Five fundamental constraints on theories of the origins of music

Bjorn Merker, Iain Morley and Willem Zuidema

1. Introduction

Music is a cherished art form and a daily source of inspiration and pleasure, as well as occasional irritation, for billions. It is also an extraordinarily complex phenomenon that appears to be not only uniquely human, but a human universal [1–3]. This uniqueness and universality raises the question of how and why the human ability to appreciate and produce music evolved. However, as is the case for language and other aspects of human cognition, it is not obvious how to properly constrain our theorizing so as to avoid producing no more than ‘just-so stories’. Evolutionary biologist Richard Lewontin [4] has warned against ‘the childish notion that everything that is interesting about nature can be understood. History, and evolution is a form of history, simply does not leave sufficient traces, especially when it is the forces that are at issue. Form and even behaviour may leave fossil remains, but forces like natural selection do not. It might be interesting to know how cognition (whatever that is) arose and spread and changed, but we cannot know. Tough luck.’

Against this blunt pessimism stand those who hold, with Richard Byrne, that ‘comparative analysis of the behaviour of modern primates, in conjunction with an accurate phylogenetic tree of relatedness, has the power to chart the early history of human cognitive evolution’ [5, p. 543]. With regard to human music, we suspect that neither side of this conceptual divide has rendered good advice to those who would explore its evolutionary origins. Perhaps the pessimism of Lewontin might be overcome by casting the comparative and inferential net wide enough. However, to do so we can no longer, as Byrne does, restrict ourselves to the study of primate homologies, but must explore analogies wherever they are found in the animal kingdom. Some traits do after all evolve de novo in a lineage. To understand such novelties, analogous developments in unrelated animals provide invaluable information regarding potential selection pressures and
ecological conditions favouring their evolution. The fruitfulness of such exercises, whether pursuing homologies or analogies, depends on the extent to which they can be constrained by stubborn facts regarding the phenomenon in search of an evolutionary explanation.

Here we focus on a small set of characteristics of human music that should help constrain accounts of its origins. They can be conceived of as basic hurdles that must be cleared along the way to a comprehensive theory of the origins of human music. They were chosen above all for their generality, with the additional desideratum of involving mechanisms that generate consequences for the structural content of music. At present a number of these constraints are difficult to meet, which means that besides their potential bearing on already proposed theories, they pose challenges for and may perhaps even inspire future ones.

2. Constraint no. 1: cultural transmission

Music, like language, is a complex product of cultural history. Its present-day patterns rest on traditions extending back over many thousands of years of inter-generational transmission of learned cultural lore [6,7]. This simple fact, so obvious that it typically is taken for granted in theories of music origins, nevertheless has profound consequences for any attempt to reconstruct the biological background to human music.

If patterns of cultural goods were only matters of human tastes and preferences—a common misconception regarding the nature of culture—the cultural transmission of musical lore would have no systematic or principled bearing on the reconstruction of music origins. However, when sustained over many generations, inter-generational transmission itself exerts profound and predictable effects on the contents of the transmitted lore, even in the complete absence of natural selection or any differential reinforcement of outcomes [8–12]. Thus, to go in search of evolutionary explanations for aspects of music that result from such a cultural process would be a serious mistake. As we shall see, major structural features of music are likely to be shaped by the cultural transmission process itself.

The key insight here is that with each generational transfer, the cultural lore (be it language, music, or any similar system transmitted culturally through learning) has to pass the so-called ‘learner bottleneck’. Any given learner is only exposed to a portion of the cultural lore extant in the population into which they are born and has, moreover, a limited capacity to absorb even the portion to which they are exposed. This means that the many items that make up that lore compete with each other for passage to the next generation. Through this competitive filtering process any and all aspects of the lore that bear on transmittability, including small differences in learnability and ease of processing, come to transform the cultural corpus in predictable ways, amounting to a cumulative process of informational ‘compression’ over many generations. This tends to issue in a tight fit between properties of the cultural lore and properties of the learner, introducing commonalities across the lore of different, separated, populations, all without the agency of biological selection.

In the field of language evolution, this mechanism which we refer to as ‘cultural evolution’ has been extensively studied over the past two decades (e.g. [13–16]). This work has led to a growing consensus (i) that cultural evolution is a powerful mechanism, (ii) that many features of languages are potentially best understood as resulting from cultural adaptation to (pre-existing) hominin cognitive and physiological features, and (iii) that theorizing about the evolution of the biological basis of language can only sensibly proceed if we explicitly take into account the possibility that cultural evolution has shaped the linguistic phenotype. There is no reason to believe that any of this is any less applicable to the cultural transmission of music than it is to that of language.

Cultural evolution is a gradual, unconscious and obligatory process that extends over many generations, and restructures the cultural corpus in ways that increase its salience, expressive economy, communicative generality and grammatical power, all of which turn on enhanced communicability and learnability in various ways [15,16]. This allows for learners to manage ever larger amounts of cultural content without change in the neural resources devoted to it (through data compression) and lets the cultural products ‘exploit’ existing peculiarities of neural organization. For instance, Zuidema [17] discusses the finding of Smith & Lewicki [18] that the neural code in the auditory nerve of cats appears to be optimized for human speech sounds and argues that this finding only makes sense if the direction of causality is inverted: speech sounds have evolved in a process of cultural evolution to exploit features of a pre-existing general mammalian neural code, i.e. to achieve maximum discriminability under noise and time constraints.

Turning then to music, some major structural features of music widely distributed across cultures might likewise be a consequence of cultural evolution. Until recently, the failure of most musical tuning systems to conform to the mathematics of small integer ratios was grounds for rejecting Pythagoras’s proposal that small integer frequency ratios account for the perception of musical consonance and harmonicity [19]. However, recent modelling of the cumulative effects of physiological nonlinearities at each way-station of the ascending auditory pathway has disclosed the presence of ‘resonance neighbourhoods’ at whole integer ratio spacings on the tonotopic maps of the auditory system [20–22]. This finding not only accommodates a wide range of tuning systems and musical scales found worldwide, but appears capable of accounting for human judgements of consonance, dissonance and tonal stability/attractiveness terms of inherent organizational features of our auditory system [20]; see also [23–27], as follows.

The pattern of ‘resonance neighbourhoods’ in auditory system tonotopy is likely to be shared by all mammals, being a product of quite elementary properties of the neural circuitry in question. It did not evolve for purposes of music, in other words, but as an incidental by-product of the interaction of excitation and inhibition in a neural system evolved to process natural sounds efficiently [20]. Not being confined to humans [28,29], it is not likely to represent an adaptation to music. The presence in humans of cultural patterns of musical practice conforming to these subtle resonances in auditory physiology accordingly requires an explanation.

The fact that those musical practices are, indeed, cultural patterns formed by inter-generational transmission may supply the explanation. In principle, the formal powers of cultural evolution should suffice to allow musical practice to eventually find its way to the ubiquitous and inherent resonant biases of the auditory system, given a long running cultural tradition of song [30]. Through the external loop of inter-generational transmission of learned musical lore, the
production of musical patterns would pass through the ‘learner bottleneck’ to be shaped by pre-existing biases on the part of learners, specifically the purely perceptual resonant biases just invoked. This assumes not only a capacity for vocal learning (see ‘Constraint no. 3’), but one emancipated from innate song templates, for which there is precedence in the true mimics among the birds [31–33]. From such a starting point, devoid of scales, tonality and small integer ratio consonances, thousands of generations of cultural transmission could eventually externalize even subtle biases in auditory perception in the musical practices of human cultures (see also [28]).

The principal constraint this process imposes on theories of music origins is that they must provide a non-arbitrary reason for our distant forebears to have engaged in persistent inter-generational transmission of vocal lore lacking the tonal organization of music as we know it for long enough to allow the transmission mechanism to find its way to the resonances already embedded in basic auditory physiology. Since thousands of generations may be required for this to happen [14,37], the constraint is a real one. Perhaps our forebears, like many bird species, maintained cultural traditions of learned song, and became vocal learners as part of the cluster of changes that define the emergence of Homo some 2 Myr [38,39]. An increasingly refined vocal communication system for the accurate communication and extraction of emotional information from vocal prosody is likely to also have contributed to this process ([7,40,41] and [42, note 3]).

More generally, placing the historical nature of the pattern-content of human music at the head of the effort to understand its biology greatly facilitates the reconstruction of its evolutionary background. Because its patterns are learned from a corpus of cultural models subject to the transformative dynamics of the inter-generational ‘learner bottleneck’, there is no need to ask evolutionary selectional mechanisms to equip us with those pattern specifics, even when they happen to be cross-culturally widespread, as in the ‘auditory system resonance’ example.

3. Constraint no. 2: generativity or infinite variety by finite means

Music, like language, is generative, i.e. it produces infinite pattern variety by finite means [1]. The key to that variety in both music and language is of course not recursion [43] but combinatorics [1,44,45]. By combining a finite set of elements—discrete pitches and durations—music creates composite patterns without limit. For this to be possible, the combining elements must be non-blending in the sense of not producing an average when combined [46], i.e. they must retain their individuality on combining (figure 1a,b). When that is the case, each such combination ‘creates something which is not present per se in any of the associated constituents’ [1, p. 67], making infinite pattern variety possible (figure 1c). There is a total of four major such open-ended generative systems in existence, two of which are natural ones (chemistry and genetics), while two are cultural (music and language; table 1).

In the cases of music and language, the combining elements are conventional, the musical ones arising through a radical reduction in the degrees of freedom available to vocal or instrumental sound production [45]. This is accomplished by discretizing two continua, those of pitch and duration, to yield musical notes with determinate pitch and—in all rhythmic music—proportional durations based on discretizing time through an isochronous pulse (see ‘Constraint no. 4’).

In other words, musical notes are not simply pitches. Rather, they are individuated pitch locations within a discretized pitch continuum. They are fixed reference points on that continuum, between which even glissandi must travel with the same necessity as does any ordinary note if they are to be musical. The designation of a specific location on the pitch continuum as a ‘note’ by a culturally determined ‘pitch standard’—applied with a conventionalized margin of tolerance—lifts that location out of its relation of equivalence to its infinitude of pitch neighbours. It breaks its ‘anonymity’, as it were, and turns it into an individuated and specific musical note to which a musical
This discretization of the pitch continuum into determinate ‘pitch sets’ supplies music with combinable pitch elements featured in musical melodies and chords [19,47,48]. Pitch sets thus supply the ‘particulates’, the individuated elements, needed for its combinatorial mill. They are found in all musical traditions cross-culturally. Indeed, one of the distinguishing marks of musicianship anywhere is adherence to the pitch locations designated by a pitch standard during musical performance. Not to do so is to sing or play ‘out of tune’, the quintessential demarcation line between musical and other employments of human capacities.

The constraint imposed on theories of music origins by the generativity of music is that no such theory can account for the genesis of music as we know it without giving a credible account of how we came to conquer for ourselves the discretized (‘particulate’) elements without which there can be no open-ended generativity of music. In light of what has been covered under Constraint no. 1, these elements may of course be prime products of a protracted process of cultural evolution exploiting the tonal scaffolding of auditory system resonances along with factors such as the convenience of dividing the octave into steps that maximize our closest living relatives included. The vocal patterns of song and speech are acquired through motor learning on the basis of heard, culturally transmitted models through a process requiring intact hearing and feedback from one’s own voice [51–56]. The process by which they are acquired is technically known as vocal production learning [57,58], a dedicated and highly specialized capacity that has no other common uses in our lives besides song and speech.

Since there can be no human singing without it, the origin of our capacity for vocal production learning bears directly on scenarios for the origins of music. The issue is an acute one, since the fact that other primates lack this capacity [57,58] means that we became vocal learners at or after our divergence from the common ancestor we share with chimpanzees. One limb of the comparative method—the tracing of continuities (homologies) with our close evolutionary relatives—is therefore unavailable for reconstructing its origin in this particular case, Byrne’s assertion to the contrary quoted in our introduction notwithstanding.

There are a variety of context- or learning-based modifications of vocal output that do not involve the mechanism of vocal production learning in the technical sense applicable to human song and speech. They include contextual modulation of vocal behaviour, socially or environmentally contingent selection among innate calls and their variants, and their learned modification, as detailed by Janik & Slater [58]. There is no dearth of evidence for such vocal phenomena among primates. They are an integral part of the vocal expressiveness primates share with many other mammals, but occur without reliance on the specialized mechanism of vocal learning.

Vocal learning proper, by contrast, is the ability to convert heard sound patterns that are not in the species-specific innate vocal repertoire into vocal output, using feedback from one’s own voice to achieve the match [55]. It has been studied in detail above all in birds [51,59], among whom there is a rich assortment of vocal learning phenomena, by no means all the same. They differ along at least six major dimensions of classification, as reviewed by Beecher & Brenowitz [60]. Human vocal learning occupies the more advanced end of several of these dimensions in that it is open ended, allowing new patterns to be added throughout life (though with diminished accuracy after puberty), as well as being emancipated from dependence on a species-specific vocal template. Given that the human capacity is an advanced one, comprehensive studies of the true mimics among birds (mynahs, many species of parrots, lyrebirds, butcherbirds, mockingbirds, etc. [33]) are needed to supplement the invaluable knowledge about mechanisms of vocal learning supplied by the bird species typically employed in the study of vocal learning in the laboratory.

As far as is currently known, vocal production learning proper is found only in humans, cetaceans, pinnipeds,
elephants, bats, oscine songbirds, parrots and hummingbirds [57,61,62]. Given its absence in non-human primates, the process by which our ancestors were equipped with this capacity is a major evolutionary event intervening between the last common ancestor and the first singing or speaking humans. It supplies a major biological constraint or ‘evolutionary bottleneck’ [63] on the path to human music. Its origin in our lineage could have been driven by either song, speech or other factors. This virtually forces the theorist to come to grips with the order of precedence of song and speech in our ancestry ([64, figure 21.1 and accompanying text]).

There is currently no good account of how humans evolved the capacity for vocal production learning. As Nottebohm noted many years ago: ‘you might find it much harder to explain this first step, vocal learning, than the latter acquisition of language’ [63, p. 645]. Nottebohm’s warning, we submit, applies to music no less than to language. One possibility in this regard is that we acquired vocal learning, like some of the songbirds, as a means to sustain cultural traditions of learned song (see ‘Constraint no. 1’). Another might be that vocal learning built upon comparable abilities for manual imitative learning and variation that were already developed, or developing. However conceived, our possession of vocal production learning is a fact, and one that any theory of the origins of music leaves unaccounted for at its peril.

5. Constraint no. 4: entrainment

The constraints considered so far apply both to human music (song) and to language (speech), and therefore cannot help us home in on evolutionary factors unique to music as such. This is no longer so for the final two constraints we shall consider, beginning with our capacity to entrain our behaviour to one another with perfect synchrony.

The type of temporal coordination of inter-individual behaviour that is most distinctively musical, hardly occurring outside the domains of music, dance and drill in the human case, and certainly not in speech, is the type of entrainment that features what Ermentroun has dubbed ‘perfect synchrony’ [65]. It consists of mutual phase-locking with zero (or even slightly negative) phase lag between the periodic signals of two or more signallers sustained consistently at a given tempo, with a capacity to do so at different tempos.

As members of a species in possession of such an entrainment capacity, we tend to take it for granted. Doing so may obscure from us not only its key characteristics but also the reason for the exceedingly sparse distribution of this capacity among animal species (see below). We will therefore endeavour to make its phenomenology explicit so as to avoid confusing it with unrelated forms of temporal coordination commonly occurring in the animal kingdom.

The human capacity for perfect synchrony has been well established and explored through more than a century of sensorimotor synchronization studies [66,67]. The zero (or even slightly negative) phase lag of such entrainment means that its timing mechanism is predictive. A punctate behaviour that coincides with (or even slightly leads) a given beat in an isochronous sequence cannot be caused by a reaction to that beat because of reaction time limitations. Predictive timing is made possible by the regular periodicity of the entraining signal, its isochrony, also known as tactus or ‘pulse’ in music [68,69]. It makes an upcoming beat in the sequence perfectly predictable, and therefore targetable by the predictive timing mechanism [69,70]. Positive asynchronies large enough to come within reach of auditory reaction time, such as those reported for macaques in a synchronization task [71,72], are thus automatically excluded as evidence for entrainment of a kind relevant to the human capacity.

In inter-individual entrainment, the isochrony needed for synchrony must be motorically produced by the entraining individuals on an endogenous (generative) basis. To sustain phase-locking at an average of zero phase lag between such periodic outputs under conditions of a variety of biologically inevitable local perturbations and drift requires mechanisms of phase correction as well as period adjustment [65]. Both are well documented for human sensorimotor synchrony [67]. A mechanism equipped with these features latches on to the regular beat and stays on it as long as that beat stays reasonably steady and lies within the operational tempo range of the entrainment mechanism. In humans that range centres on 2 Hz, which is also the human locomotor tempo [70,73]. Entrainment precision is dependent on predictability, so variance in period length has to be small, typically exhibiting a standard deviation of a few (2–5) percentage points in human tapping performance [66,67].

Entrainment between two or more motoric time series by such a mechanism establishes an unequivocal, unique and rather precise correspondence between the individual events making up the simultaneously unfolding sequences. That correspondence is either one to one or related by small whole integer ratios for harmonically related tempos [74]. It is from this unique correspondence between the individual events of separate time series that asynchronies and their variability are calculated as a measure of synchronization skill [67]. That is, the achievement of beat matching is presupposed by these measures, which assess only how precisely in time that matching occurs. One cannot therefore—as was done in a study of purported entrainment in macaques [75]—take the time series of the animal with the slower movement pace as a reference, and for each of its events select the closest match in time from the other animal’s record as a basis for calculating asynchrony. There is always such a ‘closest event’ irrespective of entrainment, and when as in this case (see tempo means and variances in fig. 2 of that study) sequences drift in phase relative to one another some of these events will occur before and some after the reference event, and these will average out to small asynchronies, again irrespective of entrainment.

The study of cockatoo dancing to human music by Patel et al. ([76], see also [77]) helps define the entrainment phenomenon further by way of contrast with human performance. The bird’s episodic stretches of on-the-beat synchrony emerged from intervening phase drifts over the full phase range in an erratic pattern, while the musical beat remained steady all along. This is not the behaviour expected of a mechanism dedicated to entrainment through provisions for phase and period correction. As already mentioned, such a mechanism locks on to a steady entraining signal and stays locked with only minor asynchronies as long as that signal remains reasonably steady. We do not at this point know why the bird does not do so, but one possibility is that producing dance-like movements mimicking those of its human keepers takes precedence over sustained and precise behavioural matching of the musical beat in the bird’s performance. There are no indications that these birds engage in pulse-based synchrony in their natural
habitats, but many monogamous parrots do engage in joint cooperative pair displays, some of which are learned by imitation [78]. Perhaps, then, the imitative capacities of these highly intelligent birds may help account for their imperfectly entrained dancing to rhythmic music in a human environment.

There are, however, animals who, like us, produce steady isochronous signal sequences on an individual basis and mutually entrain to such signals on a group basis with sustained and consistent on-the-beat matching across individuals as part of their natural behaviour in the wild. These species of ‘natural synchronizers’ are few in number and far removed from us in evolutionary terms, being found among species of fireflies, crickets, cicadas and fiddler crabs [79,80]. To these may be added some marine bioluminescent crustaceans [81], and the rattan ant [82], the latter virtually unstudied (see below).

The champions among these non-human synchronizers are three species of synchronously flashing fireflies and possibly a few species of synchronously chorusing crickets [65]. They entrain their behaviour to one another with a sustained pulse-based rhythmic precision featuring both phase and period correction that equals or exceeds that of human mutual synchrony in music, dance and drill [65,83,84].

That is not to say that the mechanism by which these insects achieve their impressive synchrony is the same as the human mechanism. They only share with us those features of it needed to achieve behavioural entrainment with perfect synchrony. We can easily entrain to a synchronous cicada chorus, but cicadas are unlikely to entrain to our favourite dance music, given the limited auditory scene analysis performed by their fraction-of-a-milligram brains. It is to say, however, that in these animals we have the only documented instances besides our own of genuine beat-based group synchrony that plays a role in the natural behaviour of the species in question. These species therefore may provide us with invaluable hints regarding our own path to this rare behaviour by comparative scrutiny of its functions and evolution in these insects.

The reason for the sparse distribution of the capacity for mutual beat-based entrainment in nature is not far to seek. It resides in the apparent lack of its general biological utility, being virtually useless, with a very few narrowly constrained exceptions. As far as is presently known, the functions it serves where it features in the behavioural repertoire of non-human species in the wild are confined to special cases of mate attraction and predation defence (reviewed in [79,80]). Among the former the so-called ‘beacon effect’ [85] takes pride of place, featuring thousands of synchronously signalling male fireflies whose entrainment precision falls at the more skilled end of the human range, with a standard deviation in period length of less than 3% [83]. Through their entrained luminescent signalling, single trees of permanent male congregations are converted to flashing ‘beacons’ visible from all directions despite the foliage that obstructs single lines of sight in the tropical rainforest. Only male fireflies synchronize their flashing, and females are attracted to these displays. The synchronizing males, of course, are in competition for the females who arrive, and at close quarters females prefer more luminous males, who are also bigger in size.

Defensive uses of synchrony are of two principal kinds: evasive and deterrent. Synchronous calling among neighbouring callers may confuse a predator’s auditory ability to localize any given caller in the chorus, as appears to be the case for a species of treefrog preyed upon by bats [86]. These amphibians do not, however, call rhythmically, but achieve collective superposition of calls by calling at very short latency following the first individual to call spontaneously. This renders their behaviour a special case of reaction-time-limited calling and is therefore irrelevant to our topic.

A deterrent use of synchrony is that of the rattan ant, which lives in symbiosis with the rattan vine [82]. When a vine is disturbed by a sudden external impact, it emits an unexpected and potentially unsettling audible rattle at sound levels far beyond what any single ant can produce. The rattle is a result of entrainment of the alarm behaviour of the ants, which consists of rhythmic beating of their gasters against the vine surface. Local clusters of ants do so in synchrony, with lack of entrainment to more distant ants. Many such locally synchronized clusters produce the unexpected rattle. Amplitude summation (auditory ‘beacon effect’) is the key to this defensive use of inter-individual synchrony.

The narrow compass of behavioural and ecological conditions under which the otherwise useless capacity for entrainment with perfect synchrony has evolved among animals imposes an exceedingly stringent constraint on theorizing about its origin in our case. This is all the more noteworthy in that in the human case the constraint pertains specifically to music and little else in our behavioural repertoire except the music-related disciplines of dance and drill. As such it provides an invaluable asset in evolutionary scenario building. This fortunate circumstance has been exploited in an account of the origin of the human entrainment capacity proposed by Merker [42,69,87], briefly summarized in the next section.

6. Constraint no. 5: motivational basis

Wherever humans live, and however they have organized their societies, they exhibit a behavioural peculiarity of gathering from time to time to sing and dance together in a group [1–3]. By featuring both human song (Constraint no. 3) and entrainment (in the dancing movements and perhaps clapping performed in synchrony with the singing/music, Constraint no. 4), such behaviour qualifies as human music. Indeed, the fact that it occurs in every human culture, and indeed subculture, without exception, unless deliberately suppressed by severe sanctions against it, marks this phenomenon as the most universal human behaviour of a musical kind on record.

In its ubiquity, this human propensity for occasional group singing and dancing would seem to constitute a prototypical musical behaviour, all the more so as it can be staged entirely without musical instruments (as in the traditional trance dance of the hunter–gatherers of the Kalahari Desert [88], see also [89]). It may in fact represent the motivational core of the human capacity for music from which its many other manifestations may have developed by differentiation, elaboration and specialization. One is assisted in becoming aware of the peculiarity and specificity of this behavioural propensity by imagining that in exactly those circumstances in which we typically gather to sing and dance together in a group, another human culture would gather in groups to draw pictures together instead.

The ubiquity and specificity of this putatively prototypical musical behaviour would seem to require an explanation. In searching for one we enter for the first time onto the grounds of a possible homology, because certain social displays of our closest living primate relatives may provide a biological background to the human tendency.
As pointed out by Geissmann [90] there is an association between ‘loud calls’ (‘distance calls’) and physical display among our closest living primate relatives, the apes. The loud calls used by apes in distance signalling (‘long call’, ‘pant–hoot’, ribbon pair duet, etc.) tend to be accompanied by vigorous physical displays such as locomotor excitement, branch shaking, chest beating and other forms of noise-making called ‘drumming’ by Fitch [91], although lacking the pulse-based rhythmicity of drumming in the musical sense. These displays do not feature any metrical structure resembling isochrony, nor any pulse-based rhythmic entrainment between individuals, but they do provide a precedent for the linkage between vocalization and bodily movement that occurs in human group singing and dancing.

Chimpanzees exhibit a social elaboration of this coupling between voice and physical display into an occasional group frenzy called the ‘carnival display’. On irregular occasions, typically when a foraging subgroup discovers a ripe fruit tree or when two subgroups of the same territory meet after a period of separation, the animals launch an excited bout of loud calling, stomping, bursts of running, slapping of tree buttresses and other means of chaotic noise-making. There are no indications that any kind of inter-individual coordination, let alone rhythmic synchrony, forms part of these chimpanzee group displays. They may last for hours, even a whole night, and induce distant subgroups and individuals on the territory, both male and female, to approach and join the fray [92–97].

Our social–emotional propensity to occasionally gather for excited group displays appears to be shared, in other words, with our closest living relative among the apes, the chimpanzee. We are not alone in sensing a possible connection in this regard. BaYaka pygmy hunter–gatherers inhabit the Congo-Brazzaville rainforest, which they share with chimpanzees. Mokondi Massana ‘spirit plays’ featuring ritual singing and dancing are a significant aspect of BaYaka culture.

When BaYaka... hear a chimpanzee ‘carnival display’ from their camp it provokes great hilarity among camp members as one or two of them begin imitating the frenetic actions of the chimpanzees as they pound buttress roots or shriek at the canopy. The camp is launched into laughter as they explicitly ridicule the chimpanzees as they pound buttress roots or shriek at the canopy. The camp is launched into laughter as they explicitly ridicule the chimpanzees in the course of detailing the foregoing constraints, we have noted that some distinguishing aspects of music (e.g. scale systems) require no Darwinian explanation for their widespread yet unique occurrence in humans; equifinality can occur as a consequence of characteristics of learning mechanisms and existing constraints providing the necessary frameworks for such development. Other underlying abilities required for musical behaviours, however (e.g. vocal learning, entrainment), are likely to be the product of forms of Darwinian selection. In these cases, the question then arises regarding what modes of selection might be operative, and on what, specifically, they might be operating.

There are some general lessons from evolutionary theory that are relevant, but often ignored, in constructing evolutionary scenarios for music (and language).

The first point is that biological evolution is always about genetic change. Even though very few genes involved in music have been identified [101], it is important to recognize that evolutionary scenarios, implicitly or explicitly, assume a sequence of changes in gene frequencies in a population, including the appearance of new genetic variants. Making this assumption explicit helps in avoiding the common fallacies of assuming (implicitly) unrealistic amounts of genetic changes (although that is difficult to quantify), assuming instantaneous adoption of new variants, or ignoring the fact that new variants, arising from mutation, are initially always rare.
A second point is that although evolution involves non-adaptive mechanisms such as random mutation and drift, a series of non-adaptive genetic changes leading to a complex new phenotype is exceedingly improbable. To establish the plausibility of a scenario for a trait shared by all humans, we thus have to show that each new variant conveys a fitness advantage both when it is rare in a population and when it has already become quite common. Moreover, we must show that this advantage applies to the individual that carries it, rather than to the group as a whole (simply assuming selection for the benefit of the group is widely considered a fallacy). Traits that benefit the group rather than the individual can only evolve under quite specific circumstances described by kin selection and social evolution theory sensu Frank [102].

A third point is that we need to be aware of the fact that the fitness advantage of a trait might not, or not only, come from its contribution to increased success in reproduction through increased survival (natural selection in the narrow sense, though including benefits associated with individuals’ ability to establish effective social alliances), but may also come from the trait’s effects on increased success in reproduction via attractiveness to potential partners (Darwinian sexual selection [103]). This could be particularly relevant for the evolution of music, as sexual selection is invariably invoked in understanding the evolution of elaborate animal aesthetic displays (where the connection between display and fitness can be very indirect). Music is nothing if not an aesthetic display (although possibly much else besides). Darwin treated it as such, and proposed sexual selection as the mechanism behind it.

Finally, a fourth point is that in order to confer a selective advantage, a trait or behaviour need not be essential for survival, but need only confer a slightly improved likelihood of survival to procreation, and/or a greater rate of procreation—thus perpetuating and increasing the frequency of that trait or behaviour—than would otherwise be the case. There is thus no justification for the common observation that music could hardly be a product of evolution by selection as it is hardly essential for survival. The former observation does not in fact follow from the latter. Furthermore, it does not rule out the possibility that various of the abilities that are used in musical activities may have been initially selectively favoured as a consequence of their fulfillment of other purposes (e.g. interpersonal communication and the establishment of interpersonal relationships), and that musical practices may have developed within the context of those uses; musical behaviours have the potential to fulfill some of those same purposes, or other purposes, potentially in even more effective ways. The co-use of these underlying abilities could lead to increasing interdependence between them, uniting them functionally in this new behavioural system, and potentially leading to further selective processes acting upon those underlying abilities and the behaviours that use them.

Some of the traits that are essential for musical activities may have been a product of biological (natural or sexual) selection, and this could be by conferring a fitness advantage either in the context of their use in musical activities themselves, or in a different context of use. Meanwhile, as observed in the preceding sections, certain properties of music and the traits that support them need not have been the product of biological selection at all.

The constraints outlined in the preceding sections indicate that some of the abilities prerequisite for music (e.g. entrainment and vocal learning) would appear to have arisen in our lineage, or at the least adapted from existing mechanisms to take on essentially novel form, in the period between our last common ancestor with chimpanzees (approx. 6–7 Myr) and the appearance of our own species (approx. 200 000 years ago). Proposals regarding the emergence of these abilities should be complementary to, and tested against, what we know of the physiology, behaviour and ecology of the hominins in that intervening period. This is no small task as clearly the knowledge both in palaeoanthropology and in the study of primate and human cognition is in a state of constant flux. Nevertheless, some aspects of our understanding in both areas are well-enough established that we should undertake to ensure that proposals regarding the evolution of these capacities do not contradict core understanding in hominiv evolution.

For example, hypotheses regarding the development of collective bodily and vocal display behaviours in early hominins from those exhibited by chimpanzees (and presumably our last common ancestor with them) should be framed in the context of changes in the habitat and group size of successive hominin species. It is now well established that gracile australopithecines exploited more open environments and a more omnivorous diet than higher primates of today, but nevertheless continued to exploit wooded environments for shelter and some aspects of subsistence [104,105]. Meanwhile the physiological characteristics and ecological contexts of early Homo (H. habilis, early African H. erectus and their descendants such as H. heidelbergensis) indicate that they were exploiting a far greater range of more open environments, lacustrine and riverine habitats, and that carnivory had increasingly taken the place of arboreal frugivory in their subsistence resource exploitation [104,105]. The efficacy of any proposed alterations to ancestral ‘carnival displays’ [69], coalitional displays [100] or size-exaggeration vocalizations [106], for example, needs to be situated within the context of these changes (see also [7]).

The mating strategies of human ancestors, and the social organization that arises from them, are also relevant to assessing the ecological validity of models regarding foundations of musical behaviours in interpersonal communication and display behaviours. This is because strategies for interpersonal communication, alliance and pair-bond formation, and display behaviours, will vary according to whether, for example, populations are monogamous, polygynous or polyandrous. In polygynous species, for example, males compete with other males for access to multiple females, with little or no long-term alliance commitment to any one female. By contrast, monogamous species form pair-wise long-term cooperative bonds between males and females. In each case the types of cooperation and competition, and with whom cooperation and competition occur, vary, and the behaviours leading to success in negotiating alliances and long-term bonds, and in display directed at the same sex, and directed at the opposite sex, vary accordingly (e.g. [107]).

In the case of human ancestors, high levels of sexual dimorphism and rapid developmental life history in australopithecines comparable to that of chimpanzees [108] have been taken to indicate broadly similar mating strategies and male–female relations to those exhibited by chimpanzees (e.g. [105], and references therein). Meanwhile, trends towards a reduction in sexual dimorphism and increased altriciality in the infants of early Homo indicate the development of increased cooperative long-term pair-bonding (e.g. [105] and references


therein). As noted in the previous paragraph, these developments have a direct influence upon the forms and efficacy of behaviours related to display, sexual selection, pair-bonding and vocalization between adults (see also [39]). The development of greater altriciality in infants, a longer developmental process and greater dependence upon adult care also have direct impacts upon the vocal behaviours between adults and infants (e.g. [109]), and the learning opportunities of infants and juveniles (e.g. [108, 110, 111]).

Similarly the potential value and form of vocal learning capabilities should be tested and understood against the backdrop of physiological constraints for vocalization capabilities. For example, MacLarnon & Hewitt [112] studied the size of the nerve canal in the thoracic vertebrae of australopithecines, early African H. erectus (H. ergerst) and its later descendant H. heidelbergensis. These dimensions provide an indication of the level of fine control of breathing musculature present in the species, allowing controlled vocalizations of extended duration, with greater control over intensity and pitch contour. They concluded that the level of fine control of breathing musculature in early African H. erectus (ca 1.8 Myr) was not increased relative to that of chimpanzees or the earlier australopithecines, but that by the time of its immediate descendant, H. heidelbergensis (from approx. 5 to 600,000 years ago), the level of control would have been equivalent to that of modern H. sapiens. These changes could have served either song or speech, and it is worth noting that on a number of measures such as tidal volume, range of sub-glottal pressure and muscular control, the biomechanics of human song are more demanding than those of conversational speech [113].

While we are able to conclude that neural connections allowing deliberate planning, fine control and integration of laryngeal, oro-facial and respiratory musculature used in vocalizations (as would be required for vocal learning) emerged in the lineage since our last common ancestor with chimpanzees, the hominin fossil record does not preserve neural changes internal to the brain. However, if the development of the thoracic neurology for voluntary extended breathing control was to be useful for vocalization, it would have had to have been accompanied by, or preceded by, the development of the neurological connections in the brain allowing the planning and control of these aspects of vocalization. Conversely, the usefulness of the neurological connections in the brain allowing integrated voluntary control over the larynx, articulation and breathing would have been increased by the development of greater thoracic breathing control. It would seem likely that these two neurological systems, in the brain and the body, would have evolved in tandem, during the 1 million-or-so-year period between early African H. erectus and H. heidelbergensis (see also [7]).

The many ways in which evolutionary changes in traits and behaviours relevant to musical behaviours can occur, by biological selection or otherwise, are not mutually exclusive; by contrast, they can interact in important and complex ways, and any or all of them could have operated at various times in the course of human evolution. However, the distinctions between them have not always been clearly made in the literature discussing evolutionary rationales for musical behaviours. It is important that any future proposals do so, and clearly situate such mechanisms within what we know of the social and ecological contexts of human ancestors, and their physiological and neurological capabilities.

Acknowledgements. We thank Henkjan Honing for organizing, and the Lorentz Center, The Netherlands, for hosting the workshop that resulted in this paper. Our thanks also go to our editor Carel ten Cate, Guy Madison and an additional anonymous reviewer, for numerous comments and suggestions that have helped us improve its contents.

Endnote

1This sketch leaves out the many critical developments a vocal learning tradition must traverse in order to do what we propose it to have done in our case. The vocal learning capacity must first be emancipated from dependence on an innate song template, as in the bird mimics cited in the text. It must also abandon exclusive reliance on the tiny ‘vocal gestures’ that supply the song elements for most bird-song, learned and unlearned, to include elements more akin to musical notes, i.e. notes sustained at a given pitch with spectral energy concentrated to the fundamental. There is precedent for this in birds such as the pied butcherbird of Australia, a mimic and virtuoso singer [34–36]. Only on the basis of producing such music-like notes is vocal production learning likely to engage auditory system resonances with enough strength and reliability to become a factor in cultural transmission. The requirement that all this be in place if the process we have postulated is to get a start may help explain the rarity of actual tonal phenomena in animal song.

References


