python script 'vcfparser.py vcf2fasta' (https://gitlab.gwdg.de/evolgen/introgression; -type refmajorsample -R mm10.fasta -cov2N 4) to obtain the major allele per variant and to mask all regions with a coverage smaller than 5. This step was repeated for individuals using the individual masking regions resulting in CONSENSUS and INDIVIDUAL FASTA files used for the inference of introgression.

Simulations

Simulations were performed to evaluate dK80 under a phylogenetic scenario mimicking the real data without introgression, as well as to ensure that the mapping procedure does not lead to artefacts, especially in the copy number variable regions (see Figure 1-Figure supplement 6 for the simulation design).

First, to provide the appropriate distance framework, all pair-wise polymorphic sites and all pair-wise informative sites (excluding pair-wise missing sites) of all autosomes were counted between the investigated populations using the CONSENSUS FASTA files. The resulting pair-wise distances defined as:

$$d_{xy} = \frac{\sum_{n=chr1}^{chr19} \text{polymorphic sites}_{xy}}{\sum_{n=chr1}^{chr19} \text{informative sites}_{xy}} \quad (2)$$

where d_{xy} is the pair-wise distance between the consensus sequence of population x and y, were then used as a distance matrix to calculate an "Unweighted Pair Group Method with Arithmetic Mean" (UPGMA) tree in R. The resulting UPGMA tree distances were used as a proxy to simulate chromosome 1 with the python script 'simdiv.py' taking the phylogenetic context into account.

Second, the simulated sequences representative of each population (SPRE, CAS, MUS, Ira, Ger, Fra) were used to generate artificial Illumina reads to test the influence of possible sequencing errors with 'ART' (***Huang et al., 2011***, v2.5.8) and to mimic the original sequence libraries (art_illumina: -sam -na -ss HS20 -f 20 -l 100 -m 250 -s 80 -p). Subsequently, the artificial Illumina reads were mapped against the simulated reference with 'bwa mem' (***Li and Durbin, 2009***, 0.7.12-r1039), followed by sorting, marking and removing duplicates with the picard software suite (https://broadinstitute.

github.io/picard/) and an indel realignment step with 'GATK' (***McKenna et al., 2010***, v3.7) as described in ***Harr et al.*** (***2016***). Masking, SNP calling, FASTA sequence construction and dK80 calculation was performed as described above.

Data visualization and availability

Data visualization was done with the UCSC genome browser for the mouse assembly mm10. Custom tracks were generated and and are available as "Public Session" under the term "wildmouse-introgression". Individual and population specific consensus sequences can be accessed via ftp http://wwwuser.gwdg.de/ evolbio/evolgen/wildmouse/introgression/.

Amylase purification and quantification

Approximately equal weights of pancreas per sample were used for purification. Tissues were homogenised in PBS using a TissueLyser II (Qiagen), centrifuged at 13,000 x g at 4°C for 10 minutes, and the crude lysate was collected. Ethanol was added to a final concentration of 40%, centrifuged at 10,000 x g for 10 minutes at 4°C, and the supernatant was collected. Amylase was precipitated by addition of 1mg of oyster glycogen (Sigma-Aldrich) according to ***Schramm and Loyter*** (***1966***), followed by shaking on ice for 5 minutes. It was then pelleted by centrifugation at 5,000 x g for 3 minutes at 4°C. The samples were washed, re-suspended in PBS, and glycogen digested by incubation at 30°C for 20 minutes (***Hjorth, 1979***). Samples were stored at -80°C in aliquots to avoid repeated freeze/thaw cycles.

Protein concentration was determined using Thermo Scientific's Coomassie Plus™ (Bradford) Assay kit according to the manufacturer's instructions as follows. A standard curve was generated using the BSA provided at 2000, 1500, 1000, 750, 500, 250, 125 and 25μg/mL. All unknown samples were diluted 1:2, and 10μL of each sample and standard was added to a 96 well plate in duplicate. A PBS blank was also included. 300μL of Coomassie Plus Reagent was added to each well and mixed on a plate mixer for 30 seconds. The plate was incubated at room temperature for 10 minutes, and then measured on a Tecan Infinite® M200 PRO plate reader at 595nm. The blank measurement was subtracted from all other readings, and then the concentration of the unknown samples determined using the standard curve, and multiplied by the dilution factor.

Native PAGE

Amylase extracts were separated on 7.5% Mini-PROTEAN® TGX™ gels (Bio-Rad), but in their native form (no boiling, no SDS). Gels were loaded with 1.5μg of each sample (using Bradford assay measurements) and run for 45 minutes at 100V, followed by 3 hours 45 minutes at 300V, with the tank immersed in ice throughout (adapted from Hjorth, 1979). Western blot analysis was done by semi-dry transfer to PVDF membrane for one hour at 20V, followed by blocking the membrane with 5% milk powder in PBS supplemented with 0.1% tween for one hour at room temperature. The membrane was probed for amylase using anti-α -amylase antibody from Cell Signalling (#4017) at 1:2000 in 2% milk in PBS tween overnight at 4°C. The primary antibody was detected using goat anti-rabbit HRP (Southern Biotech) at 1:5000 for one hour at room temperature followed by incubation with Immun-Star™ WesternC™ Chemiluminescent substrate (Bio-Rad). Blots were visualised using an Alpha Innotech FluorChem™ MultiImage™ light cabinet.

Amylase activity

The activity of each sample was determined using an amylase activity assay kit (Sigma-Aldrich), which measures a colorimetric product resulting from the cleavage of ethylidene-pNP-G7 by amylase to generate p-nitrophenol. Samples were measured with a slightly modified version of the manufacturer's instructions as follows. A nitrophenol standard curve was generated by adding 0, 2, 4, 6, 8, and 10μL of 2mM Nitrophenol Standard to a 96 well plate and adding water to a final volume of 50μL. Unknown samples were diluted 1:5000 and 50μL added to the plate in duplicate. Amylase Assay Buffer and Amylase Substrate Mix were mixed together 1:1 and 100μL added to each

well. After 3 minutes incubation (= $T_{initial}$) a reading was taken at 405nm ($A405_{initial}$). The plate was incubated at 25°C and the absorbance measured every 2 minutes (as opposed to every 5 minutes in the manufacturer's instructions). Readings were taken until the value of the most active sample exceeded the linear range of the standard curve. $A405_{final}$ was taken as the penultimate reading before this. The background was subtracted, and the change in absorbance from $T_{initial}$ to T_{final} calculated: $\Delta A405 = A405_{final} - A405_{initial}$. The amount of nitrophenol generated between $T_{initial}$ to T_{final} was calculated using the standard curve. The activity was then determined using the equation:

$$\text{Amylase Activity} = \frac{Bx\text{Sample Dilution Factor}}{(\text{Reaction Time})xV} \quad (3)$$

where: B = Amount (nmole) of nitrophenol generated between $T_{initial}$ and T_{final} Reaction Time = T_{final} – $T_{initial}$ (minutes) V = sample volume (mL) added to well One unit of amylase is the amount of amylase that cleaves ethylidene-pNP-G7 to generate 1.0 m mole of p-nitrophenol per minute at 25°C.

Acknowledgments

We thank Emre Karakoc for the initial exploration of algorithms to detect introgression patterns in the mouse dataset, Ellen McConnell for the alpha-amylase experiments and the members of our lab for comments on the manuscript.

Additional information

Competing interests

DT: Senior editor, *eLife*. The other authors declare that no competing interests exist.

References

Abi-Rached L, Jobin MJ, Kulkarni S, McWhinnie A, Dalva K, Gragert L, Babrzadeh F, Gharizadeh B, Luo M, Plummer FA, et al. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science. 2011; 334(6052):89–94.

Chavarría-Smith J, Vance RE. The NLRP1 inflammasomes. Immunological reviews. 2015; 265(1):22–34.

Cucchi T, Vigne JD, Auffray JC. First occurrence of the house mouse (Mus musculus domesticus Schwarz & Schwarz, 1943) in the Western Mediterranean: a zooarchaeological revision of subfossil occurrences. Biological Journal of the Linnean Society. 2005; 84(3):429–445.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158.

Didion JP, Morgan AP, Yadgary L, Bell TA, McMullan RC, Ortiz de Solorzano L, Britton-Davidian J, Bult CJ, Campbell KJ, Castiglia R, et al. R2d2 drives selfish sweeps in the house mouse. Molecular biology and evolution. 2016; 33(6):1381–1395.

Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Molecular biology and evolution. 2011; 28(8):2239–2252.

Fraser SD, Sadofsky LR, Kaye PM, Hart SP. Reduced expression of monocyte CD200R is associated with enhanced proinflammatory cytokine production in sarcoidosis. Scientific reports. 2016; 6:38689.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al. A draft sequence of the Neandertal genome. science. 2010; 328(5979):710–722.

Grossen C, Keller L, Biebach I, Croll D, Consortium IGG, et al. Introgression from domestic goat generated variation at the major histocompatibility complex of alpine ibex. PLoS genetics. 2014; 10(6):e1004438.

Guénet JL, Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. Trends in Genetics. 2003; 19(1):24–31.

518 **Hamid HS**, Darvish J, Rastegar-Pouyani E, Mahmoudi A. Subspecies differentiation of the house mouse Mus
519 musculus Linnaeus, 1758 in the center and east of the Iranian plateau and Afghanistan. Mammalia. 2017;
520 81(2):147–168.

521 **Hardouin EA**, Orth A, Teschke M, Darvish J, Tautz D, Bonhomme F. Eurasian house mouse (Mus musculus
522 L.) differentiation at microsatellite loci identifies the Iranian plateau as a phylogeographic hotspot. BMC
523 evolutionary biology. 2015; 15(1):26.

524 **Harr B**, Karakoc E, Neme R, Teschke M, Pfeifle C, Pezer Ž, Babiker H, Linnenbrink M, Montero I, Scavetta R,
525 et al. Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus
526 spretus. Scientific data. 2016; 3:160075.

527 **Hasenkamp N**, Solomon T, Tautz D. Selective sweeps versus introgression-population genetic dynamics of
528 the murine leukemia virus receptor Xpr1 in wild populations of the house mouse (Mus musculus). BMC
529 evolutionary biology. 2015; 15(1):248.

530 **Hedrick PW**. Adaptive introgression in animals: examples and comparison to new mutation and standing
531 variation as sources of adaptive variation. Molecular ecology. 2013; 22(18):4606–4618.

532 **Helleu Q**, Gérard PR, Montchamp-Moreau C. Sex chromosome drive. Cold Spring Harbor perspectives in biology.
533 2015; 7(2):a017616.

534 **Hjorth JP**. Genetic variation in mouse salivary amylase rate of synthesis. Biochemical genetics. 1979; 17(7):665–
535 682.

536 **Hocking AM**. The role of chemokines in mesenchymal stem cell homing to wounds. Advances in wound care.
537 2015; 4(11):623–630.

538 **Huang W**, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2011;
539 28(4):593–594.

540 **Ihle S**, Ravaoarimanana I, Thomas M, Tautz D. An analysis of signatures of selective sweeps in natural populations
541 of the house mouse. Molecular Biology and Evolution. 2006; 23(4):790–797.

542 **Imbeault M**, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory
543 networks. Nature. 2017; 543(7646):550–554.

544 **Jacobs FM**, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. An
545 evolutionary arms race between KRAB zinc finger genes 91/93 and SVA/L1 retrotransposons. Nature. 2014;
546 516(7530):242.

547 **Janoušek V**, Munclinger P, Wang L, Teeter KC, Tucker PK. Functional organization of the genome may shape the
548 species boundary in the house mouse. Molecular biology and evolution. 2015; 32(5):1208–1220.

549 **Janoušek V**, Wang L, Luzynski K, Dufková P, Vyskočilová MM, Nachman MW, Munclinger P, Macholán M, Piálek J,
550 Tucker PK. Genome-wide architecture of reproductive isolation in a naturally occurring hybrid zone between
551 Mus musculus musculus and M. m. domesticus. Molecular Ecology. 2012; 21(12):3032–3047.

552 **Kauzlaric A**, Ecco G, Cassano M, Duc J, Imbeault M, Trono D. The mouse genome displays highly dynamic
553 populations of KRAB-zinc finger protein genes and related genetic units. PloS one. 2017; 12(3):e0173746.

554 **Keyes BE**, Liu S, Asare A, Naik S, Levorse J, Polak L, Lu CP, Nikolova M, Pasolli HA, Fuchs E. Impaired epidermal to
555 dendritic T cell signaling slows wound repair in aged skin. Cell. 2016; 167(5):1323–1338.

556 **Kimura M**. A simple method for estimating evolutionary rates of base substitutions through comparative
557 studies of nucleotide sequences. Journal of molecular evolution. 1980; 16(2):111–120.

558 **Kumar V**, Lammers F, Bidon T, Pfenninger M, Kolter L, Nilsson MA, Janke A. The evolutionary history of bears is
559 characterized by gene flow across species. Scientific Reports. 2017; 7.

560 **Laurie CC**, Nickerson DA, Anderson AD, Weir BS, Livingston RJ, Dean MD, Smith KL, Schadt EE, Nachman MW.
561 Linkage disequilibrium in wild mice. PLoS Genetics. 2007; 3(8):e144.

562 **Li H**. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical
563 parameter estimation from sequencing data. Bioinformatics. 2011; 27(21):2987–2993.

564 **Li H**, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;
565 25(14):1754–1760.

566 **Linnenbrink M**, Wang J, Hardouin EA, Künzel S, Metzler D, Baines JF. The role of biogeography in shaping
567 diversity of the intestinal microbiota in house mice. Molecular ecology. 2013; 22(7):1904–1916.

568 **Liu KJ**, Dai J, Truong K, Song Y, Kohn MH, Nakhleh L. An HMM-based comparative genomic framework for
569 detecting introgression in eukaryotes. PLoS computational biology. 2014; 10(6):e1003649.

570 **Liu KJ**, Steinberg E, Yozzo A, Song Y, Kohn MH, Nakhleh L. Interspecific introgressive origin of genomic diversity
571 in the house mouse. Proceedings of the National Academy of Sciences. 2015; 112(1):196–201.

572 **Martin SH**, Davey JW, Jiggins CD. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Molecular
573 biology and evolution. 2014; 32(1):244–257.

574 **McKenna A**, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S,
575 Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
576 sequencing data. Genome research. 2010; 20(9):1297–1303.

577 **Morales J**, Homey B, Vicari AP, Hudak S, Oldham E, Hedrick J, Orozco R, Copeland NG, Jenkins NA, McEvoy
578 LM, et al. CTACK, a skin-associated chemokine that preferentially attracts skin-homing memory T cells.
579 Proceedings of the National Academy of Sciences. 1999; 96(25):14470–14475.

580 **Moretti C**, Vaiman D, Tores F, Cocquet J. Expression and epigenomic landscape of the sex chromosomes in
581 mouse post-meiotic male germ cells. Epigenetics & chromatin. 2016; 9(1):47.

582 **O'Reilly MA**, Yee M, Buczynski BW, Vitiello PF, Keng PC, Welle SL, Finkelstein JN, Dean DA, Lawrence BP. Neonatal
583 oxygen increases sensitivity to influenza A virus infection in adult mice by suppressing epithelial expression
584 of Ear1. The American journal of pathology. 2012; 181(2):441–451.

585 **Pages H**, Aboyoun P, Gentleman R, DebRoy S. String objects representing biological sequences, and matching
586 algorithms. R package version. 2009; 2(2).

587 **Paradis E**, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics.
588 2004; 20(2):289–290.

589 **Parham P**, Ohta T, et al. Population biology of antigen presentation by MHC class I molecules. SCIENCE-NEW
590 YORK THEN WASHINGTON-. 1996; p. 67–73.

591 **Pease JB**, Hahn MW. Detection and polarization of introgression in a five-taxon phylogeny. Systematic biology.
592 2015; 64(4):651–662.

593 **Pezer Ž**, Harr B, Teschke M, Babiker H, Tautz D. Divergence patterns of genic copy number variation in natural
594 populations of the house mouse (Mus musculus domesticus) reveal three conserved genes with major
595 population-specific expansions. Genome research. 2015; 25(8):1114–1124.

596 **Phifer-Rixey M**, Nachman MW. The Natural History of Model Organisms: Insights into mammalian biology from
597 the wild house mouse Mus musculus. Elife. 2015; 4:e05959.

598 **Quach H**, Rotival M, Pothlichet J, Loh YHE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al.
599 Genetic adaptation and neandertal admixture shaped the immune system of human populations. Cell. 2016;
600 167(3):643–656.

601 **Quinlan AR**, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;
602 26(6):841–842.

603 **Sage RD**, Heyneman D, Lim KC, Wilson AC. Wormy mice in a hybrid zone. Nature. 1986; 324(6092):60–63.

604 **Saitou N**, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular
605 biology and evolution. 1987; 4(4):406–425.

606 **Salazar C**, Baxter SW, Pardo-Diaz C, Wu G, Surridge A, Linares M, Bermingham E, Jiggins CD. Genetic evidence
607 for hybrid trait speciation in Heliconius butterflies. PLoS genetics. 2010; 6(4):e1000930.

608 **Schramm M**, Loyter A. Purification of α-amylase by precipitation of amylase-glycogen complexes. Methods
609 Enzymology. 1966; 8:533–537.

610 **Snelgrove RJ**, Goulding J, Didierlaurent AM, Lyonga D, Vekaria S, Edwards L, Gwyer E, Sedgwick JD, Barclay AN,
611 Hussell T. A critical function for CD200 in lung immune homeostasis and the severity of influenza infection.
612 Nature immunology. 2008; 9(9):1074–1083.

Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, Shih CH, Nachman MW, Kohn MH. Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Current Biology. 2011; 21(15):1296–1301.

Spiess AN, Walther N, Mueller N, Balvers M, Hansis C, Ivell R. SPEER-a new family of testis-specific genes from the mouse. Biology of reproduction. 2003; 68(6):2044–2054.

Staubach F, Lorenc A, Messer PW, Tang K, Petrov DA, Tautz D. Genome patterns of selection and introgression of haplotypes in natural populations of the house mouse (Mus musculus). PLoS Genetics. 2012; 8(8):e1002891.

Stavrou S, Blouch K, Kotla S, Bass A, Ross SR. Nucleic acid recognition orchestrates the anti-viral response to retroviruses. Cell host & microbe. 2015; 17(4):478–488.

Storz JF, Baze M, Waite JL, Hoffmann FG, Opazo JC, Hayes JP. Complex signatures of selection and gene conversion in the duplicated globin genes of house mice. Genetics. 2007; 177(1):481–500.

Stukenbrock EH. The role of hybridization in the evolution and emergence of new fungal plant pathogens. Phytopathology. 2016; 106(2):104–112.

Taylor K, Rolfe M, Reynolds N, Kilanowski F, Pathania U, Clarke D, Yang D, Oppenheim J, Samuel K, Howie S, et al. Defensin-related peptide 1 (Defr1) is allelic to Defb8 and chemoattracts immature DC and CD4+ T cells independently of CCR6. European journal of immunology. 2009; 39(5):1353–1360.

Teeter KC, Payseur BA, Harris LW, Bakewell MA, Thibodeau LM, O'Brien JE, Krenz JG, Sans-Fuentes MA, Nachman MW, Tucker PK. Genome-wide patterns of gene flow across a house mouse hybrid zone. Genome research. 2008; 18(1):67–76.

Teschke M, Mukabayire O, Wiehe T, Tautz D. Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. Genetics. 2008; 180(3):1537–1545.

Tu S, Shin Y, Zago WM, States BA, Eroshkin A, Lipton SA, Tong GG, Nakanishi N. Takusan: a large gene family that regulates synaptic activity. Neuron. 2007; 55(1):69–85.

Turner LM, Harr B. Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. Elife. 2014; 3:e02504.

Usami Y, Wu Y, Göttlinger HG. SERINC3 and SERINC5 restrict HIV-1 infectivity and are counteracted by Nef. Nature. 2015; 526(7572):218.

Wang L, Luzynski K, Pool JE, Janoušek V, Dufkova P, Vyskočilová MM, Teeter KC, Nachman MW, Munclinger P, Macholán M, et al. Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. Molecular ecology. 2011; 20(14):2985–3000.

Wu Y, Guo Z, Yao K, Miao Y, Liang S, Liu F, Wang Y, Zhang Y. The transcriptional foundations of Sp110-mediated macrophage (RAW264. 7) resistance to Mycobacterium tuberculosis H37Ra. Scientific reports. 2016; 6.

Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS genetics. 2012; 8(4):e1002660.

Zhang W, Dasmahapatra KK, Mallet J, Moreira GR, Kronforst MR. Genome-wide introgression among distantly related Heliconius butterfly species. Genome biology. 2016; 17(1):25.

Tables and Figures (not placed within the main text)

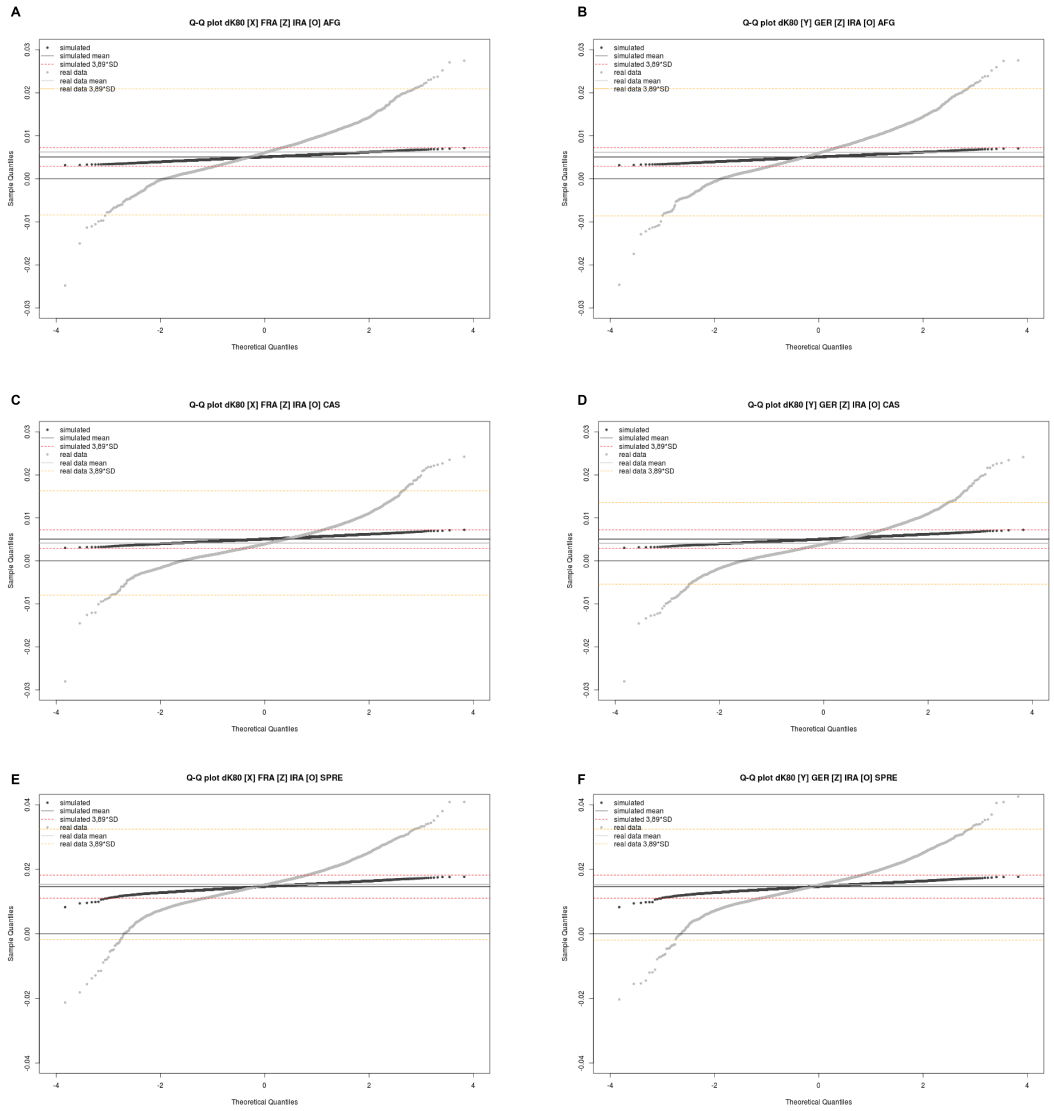

Figure 1–Figure supplement 1. dK80 q-q plots for different population combinations. (A) [X]: Fra [Z]: IRA [O_1]: AFG, (B) [Y]: Ger [Z]: IRA [O_1]: AFG, (C) [X]: Fra [Z]: IRA [O_2]: CAS, (D) [Y]: Ger [Z]: IRA [O_2]: CAS, (E) [X]: Fra [Z]: IRA [O_3]: SPRE, (F) [Y]: Ger [Z]: IRA [O_3]: SPRE. Mean of dK80 distribution is highlighted as solid lines (black: simulated data; grey: real data), 3.89 standard deviations are highlighted by dashed lines (red: simulated data; orange: real data).

	A	B	C	D	E	F	G
1	dK80 values and total tree length for x (FRA), y(GER), z(IRA) and O(MUS)						
2	CHR	START	END	MISSING	dK80_FRA	dK80_GER	TREELENGTH
151	chr1	3700001	3725000	153	0.0103784	0.011168	0.01371579
152	chr1	3725001	3750000	362	0.00533476	0.00853908	0.01234405
153	chr1	3750001	3775000	567	0.01707585	0.01784546	0.01993973
154	chr1	3775001	3800000	1385	0.00721186	0.00741082	0.01598245
155	chr1	3800001	3825000	1380	0.00333998	0.00350462	0.01025122
156	chr1	3825001	3850000	40	0.00360553	0.00343774	0.01058511
157	chr1	3850001	3875000	61	0.00520741	0.00512491	0.00773203
158	chr1	3875001	3900000	2145	0.00705015	0.00695451	0.00885893
159	chr1	3900001	3925000	1266	0.0062787	0.00669867	0.00887096
160	chr1	3925001	3950000	179	0.01292228	0.01267463	0.01586174
161	chr1	3950001	3975000	68	0.01706809	0.01648653	0.01775725
162	chr1	3975001	4000000	829	0.0157333	0.01567397	0.01794409
163	chr1	4000001	4025000	52	0.00560981	0.00617481	0.00916173
164	chr1	4025001	4050000	45	0.00190678	-0.0006099	0.01736611
165	chr1	4050001	4075000	10	0.00281646	0.00168999	0.00687375
166	chr1	4075001	4100000	359	0.00313861	0.00346266	0.0055956
167	chr1	4100001	4125000	40	0.00080782	0.00096833	0.00876572

Figure 1–Figure supplement 2. Preview of supplementary Table 1. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: AFG.

	A	B	C	D	E	F	G
1	dK80 values and total tree length for x (FRA), y(GER), z(IRA) and O(CAS)						
2	CHR	START	END	MISSING	dK80_FRA	dK80_GER	TREELENGTH
151	chr1	3700001	3725000	47	0.00756217	0.00835253	0.0108871
152	chr1	3725001	3750000	310	0.00536545	0.00774192	0.01190855
153	chr1	3750001	3775000	296	0.01892805	0.01985318	0.02186232
154	chr1	3775001	3800000	4	-0.0010885	0.00217617	0.00981836
155	chr1	3800001	3825000	1378	-0.0024733	-0.0019156	0.00771982
156	chr1	3825001	3850000	64	0.00239069	0.00263015	0.01150338
157	chr1	3850001	3875000	7	0.00539885	0.00523547	0.00771566
158	chr1	3875001	3900000	44	0.00329995	0.00329684	0.00518171
159	chr1	3900001	3925000	6	0.00689359	0.00720878	0.00949032
160	chr1	3925001	3950000	1	0.00373886	0.00373845	0.00688184
161	chr1	3950001	3975000	1	0.01313399	0.01255869	0.01380397
162	chr1	3975001	4000000	9	0.00705747	0.00737391	0.00957724
163	chr1	4000001	4025000	2	0.00733942	0.00668748	0.0102754
164	chr1	4025001	4050000	2	-0.0010518	-0.0035599	0.0143271
165	chr1	4050001	4075000	0	0.00108464	-0.00012	0.00515088
166	chr1	4075001	4100000	40	0.00338066	0.00370063	0.00582645
167	chr1	4100001	4125000	0	0.00375977	0.00383547	0.01018714

Figure 1–Figure supplement 3. Preview of supplementary Table 2. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: CAS.

	A	B	C	D	E	F	G
1	dK80 values and total tree length for x (FRA), y(GER), z(IRA) and O(SPRE)						
2	CHR	START	END	MISSING	dK80_FRA	dK80_GER	TREELENGTH
151	chr1	3700001	3725000	73	0.01075072	0.01182333	0.014129456
152	chr1	3725001	3750000	99	0.01035236	0.01301701	0.017176465
153	chr1	3750001	3775000	219	0.01969981	0.02024767	0.022397307
154	chr1	3775001	3800000	447	0.01532601	0.01662267	0.024770115
155	chr1	3800001	3825000	7859	0.0110984	0.01136991	0.018444923
156	chr1	3825001	3850000	845	0.01531886	0.01547036	0.022584419
157	chr1	3850001	3875000	62	0.0186267	0.01853943	0.020934784
158	chr1	3875001	3900000	35	0.01369801	0.01368472	0.015527894
159	chr1	3900001	3925000	91	0.01455824	0.01495146	0.017218374
160	chr1	3925001	3950000	52	0.01576627	0.01531267	0.018615489
161	chr1	3950001	3975000	174	0.01701792	0.01647532	0.017720405
162	chr1	3975001	4000000	87	0.01690763	0.01726162	0.019451045
163	chr1	4000001	4025000	88	0.01547185	0.01447689	0.018244756
164	chr1	4025001	4050000	12	0.00933893	0.00820952	0.022272263
165	chr1	4050001	4075000	240	0.01768169	0.01711415	0.02175166
166	chr1	4075001	4100000	2184	0.01352769	0.01432316	0.016338947
167	chr1	4100001	4125000	57	0.01441057	0.01463817	0.021437542

Figure 1–Figure supplement 4. Preview of supplementary Table 3. dK80 values for all genomic 25kb windows for the quartet [X]: Fra [Y]: Ger [Z]: Ira [O_1]: SPRE.

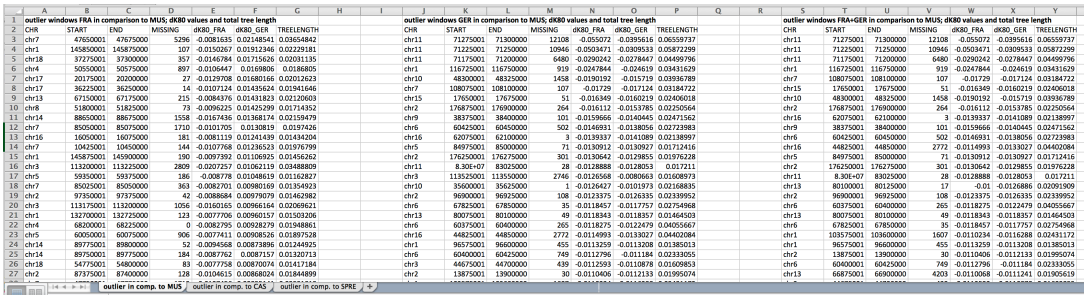

Figure 1–Figure supplement 5. Preview of supplementary Table 4. Lists of outlier windows for the three quartet comparisons.

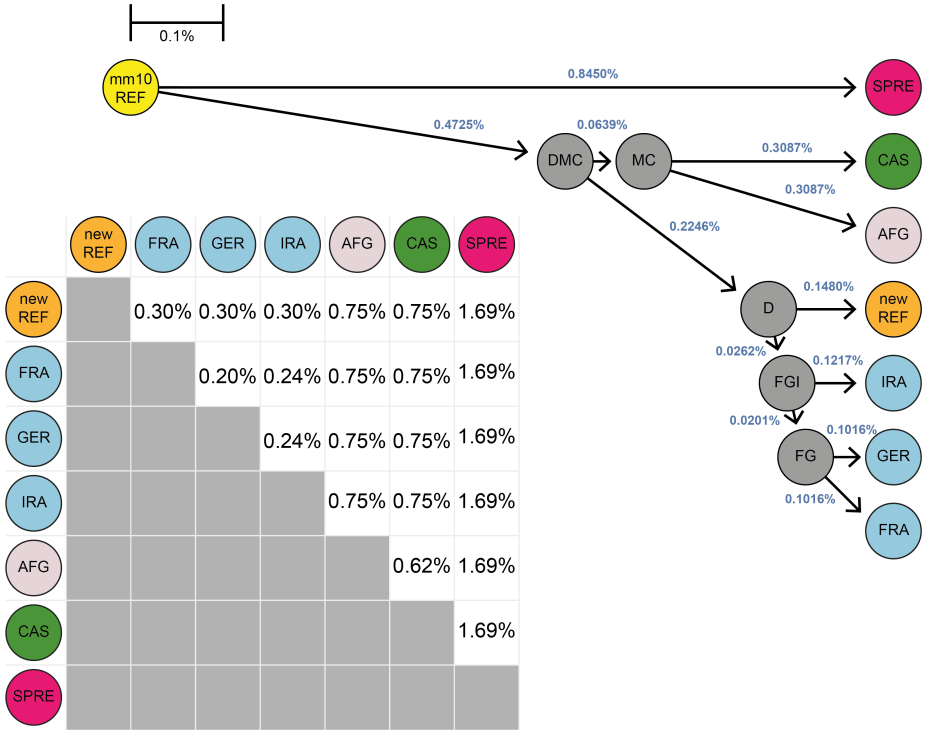

Figure 1–Figure supplement 6. Scheme of the simulation approach. Taking the mm10 reference sequence (yellow) as a start point, genomes were constructed in a phylogenetic context mimicking the real data including the construction of a 'new' reference (orange). Nucleotides were randomly altered given a percentage divergence value including ancestral states (grey). The resulting distances represents the phylogenetic context obtained as described in the method section.

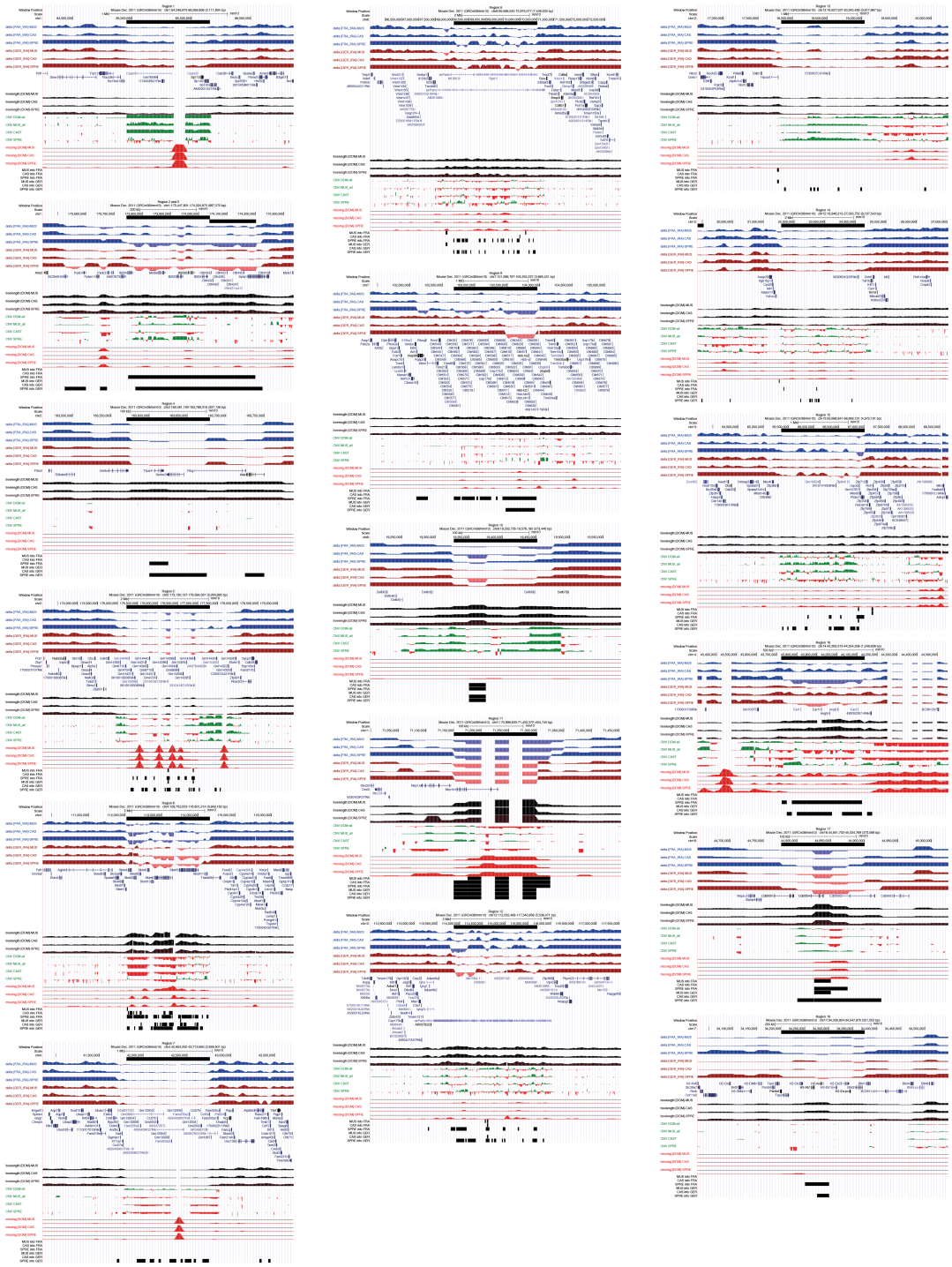

Figure 4–Figure supplement 1. Collection of screen shots of browser windows for all mutual introgression regions listed in Table 2. For full track description see legend of Figure 2.

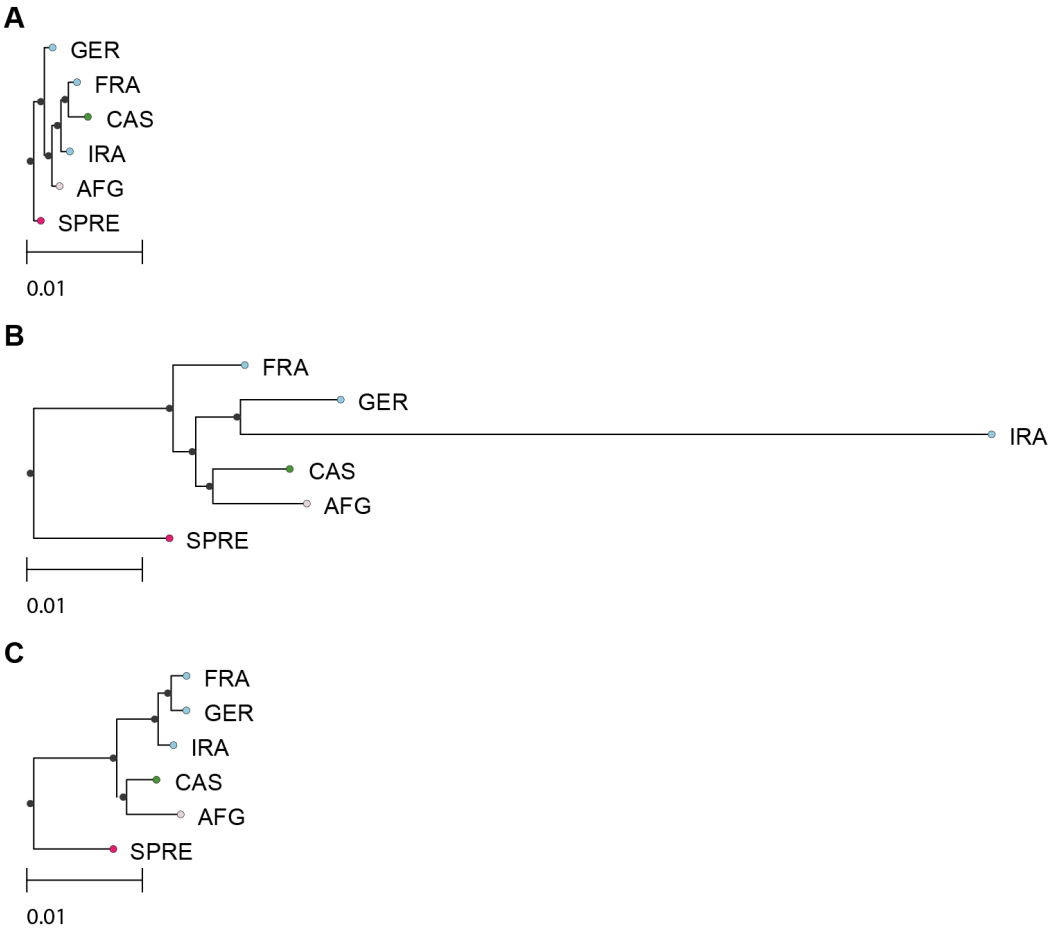

Figure 4–Figure supplement 2. Neighbor-joining (NJ) tree comparisons with consensus sequences in extreme introgression regions. (A) Region 11 from Table 2 (tree constructed for chr11:71,150,213-71,301,792) including *Nlrp1b*. (B) Region 7 from Table 2 (tree constructed for chr4:41,807,517-42,760,683) including a chemokine ligand cluster. (C) Standard tree structure represented by a tree of whole chr19. NJ (***Saitou and Nei, 1987***) trees based on pair-wise K80 distances (***Kimura, 1980***) with the R ape package (***Paradis et al., 2004***). Prior the NJ tree calculation all masked sites were removed from all included sequences.

	A	B	C	D	E	F	G	H
1	overlap with regions in Table2	genomic position	size (MB)	dK80	tree length	significant windows	annotated coding genes	general classification
2		chr1:116,581,664-117,036,128	0.45	-3.99	12.94	yes	none	non-coding
3	2	chr1:173,615,452-174,039,180	0.42	0.37	17.85	yes	Pyhin1, Pydc3, AI607673, Ifi204, Mndal, Ifi203, Ifi202b, Ifi205	immunity (innate)
4	3	chr1:174,036,870-174,202,341	0.17	0.39	26.04	yes	Olfr receptor cluster	sensory
5		chr1:49,248,596-49,400,827	0.15	-0.62	9.99	no	C230029F24Rik	unknown
6	1	chr1:84,950,951-85,654,928	0.7	1.01	2.80	no	Csprs, Sp110, Sp140	immunity (innate)
7		chr2:126,649,041-126,777,659	0.13	0.31	8.44	no	Usp8	immunity (adaptive)
8	4	chr2:163,570,927-163,750,004	0.18	0.15	17.60	yes	Ttpal, Serinc3; Pkig	immunity (adaptive)
9	5	chr2:174,938,301-177,965,862	3.03	-0.02	2.67	yes	cluster of Zf-C2H2 predicted proteins	immunity (transposon)
10		chr2:77,872,657-78,026,277	0.15	8.31	9.32	yes	CWC22	meiotic drive
11		chr3:105,996,817-106,452,618	0.46	0.04	4.75	no	Pifo and Chil3 - gene cluster	immunity (innate)
12		chr3:112,991,969-113,581,063	0.59	-2.91	16.26	yes	amylase gene cluster	metabolism
13		chr3:134,000,770-134,100,665	0.1	-3.38	10.80	yes	none	non-coding
14		chr3:144,725,770-145,097,656	0.37	1.52	13.74	yes	Clca gene cluster	immunity (innate)
15		chr3:48,847,480-49,123,658	0.28	0.32	9.57	yes	none	non-coding
16		chr3:92,224,164-92,375,422	0.15	-2.37	11.70	no	Sprr gene cluster	immunity (innate)
17	6	chr4:112,035,107-114,318,160	2.28	-1.63	26.70	yes	Skint gene cluster	immunity (adaptive)
18		chr4:145,437,109-147,825,549	2.38	0.56	5.53	no	cluster of Zf-C2H2 predicted proteins	immunity (transposon)
19	7	chr4:41,807,517-42,760,683	0.95	0.18	1.21	yes	Ccl - chemokine ligand gene cluster	immunity (adaptive)
20		chr4:59,945,328-62,187,387	2.24	1.91	3.78	no	MUP gene cluster	pheromone
21		chr5:104,450,295-104,701,942	0.25	-0.82	10.73	yes	Pkd2, BC005561, AK049668	metabolism
22		chr5:109,000,917-109,476,050	0.47	0.98	5.61	yes	vomeronasal receptor cluster	sensory
23		chr5:109,500,815-109,604,610	0.1	-1.95	10.58	yes	Crlf2	immunity (adaptive)
24		chr5:109,819,062-110,070,047	0.25	-1.40	12.20	no	Zfp932	regulatory
25		chr5:14,904,297-15,721,130	0.82	2.15	6.20	no	Takusan domain gene cluster	testis
26		chr5:8,172,153-8,329,824	0.16	-1.31	8.19	no	Adam22 (partial)	unknown
27		chr5:93,448,126-95,827,540	2.38	0.96	1.55	no	cluster of predicted genes	unknown
28		chr6:131,023,045-131,276,322	0.25	-0.33	6.68	no	Klra2	immunity (adaptive)
29		chr6:60,348,275-60,452,556	0.1	-10.10	31.29	yes	none	non-coding
30	8	chr6:68,036,056-70,505,566	2.47	1.54	15.96	yes	abParts	immunity (adaptive)
31		chr7:102,099,272-102,276,909	0.18	1.79	14.84	yes	Chrna10, Nup98, Pgap2, Rhog	immunity (innate)
32	9	chr7:102,472,207-103,966,610	1.49	0.88	15.79	yes	Olfr receptor cluster, Hbb genes	sensory
33		chr7:105,869,048-106,584,033	0.72	0.26	6.64	yes	Gvin1 gene cluster	immunity (adaptive)
34		chr7:107,968,504-108,106,027	0.14	-0.84	10.34	yes	Olfr receptor cluster	sensory
35		chr7:20,013,374-23,334,034	3.32	1.63	3.05	no	vomeronasal receptor cluster	sensory
36		chr7:48,146,207-48,278,851	0.13	-3.58	16.35	yes	Mrgprb4, Mrgpr5	sensory
37	10	chr8:19,327,217-19,451,698	0.12	-3.77	25.19	yes	Defb8	immunity (innate)
38		chr8:19,676,960-20,814,514	1.14	1.88	3.29	no	predicted genes	unknown
39		chr8:24,627,009-24,885,472	0.26	0.76	5.84	no	ADAM gene family	testis
40		chr10:129,673,649-129,776,654	0.1	-2.73	9.56	yes	Olfr receptor cluster	sensory
41		chr10:81,700,894-81,975,827	0.27	0.86	4.29	no	cluster of Zf-C2H2 predicted proteins	immunity (transposon)
42		chr11:54,494,130-54,704,357	0.21	1.55	2.93	no	Rapgef6	metabolism
43	11	chr11:71,142,650-71,347,510	0.2	-26.17	44.11	yes	Nlrp1b	immunity (innate)
44		chr11:82,942,065-83,247,359	0.3	-0.04	7.14	yes	Slfn gene cluster	immunity (adaptive)
45	12	chr12:113,866,694-115,905,783	2.03	0.89	18.74	yes	abParts	immunity (adaptive)
46	13	chr12:17,833,516-19,039,504	1.2	0.83	3.29	yes	5730507C01Rik	immunity (transposon)
47	14	chr12:21,575,393-24,304,572	2.73	1.33	4.85	yes	9030624G23Rik	unknown
48		chr12:37,121,276-37,252,097	0.13	-1.32	9.58	no	Meox2, Agmo (both partial)	regulatory
49		chr13:33,206,250-34,004,632	0.8	-0.45	10.68	yes	Serpinb gene cluster, NQO2	immunity (adaptive)
50	15	chr13:65,354,338-67,011,734	1.7	1.05	8.66	yes	cluster of Zf-C2H2 predicted proteins	immunity (transposon)
51		chr13:85,949,160-86,052,577	0.1	-3.22	16.97	no	Cox7c	metabolism
52		chr14:19,467,178-19,606,600	0.14	1.09	1.56	no	Takusan domain gene cluster	testis
53		chr14:3,052,032-7,785,533	4.73	1.70	2.67	no	Takusan domain gene cluster	testis
54		chr14:41,224,160-43,667,865	2.44	2.26	4.60	no	Takusan domain gene cluster	testis
55	16	chr14:43,721,785-44,137,996	0.42	2.33	23.56	yes	Ear gene family	immunity (innate)
56	17	chr16:44,776,058-44,900,413	0.12	-3.68	25.54	yes	Cd200r genes	immunity (innate)
57		chr16:48,919,828-49,078,537	0.16	-0.20	5.49	no	Dzip3, Cip2a	immunity (adaptive)
58		chr16:56,922,954-57,099,170	0.18	-1.75	10.41	yes	Tmem45a2	unknown
59		chr16:61,970,281-62,128,764	0.16	-2.63	14.68	yes	none	non-coding
60		chr17:27,122,393-27,373,816	0.25	0.05	7.45	no	Uqcc2, Ip6k3, Lemd2, Gm10505	metabolism
61	18	chr17:33,407,557-36,293,253	2.89	1.06	7.33	yes	extended MHC region	immunity (adaptive)
62		chr17:37,574,925-37,853,356	0.28	0.55	5.51	no	Olfr receptor cluster	sensory
63		chr17:57,423,660-57,552,843	0.13	-2.55	8.10	yes	Adgre1 (partial), Vmn2r120	immunity (adaptive)
64		chr19:12,922,815-13,051,570	0.13	-0.77	9.95	no	Olfr receptor cluster	sensory
65		chrX:27,201,903-30,998,527	3.8	1.27	1.47	no	Slx gene cluster	testis
66		chrX:3,181,827-4,950,229	1.8	1.48	2.48	no	BTB domain predicted protein cluster	testis
67		chrX:31,217,092-34,617,228	3.4	1.98	2.61	no	BTB domain and Spindlin predicted protein cluster	testis
68		chrX:81,098,297-81,276,935	0.18	-1.24	10.18	no	none	non-coding
69								

Figure 4–Figure supplement 3. Preview of suppl. Table 5. This table represents an extended version of Table 2 including all identified introgression regions between subspecies and species, as well as the respective genomic locations.