Genomics 95 (2010) 196–202

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

The consequence of natural selection on genetic variation in the mouse

Eli Reuveni [a,*], Ewan Birney [b], Cornelius T. Gross [a]

[a] Mouse Biology Unit, European Molecular Biology Laboratory (EMBL), via Ramarini 32, 00015 Monterotondo, Italy
[b] European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL), Wellcome Trust Genome Campus, Hinxton, UK

ARTICLE INFO

Article history:
Received 22 October 2009
Accepted 10 February 2010
Available online 18 February 2010

Keywords:
Mouse evolution
Natural selection
Laboratory mouse strains
SNP
Phylogenetic reconstruction

ABSTRACT

Laboratory mouse strains are known to have emerged from recent interbreeding between individuals of *Mus musculus* isolated populations. As a result of this breeding history, the collection of polymorphisms observed between laboratory mouse strains is likely to harbor the effects of natural selection between reproductively isolated populations. Until now no study has systematically investigated the consequences of this breeding history on gene evolution. Here we have used a novel, unbiased evolutionary approach to predict the founder origin of laboratory mouse strains and to assess the balance between ancient and newly emerged mutations in the founder subspecies. Our results confirm a contribution from at least four distinct subspecies. Additionally, our method allowed us to identify regions of relaxed selective constraint among laboratory mouse strains. This unique structure of variation is likely to have significant consequences on the use of mouse to find genes underlying phenotypic variation.

Introduction

Mouse laboratory strains are thought to have emerged from the domestication of three wild-derived subspecies, *M. m. domesticus*, *M. m. musculus*, and *M. m. casteneus*, that diverged approximately one million years ago [1–3] into reproductively isolated populations undergoing independent speciation [4]. Some laboratory strains also include genetic material from *M. m. molossinus* that is a relatively recent natural hybrid of the *M. m. casteneus* and *M. m. musculus* subspecies [5]. Several studies have shown that the genome of mouse laboratory strains is a mosaic of regions with higher or lower variation depending on their founder origin [6] with the majority of variation contributed by *M. m. domesticus* [6–9].

The recent re-sequencing of over 14 laboratory and three wild-derived mouse strains [7,10], helped to identify SNPs that are likely to represent variation contributed by intra-subspecific (genetic variation contributed by individuals from the same population) and inter-subspecific (genetic variation contributed by individuals from distinct populations) origin. Using this dataset Yang et al. [9] estimated that 92% of the variation in laboratory mouse strains derives from variation within the *M. m. domesticus* founder sub-species, while Frazer et al. [7] estimated that only 68% derive from this sub-species and the remaining variation derives from up to three other founders. The discrepancy between the two studies could be explained by the fact that both studies made different *a priori* assumptions about the likelihood number of founder sub-species. While Frazer et al. [7] assumed four founders (*M. m. domesticus, M.m. musculus, M. m. casteneus,* and M. m. *molossinus*), Yang et al. [9] assumed three founders (*M. m. domesticus, M. m. musculus,* and *M. m. casteneus*). In addition, Frazer et al. [7] estimated that 20% of the SNPs present in laboratory mouse strains are invariant in the sequenced wild-derived strains. This estimate would suggest that the origin of a large proportion of laboratory mouse SNPs are unaccounted for in the sequenced wild-derived strains.

Several origins for the unaccounted for SNPs are possible, including 1) new SNPs emerged since the creation of laboratory strains, 2) additional founder sub-species contributed SNPs to laboratory strains, and 3) rare SNPs from founder species became incorporated into laboratory strains. The first possibility is unlikely given the relatively short time since the establishment of laboratory strains. So far, however, no study has attempted to examine the evolution of mouse SNPs in order to help distinguish between the second and the third possibilities. SNPs that have been present within a founder sub-species for a significant period of time should bear the hallmarks of natural selection, in that deleterious mutations are far less likely to be fixed, whereas neutral and beneficial mutations will be fixed at a rate proportional to the population size. Several tools have been used to assess the relative rate of fixation of deleterious compared to neutral mutations over population genetics time scales, such as the McDonald-Kreitman [11] test, and, over evolutionary time scales, the relative rate of non-synonymous compared to synonymous changes in protein coding genes [12].

In this work we have relied on the ratio between synonymous (d_N) and non-synonymous (d_S) substitution rates, $\omega = d_N/d_S$ [13]. The d_N/d_S ratio can be used as an indication of the relative rate of selection on a protein coding gene, with $d_N/d_S = 0$ indicating that no non-synonymous mutations occurred in this gene and $d_N/d_S = 1$ indicating

[*] Corresponding author.
E-mail address: reuveni@embl.it (E. Reuveni).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.02.004

that non-synonymous mutations occurred at the same rate as synonymous mutations [14]. In the context of a mouse population, if a SNP were to have been introduced in a founder sub-species early during the divergence to distinct populations, natural selection will have been effective in removing or favoring, respectively, disadvantageous or advantageous SNPs. If, on the other hand, a SNP were to have been recently introduced in a founder sub-species, in many cases insufficient time would have elapsed for negative selection to remove the mutation from the population. In the first case, we would expect to find reduced d_N/d_S levels compared to the second scenario. Thus, an analysis of the d_N/d_S ratio for laboratory mouse SNPs could help to discriminate between haplotypes coming from within a founder sub-species and those coming from different founders.

The existence of genomic sequences for two laboratory rat inbred strains offers the possibility to examine SNP evolution within a closely related rodent species [15,16]. Because rat inbred strains are thought to derive from a single population, these strains should not show separable inter and intra-subspecific genetic variation. Moreover, rats are likely to show a different distribution of high and low d_N/d_S ratio given their populations origin and a comparison of SNP variation between mouse and rat laboratory strains could help to identify ancestral genetic variation and test the hypothesis that three founder sub-species contributed to laboratory strains.

Here we use this novel, evolution-based approach to predict the founder origin of laboratory mouse SNPs. The method uses clustering approaches to distinguish SNPs that differ between and among putative mouse founder sub-species and consider bias on SNP frequency due to natural selection mechanisms. This approach differs from those of previous published mouse SNP analyses because it does not make prior assumptions regarding the number of parental strains and thus allowed us to draw conclusions about the 20% monomorphic SNPs [7] in the ancestral origin. Our analysis is consistent with the mosaic model and supports the existence of at least four sub-species origin. Next, we demonstrate that SNPs deriving from inter-subspecific variation show evidence of purifying selection, suggesting that the majority of them were introduced to laboratory mouse early during the divergence into the founder populations while small fraction points to a recent expansion of polymorphisms which are more relaxed to adaptation. These findings reconcile discrepancies about the origin of laboratory mouse genetic variation deriving form haplotype analyses and support the existence of significant inter-subspecific genetic diversity in the laboratory mouse.

Materials and methods

Dataset assembly

Mouse and rat coding sequences were selected from the Ensembl annotation system (v50, July 2008 and dbSNP v126). For each transcript a non-redundant Multiple Sequence Alignment (MSA) was reconstructed according to the variation information assigned separately for each one of the 14 mouse laboratory haplotypes re-sequenced by Celera [10] (4 laboratory inbreds)(129X1/SvJ, 129S1/SvImJ, A/J, DBA/2J) and Perlegen [7] (10 laboratory inbreds) (BTBR T + tf/J2, A/J, KK/HlJ3, AKR/J, NZW/LacJ4, BALB/cByJ, C3H/HeJ, DBA/2J, FVB/NJ, NOD/LtJ) and 3 wild-derived haplotypes (CAST/EiJ [*M. m. castaneus*], WSB/EiJ [*M. m. domesticus*], PWD/PhJ [*M. m. musculus*]). We have mapped nucleotide variation from laboratory and wild-derived strains to the reference strain C57BL/6J [17] excluding strains for which no re-sequencing information was available at that position. Rat coding sequences and SNPs were obtained from the Ensembl annotation system (v47, July 2008 and dbSNP 126). We have used the reference strain Brown Norway (BN) as a template for the assignment of nucleotide variation with the Sprague Dawley (SD) strain. In order to avoid bias, only transcripts with at least two SNPs were included in the analysis. All data were stored and annotated using a MySQL database with custom Perl and Java programs.

Reconstruction of mouse ancestry using single-linkage clustering

We used single-linkage clustering to map ancestry for each gene in the following manner. First, we calculated pair-wise π values for each gene between haplotypes and defined a threshold value of $\pi = 0.0005$ (0.05% difference between sequences) as a hallmark to test divergence (e.g., pairwise sequences with $\pi < 0.0005$ were considered to be derived from intra-subspecific variation while pair-wise sequences with $\pi > 0.0005$ was considered to be inherited from different founders), see Results and Discussion. Next, we applied single-linkage clustering individually for each gene and merge each pairwise distance between segregate haplotype into a single cluster until a maximum distance of $\pi_t = 0.0005$ (Fig. 1). Single-linkage clustering was performed in the following manner: for each gene we calculated the minimum distance, $\min(d[h1, h2])$, between haplotypes $h1 \in X$ and $h2 \in Y$ according to the rule that $\{h1 \in X \text{ and } h2 \in Y{:}\min(d(h1, h2)) > \pi t\}$. The haplotypes *h1* and *h2* were merged into a single cluster, *X*, when $\{h1 \in X \text{ and } h2 \in X{:}\ \min(d(h1, h2)) < \pi t\}$, where *X* and *Y* are different clusters for the same gene.

Evolutionary analysis

PhyML program [18] was used for the reconstruction of polygenetic trees for each mouse gene with more than two clusters. In order to check whether a selected gene was subjected to evolutionary constraints we used the non-synonymous (d_N) and synonymous (d_S) substitution rate and the ratio (ω) calculated in the PAML package V3.15 [13,19], using the yn00 program in the case of pairwise comparisons and the codeml program in the case where more than two sequences were available. For the codeml analysis a single ω ratio for all sites (NSsites = 0) and a single ω ratio for all lineages (Model = 0) was used. We calculated the d_N/d_S statistic for genes within three different groups of strains, 1) laboratory mouse inbred strains (to avoid redundancy, calculations were performed on one randomly chosen representative haplotype from each single-linkage clusters), 2) wild-derived mouse inbred strains, and 3) laboratory rat inbred strains. Genes with d_N or $d_S = 0$ were excluded. For genes with

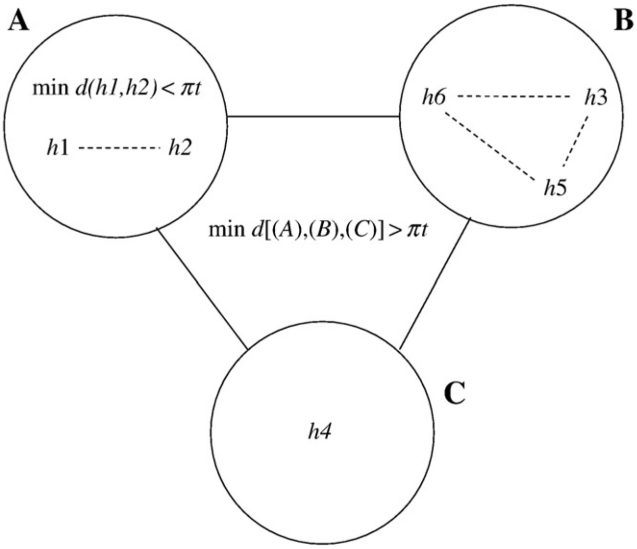

Fig. 1. Single-linkage clustering of mouse gene haplotypes. The figure illustrates a hypothetical single-linkage clustering for a gene with six haplotypes (*h*1-6) and a given π cutoff (πt = 0.0005). Each haplotype represents one inbred strain. Haplotypes with differences below the cutoff are assigned to the same cluster (circle). Dashed and straight lines represent distance among (intra-subspecific) and between (inter-subspecific) clusters correspondingly. In this hypothetical example the six haplotypes of this gene mapped to three clusters (*A*, *B* and *C*) one of which contains a single haplotype (*h*4).

alternatively spliced transcripts we calculated the average d_N/d_S for all transcripts.

For admixture analysis, we calculated d_N/d_S distributions for 1000 composite sets of genes generated by the union of iterated sampling of wild-derived mouse (f_{WD}) and rat (f_{RAT}) genes. An admixture vector, $f_\theta = \alpha f_{WD(c)} + \beta f_{RAT}$ (where C = number of clusters, C = 2, C = 3, or C>3, and α and β are the proportions of wild-derived and rat genes in increments of 0.5% where α + β = 100%) was calculated. Statistical significance was tested by comparing each iterated admixture gene set (f_θ) with the genes set of the laboratory strains using the Kolmogorov-Smirnov test. α and β values returning the highest P value (P_{MAX}) were selected.

Results

Distribution of genetic variation among mouse and rat inbred strains

In order to assess the distribution of genetic variation on a gene-by-gene basis, we calculated the average number of pairwise SNP differences between inbred strains for each transcript (π, [20]) within three populations: 1) laboratory mouse inbred strains (14 strains), 2) wild-derived mouse inbred strains (3 strains), and 3) laboratory rat inbred strains (2 strains). Our initial analysis was restricted to synonymous SNPs in order to avoid bias due to the effects of natural selection. We expected that variation between laboratory mouse inbred strains would be less than between wild-derived mouse inbred strains and more than between laboratory rat inbred strains. Frequency plots of the distribution of genetic variation in these three populations were generally consistent with our hypotheses. While laboratory mouse inbred strains illustrated a bimodal distribution of pair-wise SNP variation with two identifiable peaks (called π_1 and π_2), wild-derived mouse inbred strains showed a single prominent peak roughly overlapping with π_2 of the laboratory strains (Figs. 2AB). The bimodal distribution of laboratory mouse inbred strain variation is possibly due to the contributions of intra and inter-subspecific SNPs, represented by π_1 and π_2, respectively. This hypothesis is supported by the apparent existence of only a single dominant peak corresponding to π_2 in laboratory rat strains that derive from a single founder. Similarly, the absence of a peak corresponding to π_1 in wild-derived mouse strains supports an intra-subspecific origin for this peak. Furthermore, the cutoff that distinguishes the bimodal distribution in our mouse data, $\pi_t = 0.5 \times 10^{-3}$, is identical to the frequency cutoff found to distinguish low and high polymorphic genomic regions in other studies (1 SNP each 200 bases, [6–8]). Finally, we also reached a similar conclusion when using only genes harboring precisely three SNPs, suggesting no bias due to variation in the number of SNPs per transcript (Fig. S1).

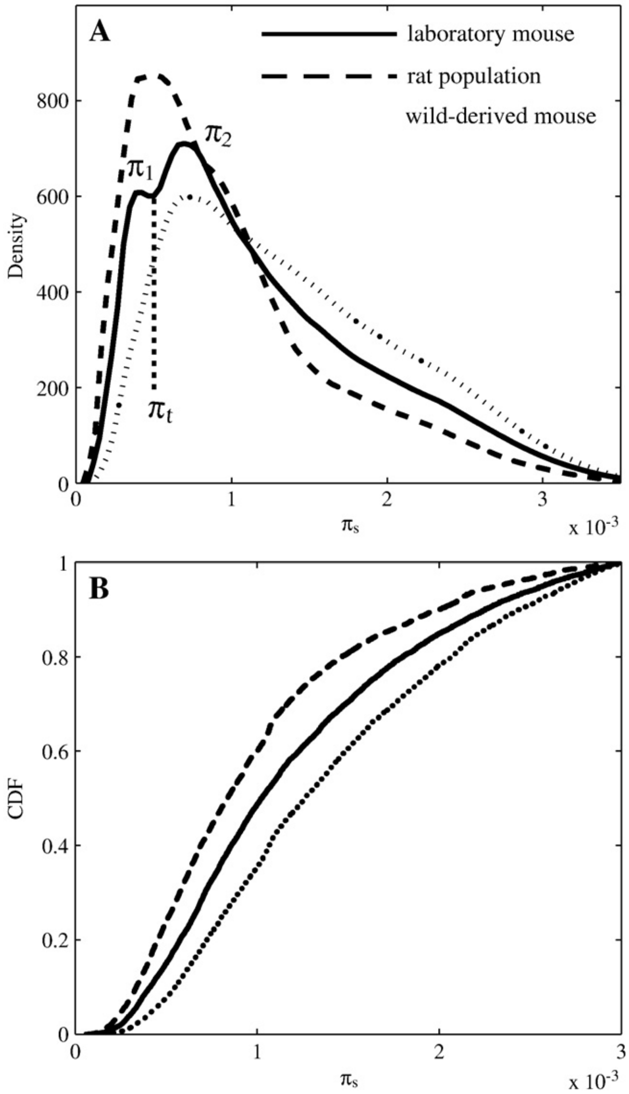

Fig. 2. Distribution of genetic variation within mouse and rat populations as assessed by pair-wise SNP differences. Graphs represent frequency distributions of gene-by-gene pairwise SNP differences within laboratory mouse inbred (solid line), wild-derived mouse inbred (dotted line), and laboratory rat inbred (dashed line) strains. Frequency (A) and cumulative distribution (B) of variation measured using signals neutral (synonymous) SNPs.

Single-linkage clustering identifies mouse inter-subspecific SNPs

The distribution of pair-wise SNP differences in laboratory mouse inbred strains suggests that variation within and between populations origin can been classified using a cut-off of $\pi_t = 0.5 \times 10^{-3}$ (Fig. 2A, π_t = 1 SNP every 200 nt). To distinguish intra from inter-subspecific SNPs we performed single-linkage clustering of gene haplotypes. This type of clustering ensures that all gene haplotypes where at least one strain pair differs by less than the π_t are grouped together, while those that differ by more than π_t are grouped in different clusters (Fig. 1). Given our data from pair-wise comparisons above, haplotypes classified to be within a single cluster could be assumed to reflect intra-subspecific variation. For this analysis we selected 2980 mouse genes with at least 2 SNP's from the Ensembl database. In order to make sure that the analysis was not biased by gene diversity we chose only genes for which re-sequencing information was available for both laboratory and wild-derived strains. This gene set was used in all subsequent analyses.

Single-linkage clustering revealed that 27% of genes were singletons (all haplotypes collapsed to only one cluster), 65% were mapped to two or three clusters, and 8% were mapped to more than 3 clusters (Table 1). The observation that the majority of genes (92%) contained three or fewer clusters is consistent with our hypothesis that the $\pi_t = 0.5 \times 10^{-3}$ cutoff can be used to identify haplotypes deriving from the three putative founder sub-species. However, the origin of the 8% of genes that contained more than three clusters is not clear. Several possible origins exist for the large genetic variation in these genes. One possibility is that additional, as yet undescribed founder sub-species contributed to genetic variation in laboratory strains. A second possibility is that within-founder variation for this set of genes is higher than expected and represents a rapidly expanding population of SNPs. A third possibility is that these high cluster genes represent false-positives that cannot be collapsed into 3 or fewer clusters due to missing haplotypes not captured in sequenced laboratory inbred strains.

Table 1
Single-linkage cluster dimensions of genes from mouse inbred strains. Only genes with at least 2 haplotypes and for which re-sequencing information was available for both laboratory and wild-derived strains (N = 2980) were subjected to single-linkage clustering. (Top) Number of genes having 2, 3, and >3 haplotypes. (Bottom) Number of genes showing 1 (singleton), 2, 3, and >3 clusters.

Prior to clustering		
# of haplotypes	# of genes	% from the total
2	1762	59%
3	948	31%
>3	270	9%
Post clustering		
Cluster size	# of genes	% from the total
Singleton	778	27%
2	1261	42%
3	744	23%
>3	197	8%

Mouse SNPs show signs of purifying selection

One distinguishing feature of genetic variation in isolated populations is its exposure to negative, or purifying, selection. Thus, mouse inter-subspecific genetic variation should show signs of purifying selection while those from the same population origin such as mouse intra-subspecific haplotypes or rats should show fewer signs of natural selection. In general, variation between isolated populations show strong purifying selection and low d_N/d_S values, while strains from the same population harbor a significant number of variants that have not yet undergone selection and show higher d_N/d_S values. Thus, we reasoned that the d_N/d_S ratio might provide additional information about whether the variation seen among mouse inbred strains derives from variation between or within populations.

We calculated the d_N/d_S ratio for the same set of genes as used above for the single-linkage clustering. For the laboratory mouse (inter-cluster), we calculated d_N/d_S using one randomly selected, representative haplotype from each cluster. In this way, we examined evidence for natural selection in putative inter-subspecific variation (see Materials and Methods). Cumulative frequency plots demonstrated that both wild-derived mouse strains and laboratory mouse clusters have d_N/d_S distributions that are significantly shifted toward smaller values than the rat (laboratory clusters vs. rat: $P = 2\times10^{-75}$, wild-derived mouse vs. rat: $P = 1\times10^{-66}$, Kolmogorov-Smirnov test) suggesting greater negative selection in the mouse (Fig. 3). The absence of a large difference in d_N/d_S distribution between laboratory clusters and wild-derived mouse strains is consistent with their common origin from isolated populations of mouse under similar selective pressure. Our results suggest that the rat contains a significantly greater number of genes containing recent mutations that have not yet been eliminated by natural selection and are

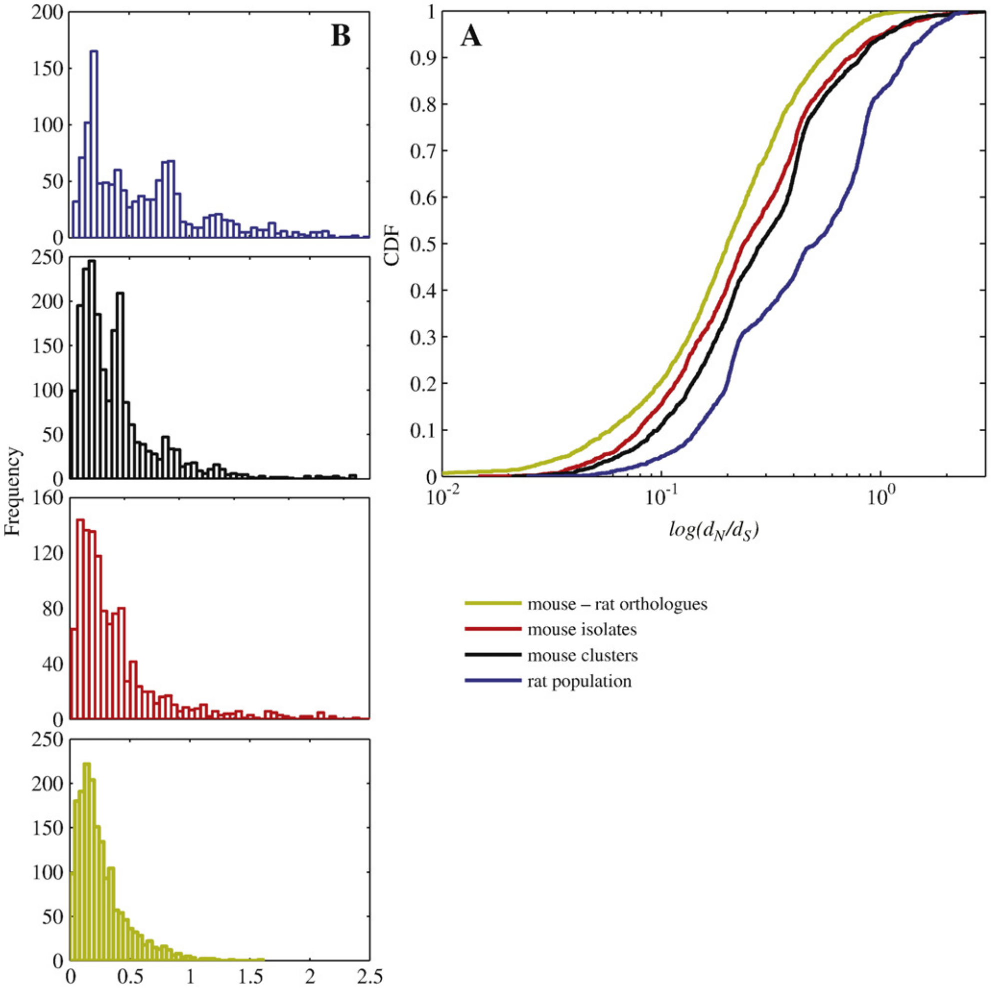

Fig. 3. Evidence for purifying selection in the mouse. Cumulative fraction plot (A) and frequency histograms (B) of d_N/d_S distributions for genes of laboratory inter-clusters and wild-derived mouse inbred strains compared with laboratory rat strains and mouse-rat orthologuous gene pairs (laboratory clusters vs. rat: $P = 1\times10^{-66}$, wild-derived mouse vs. rat: $P = 2\times10^{-75}$, laboratory clusters vs. wild-derived mouse: $P = 6\times10^{-6,}$ wild-derived vs. mouse-rat orthologues: $P = 1\times10^{-13}$, Kolmogorov-Smirnov test). The mouse d_N/d_S distributions tended to be more uniform shifted to the left, while rat d_N/d_S distribution exhibited a bimodal distribution with a large fraction of genes with $d_N/d_S \approx 1$.

consistent with a recent origin of laboratory rat strains from a single population origin. Low d_N/d_S among mouse strains, in turn, reflects their origin from isolated populations. However, although the laboratory and wild-derived mouse distributions were similar, a statistical comparison revealed that the laboratory clusters showed a small, but significant shift toward higher d_N/d_S values ($P=6\times10^{-6}$; Kolmogorov-Smirnov test). This shift may reflect contamination of inter-cluster variation by intra-subspecific variation as discussed above. To rule out possible confounds due to our use of a set of non-overlapping rat and mouse genes we repeated the analysis on a smaller subset of rat and mouse orthologues (N = 200). This analysis revealed similar d_N/d_S distributions as reported above and argued against a bias in our selection of genes (Fig. S2 and Table S1). Finally, we found that d_N/d_S calculated for SNP differences between rats and mice (across species d_N/d_S ; Fig. 3; N = 2202 orthologous genes) was significantly shifted toward lower values. This finding is consistent with the relatively long time since divergence of these species (~20 million years) and confirms previous studies showing that d_N/d_S correlates with the time of divergence of populations [14,21].

Large clusters show signs of purifying selection

The existence of strong signs of purifying selection in the mouse suggests that the vast majority of variation in laboratory inbred strains derives from inter-subspecific variation. This conclusion has important implication for the origin of genetic variation in the 8% of genes that show greater than three clusters (Table 1). If this variation were to derive from a contribution of additional founder sub-species, d_N/d_S should remain low for this group of genes. If, however, this variation were to derive from rapidly expanding intra-subspecific variation, d_N/d_S should be high.

In order to quantify d_N/d_S distributions for laboratory mouse clusters of different dimensions, we used a mixed-model reiteration technique (see Materials and Methods). We attempted to model the d_N/d_S distribution of laboratory clusters by a variable admixture of rat and wild-derived d_N/d_S distributions. For each series of rat and wild-derived distributions (from 0.5% rat/99.5% mouse genes to 99.5% rat/0.5% mouse genes in 0.5% increments) we calculated the maximum *P* value (P_{MAX}) for its fit to the laboratory cluster distribution. Table 3 shows that for genes with two clusters the d_N/d_S distribution is poorly modeled by an admixture ($P_{MAX}=0.06$), while for genes with 3 or more clusters the admixture model is good ($P_{MAX}=0.95$) and the best fit is achieved with 85–90% wild-derived genes. This modeling demonstrates that the d_N/d_S distribution of high (3 or more) cluster genes is indistinguishable from wild-derived strains and suggests that these high cluster genes do not show signs of rapidly expanding, intra-subspecific variation.

To rule out bias due to the use of different subsets of genes, we also calculated the *P* value between d_N/d_S distributions for genes of each cluster dimension and the same set of genes from wild-derived strains (Table 2). The d_N/d_S distribution of 2 cluster genes was significantly different compared to the wild-derived d_N/d_S distribution ($P=2\times10^{-5}$, Table 2) while those of higher cluster genes were much less significantly different. This observation supports our finding that variation among haplotypes of highly variant mouse genes (those showing 3 or more clusters) show similar marks of natural selection as wild-derived variation and thus are not likely to be the result of a recent expansion in intra-founder variation. This observation is further supported by the fact that the distribution of d_N/d_S between cluster groups shows a tendency to be reduced as the number of clusters per gene increases (Fig. 4, Table 2).

Finally, we also examined d_N/d_S for variation within the singleton group. In order to gather evidence that haplotypes in the singleton group represent intra-subspecific variation we calculated intra-cluster d_N/d_S and compared them with the d_N/d_S distribution obtained for the same genes in the inter-cluster and wild-derived groups. Our results show that the singleton group had elevated d_N/d_S values compared to wild-derived strains ($N=89$; $P=6\times10^{-6}$) and inter-cluster groups ($N=64$; $P=0.05$). Moreover, intra-cluster

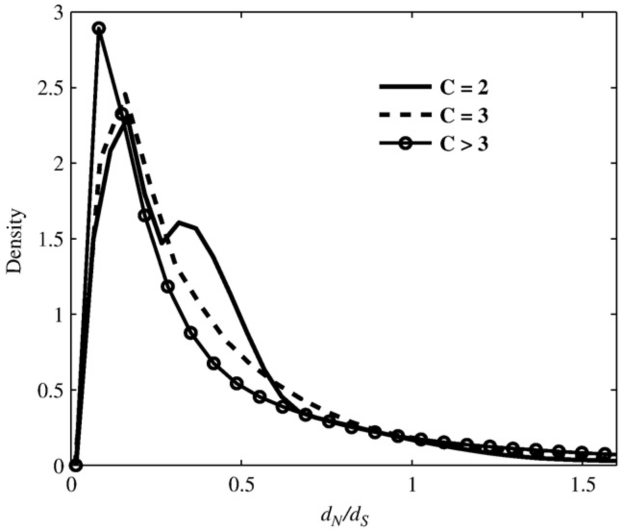

Fig. 4. Distribution of d_N/d_S for subsets of genes categorized by cluster dimension. Comparison of the distribution of d_N/d_S for genes with 2, 3, or >3 clusters as determined by single-linkage analysis showed a trend for a shift toward lower values as cluster size increased.

Table 2
d_N/d_S distribution of for genes with 3 clusters and more are similar to that of wild-derived strains. Modeling of laboratory mouse cluster d_N/d_S distribution with an admixture of wild-derived mouse and laboratory rat genes. P_{MAX} indicates the significance of fitting (Kolmogorov-Smirnov test) of the best model ($P_{MAX}>0.90$ suggests good fit) of all possible admixtures. *P* denotes the significance of fitting of the d_N/d_S distribution for genes of each cluster dimension to the d_N/d_S distribution of the same set of genes from wild-derived strains. For good fitting admixture models ($P_{MAX}>0.90$) the contribution of wild-derived d_N/d_S distribution is given (α).

Cluster size	P_{MAX}	p[a]	α[b]	Admixture
2	0.06	2×10^{-5}	-	yes
3	0.95	0.004	0.85	no
>3	0.95	0.02	0.90	no

[a] *P* - P value obtained from the intersection of cluster group and equivalent wild—derived genes.
[b] α - the fraction of the wild-derived population after admixture modeling.

Table 3
Enrichment of functional gene classifications according to cluster dimension. Significant over-representation of selected functional *Gene Ontology* (GO) categories was observed for laboratory mouse inbred strain genes showing either one or two single-linkage clusters. No enrichment was found in genes with three or more clusters (calculated using FatiGO).

	% genes enriched	*P*
2 clusters	**15**	
neurological process		0.02
sensory perception		8×10^{-5}
rhodopsin-like receptor activity		2×10^{-7}
olfactory receptor activity		7×10^{-8}
Singleton	**50**	
regulation of biological process		1×10^{-5}
multicellular organismal development		1×10^{-5}
cellular component organization and biogenesis		1×10^{-5}
anatomical structure development		4×10^{-5}
localization of cell		6×10^{-3}
macromolecule metabolic process		7×10^{-3}
cell adhesion		8×10^{-3}
cellular developmental process		8×10^{-3}
primary metabolic process		2×10^{-2}
regulation of a molecular function		2×10^{-2}

cumulative fraction of d_N/d_S was significantly lower compared to rat orthologues ($N=28$, $P=1\times10^{-4}$) suggesting that laboratory mouse strains derived from a relatively small pool of genetic variation. However due to the restricted number of equivalent genes in the inter-cluster and wild-derived gene list ($N=40$) we could not use it for further statistical analysis.

No functional enrichment in large cluster genes

One assumption of the multi-founder origin of genetic variation in laboratory mouse strains is that the genes contributed by each founder should be random and should not show any particular enrichment for a specific functional class. If, on the other hand, laboratory strain variation were to derive from intra-subspecific origin, we might expect functional enrichment. To test this hypothesis, we calculated the significance of enrichment of functional classes of genes using Fatigo [22]. We found that only genes with one (singleton) or two clusters were significantly enriched in specific functional classes (Table 3). Three conclusions can be drawn from these data. First, they confirm that singleton haplotypes are likely to represent intra-subspecific variation that has undergone functional selection. Second, genes with two clusters are likely to contain a significant number of haplotypes that may be clustered together and represent intra rather than inter-subspecific variation (Fig. 4, Table 2). This group of genes may represent recent expansion of SNPs cluster due to more relaxed evolutionary constraints. The fact that this cluster is enriched with olfactory receptor genes (Table 3) a family which is already known to evolve rapidly in the mouse lineage [23] make our results more reliable. This second finding is similar to the conclusion drawn from our analysis of d_N/d_S distributions (Table 2). Third, these data suggest that haplotypes showing 3 or more clusters are likely to represent a random mixture of inter-subspecific variation. Taken together, these results argue against a homogenous origin for haplotypes with more than three clusters and open the possibility of a contribution of additional founders to laboratory mouse variation.

Discussions

As noted by others, genetic variation in the mouse (3.1%, [24] is higher than that found in all other mammals studied to date and is similar to the variation found between primate species (e.g. orangutans vs. human: 3.08%, chimpanzee vs. human: 1.24%; [25]. The ability to breed mice carrying genetic variation deriving from reproductively isolated populations is unique and has proven to be a powerful tool to map phenotype-genotype associations. However, although the multi-founder origin of genetic variation in the mouse has been well documented, until now no study has examined the consequences of natural selection in the mouse.

Our findings using pairwise SNP comparisons and single-linkage clustering within coding sequences are consistent with previous findings from genomic haplotype analyses that estimated significant contributions from at least three founder sub-species [6–8]. however our results refute similar analyses that proposed a more homogenous, single founder (population) structure [9]. We also found that a majority of genes (92%) showed variation deriving from 3 or fewer founders (Table 1). Our study allowed us to draw several novel conclusions about the genetic history of the mouse. First, mouse sequences show signs of strong purifying selection. The strong downward shift in d_N/d_S distributions for both wild-derived and laboratory strains compared to the laboratory rat is consistent with their common origin from distinct populations. Notably, the low d_N/d_S distribution was particularly reflected in genes showing 3 or more clusters in our single-linkage analysis (Table 2). Together with the fact that genes in these clusters were not enriched for any functional classes, argues strongly that variation between clusters in these classes derives from between-populations variation. In contrast, genes with two clusters show evidence of recent expansion of polymorphisms with more relaxed evolutionary constraints given the following: 1) relatively high d_N/d_S distribution, 2) the fact that after modeling, this group shows an admixture behavior of inter-subspecific and population origin, 3) the observed enrichment of olfactory genes.

Notably, our data also raise the possibility that laboratory mice harbor haplotypes from more than three populations of sub-species. Over 8% of mouse genes showed more than three clusters (1.4% showed more than 4 clusters). One possible origin of these extra clusters is excess variation due to the recent introduction of novel haplotypes associated within a population origin. This is unlikely, however, because the low d_N/d_S values (Fig. 3) and absence of enrichment in functional classification of genes with more than 3 clusters (Table 3) reflects a within, and not between-founder origin. Interestingly, several other mouse species exist that live sympatrically with *M. musculus* sub-species but appear to be prevented from interbreeding with them (e.g., *M. spretus*, *M. spicilegus*, and *M. macedonicus* [26]. It may be that these or other species managed to contribute to the genetic variation that we see fixed in *M. musculus* laboratory strains. The existence of additional founders is also supported by genomic haplotype analyses. For example, Frazer et al. [7] found that 20% of laboratory mouse SNPs and 10% of the genome is monomorphic in the wild-derived strains. One potential source of genetic variation is the *M. m. molossinus* sub-species. However, this strain appears to be a recent hybrid of *M. m. casteneus* and *M. m. musculus* and haplotypes deriving from this sub-species would be expected to overlap with these subspecies in our clustering analysis.

Because rat and mouse may show different mutation spectra, the use of rats to model variation within a population may have biased our findings. However, it is likely that the mutation spectra of rat and mouse are similar, given their similar natural history, and the d_N/d_S based statistic, with the use of a localized synonymous rate should be able to factor out different mutation spectra. More problematic is our reliance on a subset of genes to perform clustering and d_N/d_S calculations owing to the limited SNP data available for the strains involved [9]. This confound is more difficult to assess and will await the sequencing of additional laboratory and wild-derived mouse lines.

Conclusions

In conclusion, we present a novel, evolutionary-based approach to reconstruct the historical relationships between closely and distantly related species and to reveal the specific origin of organisms with a mosaic structure. Our findings obtained from mouse support a predominantly four founder origin of laboratory mouse strains and resolve inconsistencies between previous analyses of the origin of mouse SNP variation using haplotype analysis [6–9]. Moreover, we found signs that natural selection has played a major role in purifying genetic variation in mouse subspecies and that these effects have been inherited by modern laboratory strains. Genetic variation in laboratory rat strains, on the other hand, showed few signs of natural selection and was consistent with their derivation from a population origin.

These findings demonstrate a unique spectrum of genetic variation in the laboratory mouse. It is interesting to consider how this might affect genotype-phenotype studies in this species. For example, the presence in laboratory mouse populations of many variants that survived selective removal may well be the explanation for the large phenotypic variation seen in laboratory inbred mice. At the same time, functional variations in the mouse may be less relevant to the type of functional variations seen in a normal outbreds and relatively unstructured population, such as humans. The availability of parallel large-scale QTL mapping studies in mouse and rat [15,27,28] will allow a direct comparison of genetic mapping in a highly selected and relatively non-selected population, respectively. In conclusion, the highly purified spectrum of genetic variation in mouse laboratory strains is likely to continue to provide a unique source of biological variation that can be leveraged to identify disease risk genes.

Acknowledgments

We thank Nadia Rosenthal for financial support. This work was supported in part by a grant from NARSAD (C.T.G.).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2010.02.004.

References

[1] J.A. Beck, S. Lloyd, M. Hafezparast, M. Lennon-Pierce, J.T. Eppig, M.F. Festing, E.M. Fisher, Genealogies of mouse inbred strains, Nat. Genet. 24 (2000) 23–25.
[2] J.C. Auffray, P. Boursot, J. Britton-Davidian, F. Bonhomme, The Evolution of House Mice, Annu Rev Ecol Syst 24 (1993) 119–152.
[3] J.L. Guenet, F. Bonhomme, Wild mice: An ever-increasing contribution to a popular mammalian model, Trends Genet. 19 (2003) 24–31.
[4] B.A. Payseur, H.E. Hoekstra, Signatures of reproductive isolation in patterns of single nucleotide diversity across inbred strains of mice, Genetics 171 (2005) 1905–1916.
[5] K. Abe, H. Noguchi, K. Tagawa, M. Yuzuriha, A. Toyoda, T. Kojima, K. Ezawa, N. Saitou, M. Hattori, Y. Sakaki, K. Moriwaki, T. Shiroishi, Contribution of Asian mouse subspecies Mus musculus molossinus to genomic constitution of strain C57BL/6J, as defined by BAC-end sequence-SNP analysis, Genome Res. 14 (2004) 2439–2447.
[6] C.M. Wade, E.J. Kulbokas III, A.W. Kirby, M.C. Zody, J.C. Mullikin, E.S. Lander, K. Lindblad-Toh, M.J. Daly, The mosaic structure of variation in the laboratory mouse genome, Nature 420 (2002) 574–578.
[7] K.A. Frazer, E. Eskin, H.M. Kang, M.A. Bogue, D.A. Hinds, E.J. Beilharz, R.V. Gupta, J. Montgomery, M.M. Morenzoni, G.B. Nilsen, C.L. Pethiyagoda, L.L. Stuve, F.M. Johnson, M.J. Daly, C.M. Wade, D.R. Cox, A sequence-based variation map of 8.27 million SNPs in inbred mouse strains, Nature 448 (2007) 1050–1053.
[8] K. Lindblad-Toh, E. Winchester, M.J. Daly, D.G. Wang, J.N. Hirschhorn, J.P. Laviolette, K. Ardlie, D.E. Reich, E. Robinson, P. Sklar, N. Shah, D. Thomas, J.B. Fan, T. Gingeras, J. Warrington, N. Patil, T.J. Hudson, E.S. Lander, Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse, Nat. Genet. 24 (2000) 381–386.
[9] H. Yang, T.A. Bell, G.A. Churchill, F. Pardo-Manuel de Villena, On the subspecific origin of the laboratory mouse, Nat. Genet. 39 (2007) 1100–1107.
[10] T. Wiltshire, M.T. Pletcher, S. Batalov, S.W. Barnes, L.M. Tarantino, M.P. Cooke, H. Wu, K. Smylie, A. Santrosyan, N.G. Copeland, N.A. Jenkins, F. Kalush, R.J. Mural, R.J. Glynne, S.A. Kay, M.D. Adams, C.F. Fletcher, Genome-wide single-nucleotide polymorphism analysis defines haplotype patterns in mouse, Proc. Natl. Acad. Sci. USA 100 (2003) 3380–3385.
[11] J.H. McDonald, M. Kreitman, Adaptive protein evolution at the Adh locus in Drosophila, Nature 351 (1991) 652–654.
[12] Z. Yang, J.P. Bielawski, Statistical methods for detecting molecular adaptation, Trends Ecol. Evol. 15 (2000) 496–503.
[13] Z. Yang, W.S. Wong, R. Nielsen, Bayes empirical bayes inference of amino acid sites under positive selection, Mol. Biol. Evol. 22 (2005) 1107–1118.
[14] T.K. Seo, H. Kishino, J.L. Thorne, Estimating absolute rates of synonymous and nonsynonymous nucleotide substitution in order to characterize natural selection and date species divergences, Mol. Biol. Evol. 21 (2004) 1201–1213.
[15] K. Saar, A. Beck, M.T. Bihoreau, E. Birney, D. Brocklebank, Y. Chen, E. Cuppen, S. Demonchy, J. Dopazo, P. Flicek, M. Foglio, A. Fujiyama, I.G. Gut, D. Gauguier, R. Guigo, V. Guryev, M. Heinig, O. Hummel, N. Jahn, S. Klages, V. Kren, M. Kube, H. Kuhl, T. Kuramoto, Y. Kuroki, D. Lechner, Y.A. Lee, N. Lopez-Bigas, G.M. Lathrop, T. Mashimo, I. Medina, R. Mott, G. Patone, J.A. Perrier-Cornet, M. Platzer, M. Pravenec, R. Reinhardt, Y. Sakaki, M. Schilhabel, H. Schulz, T. Serikawa, M. Shikhagaie, S. Tatsumoto, S. Taudien, A. Toyoda, B. Voigt, D. Zelenika, H. Zimdahl, N. Hubner, SNP and haplotype mapping for genetic analysis in the rat, Nat. Genet. 40 (2008) 560–566.
[16] B.M. Smits, V. Guryev, D. Zeegers, D. Wedekind, H.J. Hedrich, E. Cuppen, Efficient single nucleotide polymorphism discovery in laboratory rat strains using wild rat-derived SNP candidates, BMC Genomics 6 (2005) 170.
[17] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S.E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M.R. Brent, D.G. Brown, S.D. Brown, C. Bult, J. Burton, J. Butler, R.D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A.T. Chinwalla, D.M. Church, M. Clamp, C. Clee, F.S. Collins, L.L. Cook, R.R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K.D. Delehaunty, J. Deri, E.T. Dermitzakis, C. Dewey, N.J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D.M. Dunn, S.R. Eddy, L. Elnitski, R.D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G.A. Fewell, P. Flicek, K. Foley, W.N. Frankel, L.A. Fulton, R.S. Fulton, T.S. Furey, D. Gage, R.A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T.A. Graves, E.D. Green, S. Gregory, R. Guigo, M. Guyer, R.C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D.B. Jaffe, L.S. Johnson, M. Jones, T.A. Jones, A. Joy, M. Kamal, E.K. Karlsson, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[18] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[19] J. Zhang, R. Nielsen, Z. Yang, Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level, Mol. Biol. Evol. 22 (2005) 2472–2479.
[20] G.A. Watterson, On the number of segregating sites in genetical models without recombination, Theor. Popul. Biol. 7 (1975) 256–276.
[21] E.P. Rocha, J.M. Smith, L.D. Hurst, M.T. Holden, J.E. Cooper, N.H. Smith, E.J. Feil, Comparisons of dN/dS are time dependent for closely related bacterial genomes, J. Theor. Biol. 239 (2006) 226–235.
[22] F. Al-Shahrour, R. Diaz-Uriarte, J. Dopazo, FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics 20 (2004) 578–580.
[23] R. Hoppe, H. Breer, J. Strotmann, Organization and evolutionary relatedness of OR37 olfactory receptor genes in mouse and human, Genomics 82 (2003) 355–364.
[24] F.Y. Ideraabdullah, E. de la Casa-Esperon, T.A. Bell, D.A. Detwiler, T. Magnuson, C. Sapienza, F.P. de Villena, Genetic and haplotype diversity among wild-derived mouse inbred strains, Genome Res. 14 (2004) 1880–1887.
[25] F.C. Chen, W.H. Li, Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees, Am. J. Hum. Genet. 68 (2001) 444–456.
[26] T. Hochepied, L. Schoonjans, J. Staelens, V. Kreemers, S. Danloy, L. Puimege, D. Collen, F. Van Roy, C. Libert, Breaking the species barrier: derivation of germline-competent embryonic stem cells from Mus spretus x C57BL/6 hybrids, Stem Cells 22 (2004) 441–447.
[27] J. Flint, W. Valdar, S. Shifman, R. Mott, Strategies for mapping and cloning quantitative trait genes in rodents, Nat. Rev. Genet. 6 (2005) 271–286.
[28] R. Mott, J. Flint, Simultaneous detection and fine mapping of quantitative trait loci in mice using heterogeneous stocks, Genetics 160 (2002) 1609–1618.