HHS Public Access

Author manuscript

Neuron. Author manuscript; available in PMC 2018 May 08.

Published in final edited form as:

Neuron. 2017 May 03; 94(3): 465–485.e5. doi:10.1016/j.neuron.2017.04.005.

MUPET – Mouse Ultrasonic Profile ExTraction: A signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations

Maarten Van Segbroeck[1,*], **Allison T. Knoll**[2,*], **Pat Levitt**[2,3,#], and **Shrikanth Narayanan**[1,#]

[1]Department of Electrical Engineering, University of Southern California, Los Angeles, California 90089, USA

[2]Institute for the Developing Mind, Children's Hospital Los Angeles, Los Angeles, California 90027, USA

[3]Department of Pediatrics, Keck School of Medicine of University of Southern California, Los Angeles, CA 90089 USA

SUMMARY

Vocalizations play a significant role in social communication across species. Analyses in rodents have used a limited number of spectro-temporal measures to compare ultrasonic vocalizations (USVs), which limits the ability to address repertoire complexity in the context of behavioral states. Using an automated and unsupervised signal processing approach, we report the development of MUPET (Mouse Ultrasonic Profile ExTraction) software, an open access MATLAB® tool that provides data-driven, high-throughput analyses of USVs. MUPET measures, learns, and compares syllable types and provides an automated time-stamp of syllable events. Using USV data from a large mouse genetic reference panel and open source datasets produced in different social contexts, MUPET analyzes the fine details of syllable production and repertoire use. MUPET thus serves as a new tool for USV repertoire analyses, with the capability to be adapted for use with other species.

IN BRIEF

Van Segbroeck et al. present open-access software that uses signal-processing techniques to perform rapid, unsupervised analysis of mouse ultrasonic vocalization repertoires, including unbiased syllable discovery, new metrics to compare syllable production and use, and syllable time-stamp enabling next-step behavioral analyses.

Keywords

communication; repertoire; machine learning; software; social behavior; syllable

Corresponding Author and Lead Contact: Pat Levitt, Ph.D., Children's Hospital Los Angeles, Mail Stop #135, 4650 Sunset Boulevard, Los Angeles, CA 90027, plevitt@med.usc.edu.

*These authors contributed equally

#Senior author

Author Contributions: Conceptualization, M.V.S., A.T.K., P.L., and S.N.; Software, M.V.S.; Validation, Formal Analysis, Investigation, A.T.K.; Visualization, M.V.S. and A.T.K.; Writing, M.V.S., A.T.K., P.L., and S.N.

INTRODUCTION

Vocalizations play a significant role in social communication across species (Bradbury and Vehrencamp, 2011). Although the complexity of communication varies considerably between vocal learning and innate vocalizing species, there is broad conservation of the use of frequency and amplitude modulation, multisyllabic patterns, and variable syllable durations and rates to signal caller identity, sex, intentionality, and affective state (Bradbury and Vehrencamp, 2011; Doupe and Kuhl, 1999; Fischer and Hammerschmidt, 2011; Sales and Pye, 1974). Auditory processing mechanisms for natural vocalizations also are broadly conserved across vocal learning and innate vocalizing species (Arriaga and Jarvis, 2013; Bennur et al., 2013; Woolley and Portfors, 2013). Studies of the underlying neurobiological basis and heritable nature of these mechanisms have identified promising genes and neural networks that support vocal production, auditory processing, and social communication (Arriaga et al., 2012; Konopka and Roberts, 2016).

Comparative analyses in mice serve as effective strategies to examine genetic and environmental factors that influence vocal communication (Fischer and Hammerschmidt, 2011; Konopka and Roberts, 2016). Mice generate complex multisyllabic ultrasonic vocalizations (USVs, >25 kHz) throughout development and in diverse social and motivational contexts (Chabout et al., 2015; Liu et al., 2003; Portfors, 2007; Sewell, 1970). Mouse syllables are defined as units of sound that are composed of one or more notes and which are separated by silent pauses and occur as part of sequences (vocalization bouts) (Arriaga and Jarvis, 2013; Holy and Guo, 2005; Portfors, 2007; Scattoni et al., 2010). Mice generate a diversity of syllable types and a syllable repertoire is composed of the full collection of syllable types used by a specific mouse (or strain) in a particular condition. USVs support essential social behaviors across development (Chabout et al., 2012; Hammerschmidt et al., 2009; Hanson and Hurley, 2012; Holy and Guo, 2005; Pomerantz et al., 1983; Sales and Pye, 1974; Sewell, 1970). Although mice are innate vocalizers (Arriaga and Jarvis, 2013; Hammerschmidt et al., 2015; Mahrt et al., 2013), the acoustic structure (e.g., mean frequency, amplitude) and contextual use of their syllable repertoires varies considerably across genetic strains (Panksepp et al., 2007; Scattoni et al., 2010; Sugimoto et al., 2011; Thornton et al., 2005; Wohr et al., 2008), behavioral and social environments (Chabout et al., 2015; Chabout et al., 2016; Hanson and Hurley, 2012; Liu et al., 2003; Yang et al., 2013), and development (Grimsley et al., 2011; Liu et al., 2003). Understanding complex vocalization structure of mice will be key to advancing vocal and social communication research.

There is a rich history of efforts to assess the social meaning of USVs, yet despite evidence that they serve important communicative functions in mice, it remains unclear which acoustic features and types relate to specific biological states or affect social outcomes. To address the social meaning of USVs, the field has used syllable classification schemes, among the most popular being manual or semi-automated classification of syllables into ~9–12 broad categories based on spectral shapes (e.g., chevron, upward, frequency-step) (Grimsley et al., 2011; Portfors, 2007; Scattoni et al., 2008; Scattoni et al., 2010), or classification into ~4–15 categories based on instantaneous changes in frequency within a

syllable (‘frequency jumps’) (Arriaga and Jarvis, 2013; Arriaga et al., 2012; Chabout et al., 2015; Holy and Guo, 2005; Mahrt et al., 2013). Categorization based on cluster analyses has revealed potentially meaningful spectral features in different strains and experimental contexts (Grimsley et al., 2013; Hammerschmidt et al., 2012; Sugimoto et al., 2011; von Merten et al., 2014). Several automated and semi-automated software programs have been developed to rapidly measure USV features and apply specific syllable classification schemes [e.g., SASLab Pro (Avisoft Bioacoustics, Germany), Mouse Song Analyzer v1.3 (MSA; (Arriaga et al., 2012; Chabout et al., 2015)], VoICE (Burkett et al., 2015)). Yet, there remains a lack of consensus on which classification schemes provide the best biological insights (Arriaga and Jarvis, 2013; Grimsley et al., 2013; von Merten et al., 2014), and indeed, the most informative spectro-temporal features may vary across genetic and environmental conditions. In our initial attempts to examine genetic and environmental factors that influence USV production and syllable repertoires in a large genetic reference panel (GRP) of recombinant inbred (RI) mouse strains, we encountered theoretical and technical challenges using current categorical approaches in large datasets (>500,000 syllables). In particular, categorical (rule-based) syllable classification might not provide sufficient sensitivity to detect unique syllable types, which could have condition or strain-specific meaning. In addition, no technique included important signal detection features (noise detection and removal, time-stamp) needed to facilitate processing of large volumes of recordings across conditions or laboratories and, ultimately, to determine if differences in syllable timing and type meaningfully relate to ongoing behaviors and behavioral transitions —a long standing question in the field. One strategy to address these issues is the application of signal-processing methods used in human speech and language analysis (O’Grady and Pearlmutter, 2008; Ramanarayanan et al., 2013; Smaragdis, 2007; Van Segbroeck and Van hamme, 2009), providing a methodological balance between data-driven detection of recurring types and rapid data optimization and comparative analyses.

Here, we report a novel NeuroResouce, open access Mouse Ultrasonic Profile ExTraction (MUPET) software. This MATLAB® tool with graphical user interface (Figure S1) is inspired by human speech processing (Patterson et al., 1987; Valero and Alias, 2012) and uses a Gammatone filterbank to convert USVs into compact acoustic feature representations, or “GF-USVs” (Gammatone Filterbank Ultrasonic Vocalization features, see **Results** and **STAR Methods**). The latest version of MUPET, including a subset of the audio recordings described in this paper and an experimental tutorial, are available for download at http://sail.usc.edu/mupet.

RESULTS

Using an automated and unsupervised algorithmic approach, MUPET has five core capabilities that enable it to detect, learn, and compare syllable types and repertoires: 1) syllable detection—the isolation and measurement of basic spectro-temporal syllable parameters, 2) acoustical dataset analysis—analysis of overall vocalization features, such as syllable number, rate and duration, spectral density and fundamental frequency, 3) syllable repertoire building—the extraction of up to several hundred of the most highly represented syllable types (“repertoire units”, RUs) in individual datasets using machine learning, 4) measures of syllable repertoire similarity—rank order comparisons of the similarity of

spectral types of individual RUs across datasets, and 5) RU-cluster analysis—a centroid-based cluster analysis of RUs composing different dataset syllable repertoires in order to measure the frequency of use of different RUs (syllable types) across conditions.

MUPET builds syllable repertoires—a collection of different syllable types used by a specific mouse or strain under various experimental conditions—and compares them both independently, and as a function of, how frequently each syllable type is used. Syllables occur in patterned sequences (syntax) and communicative information in USVs is likely encoded through both syntax and several features of individual syllables, including mean frequency, amplitude, duration, and shape (Chabout et al., 2015; Holy and Guo, 2005; von Merten et al., 2014). A central feature of MUPET is to determine the specific syllable types that are present in mouse vocalization repertoires. The software accomplishes this by examining the entire frequency contour (slope, duration and frequency modulation of each note), but not as a function of fundamental frequency or amplitude. This is supported by evidence in humans and mice that syllables (phonemes) are the basic communicative units and that variations in fundamental frequency (intonation, prosody) and amplitude convey additional information about social motivation, mood, or genetic background (Adolphs et al., 2002; Lahvis et al., 2011; Narayanan and Georgiou, 2013). MUPET provides users with spectro-temporal measures for each syllable and dataset (Figure S1). MUPET is designed to build separate syllable repertoires for each strain or condition, which are used for subsequent repertoire comparisons. This strategy increases the precision of syllable repertoire builds, which facilitates the accuracy of subsequent comparisons of vocal production and repertoire use across strains and behavioral conditions (see below).

The design features for MUPET are based on techniques commonly used in automated signal processing analyses of human speech (Bregman and Campbell, 1971; Gunawan and Ambikairajah, 2004; Johnson, 1997; Rabiner and Schafer, 2010; Smaragdis, 2007) and include 1) optimization of the signal-to-noise ratio (SNR) to enable detection of vocalization activity (Ramirez et al., 2007), 2) transformation of the original sound data into a more compact representation, which retains essential acoustic shape information and facilitates rapid analysis of large volumes of data (Davis and Mermelstein, 1980; Schluter et al., 2007; Shao et al., 2009), and 3) newly developed application of unsupervised machine learning algorithms to automatically identify recurring syllable types. Detailed descriptions of each step, and the rationale for using certain technical strategies are in **STAR Methods**.

1) Optimization of SNR

The quality of automated sound analyses depends critically upon optimization of SNR to maximize syllable (and minimize noise) detection. This process can be challenging in mice due to recording environments, and the high intrinsic variability in fundamental frequency, amplitude (energy), and syllable type, which is characteristic of their vocalizations. MUPET addresses these signal detection challenges by 1) high-pass filtering of the recordings to the ultrasonic range (25–125 kHz), 2) using spectral subtraction to remove the stationary noise in the recordings originating from background noise and recording equipment distortions (Martin, 2001; Van Segbroeck et al., 2013), and 3) computing the power of spectral energy in the ultrasonic range that exceeds a noise floor threshold (Yu Song et al., 2013). MUPET

provides the user the ability to select noise to be removed from the final analyses. Users also can control 6 key features of SNR optimization—noise reduction, minimum and maximum syllable duration, minimum total and peak syllable energy, and the minimum inter-syllable interval that is needed to separate rapidly successive notes into distinct syllables. As with all automated sound detection software, visual comparison of the sonograms generated by MUPET to the original spectrograms is an important initial step in defining optimal parameters. This step is aided by time-stamp information for each detected syllable, as well as by the ability to rapidly build and compare syllable repertoires using several SNR settings.

2) Transformation of sonograms into a computationally compact format

Mouse USVs are produced on millisecond time scales and are recorded using high sampling rates (e.g., 250 kHz), capturing the frequency and amplitude modulation. This high sampling rate generates a computationally dense, highly dimensional sonogram due to the high time-frequency resolution. This high-dimensionality creates challenges for the application of unsupervised signal-processing approaches, which function by iteratively processing and reprocessing large datasets. Unsupervised data-driven analyses of human speech transform the spectral representation of speech into “low dimensional” features, which capture the spectral shape of the phonemes and maintain the key spectro-temporal features that are likely to hold the most communicative significance (Bertrand et al., 2008; Joder and Schuller, 2012; Van Segbroeck and Van hamme, 2009). This automated decomposition process involves using non-negative matrix factorization (NMF) (Lee and Seung, 2001) to identify a set of non-negative spectral bases (the fundamental computational units that compose the vocalization array) that optimally compress the data and eliminate redundancy in representing the amplitude and frequency modulation (Figure S2). These spectral bases (‘basis units’) act as a set of band-pass filters (‘filterbanks’) that can be used to represent the auditory spectrum with a smaller number of data points along the frequency axis. The biological relevance and perceptual fidelity of the transformed sounds is supported by studies showing that the spectral bases that decompose speech using NMF are highly similar to the human cochlea’s biological and perceptual time-frequency resolution (“auditory filters”) (Fletcher, 1940; Gelfand, 2009), as well as to perceptual scales, such as the Mel (Stevens et al., 1937) or Bark scales (Zwicker, 1961). These experiments in humans yield strong evidence that NMF decompositions retain the most salient and informative features of sound in biological contexts.

Here, we applied a similar strategy to generate perceptually filtered representations of mouse USVs by determining the spectral bases in which they can be decomposed. Briefly, the sonograms were first computed using a short-term Fourier transform (STFT) algorithm with an analysis window of 2 milliseconds that was shifted every 1.6 milliseconds and then normalized to unit energy to prevent the decomposition process from being dominated by high energy syllables. We then applied NMF on the normalized sonograms from all of the USV data, resulting in decomposition of the original high-resolution frequency contours into a sparse set of basis units comprised of narrow-width frequency bands along the ultrasonic frequency range in which mice vocalize (Figures 1 and S2). With this strategy, USV sonograms can be approximated by a weighted sum of the basis unit functions. Finding the

weights of each basis unit corresponds to applying a filterbank operation on the sonograms with the bases serving as band-pass filters. By applying a regression analysis on the base function's peak frequencies, a logistic curve is obtained. The logistic curves (Figures 1A and S2) are derived for each mouse strain and are centered on the mean frequency (Gaussian fit line) of their power spectral density (PSD) functions (see below; Figures 4A and S3). The filterbank covers the ultrasonic range from 25 to 125 kHz and consists of a predefined number of band-pass filters (here 64), each modeled by a gammatone band-pass function (Figure 1B). Gammatone filters exploit psycho-acoustically defined properties of the human auditory system, such as spectral resolution along the frequency axis and inherent redundancy in the spectral envelope, to obtain a compact representation of salient acoustic cues that relate to distinctive speech sound units or phones. The band-pass filters are symmetrically distributed, with the frequency region containing the highest number of relevant auditory events (determined empirically) modeled by narrow bandwidth filters of increasing density (0.5–1 kHz). The upper and lower bounds of the mouse USV frequency range contain less acoustic content and are captured by a smaller number of wider bandwidth filters (2–4 kHz) (Figure 1A,B). Thus, the sonograms are spectrally transformed into low dimensional vector representations, Gammatone Filterbank USV features (GF-USVs). The GF-USVs have the advantage of representing the mouse syllables with reduced dimensionality (here 64 frequency dimensions) and maintained saliency, facilitating further application of signal processing methods to model mouse vocalization repertoires. We show an example sonogram of 1.4 sec duration (Figure 1C) and the corresponding GF-USV representation (Figure 1D). Despite small differences in the ultrasonic NMF plots (Figure S2), we used a filterbank design (spacing and number of filters) that is equal for all mouse strains, to avoid bias in the GF-USV representation that can be attributed to different filterbank features.

3) Application of unbiased machine learning algorithms

MUPET uses k-means clustering to automatically learn the most recurring syllable types. As described above, the computationally compact GF-USV feature representation allows automatic and unsupervised grouping of large sets of recorded syllables based on spectral shape similarities. The basic analytic steps are summarized in Figure 2, with a detailed flow chart of processing steps and data files generated by MUPET shown in Figures 3 and S1. The ultrasonic recordings are first processed to detect the presence (non-presence) of syllable activity over time (Figure 2A, top panel) and the detected syllables are subsequently transformed into GF-USVs (Figure 2A, bottom panel) from which a three-dimensional tensor representation is constructed (Figure 2B). Each tensor image corresponds to exactly one syllable segmented from the audio stream using the time-stamp information provided by the syllable detector. To eliminate syllable groupings based on their position in the time-frequency plane, all images of the syllable matrix are centered according to their corresponding spectral gravity center (i.e., centered in time and frequency). This allows syllable types to be compared without respect to mean frequency, but maintains comparisons based on syllable length. Next, k-means matrix clustering is done with the syllable matrix. The clustering is applied to the vectorized images and the cosine similarity is used as a distance function to measure the orientation between two vectors independent of their magnitude. This approach prevents the grouping of syllable units according to their

amplitude, instead focusing on spectral shape. The outcome of the matrix clustering method is a syllable repertoire composed of individual “repertoire units” (RUs) (Figure 2C), which are the centroids (average shapes; exemplars) of all the individual syllables that were grouped within a specific RU based on their shape similarity. During syllable repertoire refinement, the user selects RUs (representing noise) to be removed from the repertoire, dataset and syllable information files (Figures 2D and S1). The user then builds ‘final’ syllable repertoires of various sizes (20–200 RUs) from the refined datasets and selects an optimal repertoire size(s) based on repertoire modeling scores and RU goodness-of-fit measures (see below). After syllable repertoires have been constructed for different mouse strains or experimental conditions the similarity of spectral shapes of different RUs can be compared using Pearson correlation values, which are measured between the RUs of two different repertoires in a manner that is either independent of the frequency of use of each RU (‘Cross Repertoire Similarity Matrix’; Figure 2E) or dependent upon the frequency of use of each RU (‘Cross Repertoire Similarity Boxplot’; Figure 2F). Finally, the RUs identified across strains and conditions can be grouped based on shape similarity into a ‘Master Repertoire’ using k-medoids clustering of RU centroids and ‘RU-cluster’ frequency of use compared across datasets to identify shared and unique shapes (Figure 2G).

Genetic influences on USV production

Our laboratory has an interest in understanding the environmental and heritable factors that contribute to heterogeneity in social communication. To demonstrate the utility of MUPET for rapid analysis of large datasets and identification of shared and strain-specific syllable types, we measured heterogeneity in USV production and syllable repertoires in 12 genetically related adult mouse strains from the BXD GRP—the C57BL/6 (C57) and DBA/2 (DBA) parental strains, the F1 cross (B6D2F1) and 9 RI offspring strains—during direct social interaction (DSI) with an unfamiliar juvenile male partner. We recorded USVs emitted by adult males (Figure 3), selecting BXD strains that showed moderate-to-high levels of vocalizations to maximize the number and diversity of syllable types analyzed. All USVs were produced by the adult males based on several lines of evidence from our analyses that juvenile males do not vocalize during this task (see **STAR Methods**). There is a 3.0-fold range in DSI and an 11.0-fold range in mean USV counts across the 12-strain panel (Figure 3A,B). For both DSI and USV count, the F1 cross shows an intermediate phenotype compared to the parental strains, providing evidence of incomplete dominance. We note that DSI and USV count are positively correlated across the 12 strains ($R^2 = 0.62$, $P < 0.01$).

Figure 3C depicts the ‘syllable repertoire analysis method’, which uses a combination of automated and manual steps. We compared the sonograms generated by MUPET to the original spectrographs and found that the default SNR settings provided optimal syllable detection, with minimal detection of noise (Figure 3C–E, **steps 1–4**). Using these settings, all detected events in the sonograms were automatically converted into GF-USVs and compiled into an initial (unrefined) dataset for each strain (Figure 3C, **step 2**). Datasets are available for export as CSV files (Figure S1). From the initial datasets we built repertoires of size 100 (or greater) for each strain (Figure 3C, **step 3**) in order to identify and remove RUs representing noise prior to final repertoire builds (Figure 3C, **step 4**; Figure S1). As a point

of reference, the average time to process one 6-min .wav file is 30-sec, and repertoires can be built in less than 5 min.

MUPET generates several measures graphically (Figure 4). To determine the energy distribution along the frequency axis of the vocalizations, PSD is computed for each strain, normalized by the total power over all frequencies. The normalized PSD curves are fit well by a Gaussian distribution, defining an overall mean frequency and standard deviation for each strain (Figures 4A and S3). The mean frequency ranges from 65.5 to 95.1 kHz ($F_{(11,229)} = 4.87$, $P < 0.001$; Figures 4A,B and S3). To derive meaningful statistics defining syllable bouts, we used syllable on- and offset times to define syllable duration, syllable rate, and inter-syllable interval (Figure 4C–E) for syllable bouts that contain less than 200 milliseconds of silence. Bout length was determined empirically based on the value that eliminated bout-length outliers. When all syllable types generated by a strain are combined, only the mean frequency and frequency bandwidth show significant differences across strains. For all measures, the DBA and B6D2F1 strains appear more similar to each other than to the C57 strain, suggesting a dominant inheritance pattern.

Strain differences in syllable repertoires

The goal of the repertoire build step (Figure 3C, **steps 5–6**) is to learn the full diversity of syllable types within each dataset. MUPET builds syllable repertoires ranging in size from 20 to 200 RUs, in increments of 20 units. In signal processing approaches, determining an optimal ‘build size’ is a tradeoff between minimizing model complexity (i.e., minimizing RU number) and maximizing model accuracy (i.e., maximizing Pearson correlations between the shape of individual syllables and their RU centroid). These choices depend upon the size of the dataset and the diversity of syllable types it contains. To aid the user in selecting an appropriate repertoire size, MUPET provides 4 measures of model strength for each repertoire size: 1) the **Bayesian information criterion (BIC)**, which measures model accuracy as a function of model complexity, 2) the **average log likelihood** of the RU centroid measure, 3) the **overall repertoire modeling score**, which gives the global average of the normalized cosine distance between each syllable and its corresponding RU centroid, and 4) the **RU goodness-of-fit**, which provides the average Pearson correlation for all syllables within an RU to the RU-centroid. In **step 5**, the user builds repertoires of various sizes and then selects a range (or specific) repertoire size that minimizes BIC and maximizes the average log likelihood, overall repertoire modeling score, and RU goodness-of-fit across datasets. Note that the subsequent repertoire similarity and clustering steps require same-size repertoires across datasets. This is readily achievable for similarly sized datasets, as the combined modeling scores typically indicate a range of optimal repertoire sizes. Users can rapidly compare repertoires of several sizes to assess how different priorities (e.g., model accuracy vs. complexity) impact the repertoire similarity comparisons across datasets. To illustrate the use of the modeling measures in **step 6**, syllable repertoires of every size (20–200) for each of the 12 strains were built. The modeling measures across strains are presented in Figure 5 and in Figure S4. Dataset sizes varied across strains and this affected the repertoire size needed to minimize BIC. For all strains, BIC was minimized at ~40 RUs with the exception of DBA/2 (the largest dataset; ~51,000 syllables), which had a minimal BIC at 140 RUs. Despite this gap in minimal BIC, the rate of increase in BIC with larger

repertoire sizes was modest for the 11 other datasets. Visual inspection of the BIC data suggested an optimal size of 100–140 RUs across strains (Figure 5A). We next inspected the average log likelihood and overall repertoire modeling scores, which are always maximal at the largest repertoire size due to the increase in shape correlations as model complexity increases (i.e., an RU containing a single syllable has syllable-to-RU centroid Pearson correlation of 1.0). Here again, the rate of change of these measures slows as repertoire size increases and inspection of the repertoire sizes past the inflection points suggested an optimal repertoire size of 100–140 RUs across strains, consistent with the BIC (Figure 5B–C). These results are consistent with the RU goodness-of-fit measures, which also increase as a function of repertoire size, but improvement plateaus at 100–140 RUs across all strains (Figure 5D–F and S4). Finally, the number of RUs that contain a small number of syllables (e.g., <10) increases rapidly for repertoires larger than 140 (Figure S4) and this was used as an additional measure of model complexity. We performed subsequent comparisons with repertoires of size 100–140, with no substantive differences in the results (data not shown). Building larger repertoires will increase RU goodness-of-fit and maximize the ability to resolve rare syllable types, which otherwise may be condensed into spectrally impure units. Thus, selecting a suitable repertoire size also depends upon user goals. For clarity, we present the results for repertoires of size 100 generated during **step 6** (Figure 5G–I and S5). Many of the RUs learned by the algorithm can be associated with one of the syllable categories reported previously (Arriaga et al., 2012; Grimsley et al., 2011; Scattoni et al., 2008), but it is apparent that the algorithm is able to differentiate many forms of each canonical syllable (e.g., upward and chevron-shaped syllables with different slopes, durations, and magnitudes of frequency modulation). MUPET generates CSV files with 8 spectro-temporal measures for each RU (Figure S1).

Repertoire comparison between mouse strains – 1) Cross Repertoire Similarity Matrix

To compare vocal production between syllable repertoires, in **step 7** MUPET computes a Cross Repertoire Similarity Matrix of Pearson correlations (as well as the sorted syllable repertories, see below), which can be used for inspecting shape similarity between pairs of RUs. The algorithm generates the matrix by progressively pairing RUs with highest-to-lowest shape similarity in descending order (Figure 6A, left panel). The algorithm performs this process by identifying the RU-pair with the highest Pearson correlation, which then occupies the first position of the matrix diagonal. Each RU is paired only once and the similarity of RU shapes is considered independently of how often the RU is used by each strain. MUPET generates new images of each repertoire with RUs sorted from high-to-low similarity. Visually inspecting the matrix diagonal and sorted repertories provides the user with an initial assessment of the most and least similar RU shapes between two distinct repertoires (Figure 6A, right panels; Figure S7). Because RUs are learned without respect to mean frequency, the similarity metric determined here does not take mean frequency into account. The spectral range of frequencies in which mice vocalize is measured during syllable detection and hence can be used as a separate metric when comparing syllable repertoires.

Repertoire comparison between mouse strains – 2) Cross Repertoire Similarity Boxplot

In **step 7**, MUPET uses a boxplot representation to automatically display the similarity between a reference and to or more comparison repertoires as a function of how frequently each RU is used (Figures 6B,C and S6). The algorithm calculates the Pearson correlations for shape similarity for the top 5, 25, 50, 75 and 95% of most frequently produced RUs. Note that the repertoire similarity scores are not symmetric between the C57 and DBA reference repertoires (see boxplots in Figure 6B, C) due to the integration of RU activity. The boxplot can be used to rapidly assess and quantify overall repertoire similarity based on the 75% or 95% similarity scores, and the similarity of the most frequently produced units based on the 5% and 25% similarity scores. Upward (or downward) skew in the interquartile range with respect to the median indicates a repertoire in which the similarity scores change more considerably in the top (or bottom) half of most frequently produced RUs. In the same way, repertoires with large interquartile (or overall) ranges indicate repertories in which there are RUs that are both highly similar and dissimilar from the reference repertoire. In contrast, small interquartile (or overall) ranges, signal repertoires that are consistently similar (or dissimilar) from the reference repertoire, depending on their median similarity score. In comparison to the C57 repertoire (Figure 6B), the median similarity scores for the other strains range from 0.79 to 0.90. Examining the overall (95%) repertoire similarity scores reveals that BXD77 (0.86) and BXD43 (0.71) are the most highly similar and dissimilar to the C57 repertoire, respectively. The DBA and F1 cross have among the least similar repertories compared to C57 based on median similarity scores, yet all strains show similarity scores ≥0.9 for the top 5% of syllables (range 0.90–0.96), revealing that the most frequently used syllable types tend to be highly similar. To obtain summary statistics from activity-based Pearson correlations, MUPET generates repertoire comparison scores in 1% increments of RU activity (Figure S1). The average similarity of the 11 comparison strains to the C57 reference repertoire differs significantly ($F_{(10,160)} = 3.30$, $P < 0.001$; one-way ANOVA across 11 strains; Figure 6B). To identify the strains with significant differences in repertoire similarly compared to the C57 strain, we determined the average similarity of replicate C57 studies (Figure 7) and used a 12-strain ANOVA with Dunnett's post tests to compare the mean similarity of each strain to this hypothetical C57 comparison repertoire. The average similarity of replicate C57 studies (size 100) is 0.91 ± 0.03, with average similarly scores of 4 strains being significantly different: BXD48 ($P < 0.001$), DBA ($P < 0.01$) and B6D2F1 and BXD16 ($P < 0.05$).

In comparison to the DBA reference repertoire (Figure 6C), the median similarity scores range from 0.88 to 0.92. Overall, the top 5% of most frequently produced RUs across strains are highly similar to the DBA reference repertoire (range: 0.91–0.98). Of all 11 strains, the overall B6D2F1 repertoire is most similar to the DBA repertoire, consistent with the high similarity between these strains for other vocalization features. The average similarity of the 11 comparison strains to the DBA reference repertoire differs significantly across strains ($F_{(10,158)} = 2.80$, $P < 0.01$; Figure 6C). As above, we determined the average similarity of replicate DBA studies (Figure 7) to be 0.94 ± 0.03 and used this value in a 12-strain ANOVA. Five strains have average similarly scores that are significantly different from the hypothetical DBA repertoire: C57 ($P < 0.001$), BXD48, BXD62 and BXD77 ($P < 0.01$) and BXD6 ($P < 0.05$). In addition, the boxplot comparisons reveal that the BXD48 repertoire is

significantly different from both parental repertories, suggesting divergence from the parental strains.

The similarity metrics also provide an opportunity to assess syllable repertoire stability across studies or conditions. The C57 and DBA datasets were generated from recordings collected during 7 replicate studies performed with a C57 juvenile partner across 3 years. We also recorded vocalizations when the parental strains were paired with a juvenile male of a different genetic background (129S1). The Cross Repertoire Matrix and Boxplot similarity metrics were used to examine the similarity of syllable repertoires across studies. Four C57 studies had a small number of vocalizations (<600). Thus, to avoid potential bias in similarity due to incomplete repertoire sampling, we only performed the comparisons with the 4 C57 studies (3 C57 partner, 1 129S1 partner) and 8 DBA studies, which had sufficient syllables (>1300) to build repertoires of size 80. Figure 7 shows the matrix diagonal with specific Pearson correlations for the combined-C57 and DBA datasets and individual studies generated with juvenile partners. Together with the boxplot comparisons, the analyses reveal that within each parental strain, there are no significant differences in syllable repertoire usage across replicate studies or across studies with different strain partners. The data are consistent with the hypothesis of a strong genetic influence on syllable repertoires (Arriaga and Jarvis, 2013; Hammerschmidt et al., 2015; Mahrt et al., 2013).

Master Repertoire: RU-cluster identification using k-medoid clustering across strains

A final level of analysis is to determine the proportion of similar and unique RU types present across datasets. To do so, MUPET uses k-medoids clustering to build master repertoires composed of RU-clusters, which are groups of RUs with similar shapes, learned from all the RUs present in different datasets. In **step 8** (Figure 3C), the user determines the optimal master repertoire size, based on Pearson correlations for the shape similarity between the RUs and their RU-cluster, as well as data on the total number of RUs and syllables in each cluster. These data guide decisions on optimal master repertoire size. We note that RU-clusters containing a single RU will have a Pearson correlation of 1.0. We applied a k-medoids clustering on the combined set of 1200 RUs learned from the repertoires of each of the 12 strains (Figures 8 and S5). To illustrate the use of the Pearson correlations, in **step 9** we built master repertoires from 5–100, in 5 unit increments. A master repertoire of size 45 (Figure 8B) maximizes the proportion of RU-clusters that have an average Pearson correlation greater than 0.8, yet minimizes the proportion of RU-clusters containing a relatively small number of RUs (5 or less out of a total of 1200 RUs; Figure 8A). Master repertoire modeling can also be considered in terms of the proportion of RUs and syllables that are contained within RU-clusters with different levels of goodness-of-fit (Figure S8). The k-medoids clustering approach is similar to the k-means approach used during repertoire learning to identify and represent similar shapes, with the exception that the displayed unit representing the ‘RU-cluster’ is an actual RU from an individual dataset repertoire and not the mean (weighted combination) of multiple RUs. For example, the image shown for RU-cluster #28 from master repertoire size 45 (Figure 8B) is C57BL/6 RU #68 (see Figure 5G). Thus, the RU that is displayed is the cluster medoid and represents the shape of the RUs in the cluster. Medoid-based clustering was selected for this stage to avoid excessive blurring of the RU-cluster shapes, which could occur if the shapes of a relatively

small number of RUs were averaged. In this context, the Pearson correlations for the RU-clusters provide an important indicator of goodness-of-fit. Nevertheless, there are possible limitations in the display of a single RU, rather than an RU average, in facilitating the interpretation of the overall type differences between RU-cluster categories.

We determined the ‘strain of origin’ of different RU-cluster types based on whether the RUs (and syllables) within the RU-cluster were: 1) shared by both parental strains, and could be present in offspring strains, 2) unique to one parental-strain, but could be present in offspring strains, and 3) unique to the offspring strains, meaning observed in the F1-cross or BXD offspring strains, but not in the parental strains. Thus, the emergence of unique syllable types in the F1 cross and BXD strains can be identified. Figure 8C shows the assignment of each of the RU-clusters to the strain of origin categories. Visual examination of the syllable types contained in the parental, F1 cross, and BXD strains of origin reveal a diversity of types, including chevrons, chevrons with tails (‘waves’, ‘complex’ calls) and chevrons with a low frequency base note, two-component calls, short and upward shapes. BXD16 generated all the “Single BXD” RU-clusters shown in Figure 8C, revealing high repertoire diversity in this strain. Figure 8D–E summarize the distribution of syllable types as a function of the percentage of syllables (Figure 8D) and the percentage of RUs (Figure 8E) that are present within RU-clusters from each strain of origin. Examining the ‘Both Parental’ columns in Figure 8D–E reveals that for each offspring strain, a high proportion of syllables and RUs are contained within RU-clusters that include syllables from both parental strains. Here, BXD16 and BXD48 show the greatest syllable repertoire diversity, while BXD79 shows the least. Future application of MUPET in studies in unrelated inbred and mutant strains, or under different social conditions, may reveal greater syllable diversity or all-or-none patterns of syllable usage compared to a GRP, which tends to show continuous trait variation.

Comparison of selected software programs

MUPET is one of several software programs currently available to analyze mouse USVs. Each takes a different approach to syllable detection, classification, and repertoire comparison. To aid the user in selecting a program(s) that will best address their experimental questions and goals, Table 1 provides a summary of the theoretical approach, key design features, input and output materials, and the syllable and syllable repertoire analyses provided by three automated software programs: MUPET, MSA (Arriaga et al., 2012; Holy and Guo, 2005), and VoICE (Burkett et al., 2015). Each program classifies syllables based on different criteria—MUPET employs unbiased discovery of hundreds of unnamed syllable patterns, while MSA and VoICE generate a smaller number of named categories based on pre-defined rules (Arriaga et al., 2012; Holy and Guo, 2005; Scattoni et al., 2010). MUPET and MSA are high-throughput and provide similar spectro-temporal measures for each syllable, while VoICE can be challenging to apply to very large datasets due to .wav file pre-processing requirements and manual classification steps, especially if vocalizations are highly variable. MUPET is the only tool to provide automated repertoire similarity comparisons and repertoire clustering for mouse vocalizations. MUPET is uniquely sensitive to novel syllable types and advances new signal detection features (noise removal, time-stamp), which will facilitate relating syllables types to behavior states and

transitions. Syntax analysis using separate software is advanced by MUPET and MSA, providing exportable files with syllable sequences and categories. We note that the large number of syllable types identified as RUs and RU-clusters by MUPET will require advanced sequence analysis approaches to analyze syntax. MSA provides an option of generating 4–15 syllable categories, enabling a simplified syntax analysis that has been used successfully to identify syntactical changes across conditions and strains (Chabout et al., 2015; Chabout et al., 2016).

Comparison of data analyzed with MSA and MUPET

Using MUPET and MSA, we analyzed three large, open source datasets (Chabout et al., 2015) available on mouseTube (Torquet et al., 2016). We reanalyzed syllable features and repertoires from recordings of sexually-experienced B6D2F1 males vocalizing in response to female urine (UR), an anesthetized female (AF), and awake female (FE). MUPET and MSA generated similar syllable counts and spectro-temporal measures of individual syllables (data not shown). There was a high degree of similarity between dataset measures for each condition and those generated from our study of B6D2F1 males vocalizing in response to a juvenile male [e.g., PSDs are diverse in the panel (Figure 4), but were highly similar across B6D2F1 conditions; data not shown]. MSA classified syllables into 4-categories: 1) simple (S)—syllables composed of a single note, 2) up-jump (U) or 3) down-jump (D)—syllables containing one frequency jump in the up or down direction, and 4) multiple (M)—syllables containing more than one frequency jump. Chabout and colleagues found that males generate simpler syllables in response to an awake female compared to more complex (M and D) syllables in response to female urine, with the following category breakdowns: UR: 65% S, 17% M, 16% D, and 2% U; AF: 83% S, 4% M, 9% D, 4% U; FE: 80% S, 7% M, 10% D, 3% U. The syllable repertoires built by MUPET for each social condition reveals readily apparent differences in syllable types and frequency of use (Figure S9). MUPET is unique in enabling the user to visualize the full diversity of syllable types composing the repertoire. Figure S10 quantifies these differences using the Cross Repertoire Matrix and Boxplot similarity metrics, with both showing that the UR and FE repertoires are least similar and the AF repertoire is highly similar to both the UR and FE repertoires. These data corroborate the increased repertoire complexity in the UR condition and provide new evidence that the AF repertoire, in which the female is present, but unresponsive, is intermediate in complexity to the UR and FE conditions. MUPET enables the user to visualize similar and dissimilar syllable types between each condition in the sorted repertoires generated with the matrix diagonal. To further identify shared and unique RU types across repertoires we created a master repertoire of 35 RU-clusters (Figure S11). Analysis reveals that 100% of the FE repertoire contains syllable types that are also present in the UR and AF conditions, while only ~91% of the AF and ~71% of the UR repertoires are shared across conditions. MUPET also shows syllable shapes composing the non-shared portions of the AF and UR repertoires, revealing that there is an increased number of syllables whose frequency is highly modulated (large-bandwidth), including chevrons (e.g., RU-cluster 7), continuous upward notes (e.g., RU-cluster 14), and chevrons with a lower frequency base note (e.g., RU-cluster 1). Neither the 4-category classification scheme nor the criteria described in Scattoni et al., 2010 would reveal differences in the types of chevrons used (e.g., compare RU-cluster 31 and 26 in the shared condition to RU-cluster 7

in the UR condition), missing the opportunity to identify both the diversity of syllable types that are altered in different social conditions as well as the specific ways in which they are altered (e.g., altered levels of frequency modulation). Analysis with MUPET thus provides increased sensitivity to detect changes in syllable repertoires across strains or conditions.

DISCUSSION

MUPET software was developed as an open-access tool to provide advanced capabilities to generate and analyze mouse syllable repertoires. The software uses signal processing approaches similar to those applied to human speech (Rabiner and Schafer, 2010), including SNR optimization, filterbank-based transformation of sonograms into low-dimensional (compact) feature representations (GF-USVs), and machine learning algorithms to extract recurring syllable types. MUPET provides the ability to remove RUs learned from noise and rapidly generate spectro-temporal measures for syllables and datasets, such as amplitude, mean frequency, duration, and syllable rate, which may hold salient communicative information. MUPET also employs new repertoire similarity metrics and centroid-based clustering of RUs to assess differences in vocal production and repertoire use across strains and conditions. This combination of features makes MUPET well-suited for the efficient comparison of syllable types extracted from large datasets. In humans, these signal processing approaches have related subtle differences in vocal production (e.g., cadence, intonation) to underlying differences in emotionality and intention (Narayanan and Georgiou, 2013). This has not been readily attainable in mice. MUPET identifies subtle variations in syllable production and use, enabling the investigation of the communicative signification of shape variations and the influence of genetic factors and behavioral states on patterns of communication. The automated time-stamp feature of MUPET establishes a means for relating communication patterns to behavioral status across time. These have been long-standing goals (Grimsley et al., 2011; Holy and Guo, 2005; Lahvis et al., 2011; Sewell, 1970). While not done here, MUPET also has the capability to be adapted for analysis of vocalizations of species with different frequency ranges by deriving species-specific filterbanks that optimally represent the spectral information in syllables and updating software algorithms (see **STAR Methods**). Finally, as an open access software platform, the scientific community can contribute additions and improvements to MUPET's analytical capabilities.

MUPET advances data-driven syllable repertoire construction and analysis

A key feature of MUPET is the ability to automatically learn and compare the recurring syllable types (RUs) that are present in datasets containing thousands of syllables. Given the signal processing approach used by MUPET, the results depend upon three elements of the datasets: 1) number of syllables used to train the algorithm, 2) size of the repertoire build, and 3) size of the master repertoire build (RU-clustering). Because of the unbiased approach to developing the repertoire, MUPET is leveraged most effectively when trained with a modest-to-large number of syllables, which enables the software to learn the shapes of common and rare syllable types. For each experiment, determining the number of required syllables to build an accurate repertoire will depend upon the diversity of syllable production in the mouse strains and behavioral conditions being analyzed. MUPET provides four

repertoire modeling scores, which enable the user to select repertoire size(s) that balance model complexity and accuracy. It is important to note that the algorithm will extract the specified number of RUs. Thus, if a repertoire consisted solely of one syllable type (e.g., chevrons), MUPET will extract hundreds of examples of chevrons from the dataset, which may differ in specific parameters, such as duration and shape. Here, the repertoire modeling scores aid the user in assessing the number of RUs needed to capture the full diversity of syllables types, and in this example, would likely support a smaller repertoire build. Determining the biological distinguishability and significance of separate, but visually similar, RUs and RU-clusters is a high priority for future work and will be aided by psychophysical and behavioral experiments assessing the perceptual discrimination and social significance of different RUs and RU-clusters and overall repertoire variations.

An additional feature of MUPET is that the efficiency of automated analysis allows users to assess the precision of a syllable repertoire (convergence upon the true “strain” syllable repertoire in terms of syllable types and use) by empirically resampling the dataset and determining the number of syllables needed to obtain stable repertoire similarity metrics and RU-clusters. In doing this, we found that DBA and C57 master repertoires that were built from ~7K (C57) and ~50K (DBA) total syllables recorded across 7 separate studies were highly similar to the repertoires built from the analyses of the individual studies that contained only 1.3–12K syllables, suggesting a lower limit of ~1.3–4K syllables to obtain accurate strain repertoires (Figure 7). This is a modest number of syllables, given that the builds are typically based on multiple recordings from a particular strain. Additionally, mice typically produce hundreds of syllables per minute during DSI with a juvenile male, and even more when paired with an estrous female or during isolation distress. We note that MUPET also is useful for rapid assessment of types present in smaller datasets (or even from single mice) using smaller repertoire builds.

Analysis of a genetic reference panel highlights MUPET’s features

The present study analyzed syllable repertories from a subset of high vocalizing strains from the BXD RI mouse panel, which share various combinations of C57 and DBA genomes. The analyses provide evidence of heterogeneity in syllable production and use across the panel and confirm genetic regulation of syllable repertoires in the parental stains. We used two novel similarity metrics to measure the overall similarity of spectral shapes between repertoires, both dependent upon (and independent of) the frequency of RU use. Based on the RU-cluster analyses, we observed RU types that were 1) shared—observed in both parental strains and could be present in offspring strains, 2) unique to a parental-strain—observed in only one parental strain, but could be present in offspring strains, and 3) unique to the offspring strains—observed in the F1-cross or BXD offspring lines, but not in the parental strains. There was a high degree of repertoire similarity across replicate studies of the DBA and C57 strains, as well as when parental strains were paired with a C57 versus a 129S1 juvenile partner. The findings indicate that strain genetics of the adult is an important determinant of USV syllable type. The study design using RI strains allowed us to further assess the genetic architecture of vocalization in the F1 cross. We found that all parameters were most similar to the DBA parental strain, consistent with vocalization parameters being driven by DBA alleles in the F1 cross. We also observed a high degree of syllable repertoire

similarity between sexually-naïve F1 males paired with juvenile males in the present study and with sexually-experienced F1 males paired with female stimuli in Chabout et al., 2015, but with quantifiable differences for each of these social conditions (data not shown). These findings are consistent with growing evidence that mouse vocal communication is under strong genetic control (Hammerschmidt et al., 2012; Kikusui et al., 2011; Mahrt et al., 2013). The scalability of MUPET will enable complex analyses of syllable production and use in different behavioral contexts and strains, providing a new tool to address the biological significance of vocalization differences.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Pat Levitt (plevitt@med.usc.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Animals—We examined heterogeneity in mouse ultrasonic vocalization (USV) repertoires using a subset of 12 strains from the BXD genetic reference panel: C57BL/6 (C57) and DBA/2 (DBA) parental strains, F1 cross (B6D2F1), and 9 recombinant inbred offspring strains (BXD6, 16, 29, 42, 43, 48, 62, 77, and 79) (Peirce et al., 2004; Taylor et al., 1977; Taylor et al., 1999). Adult experimental mice were obtained from The Jackson Laboratory (Bar Harbor, ME) at 7–10 weeks of age and allowed to acclimate to the facility for 2–4 weeks prior to testing. Juvenile males were bred in house from breeders obtained from The Jackson Laboratory. To determine whether the genetic background of the juvenile partner influences syllable production or use, in separate studies we examined USVs generated by the parental strains during interaction with C57 or 129S1/SvImJ (129S1) juvenile males. Experimental mice were housed in same strain pairs and juvenile partner mice were group-housed (2–3/cage) following weaning at P21. Mice were maintained on a 12-hour light/dark cycle (lights on 6:00 A.M. to 6:00 P.M.) with *ad libitum* access to food and water, except during testing, and behavioral tests were conducted during the light cycle (between 8:00 A.M. and 5:00 P.M.). All procedures were approved by the Institutional Animal Care and Use Committee of the University of Southern California and conformed to National Institutes of Health guidelines. As part of a larger study of the BXD genetic reference panel, 8 cohorts of C57 and DBA parental strains (n=10/cohort) were tested across 3 years, providing an opportunity to examine the stability of USV repertoires across cohorts.

METHOD DETAILS

Detailed methods for the mouse USV studies and the development and implementation of MUPET are provided below. Our rationale for selecting specific signal processing methods and analysis strategies used by MUPET are described in detail after the methods.

Direct social interaction task: Affiliative social interaction and vocal communication were assessed using a direct social interaction (DSI) task with synchronous USV recording. In this task, adult experimental mice were acclimated to the testing chamber (30 × 19 × 19 cm, L × W × H; transparent polycarbonate) for 10-min and then allowed to freely interact with

an unfamiliar juvenile male for 6-min. Juvenile males (P26–P30) were used in order to minimize aggressive and sexual behaviors. Interaction sessions were video-taped and vocalizations were recorded using a CM16/CMPA ultrasound microphone, positioned 16 cm above the chamber floor, and an UltraSoundGate 116H recorder (Avisoft Bioacoustics). Videos were scored for the duration and frequency of sniffing the juvenile, self-grooming, aggression and sexual behavior using MOOSES Observation software (Jon Tapp, Vanderbilt Kennedy Center). For all strains the nature of the interaction was affiliative and contained negligible levels of aggression or sexual behavior. All USVs were attributed to the adult (experimental) male based on several lines of evidence that juvenile males do not vocalize during this task: 1) we did not observe overlapping vocalization streams, consistent with a single vocalizer, 2) vocalizations were highly synchronous with adult sniffing of the juvenile, 3) we observed broad heterogeneity in the number of USVs generated by different BXD strains, but consistent levels within a strain, despite all strains being paired with a C57 juvenile, and 4) MUPET provided evidence of distinct syllable types, which were driven by the identity of the experimental strain, rather than the genetic identity of the juvenile.

Audio pre-processing: MUPET automatically performs all pre-processing of audio files and currently supports .wav file formats. Audio files were each 6-min in duration and collected at a sampling rate of 250 kHz. MUPET can be used to analyze files of any length and with sampling rates larger than 90 kHz. All analyses in this study were performed using MATLAB® version 14b on a Macbook Pro running OS X Yosemite (2.5 GHz Intel Core i5 processor, 8 GB 1600 MHz DDR3 memory). The audio files are first high-pass filtered using an 8^{th} order Chebychev filter with a 25 kHz corner frequency in order to extract the ultrasonic frequency range. MUPET generates the sonograms by calculating the power spectrum on Hamming windowed data using a frame size of 500 samples (2 msec) and a frame shift of 400 samples (1.6 msec). The spectra are computed using a 512-point STFT algorithm resulting in a frequency resolution of ~0.5 kHz. The parameter settings were selected empirically to optimize the trade-off between time and frequency resolution. For subsequent feature analysis, background noise reduction was performed on the sonogram by spectrally subtracting the noise floor spectrum computed over the ultrasonic frequency range. The user can modify the degree of noise subtraction to a desired trade-off between minimizing the number of noise events and enabling the detection of faint syllables. All analyses were performed using the default SNR settings in MUPET (Noise-reduction, 5.0 [scale 0–10]; Minimum syllable duration, 8.0 msec; Maximum syllable duration, 200 msec; Minimum syllable total energy −15 dB; Minimum syllable peak amplitude −25 dB and Minimum syllable distance [hold-time], 5.0 msec).

Gammatone filterbank: Non-negative matrix factorization (NMF) decomposes the spectral data of the USVs into basis units from which meaningful information can be derived about the acoustic production of mouse syllables, which is likely related to their auditory acuity (Ehret and Haack, 1982; Holmstrom et al., 2010; Neilans et al., 2014) (Figure S2). This strategy has been used in a similar fashion to find optimal decompositions for the analysis of human speech (Stevens et al., 1937; Zwicker, 1961). The time-frequency representation of mouse vocalizations can be interpreted as a superposition of narrow frequency bands at different spectral energy. To find the frequency band decomposition we applied NMF (Lee

and Seung, 2001) on the sonograms. At each frame of the sonogram the spectral vectors are normalized to obtain unit energy. By concatenating all spectral vectors over all available data from a given mouse strain, a matrix **V** is constructed of dimension **N** × **T**, where **N** is the number of frequency bins and **T** is the total number of frames in the audio file. By applying NMF on V, and by enforcing non-negative constraints on all matrices, the matrix is approximated by

$$\mathbf{V} \approx \mathbf{WH} \text{ subject to } D(\mathbf{V}\|\mathbf{WH}) \quad (1)$$

$$\text{with } D(\mathbf{V}\|\mathbf{WH}) = \sum_{ij}\left(\boldsymbol{C}_{ij} \log \frac{\boldsymbol{C}_{ij}}{(\boldsymbol{WH})_{ij}} - \boldsymbol{C}_{ij} + (\boldsymbol{WH})_{ij}\right) \quad (2)$$

The Kullback-Leibler divergence criterion is used as a cost function to address the high dynamic range of syllable amplitude across frequencies, thus normalizing frequency data independent of amplitude. The factorization corresponds to a linear combination of basis spectra that characterize the recordings and which are found in the columns of **W**, while the corresponding rows of **H** contains activation values of these spectra. The number of basis spectra N is a design parameter of the algorithm and is a trade-off between obtaining an accurate frequency resolution per frequency band and finding multiple base spectra modeling the same frequency band. We found that a good choice for N is 64 to capture the most relevant acoustic information along the frequency axis while minimizing the number of frequency filters. The value of N defines the dimensionality of the Gammatone feature representation and permits the computational feasibility of subsequent signal processing analysis. Figure S2 shows the outcome of this approach applied on audio recordings from several mouse strains after sorting the base functions according to peak frequency. Regression analysis on the base function's peak frequencies produces a logistic curve (Figure 1A, S2), which converts the peak frequencies **f** in the ultrasonic range into the Gammotone filterbank scale:

$$n = \frac{N}{1 + \mathrm{e}^{-\gamma(f_0 - f)}} \text{ with } \gamma = 2\alpha/\mathrm{f_s}. \quad (2)$$

Here, f_s is the sampling frequency and N corresponds to the chosen number of filters in the filterbank. The midpoint frequency $\mathbf{f}_0$ and the slope variable α were derived by regression analysis applied on the basis spectra spanned by the columns of **W** of the NMF algorithm: f_0 = 68.5kHz and α = 16.2. From equation (2) we can derive the center frequencies f_n of the Gammatone scale filterbank. The associated equivalent rectangular bandwidth (ERB) of each n^{th} Gammatone filter is set equal to:

$$\mathrm{ERB(n)} = \frac{1}{2}(f_{n-1} - f_n). \quad (3)$$

The filterbank integration of the USV sonograms represents the USVs in the spectral domain as a weighted linear combination of the band-pass filter functions. The weights of each function relate to the spectral magnitude associated with the corresponding filter. This mathematical model also resembles the task of an auditory filterbank operation. To smooth out frequency peak energies of the filterbank outputs, additional post-processing was performed by applying an autoregressive moving average (ARMA) filtering on the filterbank integrated spectra. The resulting feature representation is the Gammatone Filterbank USV feature (GF-USV). Calculation of minimum, maximum, starting and ending frequency is done by selecting the minimum and maximum Gammatone filter (out of 64) and then searching for these features in the corresponding frequency band. Based on evidence that the parental and BXD strains vocalize at a similar mean frequency, and to avoid biases in shape extraction that could be caused by using filters with different resolution across the ultrasonic range, we conducted all analyses with the same filterbank. While MUPET is readily applicable to the analysis of vocalizations from other developmental ages of mice, or other species that vocalize in the ultrasonic range, one needs to establish an optimal filterbank for new species and possibly for aged or developing mice (see below).

Clustering methods for repertoire learning: The first step in the syllable repertoire building approach is the segmentation of the audio recordings into individual syllables. To this end, we applied the syllable activity detector to find the beginning and ending time for each syllable. Each segmented window was transformed into the GF-USV feature representation and padded until a window length of 200 milliseconds, which corresponds to the maximum syllable duration. These window-extended patches are subsequently centralized in both time and frequency. The latter step is required to constrain the clustering of the syllables primarily to their spectral shape. We apply image clustering on the vectorized images of these centralized time-frequency shapes by means of k-means clustering. This is accomplished by determining the cosine distance between two vectorized images. The outcome of the repertoire machine learning algorithm is a set of cluster centroids. The time-frequency representations of the centroids are repertoire units (RUs) that represent the population based on the cluster analysis and compose the syllable repertoire. To address the problem of finding an optimal repertoire size, we used the Bayesian Information Criterion (BIC), average log likelihood, overall repertoire modeling score, and RU goodness-of-fit measures that are generated for each repertoire size (see **Results**). In addition to repertoire learning, MUPET allows the user to further refine the repertoires by removing undesired units, e.g. units that model noise events. Repertoire refinement involves deleting the RUs that correspond to the undesired clusters from the dataset and regenerating the repertoire using only the desired units as cluster centroids to initialize the k-means clustering. To compare the frequency of use of similar and unique RU types across multiple datasets, MUPET uses k-medoids clustering to generate ‘master’ repertoires, which are smaller numbers of RU-clusters identified from the total number of dataset RUs. Each unit of the master repertoire therefore represents a group of spectrally similar RUs (RU-clusters)

from the individual datasets. MUPET generates RU-cluster goodness-of-fit measures to aid the user in selecting an optimal master repertoire size(s) (see **Results**).

Use of MUPET with other species: In the current version of MUPET, we have derived a filterbank that optimally represents the spectral information in mouse syllables (see **Results** for a summary of the specific processing steps). For users that are interested in using MUPET with other species we recommend the following strategy: 1) Determine the vocalization frequency range and call duration and rate for the species. Any knowledge about species-specific vocal production mechanisms can inform choices for analyzing the vocalization properties such as overall frequency ranges and temporal patterning of sounds produced. The duration of each vocalization will determine the analysis time window. For mouse USVs, we used a STFT algorithm of 512 bins to derive the sonograms and Hamming windowed frames with an approximate overlap of 75% of the frame length. 2) In the next step, the user needs to estimate the spectral bands into which this sonographic representation can be decomposed. In our work, we have used NMF to decompose the sonograms into 64 spectral base functions. Each base function is characterized by a peak frequency and a frequency band, and can be approximated by a gammatone band-pass function centered around the peak frequency with an equivalent rectangular bandwidth (ERB) (see **STAR Methods, Gammatone filterbank**). To uniquely derive a filterbank for a new species, we advise applying the NMF algorithm to sonograms from large numbers of clean audio recordings (i.e. several hours of recordings with thousands of vocalizations). This ensures that all possible calls and associated spectral shapes are represented and ensures that the spectral base functions retrieved from the NMF algorithm will represent well the full spectral range of vocalizations expected from the new species. 3) The final step in uniquely defining the filterbank is to construct a mathematical function that analytically describes the center frequencies of the gammatone filters of the filterbank. This function corresponds to the best fitting line that connects the peak frequencies of NMF base functions. One can derive the function formula either empirically or by using regression tools. Users may contact the authors to facilitate derivation of the new filterbank and for support analyzing the vocalizations of different species in MUPET.

Comments on MUPET design features

NMF strategy: When NMF is applied to the spectrogram of speech signals, the NMF base functions resemble the individual filter response of the perceptual filterbank (see Smaragdis, 2007). Most speech processing applications primarily rely on speech filtered by this filterbank in the pre-processing steps as it easily allows software to discriminate, categorize and recognize difference between phones (e.g. for machine speech recognition, speaker verification). We opted for a similar strategy when processing the mouse USVs, that is, to apply NMF on the sonograms of USVs to estimate a filterbank that optimally extracts the spectro-temporal information. We observed that the peak frequencies of the base functions resembled a sigmoid shaped function, which could be controlled by two variables: the midpoint frequency and the slope. The midpoint frequencies can physically be interpreted as the center frequencies of the vocalization, while the slope characterizes the frequency sensitivity of the acoustical information in the USV band in which the mice vocalize. There are two major reasons why a model fit using low-order polynomials will not work well for

MUPET. Whereas some calls are simple in structure (e.g., syllables comprised of single notes), and thus can be accurately modeled by low-order polynomials, others are complex (e.g., harmonic and multi-note syllables), and need to be modeled by either high-order or multiple polynomials. Therefore, it would not be possible to accurately describe the wide variety of syllable types with low-order polynomials. Secondly, polynomial based representations do not easily lend themselves to call clustering and repertoire generation. In contrast to the feature based representation, which is obtained by the gammatone-scaled spectral filtering and can be clustered using standard pattern classification methods, clustering of polynomials requires clustering strategies that rely on heuristics and are not generalizable.

We also refer readers to work in which a similar strategy was followed for human speech (and other audio) signals (Bertrand et al., 2008; Smaragdis, 2007). We are not aware of any other alternatives than NMF to extract the spectro-temporal patterns in an unsupervised manner, due to the non-negative constraint that is enforced on the data, as opposed to PCA or ICA based methods. We have reviewed more recent approaches that involved Convolutional Neural Network and Autoencoder that are able to extract base functions similar to NMF, but these involve supervised methods, i.e. syllable labels. Thus, we believe NMF on GF-USVs is an ideal choice for MUPET.

Use of NMF Loadings: The loading matrix that emerges from the factorization shows the activation of the NMF bases along the frequency axis for each time frame. The difference between the scree plots of the activations and the outputs produced by the Gammatone-filterbank is that the NMF has been applied on the matrix V, which contains spectral frames that are normalized to unit energy (see **STAR Methods**, **Gammatone filterbank**). Hence, the loadings do not contain any information regarding the energy distribution. Loadings are not used directly to prevent variability in the generated filterbank. The NMF should be applied offline and independent of the dataset to be processed. The pre-computed NMF on a large dataset of various mouse calls produces a mathematical formulation that well represents the more general USV spectrum of mouse vocalizations.

Filter number and bandwidth: The design strategy in MUPET to use GF-USVs resulted from an objective trade-off between optimally representing the spectral shape of the syllables, while minimizing computational complexity. From our experiments in the MUPET framework, and given the redundancy present in the acoustic representations, we found that USV spectral details are well preserved for values above 32 filters. In the software, we conservatively set the number of filters to double this value (64), at the cost of more computational requirements that are needed to process the data in later stages (e.g., repertoire generation and comparisons). Using a fixed bandwidth with an increased number of filters also works in MUPET. However, there are two reasons why the use of a 64-component filterbank with variable bandwidth filters is optimal. The NMF algorithm estimates base functions with different bandwidth across the ultrasonic spectrum. This suggests that in the frequency regions where the bandwidth is higher, we need less information to characterize the syllable shapes. Thus, variable bandwidth filters enable one to represent the same amount of information with a smaller number of feature components.

This design feature reduces the computational requirements without sacrificing accuracy. Another benefit comes from the higher spectral resolution in the frequency regions where most information is present, i.e. in the regions proximate to the mid-frequency. The gammatone-spectral representation with variable bandwidth filters exploits this behavior by assigning more feature components to these regions. This has a positive impact on syllable clustering and repertoire generation, because the most relevant spectral regions receive a higher weight in the decision making process.

Power spectral density: The Gaussian fit for the PSD is an approximation that is designed to provide an estimate of the frequency bandwidth, but not to optimally model the PSD (i.e., in some cases, two or more Gaussian curves would yield a better model for the PSD). Here, the use of a single Gaussian adequately estimates the frequency bandwidth and allows for an objective comparison between different datasets.

Selecting a repertoire build size: The k-means algorithm is designed to find a good fit within clusters, and high distance between them, by means of solving the Expectation Maximization problem and the distance function of choice. Picking a suitable repertoire size remains a user-defined step (guided by goodness-of-fit metrics) due to the nature of solving an unsupervised clustering problem. The lack of a priori knowledge of how many different syllable types are present in a dataset limits the ability to objectively measure the accuracy of the clustering or to train a supervised (more accurate) model. While there is no better method currently available to determine repertoire sizes, we expect that additions from open source users of the current version of MUPET, and our own efforts, will facilitate improvements in the labeling task and clustering methods.

Clustering approach: There are different categories of clustering techniques such as hierarchical, agglomerative, Bayesian and partitional clustering (to which k-means belongs). Hierarchical clustering is very slow on large datasets and makes this computationally (O(2^n)) intractable. Bayesian clustering on the other hand requires labels available for some data of the training datasets to generate a posteriori distribution over the data. These labels (knowledge of different syllable types present) are not available when clustering the USV syllables, which leaves the option of a partitional clustering method. From the different methods available in partitional clustering, MUPET utilizes the k-means algorithm due to its popularity of usage and its efficient and easy to understand implementation. We have further modified the k-means distance criterion by a cosine distance function to ensure syllables are clustered based on their space, and not their energy or amplitude (as would be the case with the conventional Euclidean distance function). There are limitations of k-means clustering, and of other clustering methods. MUPET will become fully automated with future efforts to advance the clustering steps and empirically define the distance between clusters.

QUANTIFICATION AND STATISTICAL ANAYSIS

Statistical Analyses—Data are presented as the mean ± SEM unless otherwise noted Differences in the means of three or more groups were tested using one-way analysis of variance (ANOVA) followed by Dunnett’s post hoc tests.

DATA AND SOFTWARE AVAILABILITY

The latest version of MUPET, including a subset of the audio recordings described in this paper and an experimental tutorial, are available for download at http://sail.usc.edu/mupet

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Science Foundation Grant IIS1029373 (S.N.), Autism Speaks Translational Postdoctoral Fellowship 7595 (A.T.K.), Project 2 of the Conte Center Grant P50 MH096972 and the Simms/Mann Chair in Developmental Neurogenetics (P.L.). We thank Vikram Ramanarayanan, Hanke Heun-Johnson, Ryan Kast and Kathie Eagleson for helpful discussions on the manuscript and software.

References

Adolphs R, Damasio H, Tranel D. Neural systems for recognition of emotional prosody: a 3-D lesion study. Emotion. 2002; 2:23–51. [PubMed: 12899365]

Arriaga G, Jarvis ED. Mouse vocal communication system: are ultrasounds learned or innate? Brain Lang. 2013; 124:96–116. [PubMed: 23295209]

Arriaga G, Zhou EP, Jarvis ED. Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds. PLoS One. 2012; 7:e46610. [PubMed: 23071596]

Bennur S, Tsunada J, Cohen YE, Liu RC. Understanding the neurophysiological basis of auditory abilities for social communication: a perspective on the value of ethological paradigms. Hear Res. 2013; 305:3–9. [PubMed: 23994815]

Bertrand, A., Demuynck, K., Stouten, V., Van hamme, H. Unsupervised learning of auditory filter banks using non-negative matrix factorization. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on; 2008. p. 4713-4716.

Bradbury, JW., Vehrencamp, SL. Principles of Animal Communication. 2. Sunderland MA: Sinauer Associates; 2011.

Bregman A, Campbell J. Primary auditory stream segregation and perception of order in rapid sequences of tones. Journal of Experimental Psychology. 1971; 89:244–249. [PubMed: 5567132]

Burkett ZD, Day NF, Penagarikano O, Geschwind DH, White SA. VoICE: A semi-automated pipeline for standardizing vocal analysis across models. Sci Rep. 2015; 5:10237. [PubMed: 26018425]

Chabout J, Sarkar A, Dunson DB, Jarvis ED. Male mice song syntax depends on social contexts and influences female preferences. Front Behav Neurosci. 2015; 9:76. [PubMed: 25883559]

Chabout J, Sarkar A, Patel SR, Radden T, Dunson DB, Fisher SE, Jarvis ED. A Foxp2 Mutation Implicated in Human Speech Deficits Alters Sequencing of Ultrasonic Vocalizations in Adult Male Mice. Front Behav Neurosci. 2016; 10:197. [PubMed: 27812326]

Chabout J, Serreau P, Ey E, Bellier L, Aubin T, Bourgeron T, Granon S. Adult male mice emit context-specific ultrasonic vocalizations that are modulated by prior isolation or group rearing environment. PLoS One. 2012; 7:e29401. [PubMed: 22238608]

Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1980; 28:357–366.

Doupe AJ, Kuhl PK. Birdsong and human speech: common themes and mechanisms. Annu Rev Neurosci. 1999; 22:567–631. [PubMed: 10202549]

Ehret G, Haack B. Ultrasonic recognition in house mice: Key-stimulus configuration and recognition mechanism. Journal of Comparative Physiology. 1982; 148:245–251.

Fischer J, Hammerschmidt K. Ultrasonic vocalizations in mouse models for speech and socio-cognitive disorders: insights into the evolution of vocal communication. Genes Brain Behav. 2011; 10:17–27. [PubMed: 20579107]

Fletcher H. Auditory patterns. Reviews of Modern Physics. 1940; 12:47–65.

Gelfand, S. Hearing: an introduction to psychological and physiological acoustics. 5. New York: Taylor & Francis Group; 2009.

Grimsley JM, Gadziola MA, Wenstrup JJ. Automated classification of mouse pup isolation syllables: from cluster analysis to an Excel-based “mouse pup syllable classification calculator”. Front Behav Neurosci. 2013; 6:89. [PubMed: 23316149]

Grimsley JM, Monaghan JJ, Wenstrup JJ. Development of social vocalizations in mice. PLoS One. 2011; 6:e17460. [PubMed: 21408007]

Gunawan, T., Ambikairajah, E. Speech enhancement using temporal masking and fractional Bark gammatone filters. Proceedings of the 10th Australian International Conference on Speech Science & Technology; 2004. p. 8-10.

Hammerschmidt K, Radyushkin K, Ehrenreich H, Fischer J. Female mice respond to male ultrasonic ‘songs’ with approach behaviour. Biol Lett. 2009; 5:589–592. [PubMed: 19515648]

Hammerschmidt K, Reisinger E, Westekemper K, Ehrenreich L, Strenzke N, Fischer J. Mice do not require auditory input for the normal development of their ultrasonic vocalizations. BMC Neurosci. 2012; 13:40. [PubMed: 22533376]

Hammerschmidt K, Whelan G, Eichele G, Fischer J. Mice lacking the cerebral cortex develop normal song: insights into the foundations of vocal learning. Sci Rep. 2015; 5:8808. [PubMed: 25744204]

Hanson JL, Hurley LM. Female presence and estrous state influence mouse ultrasonic courtship vocalizations. PLoS One. 2012; 7:e40782. [PubMed: 22815817]

Holmstrom LA, Eeuwes LB, Roberts PD, Portfors CV. Efficient encoding of vocalizations in the auditory midbrain. J Neurosci. 2010; 30:802–819. [PubMed: 20089889]

Holy TE, Guo Z. Ultrasonic songs of male mice. PLoS Biol. 2005; 3:e386. [PubMed: 16248680]

Joder, C., Schuller, B. Exploring nonnegative matrix factorization for audio classification: Application to speaker recognition. Speech Communication; ITG Symposium; Proceedings of; 2012. p. 1-4.

Johnson K. The auditory/perceptual basis for speech segmentation. Ohio State University Working Papers in Linguistics. 1997; 50:101–113.

Kikusui T, Nakanishi K, Nakagawa R, Nagasawa M, Mogi K, Okanoya K. Cross fostering experiments suggest that mice songs are innate. PLoS One. 2011; 6:e17721. [PubMed: 21408017]

Konopka G, Roberts TF. Animal Models of Speech and Vocal Communication Deficits Associated With Psychiatric Disorders. Biol Psychiatry. 2016; 79:53–61. [PubMed: 26232298]

Lahvis GP, Alleva E, Scattoni ML. Translating mouse vocalizations: prosody and frequency modulation. Genes Brain Behav. 2011; 10:4–16. [PubMed: 20497235]

Lee DD, Seung HS. Algorithms for non-negative matrix factorization. Advances in Neural Information Processing Systems. 2001:556–562.

Liu RC, Miller KD, Merzenich MM, Schreiner CE. Acoustic variability and distinguishability among mouse ultrasound vocalizations. J Acoust Soc Am. 2003; 114:3412–3422. [PubMed: 14714820]

Mahrt EJ, Perkel DJ, Tong L, Rubel EW, Portfors CV. Engineered deafness reveals that mouse courtship vocalizations do not require auditory experience. J Neurosci. 2013; 33:5573–5583. [PubMed: 23536072]

Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing. 2001; 9:504–512.

Narayanan S, Georgiou PG. Behavioral Signal Processing: Deriving Human Behavioral Informatics From Speech and Language: Computational techniques are presented to analyze and model expressed and perceived human behavior-variedly characterized as typical, atypical, distressed, and disordered-from speech and language cues and their applications in health, commerce, education, and beyond. Proc IEEE Inst Electr Electron Eng. 2013; 101:1203–1233. [PubMed: 24039277]

Neilans EG, Holfoth DP, Radziwon KE, Portfors CV, Dent ML. Discrimination of ultrasonic vocalizations by CBA/CaJ mice (Mus musculus) is related to spectrotemporal dissimilarity of vocalizations. PLoS One. 2014; 9:e85405. [PubMed: 24416405]

O'Grady P, Pearlmutter B. Discovering speech phones using convolutive non-negative matrix factorization with a sparseness constraint. Neurocomputing. 2008; 72(1):88–101.

Panksepp JB, Jochman KA, Kim JU, Koy JJ, Wilson ED, Chen Q, Wilson CR, Lahvis GP. Affiliative behavior, ultrasonic communication and social reward are influenced by genetic variation in adolescent mice. PLoS One. 2007; 2:e351. [PubMed: 17406675]

Patterson R, Nimmo-Smith I, Holdsworth J, Rice P. An efficient auditory filterbank based on the gammatone function. IOC Speech Group meeting on Auditory Modeling at RSRE. 1987; 2(7)

Peirce JL, Lu L, Gu J, Silver LM, Williams RW. A new set of BXD recombinant inbred lines from advanced intercross populations in mice. BMC Genet. 2004; 5:7. [PubMed: 15117419]

Pomerantz SM, Nunez AA, Bean NJ. Female behavior is affected by male ultrasonic vocalizations in house mice. Physiol Behav. 1983; 31:91–96. [PubMed: 6685321]

Portfors CV. Types and functions of ultrasonic vocalizations in laboratory rats and mice. J Am Assoc Lab Anim Sci. 2007; 46:28–34. [PubMed: 17203913]

Rabiner, L., Schafer, R. Theory and Applications of Digital Speech Processing. Prentice Hall; 2010.

Ramanarayanan V, Goldstein L, Narayanan SS. Spatio-temporal articulatory movement primitives during speech production: extraction, interpretation, and validation. J Acoust Soc Am. 2013; 134:1378–1394. [PubMed: 23927134]

Ramirez, J., Gorriz, JM., Segura, JC. Voice Activity Detection. In: Grimm, M., Kroschel, K., editors. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding. InTech; 2007.

Sales, G., Pye, D. Ultrasonic Communication by Animals. London/New York: Chapman and Hall, distributed in the U.S. by Halsted Press; 1974.

Scattoni ML, Gandhy SU, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in the BTBR T +tf/J mouse model of autism. PLoS One. 2008; 3:e3067. [PubMed: 18728777]

Scattoni ML, Ricceri L, Crawley JN. Unusual repertoire of vocalizations in adult BTBR T+tf/J mice during three types of social encounters. Genes Brain Behav. 2010; 10:44–56.

Schluter, R., Bezrukov, L., Wagner, H., Ney, H. Gammatone features and feature combination for large vocabulary speech recognition. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on; 2007.

Sewell GD. Ultrasonic communication in rodents. Nature. 1970; 227:410.

Shao, Y., Jin, Z., Wang, DL., Srinivasan, S. An auditory-based feature for robust speech recognition. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on; 2009. p. 4625-4628.

Smaragdis P. Convolutive speech bases adn their application to supervised speech separation. IEEE Transactions on Audio, Speech, and Language Processing. 2007; 15(1):1–12.

Stevens SS, Volkmann J, Newman EB. A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America. 1937; 8:185–190.

Sugimoto H, Okabe S, Kato M, Koshida N, Shiroishi T, Mogi K, Kikusui T, Koide T. A role for strain differences in waveforms of ultrasonic vocalizations during male-female interaction. PLoS One. 2011; 6:e22093. [PubMed: 21818297]

Taylor BA, Bedigian HG, Meier H. Genetic studies of the Fv-1 locus of mice: linkage with Gpd-1 in recombinant inbred lines. J Virol. 1977; 23:106–109. [PubMed: 196096]

Taylor BA, Wnek C, Kotlus BS, Roemer N, MacTaggart T, Phillips SJ. Genotyping new BXD recombinant inbred mouse strains and comparison of BXD and consensus maps. Mamm Genome. 1999; 10:335–348. [PubMed: 10087289]

Thornton LM, Hahn ME, Schanz N. Genetic and developmental influences on infant mouse ultrasonic calling. III. Patterns of inheritance in the calls of mice 3–9 days of age. Behav Genet. 2005; 35:73–83. [PubMed: 15674534]

Torquet N, de Chaumont F, Faure P, Bourgeron T, Ey E. mouseTube - a database to collaboratively unravel mouse ultrasonic communication. F1000Res. 2016; 5:2332. [PubMed: 27830061]

Valero X, Alias F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. IEEE Transactions on Multimedia. 2012; 14(6)

Van Segbroeck M, Tsiartas A, Narayanan SS. A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. INTERSPEECH. 2013:704–708.

Van Segbroeck M, Van hamme H. Unsupervised learning of time–frequency patches as a noise-robust representation of speech. Speech Communication. 2009; 51:1124–1138.

von Merten S, Hoier S, Pfeifle C, Tautz D. A role for ultrasonic vocalisation in social communication and divergence of natural populations of the house mouse (Mus musculus domesticus). PLoS One. 2014; 9:e97244. [PubMed: 24816836]

Wohr M, Dahlhoff M, Wolf E, Holsboer F, Schwarting RK, Wotjak CT. Effects of genetic background, gender, and early environmental factors on isolation-induced ultrasonic calling in mouse pups: an embryo-transfer study. Behav Genet. 2008; 38:579–595. [PubMed: 18712592]

Woolley SM, Portfors CV. Conserved mechanisms of vocalization coding in mammalian and songbird auditory midbrain. Hear Res. 2013; 305:45–56. [PubMed: 23726970]

Yang M, Loureiro D, Kalikhman D, Crawley JN. Male mice emit distinct ultrasonic vocalizations when the female leaves the social interaction arena. Front Behav Neurosci. 2013; 7:159. [PubMed: 24312027]

Yu Song N, Nicod J, Min B, Cheung RCC, Amin MA, Yan H. Noise filtering and occurrence identification of mouse ultrasonic vocalization call. International Conference on Machine Learning and Cybernetics. 2013:1218–1223.

Zwicker E. Subdivision of the audible frequency range into critical bands (Frequenzgruppen). Journal of the Acoustical Society of America. 1961; 33:248.

HIGHLIGHTS

- Open-access software automatically generates mouse vocalization repertoires.
- New similarity metrics enable comparisons of syllable production and use.
- MUPET compares syllable repertoires across mouse strains and social conditions.

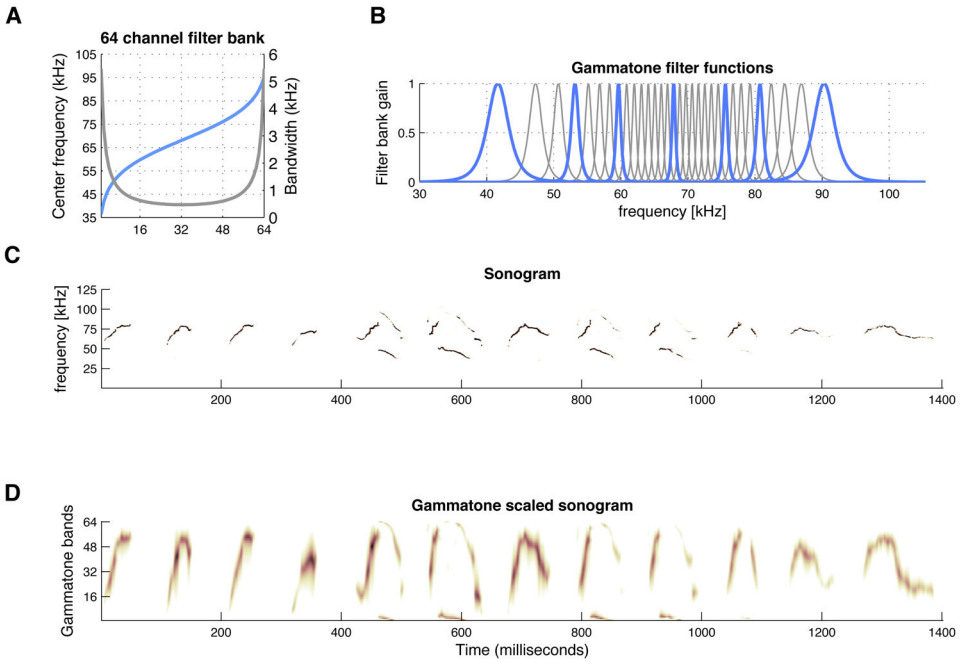


Figure 1. Generating perceptually filtered representations of mouse USVs

(A) Frequency warping curve with a logistic shape, illustrating the 64 gammatone filters (X-axis), corresponding mouse ultrasonic vocalization frequencies (blue line; left Y-axis) and gammatone filter bandwidths (gray line; right Y-axis).

(B) Gammatone filterbank composed of 64 band-pass filters. Each filter is modeled by a gamma distribution function, where the center frequency and bandwidth of the gammatone impulse responses are derived from the frequency warping curve. The band-pass filters are symmetrically distributed, with the frequency region containing the highest number of auditory events modeled by narrow filters and the upper and lower bounds of the mouse USV frequency range modeled by a smaller number of wider filters. For clarity, the image shows 32 filters, with example band-pass filters highlighted in blue.

(C) Sonogram showing frequency versus time for a 1.4 sec excerpt of USVs from a DBA/2 mouse recording.

(D) Gammatone Filterbank USV feature (GF-USV) representations of the sonogram in C, obtained from the 64-channel Gammatone filterbank. The plot illustrates how the filterbank captures the salient spectral features of the USVs. The reduced dimensionality of the GF-USV representation facilitates subsequent signal processing by lowering the computational requirements.

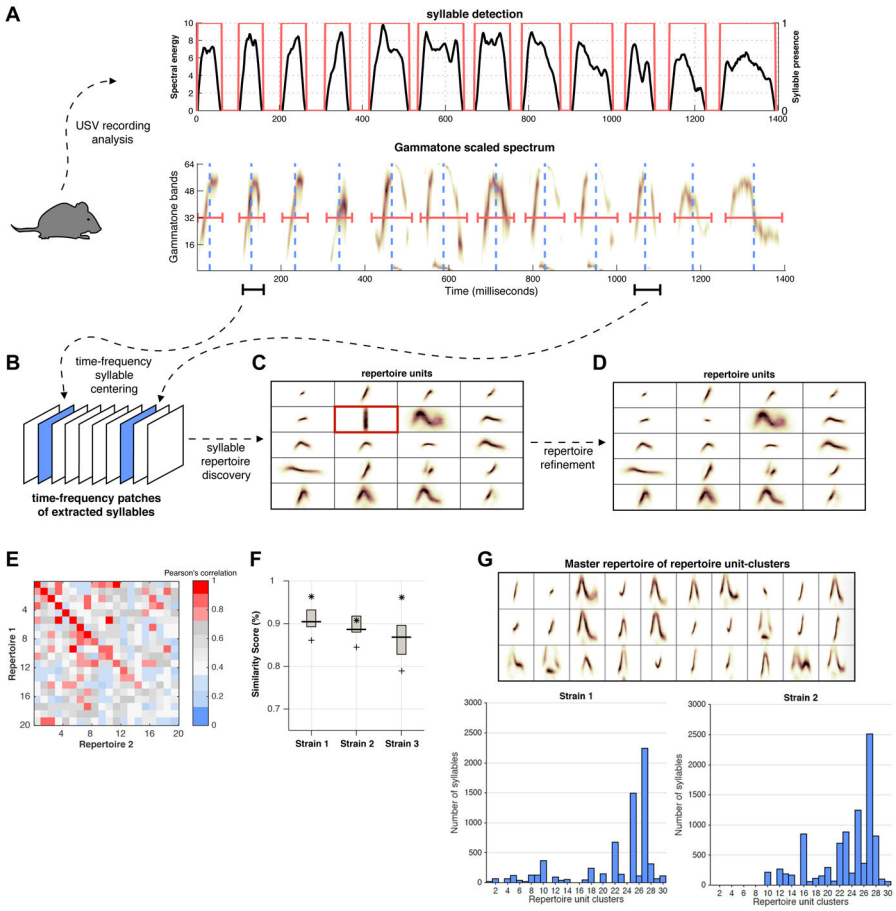


Figure 2. Computational framework for syllable repertoire learning and repertoire analysis functions

(A) The mouse USV recordings are loaded into MUPET and the syllable detector segments individual syllables by measuring the power spectrum (black lines) in the ultrasonic range and comparing it with a noise threshold. The regions of vocalized activity/non-activity (red boxes; top panel) are used to extract the syllable types from the GF-USV spectral representation (bottom panel). The center (dashed blue line) and duration (red horizontal line) of the GF-USV, and key spectro-temporal features, are automatically measured.

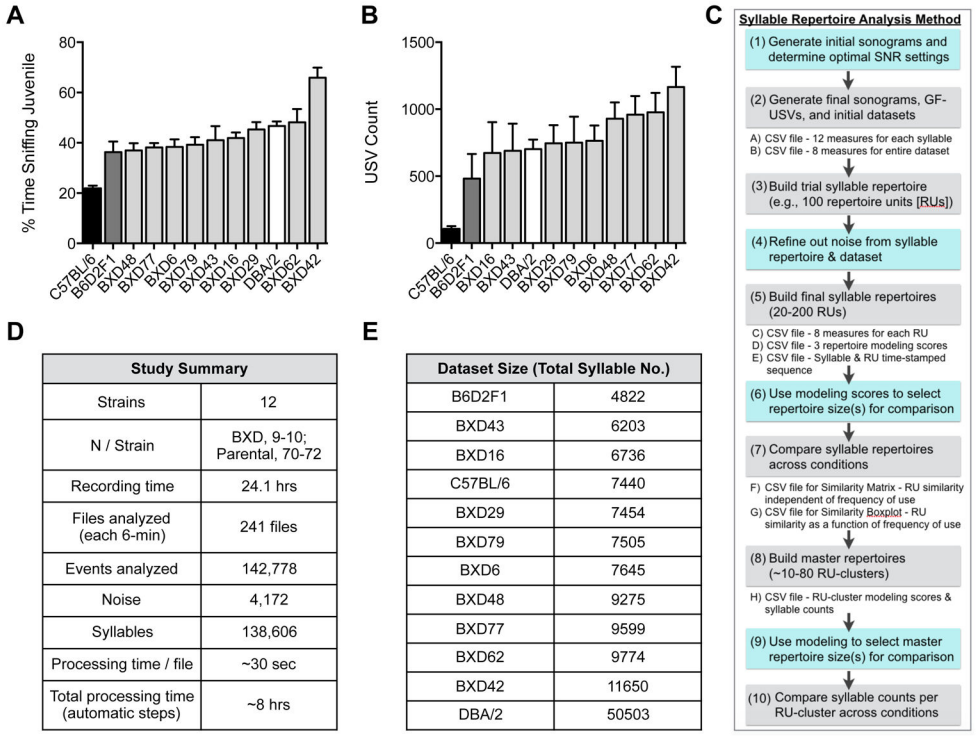
(B) During processing, the extracted syllable shapes are centered along the time and frequency axes and subsequently vectorized before stacking them into a data matrix. Iterative clustering is then performed with a k-means algorithm using the cosine distance measure, which enables the algorithm to learn the most repeated spectral shapes in the dataset.

(C–D) The algorithmic output (C) is a collection of exemplar ‘repertoire units’ (RUs; cluster centroids), which show the average shape of the different syllables that recur in the dataset. RUs learned from noise (red box in C) are removed during syllable repertoire refinement (D).

(E–F) MUPET compares the shapes of RUs from different repertoires using two similarity metrics (E) The Cross Repertoire Similarity Matrix gives the Pearson correlations between RU-pairs from two different repertoires, which are sorted from highest to lowest shape

similarity (see diagonal), irrespective of frequency of RU use in each repertoire. (F) The Cross Repertoire Similarity Boxplot gives the Pearson correlations between collections of RUs, which represent the top 5, 25, 50, 75, and 95% of most frequently used RUs in each repertoire.

(G) To compare the frequency of use of similar and unique RU types across different datasets, MUPET performs a cluster analysis of RU types in order to generate a ‘master repertoire’ of RU-clusters (top panel). MUPET provides information on the frequency of use of each RU-cluster, enabling the user to identify shared and unique RU types and usage across strains or conditions (bottom panels).



D

Study Summary	
Strains	12
N / Strain	BXD, 9-10; Parental, 70-72
Recording time	24.1 hrs
Files analyzed (each 6-min)	241 files
Events analyzed	142,778
Noise	4,172
Syllables	138,606
Processing time / file	~30 sec
Total processing time (automatic steps)	~8 hrs

E

Dataset Size (Total Syllable No.)	
B6D2F1	4822
BXD43	6203
BXD16	6736
C57BL/6	7440
BXD29	7454
BXD79	7505
BXD6	7645
BXD48	9275
BXD77	9599
BXD62	9774
BXD42	11650
DBA/2	50503

Figure 3. MUPET proof-of-principle and a workflow for the syllable repertoire analysis method

(A–B) Heterogeneity in affiliative social interaction (A) and USV count (B) in the C57BL/6 (black bar) and DBA/2 (white bar) parental strains, F1 cross (B6D2F1, dark gray bar), and 9 recombinant inbred BXD strains (mean ± SEM; N=9–10/BXD and B6D2F1 strains; N=70–72/parental strains).

(C) Syllable repertoire analysis method. MUPET learns and compares syllable repertoires in 10 steps. Fully automated steps are show in gray and steps requiring user input based on analysis results are shown in blue. MUPET generates eight exportable CSV files containing spectro-temporal and sequence measures for the syllables and repertoire units, modeling measures for the repertoire and master repertoire builds, and similarity measures for the repertoires.

(D) Summary of key study parameters and analysis time requirements in MUPET.

(F) Summary of total syllable number (excluding noise) for each of the 12 strains.

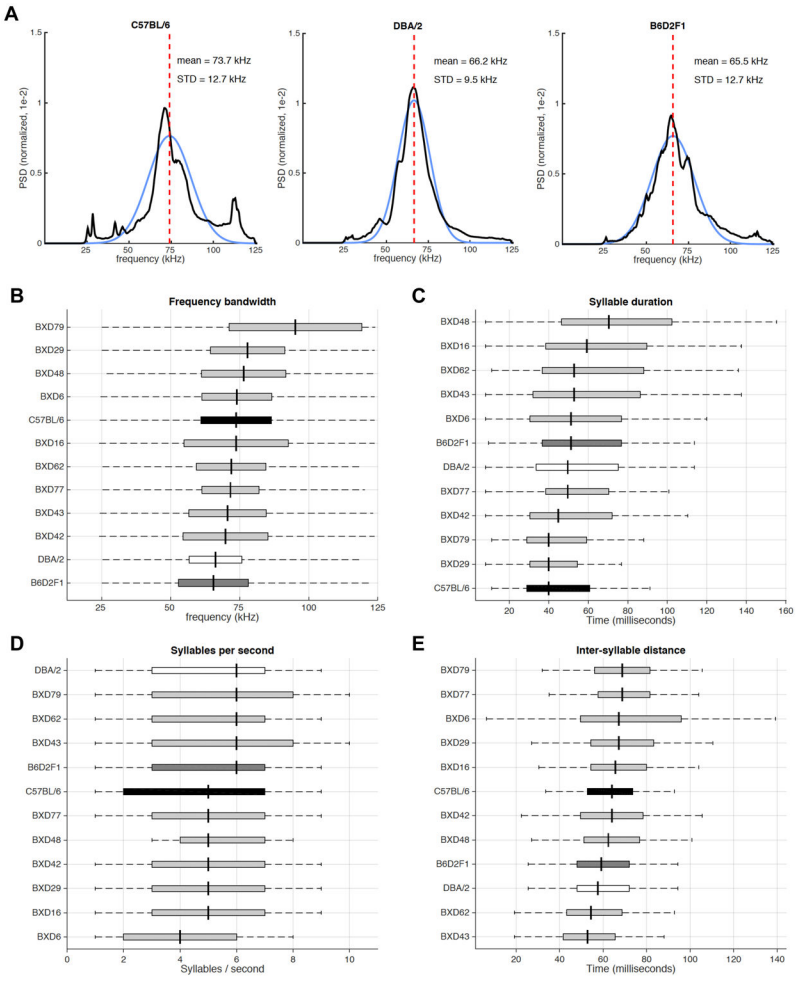


Figure 4. Spectro-temporal measurements of USV datasets

(A) Example power spectral density (PSD) functions for all USVs emitted by the C57BL/6, DBA/2, and B6D2F1 strains. PSD curves (black trace) can be fit by Gaussian functions (blue curve) with the mean frequency indicated by the dashed red line. STD, standard deviation.

(B–E) Boxplots representing (B) frequency bandwidth, (C) syllable duration, (D) syllable rate, and (E) inter-syllable interval across strains. Boxplots display the interquartile range of the Gaussian functions that were fit to each PSD, centered around the mean (black vertical line), with the entire frequency range depicted by dashed horizontal black line.

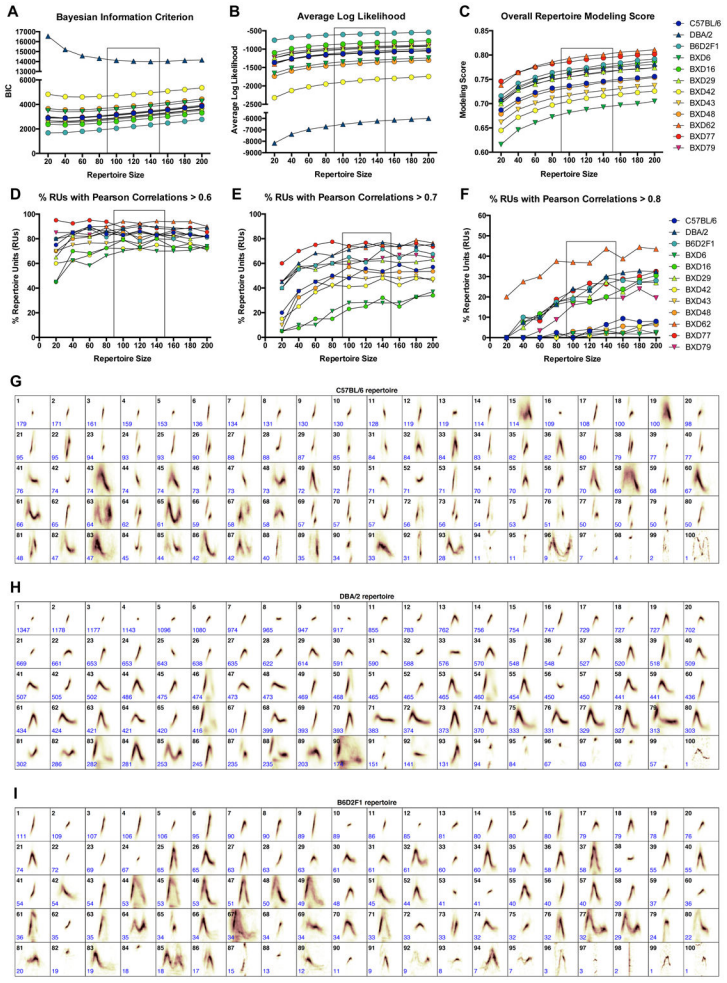


Figure 5. Analyses of syllable repertoires

(A–F) Repertoire modeling measures provided by MUPET and shown for each of the 12 strain datasets. Optimal repertoire build sizes seek to minimize the (A) Bayesian information criterion and maximize the (B) average log likelihood, (C) overall repertoire modeling score, and (D–F) the proportion of repertoire units (RUs) that have average Pearson correlations greater than (D) 0.6, (E) 0.7, and (F) 0.8. Boxes highlight a range of repertoire sizes (100–140) that optimize the repertoire modeling measures.

(G–I) Repertoires showing the top 100 syllable types (repertoire units, RUs; black numbers) learned from processing recordings from the C57BL/6, DBA/2, and B6D2F1 strains. RUs are listed in order of frequency of use from left to right (1–100), with the total number of syllables that are present in each RU given in blue.

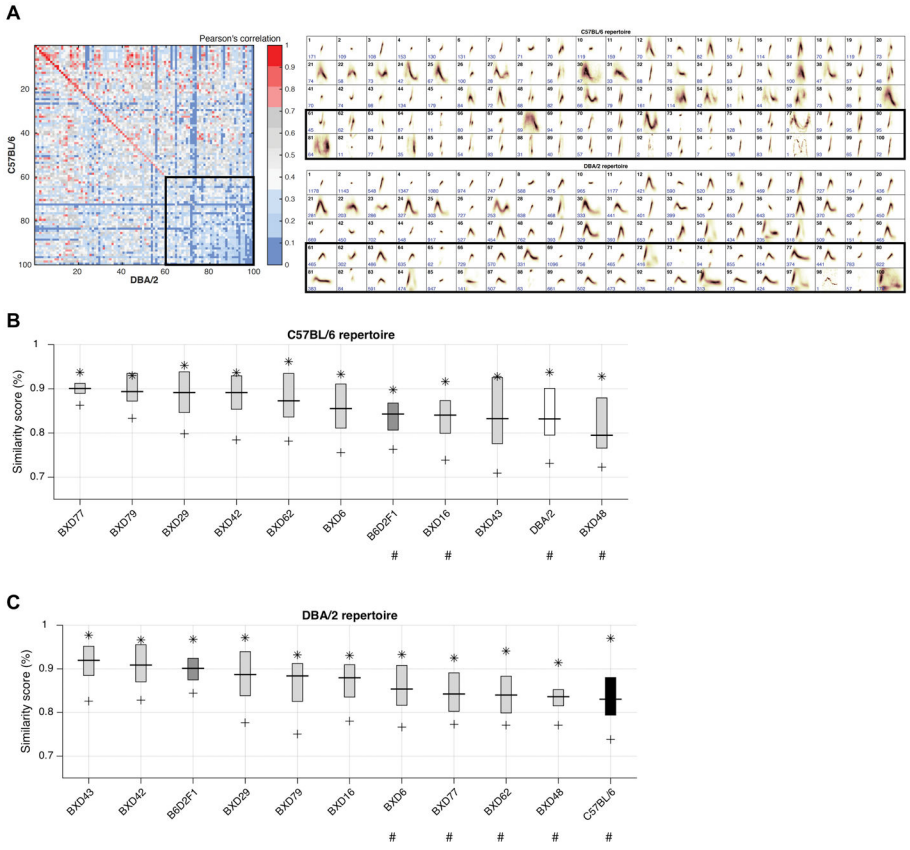


Figure 6. Similarity metrics used to compare repertoire unit type across strains

(A) Similarity Matrix (left panel) of the spectral types of pairs of repertoire units (RUs) learned from the C57BL/6 and DBA/2 datasets (right panels). The matrix diagonal gives the Pearson correlation for sequential pairs of C57BL/6 and DBA/2 RUs ranked from most to least similar (e.g., Unit 1 in both repertories are highly similar). Comparison of RU types is performed by centering RUs along the time and frequency axes and then by sequentially pairing units of greatest to least shape similarity, independent of frequency of use. The parental strains produce both highly similar (e.g., unit 1–40) and distinct (e.g., units 60–100) repertoire units. Black boxes highlight RUs that show low similarity across the parental strains.

(B–C) The Similarity Boxplots determined by comparing the similarity of RU types as a function of how frequently the RU is used by each of the 'comparison' strains (X-axis) in comparison to the C57BL/6 (B) and DBA/2 (C) 'reference' strains. The Y-axis is the % Similarity Score (average Pearson correlations) between collections of RUs compared between the reference and comparison repertoires. The star (*) denotes the Pearson correlations for the top 5% of the most frequently used RUs, where the boxplot shows the mean and interquartile range of these correlations, and the plus sign (+) shows the correlation of the top 95% of the most frequency used RUs. # Indicates strains with statistically significant ($P < 0.05$) differences in mean repertoire similarity compared to the reference repertoire.

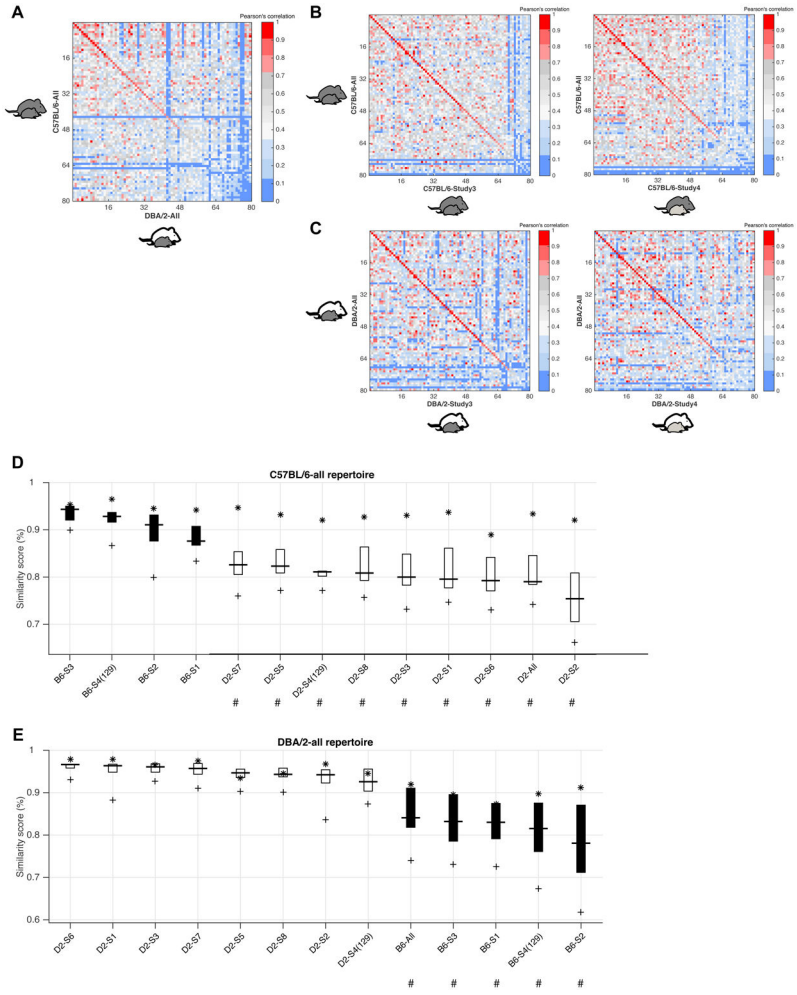


Figure 7. Syllable repertoire stability in the parental strains across replicate studies

(A) Similarity Matrix used to assess the similarity of the spectral shape of pairs of repertoire units (RUs) learned from the full C57BL/6 and DBA/2 datasets (“C57BL/6-All”, “DBA/2-All”) when these strains were paired with a C57BL/6 juvenile partner (X-axis, DBA/2 (white) with juvenile C57BL/6 (black) mouse; Y-axis, C57BL/6 (black) with juvenile (black) C57BL/6 mouse.

(B) Similarity Matrices computed between the full C57BL/6 repertoire (Y-axis) and individual studies in which the C57BL/6 strain was tested with a C57BL/6 (left panel) or 129S1 (right panel; light gray) juvenile, showing high repertoire similarity regardless of partner strain.

(C) Similarity Matrices computed between the full DBA/2 repertoire (Y-axis) and individual studies in which the DBA/2 strain was tested with a C57BL/6 (left panel) or 129S1 (right panel) juvenile, showing high repertoire similarity regardless of partner strain.

(D–E) Similarity Boxplot representation of RU similarity. Repertoires built from 4 distinct studies with C57BL/6 (B6) mice and 8 distinct studies with DBA/2 (D2) mice (individual studies are listed on the X-axis) in comparison with full repertoires built from all C57BL/6 and DBA/2 studies conducted with C57BL/6 juveniles. The B6 and D2 studies conducted with 129S1 juvenile partners (Study 4, S4) are also shown. # Indicates ‘comparison’ strains

(X-axis) with statistically significant differences in average repertoire similarity compared to the ‘reference’ repertoire (title). $P < 0.05$.

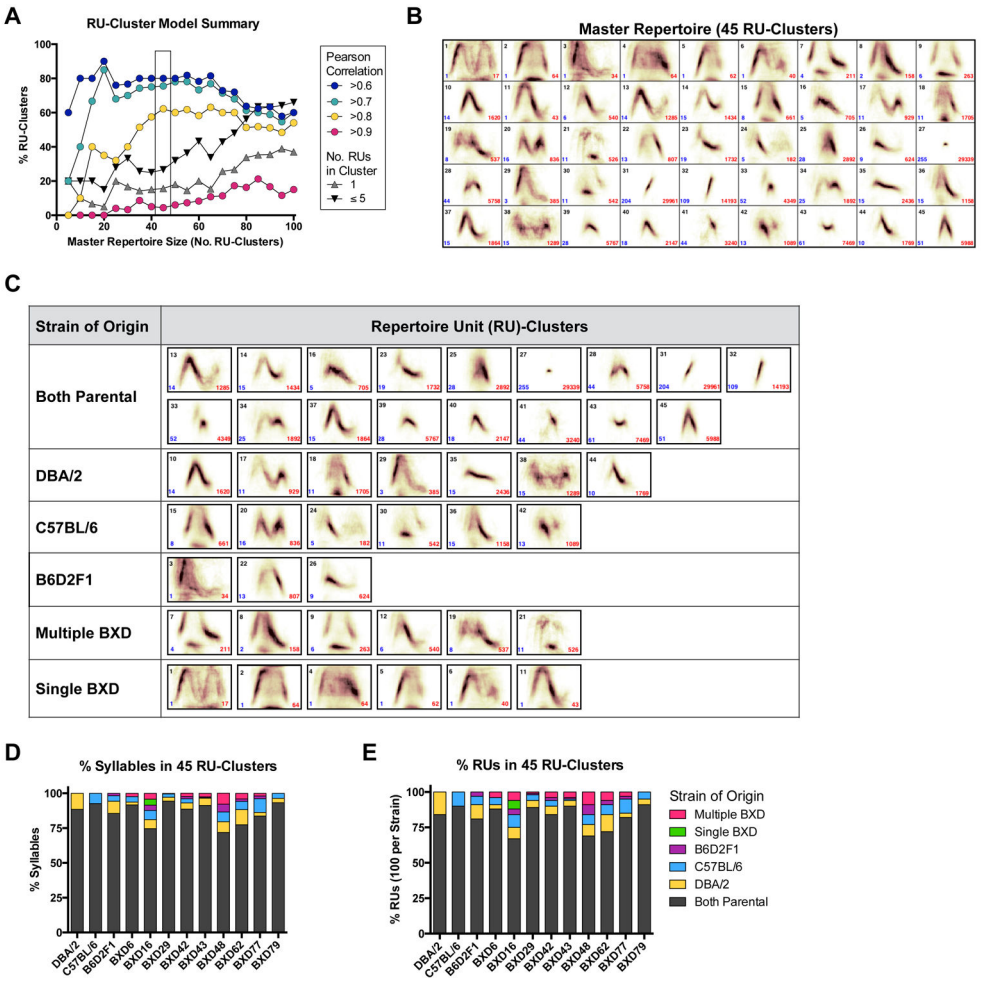


Figure 8. Master repertoire generation and comparison of repertoire unit (RU)-cluster usage and 'strain of origin'

(A) Pearson correlations for different master repertoire sizes (i.e., number of RU-clusters used to model the 12 strain repertoires). Pearson correlations are shown as a percentage of the total number of RU-clusters meeting different threshold correlation values. The box highlights a master repertoire size (45), which maximizes Pearson correlations while minimizing the proportion of RU-clusters that contain a relatively small number of RUs, which is a measure of model complexity.

(B) Master repertoire of 45 units generated using k-medoid clustering applied to the 12 individual strain repertoires (each 100 RUs; 1200 RUs total). MUPET assigns RUs learned from each strain to one of the 45 RU-clusters (black numbers), enabling determination of shared and unique RU types across strains ('strain of origin'). The total number of RUs and syllables in each cluster are shown in blue and red, respectively.

(C) Each of the RU-clusters is assigned to a strain of origin based on whether the RUs it contains are 1) observed in both parental strains, with or without presence in the offspring strains, 2) unique to a parental strain—observed in only one parental strain (DBA/2 or C57BL/6), with or without presence in the offspring strains, 3) unique to the F1 cross (B6D2F1)—observed in neither parental strain, but present in B6D2F1 and offspring strains,

4) multiple BXD strains—present in neither of the parental nor B6D2F1 strains, but present in multiple BXD strains, or 5) unique to a single BXD strain (in this analysis only BXD16 generated unique RU-clusters).

(D–E) The percentage of syllables (D) and RUs (E) present within RU-clusters that are generated by each strain of origin.

Table 1

Software	Mupet 2.0	Mouse Song Analyzer v1.3	VoICE
Primary References	Van Segbroeck et al.	Holy and Guo, 2005; Arriaga et al., 2012; Chabout et al., 2015	Burkett et al., 2015
No. of Syllables Analyzed	~200K	~60K (Chabout et al., 2015)	~8K
Analysis Platform	Matlab	Matlab	Matlab and R
Input	Original (unmanipulated) wav files	Original (unmanipulated) wav files	Independent method is needed to detect and 'clip' each syllable into a separate .wav file.
Signal Detection	Modifiable parameters to optimize syllable detection.	Modifiable parameters to optimize syllable detection.	Not supported* (see Input)
Noise Detection	Built-in noise detection and removal	Noise filtered out as unclassified syllables	Not supported (see Input)
Syllable Classification Approach	Automated, unbiased discovery of recurring syllable shapes using machine learning.	Automated, rule-based categorization of syllable shapes based on Holy and Guo, 2005 and Scattoni et al., 2010.	Automated, hierarchical clustering of similar shapes (e.g., ~70 clusters in Burkett et al., 2015) which facilitates manual, rule-based categorization of a smaller number of cluster eigencalls (and individual syllables) as described in Scattoni et al., 2010.
Syllable Categorization Dimensions	• Entire frequency contour, including duration, slope and curvature of each note. • Does not include mean frequency or amplitude.	• Multi-note syllables are classified based on the number and direction of frequency jumps, but not based on the duration, slope or curvature of each note. • Option to classify single-note syllables and harmonics as described in Scattoni et al., 2010. • Does not include mean frequency or amplitude.	• Hierarchical clustering based on mean frequency, and the slope, duration, and curvature of each note. • Final syllable categories are assigned as described in Scattoni et al., 2010.
Syllable Categories	~60–200 unnamed categories (e.g., RUs 1–200), with option to cluster RUs across datasets to identify shared and unique shapes.	A limited number (~4–15) of named categories (e.g., simple, up-jump, down-up-down jump).	A limited number (~9–12) of named categories (e.g., flat, chevron, frequency step) as described in Scattoni et al., 2010.
Output - Syllables and Repertoires	**1** **Syllable information**: Includes syllable sequence with RU designation and 10 spectro-temporal and amplitude measures. **2** **Syllable time-stamp** (start and end time)	**1** **Syllable information**: Includes syllable sequence with syllable category designation and 11	**1** **Dataset information**: Pie graphs showing the percentage of syllables assigned to each category.

Software	Mupet 2.0	Mouse Song Analyzer v1.3	VoICE
	3 **Dataset information** (averaged across all syllables): Includes PSD, syllable rate, ISI and duration. **4** **Syllable repertoire**: Visual dictionary of syllable shapes (RUs) identified in each dataset.	spectro-temporal and amplitude measures. **Notes**: • Syllable time-stamp is not provided, but an approximate time-stamp can be computed manually from the start-time of the first syllable (which can be determined from the sonogram) and from the syllable duration and ISI measurements generated by MSA. • Dataset information can be calculated manually from the syllable information.	**2** **wav files are sorted** into folders for each cluster and syllable assignment. **Notes**: • VoICE does not generate syllable time-stamp, spectro-temporal or amplitude measures. • VoICE does not generate the information required to calculate dataset information.
Output - Comparisons of Repertoires Across Datasets	**1** Compare 2 or more syllable repertoires: • **Cross Repertoire Similarity Matrix**: Compares RU shapes between 2 dataset repertoires irrespective of frequency of use. • **Cross Repertoire Similarity Boxplot**: Compares RU shapes across all dataset repertoires as a function of frequency of use. **2** **Master repertoire**: Clusters RU shapes from all datasets to generate a “master repertoire” of RU-clusters. Information	• MSA does not provide automated syllable repertoire comparisons.	• VoICE does not provide automated syllable repertoire comparisons.

Software	Mupet 2.0	Mouse Song Analyzer v1.3	VoICE
	is provided on the number of RUs and syllables from each dataset that are present in each cluster, enabling shared or unique syllable shapes to be identified across datasets.		
Repertoire Modeling Scores	1 **Repertoire modeling scores**: 3 scores summarizing model complexity and accuracy for syllable repertoires of different sizes. 2 **Goodness-of-fit measures**: Average Pearson correlations for the syllables within each RU and the RUs within each RU-Cluster, informing the selection of syllable and master repertoire sizes.	• MSA categorizes syllables based on limited specto-temporal measures rather than based on the entire frequency contour (see Holy and Guo, 2005 and Arriaga et al., 2012). There is no straightforward methodology to determine the similarity of shapes within the same category (e.g., up-jump) and categories frequently contain syllables with a diversity of note durations, slopes and contours.	• VoICE categorizes syllables based on limited specto-temporal measures rather than based on the entire frequency contour (see Scattoni et al., 2010). There is no straightforward methodology to determine the similarity of shapes within the same category (e.g., upward) and categories frequently contain syllables with a diversity of note durations, slopes and contours.
Syntax Analysis	• Syllable sequence and ISI available for syntax analysis outside of MUPET. • Note that high number of syllable categories (e.g., 100 RUs) will require higher order syntax analysis methods.	Syllable sequence and ISI available for syntax analysis outside of MSA.	Syllable category could be combined with information about ISI (generated by an independent method) and used for syntax analysis outside of VoICE.
Progress Toward Behavioral Analysis	Automated syllable time-stamp and unbiased discovery of different syllable shapes advances the field's ability to relate syllable onset and type to time-logged behaviors. Note: behaviors must be scored with separate software.	More challenging in the absence of precise syllable time-stamp information (see Output Notes).	Not supported (see Output Notes).
High-Throughput Analysis of >100K syllables	Yes	Yes	No, manual classification of eigencalls and individual syllables significantly slows processing of large datasets.
Sensitive to Novel Syllable Shapes	Yes. Unbiased discovery of recurring syllable shapes is designed to be sensitive to unique syllable shapes.	Possibly. Complex and unique shapes could be detected as a new pattern of	Miscellaneous category could capture novel shapes.

Software	Mupet 2.0	Mouse Song Analyzer v1.3	VoICE
	However, we note that because MUPET generates an average syllable shape (“RU centroid”), highly complex shapes will only be accurately detected if represented in the dataset a sufficient number of times (e.g., 10–100 syllables). More complex shapes show increased between-call variability and thus, require more example syllables to generate an RU-centroid that well represents the syllable shape.	up- and down-jumps that occurs more frequently. The absence of an ability to easily observe the frequency contours of syllables within each category challenges detection of novel syllable shapes.	
Species	Mouse. Signal processing and machine learning approach can be adapted for other species.	Mouse	Mouse, Bird