Evolutionary Bioinformatics

METHODOLOGY

A Novel Multi-Scale Modeling Approach to Infer Whole Genome Divergence

Eli Reuveni[1] and Alessandro Giuliani[2]

[1]Mouse Biology Unit, European Molecular Biology Laboratory (EMBL), via Ramarini 32, 00015 Monterotondo, Italy.
[2]Istituto Superiore di Sanita', Environment and Health Department, Roma, Italy.
Corresponding author email: reuvenieli@gmail.com

Abstract: We propose a novel and simple approach to elucidate genomic patterns of divergence using principal component analysis (PCA). We applied this methodology to the metric space generated by *M. musculus* genome-wide SNPs. Distance profiles were computed between *M. musculus* and its closely related species, *M. spretus,* which was used as external reference. While the speciation dynamics were apparent in the first principal component, the within *M. musculus* differentiation dimensions gave rise to three minor components. We were unable to obtain a clear divergence signature discriminating laboratory strains, suggesting a stronger effect of genetic drift. These results were at odds with wild strains which exhibit defined deterministic signals of divergence. Finally, we were able to rank novel and previously known genes according to their likelihood to be under selective pressure. In conclusion, we posit PCA as a robust methodology to unravel diverging DNA regions without any a priori forcing.

Keywords: principal component analysis, adaptive evolution, genetic drift, multi-scale modeling, speciation, reproductive isolation

Evolutionary Bioinformatics 2012:8 611–622

doi: 10.4137/EBO.S10194

This article is available from http://www.la-press.com.

Introduction

One of the most challenging tasks in molecular evolution is to identify genetic loci that may contribute to phenotypic and genotypic divergences. In general, genetic divergence can be classified into deterministic and stochastic parts. The deterministic part of divergence may convey the beneficial effects forward to the next generation, whereas the stochastic part of divergence may represent neutral genetic drift. For example, selective sweep describes a process in which neutral sites physically linked to an adaptive mutation can result with a skew in the distribution of the allele frequency. In general, it is possible to consider such spectra of genetic variation as being deterministic[1–5] because they can be traced back to a selective process. Tests for neutrality can be inferred by examination of genetic differentiation over short periods of time (ie, the *F* statistics[6]). Loci that are found to exceed an empirical *F* threshold are considered candidates for stabilizing or directional selection.[7,8] In contrast, there are other methods that examine divergence after a longer period of evolution and mostly relies on the relationship between the synonymous and nonsynonymous mutations (eg, the d_N/d_S substitution rates[9–11] or the McDonald–Kreitman test[12]). In principle, those methods assume that the same rate of change can be considered as neutral, while significant deviation from neutral evolution can be inferred as negative or positive selection. It is interesting to consider that all the above methods are built upon a single-gene (locus) basis without taking into explicit consideration the inherent multidimensional character of selection.

The house mouse *M. musculus* and its closely related species *M. spretus* are perfect model organisms for studying evolution.[13] *M. musculus* includes three main groups of isolated wild subspecies (*M. m. domesticus*, *M. m. musculus* and *M. m. castaneus*), which diverged from the last common ancestor around 1 million years ago and migrated into three distinct continents following allopatric separation.[14,15] The potential of using the house mouse for evolutionary studies is also attributed to the presence of extant laboratory strains. These strains are very well characterized from the phenotypic point of view, and, due to strictly controlled laboratory conditions, natural selective processes are virtually absent. The genomic background of these strains is thought to be composed of "mosaic" fragments of high and low differentiation (inter and intra sub-specific origins with correspondence) following historical domestication events of the *M. musculus* subspecies.[12,16–19] Even if their pedigree is unknown, one may postulate that since the genetic makeup of mosaic strains was not shaped by past natural events but due to random human artificial breeding, the signal derived from adaptive evolution should decrease relative to the wild strain, whereas the effect of drift should increase. This is also due to random-sampling matings (or "founder effect") of individual mice from natural breeding as well as their relative small population size.

Principal component analysis (PCA) is a multidimensional statistical technique particularly suited for optimizing the amount of biologically relevant information that may derive from genome scans of DNA polymorphism. PCA has already demonstrated its effectiveness in singling out features that are invisible to single-gene oriented techniques in gene expression analysis.[20] In addition, PCA is a common approach in the field of population genomics and is constantly used to unravel recent demographic radiation of populations.[21–23] When applied to a distance space, PCA is able to project the system into a component space in which the main "directions" of variance are easily interpreted. This is especially evident when distinct order parameters are present, directing the distances of between elements to size and shape components.[24] While the size (first principal component, all positive loadings) captures the across genome correlated distances to the reference, common to all sampled objects, shape (minor components, both positive and negative loadings correspondent to a balance among different order parameters) represents the modulation (size independent) of the distance space among the sampled objects.

There are several examples of how deterministic signal can be differentiated from stochastic noise by deconvolution of the two mutually independent modes using PCA. These include fields of plate tectonics,[25] meteorology and oceanography,[26] or global positioning (GPS).[27] In this study, we applied PCA methodology to distinguish between deterministic and stochastic drift patterns of genomic divergences using distance spaces derived from around eight million distinct patches of DNA. We found that the distantly related strain *M. spretus* could be used as an optimal

reference for the computation of genetic distances for the entire *M. musculus* genealogy. Moreover, by using the score space of the major principal components we were able to distinguish between minor fractions of deterministic signals embedded within the stochastic noise (genetic drift or metric errors).

Our results illustrate that while the first principal component (PC1) captured the speciation signal between *M. musculus* and *M. spretus* with 91% of the variance, the orthogonality between the components explained the differentiation between wild and lab strains with 3% of the variance. It will be demonstrated that wild strains that can be differentiated by PCA possess abundant fractions of deterministic signals (loci) that underscore significant association with functional groups. On the other hand, no such localized regions of divergence could be detected in the lab strains, supporting the proof of concept that selection of deterministic signals may play some significant role in evolution. In sum, the results to be described, indicate PCA as a potentially useful methodology for "deep sequencing" based evolution studies. There is thus a rational basis to explore the characteristic nature of evolutionary dynamics while minimizing the initial assumptions placed on the data.

Methods

Dataset

Mouse SNP data were obtained from the Sanger web site.[28] We used a SNP list from 13 mouse laboratory strains (129P2, 129S1/SvImJ, 129S5, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CBA/J, DBA/2J, LP/J, NOD/ShiLtJ, NZO/HiLtJ), 3 wild-derived *M. musculus* subspecies (CAST/EiJ [*M. m. castaneus*], WSB/EiJ [*M. m. domesticus*], PWD/PhJ [*M. m. musculus*]) and one wild-derived *M. spretus* species (Spretus/EiJ). All data were stored and annotated using a MySQL database with custom Perl and Java programs. A full dataset of exons and introns was obtained using the BioMart tool from Ensembl, version 61 (Ensembl, Cambridge, UK).

Data analysis

The Sanger Institute has recently established a new mouse resequencing project in order to investigate the phylogenetic origin of laboratory mouse strains and to make a comprehensive investigation of the mouse genetic variation using a list of approximately 65 million high confidence SNP calls. This is the dataset we used as a case for evolutionary study and to validate the reliability of our PCA methodology. We first partitioned the genome into equal lengths of genetic loci by using a sliding-window technique. In order to obtain the best resolution for the minimum reliable window size, we chose various windows lengths of $L \in \{50, 200, 500, 1000, 2000\}$ for all autosomes and the X chromosome. For each window length, l (or genetic loci), we calculated the average number of pairwise SNP differences (π[29]) between laboratory and *M. musculus* subspecies and *M. spretus*. We then set the data matrix, G_{jl} of $j = 17$ columns representing the different strains in terms of distance from *M. spretus* computed for the different windows as shown in formula 1.

$$G_{l,j} = \begin{bmatrix} \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,j} \\ \pi_{2,1} & \pi_{2,2} & \cdots & \pi_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{1,1} & \pi_{1,2} & \cdots & \pi_{1,j} \end{bmatrix}$$

Formula 1: Distance matrix.

Each data matrix, G_{jl}, was correspondent to $I \in \{18, 8, 4, 2, 1\}$ million statistical units (rows) of distinct measurements of the genetic distances in consecutive order along the chromosomes. For example, the data matrix of window size $l = 200$ contained $i = 8$ million genetic loci, each one represents the genetic distances between all 17 *M. musculus* with *M. spretus*. It is important to note that since *M. spretus* was the reference strain, the distance matrix, G_{jl}, does not contain *M. spretus* as a variable.

Our first intention was to determine the smallest locus size that could still show a minor decay of the between strains correlation despite the obvious discretization of the data due to small sampling of mutations. To do this, we calculated the pairwise correlation coefficients between the wild strains for each distance matrix, G_{jl}, separately for each window size and identified that the between strains correlation had a major decay when the median locus size was shifted between $l = 200$ and $l = 50$ while the correlation coefficient was stationary when $l > 200$ (Fig. 1). As expected, the X chromosome consistently showed a lower divergence than the autosomes for all strains.

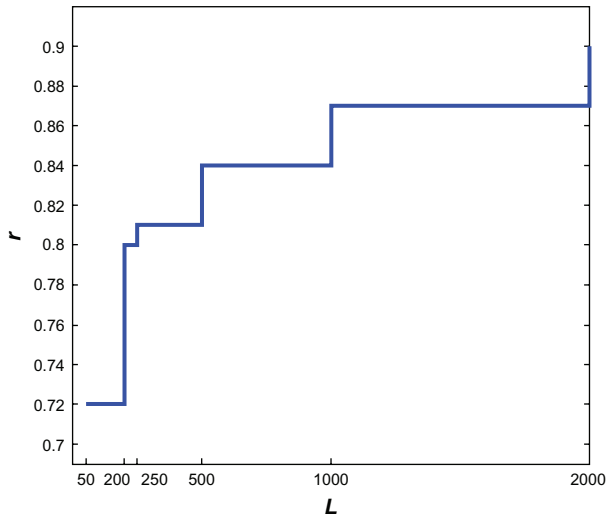

Figure 1. Correlation coefficient decay.
Notes: The graph illustrates the decay in the Pearson correlation coefficient (r) in various genetic loci length (L). The observation suggests that when $L < 200$, a major decay is obtained in the correlation coefficient.

In general, laboratory strains exhibit higher correlation coefficients than wild strains for all chromosomes, confirming less accumulation of mutations due to the short past history of human artificial breeding (S1). These ad hoc results confirm that the modulation of the within *M. musculus* mutation profile mirrors past evolutionary events even though the genetic diversity was measured between all *M. musculus* strains to *M. spretus*.

Principal component analysis

We used the distance matrix, G_{lj}, with a window size $l = 200$ bp, which corresponded to $I = 8$ million genetic loci between all 17 *M. musculus* strains and *M. spretus*. The correlation matrix (S2) reports the Pearson correlation coefficients between the studied mice strains distance vectors from reference species. PCA analysis was applied, and the relative eigenvalues from the between distances correlation matrix were extracted. Computations of the different eigenvectors (correspondent with the correlation coefficients between components and the distance vectors relative to each strain) made it possible to discriminate between the different strains in terms of relative loadings. This is explained by the fact that the PCA loadings are correspondent with the angular relationship between strains and the eigenvector (component) while at the same time a component score is attached to each locus.

We observe that the distance between *M. musculus* and *M. spretus* is represented by size and shape components.[24] The size component (major component PC1, all positive loadings) captures the across genome correlated distances among the *M. musculus* group and *M. spretus*, common to all *M. musculus* strains. The shape component (minor components PC2, PC3, PC4,-positiveandnegativeloadings)representsthemodulation (size independent) of *M. musculus—M. spretus* distances within the *M. musculus* group. The possibility of reliably discriminating between loadings (strains) provides essential knowledge of the genetic distance space represented by the eigenvector of each principal component. Thus, it becomes feasible to further classify each genetic locus into signals of high or moderated divergence without the need to rely on any a priori information.

Identifying deterministic signals of divergence

Having demonstrated the efficacy for strain discrimination of the loading space, we next concentrated on the component score space to extract regions endowed with the highest divergence signals. First, PCA was calculated for each chromosome independently, and the score profile of each principal component was converted into a discrete binary variable $b = \{0, 1\}$ where "1" substituted for any absolute score of a locus greater than 2σ ($2\sigma < |s|$). Since scores represent the magnitudes of the relationship between each datum of the manifest and latent variables, these genetic loci were considered as extreme score loci (ESL) and hotspot candidates for divergence. In contrast, "0" was assigned to the remainder of the dataset as regions of less importance. Second, each ESL was classified as intronic, exonic, or intragenic regions for further analysis. Since PCA is an unsupervised method, each ESL could be considered as a reliable classifier for divergence hotspots. In order to further assess and distinguish between genetic loci of deterministic or drift signals, the sliding windows technique for 30 consecutive loci (~6000 bp) was applied. In each step, significant clusters of ESL (cESL) were checked for by using hypergeometric distributions. Significant regions were observed when $P_R < P_B$, where P_R denotes the P value of a tested region and P_B denotes the Bonferroni threshold (see Formula 2).

$$P(X = x) = \frac{\binom{k}{x}\binom{M-K}{N-x}}{\binom{M}{N}}$$

Formula 2: Probability mass function of hypergeometric distribution.

x, the number of ESL in each one of the tested windows.

M, the total number of windows in each chromosome.

K, the number of ESL in the entire chromosome.

N, the number of windows for each calculation (constant 30 genetic loci).

From each clustered region (cESL), the corresponding ESL was extracted for further analysis whereas the rest of the genome was disregarded (ie, ESLs that were not clustered together). We assumed that those regions were hotspot candidates for evolution[30] or regions with linkage to adaptive regions.[31] We also assumed that dispersed ESLs (ESL that were not clustered together) may represent regions of past episodic selection[31] or drift. In order to test whether our native dataset contained evidence for localized regions of divergence (therefore proving strong linkage between neighbor loci), we permuted the original dataset for each chromosome so the hierarchical order between neighbor loci could not derive from their physical linkages. We then reapplied PCA on the permuted dataset relying on the robustness of PCA to keep the cumulative distribution of the scores along the eigenvector unchanged, but with different order along the chromosome. We also checked that the permutation of the native data matrix had no effect on the structure of the loadings space. Plots were then made of the permuted dataset and the native dataset with the Bonferroni threshold. It was found that the native dataset contained ESL with tide linkage compared with linkage-free relationships between loci in the simulated dataset (Fig. 2). Interestingly, while PC1 to PC3 contain more localized clusters of cESL, PC4 clusters tended to be more overdispersed across the genome.

The ability to partition the entire genome into small genetic loci (200 bp) together with the observation that some of the genetic loci are particularly more reliable and localized within specific regions allows us to classify significant genetic loci into exonic, intronic, and intergenic locations for each principal component with much better accuracy. Thus, by employing PCA strategies it is possible to gain a better association between the loading space, the distance space, and their potential functions.

Results

Size component (PC1) loading distribution interpretation

The possibility to reduce the dimensionality of the system genetic variation into orthogonal principal components allowed us to explore the flux of genetic variance that explains different trajectories of genomic evolution. All genetic distances in our model were computed by referencing various *M. musculus* strains with *M. spretus*. Since the between species distances are much bigger than those of the within species, speciation and reproductive isolation may be inferred from the presence of distinct size (between species) and shape (within species) components (see Method section).

As expected, PC1 captured by far the majority of the correlated distance between the two species (92% of the explained variance) with constant loadings scores for all *M. musculus* group (wild-derived and laboratory strains, loadings ≈ 0.3, Fig. 3). This implies that the major portion of the genome-wide divergence profile between the *M. musculus M. spretus* groups is extremely invariant. This reflects the genetic distance flux corresponding to the speciation between *M. musculus* and *M. spretus*. However, despite the fact that most of the profile of genetic distances between the two species is highly correlated, the within species distance is not coincident and can vary between genetic loci of the different mouse strains. The capturing of 92% variance in PC1 between the two species along their entire genomes confirms that *M. spretus* is an outlier species with respect to *M. musculus* strains. However, PC1appears substantially invariant among different strains (Fig. 3), implying that the within species variation has a different nature (not only a different entity) with respect to the between species one.

Minor components loading interpretation

In contrast to the invariant profile of genetic diversity carried by PC1, all minor components combined (PC2-PC4) explained only 8% of the genetic diversity, but

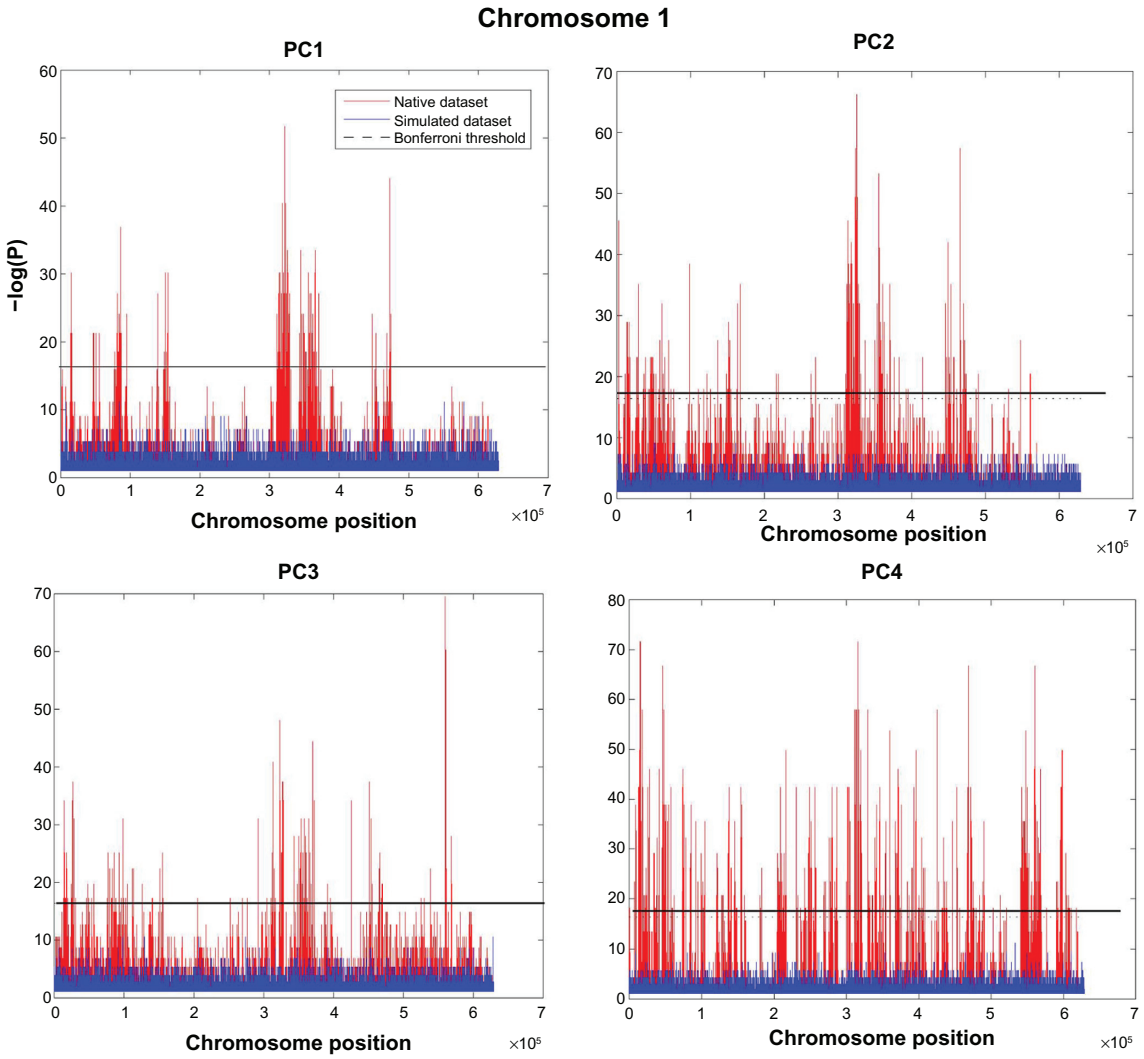

Figure 2. Simulation of ESL.
Notes: Peaks represent *P* values after hypergeometric distribution of consecutive 30 genetic loci (w = 30) for chromosome 1, PC1 to PC4. Significant peaks represent chromosomal regions with significant excess of ESL (cESL) that were found to exceed the Bonferroni threshold. The red peaks represent the native dataset while the blue peaks represent the simulated dataset with permuted location of genetic loci. While the red peaks point to spatially localized clusters, the blue peaks exhibit random distribution of ESL across the chromosome with no specific region of importance. It is clear that while PC1 to PC3 clusters are more localized in defined regions, PC4 clusters are more dispersed across the genome.

still faithfully reflected the modulation of variability within the *M. musculus* group. Regarding genetic variability, we examined only the first three minor components (PC2, PC3, and PC4), which explained about 3% of the variability, which makes a clear discrimination between the wild and the laboratory strains. The other components were disregarded as noise.

Our first observation from the loading space of the minor components revealed clear differences between wild-derived and laboratory strains, which are consistent with their varied ancestral origins and past lifestyles. PC2 (2.5% variance) explains the divergence between *M. m. musculus* and *M. m. castaneus* (loading = 0.6) from *M. m. domesticus* and the other laboratory strains (loading = −0.1, Fig. 3). Such a clade of wild and laboratory strains that are clustered together in the same principal component illustrates that the laboratory mouse strain shares a larger genetic profile with *M. m. domesticus* than with *M. m. musculus* and *M. m. casteneus*. This finding is consistent with past studies identifying *M. m. domesticus* as the most dominant founder for the laboratory strains[17,18,32] and is a proof of concept that the loading space correctly mirrors phylogenetic relationships among *M. musculus* strains.

PC3 (1.1% of variance) highlighted an evolutionary pattern similar to PC2. Whereas PC2 gave the same

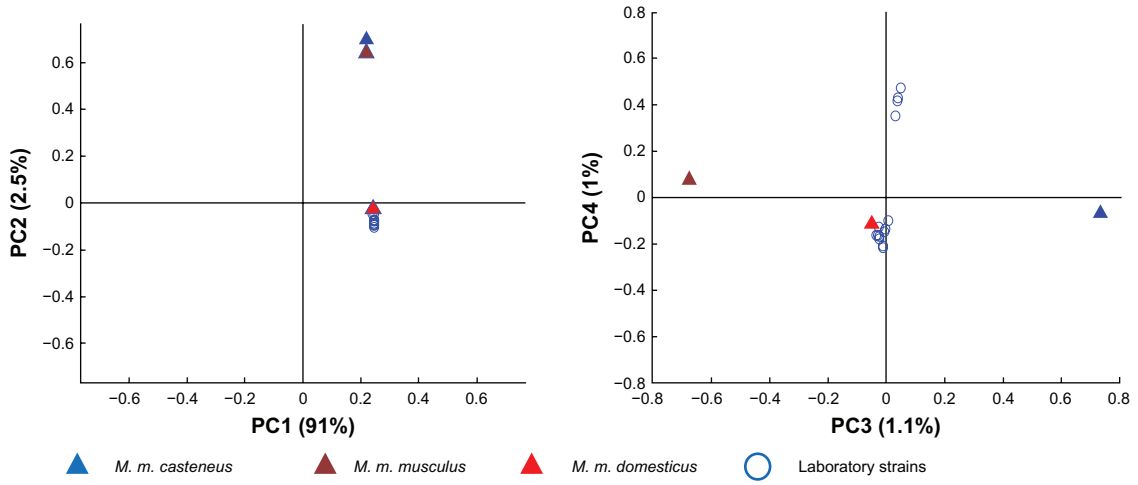

Figure 3. Loading space of PC1 to PC4.
Notes: The figure illustrates constant loading profile for all *M. musculus* strains for PC1 and hierarchical modulation of the minor components between the *M. musculus* strains. It is apparent that while PC2 and PC3 discriminate wild derived strains, PC4 discriminates two group of laboratory strains which could be segregation driven by genetic drift.

weight for the genetic variance shared by laboratory strains and *M. m. domesticus*, PC3 discriminated only wild-derived strains (*M. m. casteneus*, *M. m. musculus*, loading = 0.6 and –0.6 respectively). *M. m. domesticus* and the laboratory strains had minimal or null effects (loadings ≈ 0). Then, in contrast to PC2 and PC3, PC4 (1% variance) gave rise to a clear discrimination between two groups of laboratory strains 129 (129P2,129S1/SvImJ, 129S5) and AKR/J versus the rest (loadings ≈ 0.4, loadings ≈ –0.2). In these cases, the three wild-derived species were kept horizontal to the axes of the principal component space (loadings ≈ 0). These specific results evoke different explanatory hypotheses: (1) PC4 gives rise to a unique adaptive process that is captured neither by PC2 nor PC3, meaning that it has a low likelihood of stemming from the large signal reported from wild-derived discrimination in PC1-PC3; (2) adaptive signals are possibly derived from an unsampled ancestor (proof is beyond the scope of this study); and (3) signal arises from genetic drift (validation requires future downstream analyses).

Having demonstrated the biological relevance of both the loading spaces and extracted components, we are allowed to explore the score space, that is, the values assigned to different genetic loci by the above described components. As an aside, it is worth noting that PCA is a fully unsupervised approach and the strain discrimination emerged a posteriori from the analyses.

Underdispersion of ESL in PC4

We calculated the dispersion index of ESL for each one of the four component scores as indications of their randomness of distribution for each cluster (Fig. 4). In general, both the first component (PC1) and the minor components (PC2-PC4) included a similar number of ESLs attributed to their distribution profiles with common dispersion indexes approximating Poisson distributions $D \approx 0.97$. Even though cESL tended to cluster within specific hotspot regions of evolutionary importance, it is essential to understand the level of abundance of cESLs for the

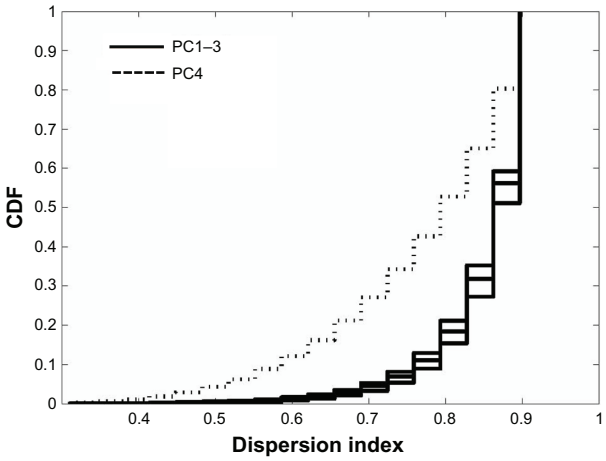

Figure 4. Cumulative distribution of dispersion index units of ESL across window groups of 30 genetic loci for the first four principal components.
Note: PC4 tends to group fewer ESL within regions than observed in PC1- to PC3, thereby affirming that those regions are under recent divergence.

different clusters. With this aim, we used sliding window approach (w = 30 genetic loci) and measured the dispersion index of cESL within each window. We then summarized the cumulative distribution of dispersion scores for all windows as a benchmark for the level of dispersion of cESL within each window. Interestingly, we found that clusters within the first three principal components are less enriched with ESL and tend to exhibit higher dispersion values. In contrast PC4 shows a significant downshift of dispersion along sequential windows. Thus, despite the fact that PC4 clusters tend to be more dispersed along the genome (Fig. 2), those regions are more enriched with cESLs. In this case, PC4 shows more dispersed regions of diversity with more regions of recent divergence compared with the other minor components (PC2 and PC3). This result is in line with the loading profile of PC4, which makes a clear discrimination between the different laboratory strains tested. Therefore, we propose PC4 as a candidate component that captures the flux of variation correspondent with genetic drift whereas PC1 to PC3 may be more affected by the selective process.

Genetic divergence analysis

While the loading space highlights latent variables explaining whole genomic evolutionary forces, the score space presents a more precise view of genomic regions involved in specific evolutionary processes. Therefore, prior knowledge of the loading profile for each principal component can be supported by detailed analyses of genetic loci as possible candidates for selection or drift. We have used the list of cESLs from each principal component (see Methods) and classified each genetic locus into exonic, intronic, and intragenic regions for further analysis. We focused on cESLs underlining exonic regions from the likelihood they may exert more effects on the function and fitness of the organism. In general, we found that the number of candidate genes increases together with the hierarchical order of the principal components. For example, the number of genes listed in PC4 was significantly higher than the one observed in PC3 and PC2 (Table 1). Those results were consistent with noncoding RNA and pseudogenes. This finding is perfectly consistent with the high dispersion levels of PC4 and opposed to PC1-3 dispersion, which tends to be localized to specific regions. In addition, we found that while exons with extreme PC4 scores were randomly scattered across functional classes, exons with higher PC1-3 scores illustrated a clear association with functional groups. The observation that candidate exons exhibit random associations with function, together with the fact that cESLs are more widely dispersed, confirms the stochastic nature of PC4 and points to the fact that the discrimination between lab strains owes more to genetic drift than selective pressure.

Functional analysis

PCA was used to functionally classify genes found in the cESL range and separate them according to mouse strains (for full list of genes see S2). In total, 503 unique candidate exon genes for divergence were found within principal components PC1, PC2, and PC3 and 970, in PC4. In general, candidate exons within PC2 were able to discriminate olfactory receptors (FDR $>$ 0.002), and those within PC3 were able to discriminate immune response genes (FDR $> 1 \times 10^{-7}$). Candidate introns within PC2 also showed positive association with synaptic transmission (FDR $>$ 0.02). Although this class usually tends to be conserved dur-

Table 1. Summary report of cESL across principal components.

	# genes	# genes exons	# genes introns	# ncRNA exons	# ncRNA introns	# pseudogenes	Exon enrichment		Intron enrichment	
							P value	q value	*P* value	q value
PC1	100	38	84	8	1	12	1×10^{-5}	[1]0.002	0.001	0.3
PC2	352	168	293	23	10	44	1×10^{-6}	[1]0.002	1×10^{-5}	[2]0.02
PC3	528	332	445	31	22	60	[3] 2×10^{-10}	[3] 1×10^{-7}	7×10^{-9}	[3] 1×10^{-5}
PC4	1432	970	1200	110	56	107	3×10^{-4}	0.5	2×10^{-3}	0.2

Notes: [1]Olfactory receptor; [2]synaptic transmission; [3]immune system genes. The number of candidate genes is significantly increased in PC4 compare to the other principal components. As expected, candidate genes within PC4 are scattered randomly across functional groups and strengthens the hypothesis that those genes are likely to be under the effect of genetic drift. In contrast, genes among the first 3 principal components are clustered among functional groups, which underlies a possible mechanism of natural selection.

ing evolution, the fact that intron enrichment occurred only within this class with no evidence for candidate exonic regions points to a neutral divergence with no imposing functional consequences. In addition, the lack of any intron enrichment of olfactory receptors is an artifact of sampling since olfactory receptors are very poor in introns. In contrast to PC2, both intron and exon candidates of PC3 show a clear association with immune system genes. The enrichment is more significant among candidate exons even if those genes are significantly more abundant with introns and suggests that these genes could undergo selection. One of the main characteristics of PCA is that individual principal components are orthogonal to each other. Because of this constraint, one would expect that genetic loci with high contributions to PC2 not be represented very well among the other principal components. However, since the dataset employed in this study focused on a single gene covered by many genetic loci, it is still expected that a certain level of superposition of candidate exon genes among the different principal components would be found. In that case, candidate genes that were found to be shared by different principal components could point to regions that have relevance for the discrimination of different lineages. Alternately, candidate exon genes specific to only one principal component could point to a lineage of specific divergence. In addition, the fact that shared candidate exons genes were represented in different principal components reveals a significant nonrandom selection of genes toward functional classes. In general, most of the candidate genes were unique to corresponding principal components (Fig. 5) with only one gene being common for PC1-3 (Skint5—T cell). We obtained 26 overlapping genes (~15%) between PC2 and PC3, most of which were histocompatible, olfactory, or C-type lectin-containing domains. This group of 26 common genes was significantly more enriched than was expected by chance alone (2 expected shared genes). This is further proof of the accuracy of PCA methodology to locate major genes playing possible roles in divergence. The fact that most of the candidate genes from both components are known to be under strong selective constrains during evolution affirms that PCA is able to decompose the genetic variation into selected regions of divergence without losing the global picture of the latent directions of divergence between various evolutionary scenarios.

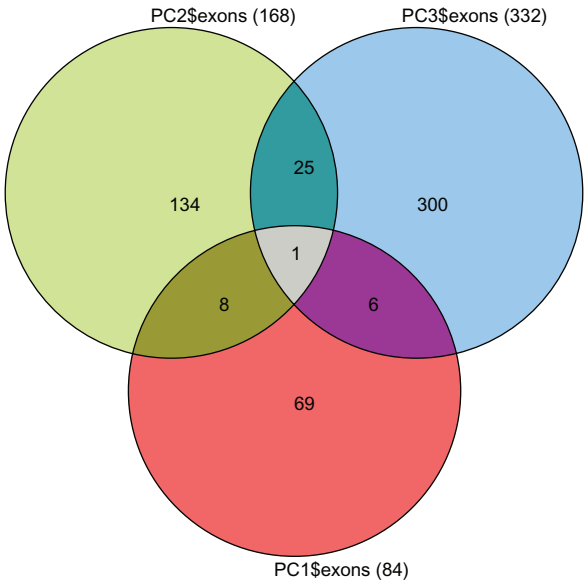

Figure 5. Venn diagram illustrates common candidate exons genes for PC1 to PC3.
Notes: The figure suggests that ~85% from the genes are unique for each one of the principal components. This is expected since the principal components are each orthogonal. However ~15% (26 observed genes) of the exon genes are shared among principal components as extremes. This number is much higher than what expected by chance (2 expected genes) and confirm their essentiality for divergence.

Discussion

Recent advances in sequencing technologies allow the acquisition of high resolution sequence information for entire genomes from which knowledge of the dynamics and trajectories of evolution between and within closely related species can be deduced. A better accuracy of the read calls[33] leads to better confidence in SNPs,[34] reducing the likelihood of making false positive and false negative predictions. Since the amount of future data is predicted to grow exponentially, new methods to handle this huge data flow are required. In addition, the fact that most of the existing methodologies are single, gene-centric oriented, they are incapable of taking a higher dimensional, more global approach on data if and when require.

In this study, we have presented a PCA-based methodology to unravel regions of important genetic divergence. We have also demonstrated that an apparent drift component could be identified in a small and artificially-bred population of laboratory animals. This finding was in direct opposition to more deterministic behavior as captured among the principal components of divergence in wild strains. Likewise, we have demonstrated that our PCA approach provided two consistent associated representations of

strains (loading space) and genetic loci (score space). Our results point to a leading common "speciation" first principal component (91% of the explained variance) which implies a very strong correlative structure of the distance between the two species measured at different loci. We found that the score distribution of the first four principal components mirrored the evolutionary trajectories drawn by the loading space following the multi-scale modulation of the genetic distance between *M. musculus* strains and *M. spretus*. It is apparent that while PC1, PC2, and PC3 discriminate wild strains, PC4 discriminates laboratory strains.

By using a clustering approach, we were able to identify hotspot regions for divergence and to characterize each one of the four principal components as potential random drift or deterministic divergence that may indicate selection processes. In contrast to PC4, the first three principal components exhibited more deterministic behaviors, which can be interpreted as: (1) discrimination of wild strains; (2) localized hotspot of divergence (Fig. 2); and (3) significant associations with function (Table 1). We conclude that PC1 to PC3 are explained more by evolution following adaptation, whereas PC4 illustrates evolution by drift with a more neutral effect.

As expected, by using high-resolution analyses of exons from candidate gene loci, enrichment was found in the second and third principal components. Interestingly and unexpectedly, each principal component was enriched with different groups of genes according to their functional groups. While PC2 was associated with the olfactory receptor, PC3 included only enrichment of immune system genes (Table 1). We found especially that the Skint family was remarkably influential by discriminating principal components with a very strong signal from PC3 (which in general was more overdominated with immune system genes, Table 1). Olfactory and immune system genes are known to be under strong selective pressure in the mouse.[13,35,36] Immune system genes are known to be under strong positive selection in *Drosophila*[37,38] and, in general, in all vertebrates.[39] The fact that different gene families exerted major effects on different principal components could be due to the fact that each orthogonal component represents different episodes of weak to high evolutionary pressure.

It will be necessary to validate our evolutionary hypothesis with pre-existing population genetics methods such as the d_N/d_S substitution rates[9–11] or the *F* statistics. However, this approach is not feasible with our present dataset due to insufficient sequenced animals either to perform the *F* test or to meet the robust statistical assumptions required by the maximum likelihood framework. However, even though validation with known tools to assess adaptive evolution cannot be performed in the context of our paper, we were nonetheless able to single out a group of genes that are known to be under strong selective constraints. And since these results were obtained without any a priori information, this validates that our PCA methodology be applied to other model systems as a suitable complement to other extant multilocus methodologies.

Aknowledgements

We wish to thank Prof. Charles Webber for his linguistic and stylistic suggestions. These data were provided by the Mouse Genomes Project and Sanger Mouse Exomes Project group at the Wellcome Trust Sanger Institute and can be obtained from http://www.sanger.ac.uk/resources/mouse/genomes/.

Author Contributions

Both authors ER and AG planned the research. ER was mostly responsible for carrying out the analyses. ER did most of the writing, which was edited by an English-language reviewer. Both authors read and approved the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria.

The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Andolfatto P. Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev.* 2001;11(6):635–41.
2. Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics.* 1995;140(2):783–96.
3. Fu YX, Li WH. Statistical tests of neutrality of mutations. *Genetics.* 1993; 133(3):693–709.
4. Schlichting CD, Mousseau TA, editors. T*he Year in Evolutionary Biology 2009*. Hoboken, NJ: Wiley-Blackwell; 2009.
5. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;123(3):585–95.
6. Wright S. Genetical structure of populations. *Nature.* 1950;166(4215): 247–9.
7. Harr B, Kauer M, Schlotterer C. Hitchhiking mapping: a population-based fine-mapping strategy for adaptive mutations in Drosophilamelanogaster. *Proc Natl Acad Sci U S A.* 2002;99(20):12949–54.
8. Schofl G, Schlotterer C. Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African D. simulans. *Mol Biol Evol.* 2004;21(7):1384–90.
9. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol.* 2000;15(12):496–503.
10. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 2005;22(4):1107–18.
11. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol.* 2005;22(12):2472–9.
12. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature.* 1991;351(6328):652–4.
13. Reuveni E, Birney E, Gross CT. The consequence of natural selection on genetic variation in the mouse. *Genomics.* 2010;95(4):196–202.
14. Boursot PAJ, Britton-Davidian J, Bonhomme F. The Evolution of House Mice. *Annu Rev Ecol Syst.* 1993;24:119–52.
15. Guenet JL, Bonhomme F. Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet.* 2003;19(1):24–31.
16. Wade CM, Kulbokas EJ 3rd, Kirby AW, et al. The mosaic structure of variation in the laboratory mouse genome. *Nature.* 2002;420(6915):574–8.
17. Beck JA, Lloyd S, Hafezparast M, et al. Genealogies of mouse inbred strains. *Nat Genet.* 2000;24(1):23–5.
18. Frazer KA, Eskin E, Kang HM, et al. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature.* 2007;448(7157): 1050–3.
19. Reuveni E. The genetic background effect on domesticated species: a mouse evolutionary perspective. *Scientific World Journal*. 2011;11:429–36.
20. Alter O. Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc Natl Acad Sci U S A.* 2006;103(44):16063–4.
21. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. *Nature.* 2008;456(7218):98–101.
22. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science.* 1978;201(4358):786–92.
23. Cavalli-Sforza LL, Menozzi P, Piazza A. Demic expansions and human evolution. *Science.* 1993;259(5095):639–46.
24. Darroch JN, Mosimann JE. Canonical and principal components of shape. *Biometrika.* 1985;72(2):241–52.
25. Kawamura M, Yamaoka K. Spatiotemporal characteristics of the displacement field revealed with principal component analysis and the mode-rotation technique. *Tectonophysics.* 2006;419(1–4):55–73.
26. Richman MB [Book review]. Principal Component Analysis in Meteorology and Oceanography by Rudolph W Preisendorfer. *Nature.* 1989;339(6227): 673–3.
27. Giuliani A, Colosimob A, Benignia R, Zbilut JP. On the constructive role of noise in spatial systems. *Physics Letters A.* 1998;247(1–2):47–52.
28. Keane TM, Goodstadt L, Danecek P, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature.* 2011;477(7364):289–94.
29. Watterson GA. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 1975;7(2):256–76.
30. Lewontin RC, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics.* 1973;74(1): 175–95.
31. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet.* 2007;8(11): 857–68.
32. Yang H, Bell TA, Churchill GA, Pardo-Manuel de Villena F. On the subspecific origin of the laboratory mouse. *Nat Genet.* 2007;39(9):1100–7.
33. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26(10):1135–45.
34. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 2009;19(6): 1124–32.
35. Hoppe R, Breer H, Strotmann J. Organization and evolutionary relatedness of OR37 olfactory receptor genes in mouse and human. *Genomics.* 2003; 82(3):355–64.
36. Harr B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* 2006;16(6):730–7.
37. Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. Dynamic evolution of the innate immune system in Drosophila. *Nat Genet.* 2007;39(12):1461–8.
38. Schlenke TA, Begun DJ. Natural selection drives Drosophila immune system evolution. *Genetics.* 2003;164(4):1471–80.
39. Hughes AL, Yeager M. Molecular evolution of the vertebrate immune system. *Bioessays.* 1997;19(9):777–86.

Supplementary Figures

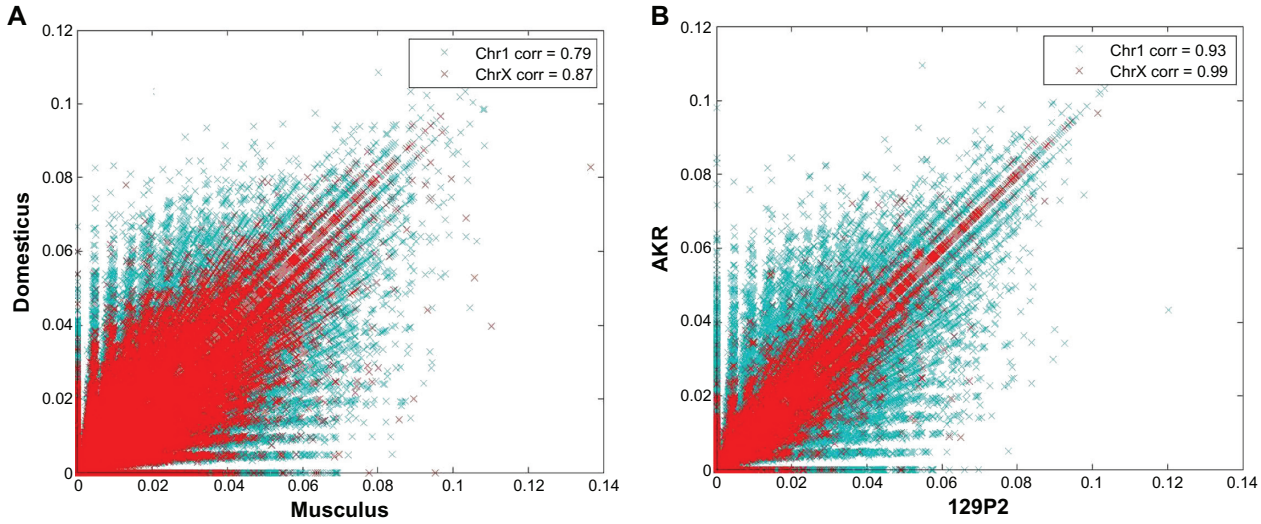

Figure S1. Correlation of genetic divergence in chromosomes 1 and X. It is notable that authosomes are over-diverged then the X chromosome due to increase amount of meiosis during evolution (**A**). In addition lab strains illustrate less divergence signals due to more homogenous genetic background (**B**).

	A_J	AKR	BALB	C3H	C57BL	CAST	CBA	DBA	LP_J	NOD	NZO	PWK	WSB	129P2	129S1	129S5
A_J																
AKR	0.94															
BALB	0.95	0.94														
C3H	0.98	0.94	0.95													
C57BL	0.95	0.93	0.96	0.94												
CAST	0.79	0.79	0.79	0.79	0.79											
CBA	0.96	0.94	0.95	0.97	0.95	0.79										
DBA	0.94	0.93	0.94	0.94	0.93	0.79	0.95									
LP_J	0.95	0.94	0.95	0.95	0.94	0.79	0.94	0.93								
NOD	0.96	0.94	0.95	0.96	0.95	0.79	0.95	0.94	0.95							
NZO	0.95	0.94	0.95	0.94	0.95	0.79	0.94	0.93	0.94	0.95						
PWK	0.81	0.81	0.82	0.81	0.81	0.80	0.81	0.82	0.81	0.81	0.81					
WSB	0.93	0.91	0.92	0.93	0.93	0.79	0.93	0.92	0.93	0.93	0.93	0.81				
129P2	0.95	0.94	0.94	0.95	0.94	0.79	0.94	0.93	0.98	0.95	0.95	0.81	0.93			
129S1	0.95	0.94	0.95	0.95	0.94	0.79	0.94	0.94	0.98	0.96	0.95	0.81	0.93	0.99		
129S5	0.95	0.94	0.95	0.95	0.94	0.79	0.94	0.94	0.98	0.96	0.95	0.81	0.93	0.99	0.99	

Figure S2. Illustrate is a matrix of Pearson correlation coefficient between studied mouse strains.

FPO

Figure S3. Table in excel format represent candidate genes under divergence for the 4 major components (PC1-PC4). (File available from article homepage).