



Phenotyping in the era of genomics: *MaTrics*—a digital character matrix to document mammalian phenotypic traits

Clara Stefen¹ · Franziska Wagner¹ · Marika Asztalos¹ · Peter Giere² · Peter Grobe³ · Michael Hiller^{4,5,6,7,8,9} · Rebecca Hofmann^{8,10} · Maria Jähde¹ · Ulla Lächele^{2,11} · Thomas Lehmann⁸ · Sylvia Ortman¹² · Benjamin Peters¹ · Irina Ruf^{8,10} · Christian Schiffmann¹² · Nadja Thier¹ · Gabriele Unterhitzberger¹² · Lars Vogt¹³ · Matthias Rudolf¹⁴ · Peggy Wehner¹⁴ · Heiko Stuckas¹

Received: 14 January 2021 / Accepted: 22 October 2021 / Published online: 7 December 2021
© The Author(s) 2021

Abstract

A new and uniquely structured matrix of mammalian phenotypes, *MaTrics* (*Mammalian Traits for Comparative Genomics*) in a digital form is presented. By focussing on mammalian species for which genome assemblies are available, *MaTrics* provides an interface between mammalogy and comparative genomics.

MaTrics was developed within a project aimed to find genetic causes of phenotypic traits of mammals using *Forward Genomics*. This approach requires genomes and comprehensive and recorded information on homologous phenotypes that are coded as discrete categories in a matrix. *MaTrics* is an evolving online resource providing information on phenotypic traits in numeric code; traits are coded either as absent/present or with several states as multistate. The state record for each species is linked to at least one reference (e.g., literature, photographs, histological sections, CT scans, or museum specimens) and so *MaTrics* contributes to digitalization of museum collections. Currently, *MaTrics* covers 147 mammalian species and includes 231 characters related to structure, morphology, physiology, ecology, and ethology and available in a machine actionable NEXUS-format*. Filling *MaTrics* revealed substantial knowledge gaps, highlighting the need for phenotyping efforts. Studies based on selected data from *MaTrics* and using *Forward Genomics* identified associations between genes and certain phenotypes ranging from lifestyles (e.g., aquatic) to dietary specializations (e.g., herbivory, carnivory). These findings motivate the expansion of phenotyping in *MaTrics* by filling research gaps and by adding taxa and traits. Only databases like *MaTrics* will provide machine actionable information on phenotypic traits, an important limitation to genomics. *MaTrics* is available within the data repository Morph-D-Base (www.morphdbase.de).

Keywords Comparative genomics · Discrete character states · Hard tissue · Museum specimens · Numeric coding · Visceral & life history traits

Introduction

Background

Knowing and understanding the organisms around us has always been important for mankind and thus describing and comparing phenotypes has a long tradition that goes beyond the emergence of academic disciplines (e.g., Pruvost et al.

2011). The phenotype of an organism refers to its observable constituents, properties, and relations. In mammalogy, morphological* and anatomical* data describing the body plan based on skeletal and visceral traits usually make up the largest part of phenotype descriptions. But features associated with physiology, behaviour, ecology, or lifestyle traits are also important to characterize intra- and interspecific differences and hence to describe biodiversity. Depending on preservation, the same traits can be studied in extinct species also via fossil remains. The phenotype of organisms and species can be considered to result from the interaction of the organism's genome with itself and its environment. Consequently, the era of genomics provides the basis to identify genomic loci that are associated with the

Handling editor: Pamela Burger.

✉ Clara Stefen
clara.stefen@senckenberg.de; heiko.stuckas@senckenberg.de

Extended author information available on the last page of the article

variety of phenotypic traits. To understand genomic bases of phenotypic diversity is not only a challenge to the field of genomics, but also to the scientific disciplines of organismic biology. To support this, a short summary of concepts underlying the discovery of genomic loci associated with phenotypic traits is given below.

Pioneering work that enabled first insight into links between genome and phenotype relied on model organisms. This required studying the molecular and phenotypic features of single species such as the fruit fly (*Drosophila melanogaster*), the zebra fish (*Danio rerio*) or the mouse (*Mus musculus*). These models provided decisive insights into the genes behind basic developmental processes, including organ function and morphogenesis (Meunier 2012). Translating developmental processes from model to a limited number of non-model organisms opened the field for evolutionary developmental biology (Evo-Devo) and explained the molecular basis of processes such as body plan evolution. Criteria and limitations in the choice of model organisms to use in Evo-Devo studies were discussed by Milinkovitch and Tzika (2007). However, there are some limitations on what model organisms can tell (Bolker 2012). Insights from experiments on a limited number of model organisms are restricted to the phenotypes present in that particular species. For example, rodents such as mice do not have canine teeth, making the mouse an inappropriate model to study the molecular mechanisms associated with these teeth. Furthermore, even if model organism research would reveal all genes necessary to develop a given phenotype (e.g., the digestive system), it would still remain unknown which of these genes played the significant role in evolution and caused specific phenotypic differences between species (e.g., adaptation to particular diets).

Given these limitations, novel approaches to explore genome–phenotype relationships were developed using the availability of an increasing number of fully sequenced genomes. In fact, with the improvement of sequencing technologies, sequencing and assembly of whole genomes became possible; the first was published in 1995 (of the bacteria *Haemophilus influenzae*, Fleischmann et al. 1995) and the mouse genome was “only” published in 2002 (Waterston et al. 2002). Due to advancements in high-throughput DNA sequencing, there is an increasing number of species for which sequenced nuclear genomes are available (e.g., Genome 10K Community of Scientists 2009; Teeling et al. 2018; Feng et al. 2020; Zoonomia Consortium 2020). This wealth of genomes provides a basis for comparative genomics (“defined as the comparison of biological information derived from whole-genome sequences” and as discipline / methodology thus only started in 1995 (de Crécy-Lagard and Hanson 2018)). While comparative genomics often aims at identifying genomic elements that are conserved across species and thus likely have an evolutionarily important

function (Nobrega and Pennacchio 2004), comparative genomics can also be used to detect differences in functional genomic elements and associate them with phenotypic differences of species of interest. For example, targeted analyses of genes associated with the formation of dentin (DSPP) and enamel (AMTN, AMBN, ENAM, AMELX, MMP20) across Mammalia and Sauropsida (including Aves, Crocodylia, Testudines, Squamata) showed an association between the loss of these genes and the loss of teeth (Meredith et al. 2009, 2013). Another example are losses of chitinase genes (CHIAs), enzymes that digest chitin, which preferentially occurred in mammalian species that have non-insectivorous diets (Emerling 2018).

The above cited studies exemplify that the application of comparative genomics to identify links between genome and phenotype requires the systematic and comparative assessment of phenotypes for many non-model organisms. The same is true for recent advances in comparative genomics which follow the idea that convergent phenotypic evolution can be associated with convergent genomic changes, e.g., gene loss (Lamichhaney et al., 2019). This assumption is one conceptual foundation of the general *Forward Genomics* approach that performs an unbiased screen for genomic changes being associated with convergent phenotypic traits (Hiller et al. 2012; Prudent et al. 2016). This approach employs phenotype matrices and genome alignments to search for associations between convergent phenotypic traits and genomic signatures. *Forward Genomics* primarily delivers candidate genes or candidate genomic signatures and their causal relationship to the phenotype of interest needs to be inferred from independent studies. This may require experimental work on gene function, e.g., using model organisms or model systems such as cell culture. But the function of candidate genes may also be described from other studies in the scientific literature. In this way, *Forward Genomics* identified new links between genomic changes in genes as well as regulatory elements and various phenotypic changes such as adaptations to fully aquatic lifestyles in cetaceans and manatees (Sharma et al. 2018a), echolocation in bats and toothed whales (Lee et al. 2018), reductions and losses of the mammalian vomeronasal system (Hecker et al. 2019a), the evolution of body armour in pangolins and armadillos (Sharma et al. 2018a), the absence of testicular descent (Sharma et al. 2018b), and the reduction of eye sight in subterranean mammals (Roscito et al. 2018; Langer et al. 2018).

Development of *MaTrics*

As demonstrated above, novel approaches in comparative genomics (including *Forward Genomics*) have proven their potential to link phenotypic differences between mammals to differences in their genomes. But these novel methods

also depend on phenotype information on species of interest. Consequently, it would be advantageous for this emerging science field to fall back to phenotype knowledge made digitally available in fully referenced data repositories. This should not only be a compendium of phenotype information on model and non-model species but should be presented in a discretized form, e.g., using a numeric code to label distinct phenotype categories. This is because currently available methods and approaches in comparative genomics (including *Forward Genomics*) cannot handle continuous data. Instead, they are best suited to explore genomic signatures underlying discrete traits such as the presence or absence of a structure or a trait (see examples and citations above). However, in contrast to genomic data, phenotypic data are not readily available in such a digitized form that it can be used by computer programs, not even for well-characterized species such as mammals with sequenced genomes. Research in zoology and related fields assembled a rich body of phenotypic knowledge. But the information assembled over centuries is usually documented using natural language and thus in the form of texts unstructured for computer programs and so the information is not machine actionable* (Vogt et al. 2010). Although this form of documentation is thorough, has proven its worth and will continue to be used effectively in zoology and related fields, it is of limited use for other disciplines. This is because substantial time investment would be required to search and extract relevant phenotypic data from published descriptions. As a result, this important cultural and scientific heritage is underutilized in scientific fields such as genomics.

Here we address the need for digitally available trait information by creating a phenotypic character matrix that summarizes the knowledge of organismic biology but meets the specific requirements of genomics. A central feature of such a data repository should be a data matrix presenting comprehensive information of many traits and where rows represent species and columns represent traits.

Constructing a comprehensive phenotype matrix poses several challenges. While “simple” phenotypes that can be compiled relatively easily across several mammals, more complex phenotypes require experienced researchers in morphology, anatomy, physiology, veterinary science or related fields. This is interpreting the collected information on phenotypes requires specialized knowledge of the terminology and taxon of interest. For example, the exact meaning of specialized terms might depend on the described taxon, the author, and the time of publication. Additionally, some terms might refer to spatio-structural properties, others to common function or presumed common evolutionary origin, or to a mixture of both. All this is well understandable to the experts, but difficult for non-experts. This also holds for information on phenotypes provided by matrices associated to and published with phylogenetic (cladistics) studies

(e.g. Horovitz and Sánchez-Villagra 2003). They represent a valuable source of information in organismic biology. But also, their use requires expert knowledge and particularly if information of different independent matrices have to be combined.

To make phenotypic information better understandable and retrievable, a Mammalian Phenotype Ontology (MPO) was developed (Smith et al. 2003; http://www.informatics.jax.org/searches/MP_form.shtml?). However, MPO is focused on “annotation of mammalian phenotypes in the context of mutations, quantitative trait loci and strains that are used as models of human biology and disease” (Smith et al. 2005). Washington et al. (2009) and Haendel et al. (2015) extended the ontology-based notation of phenotypes of human diseases to link them to animal models. So, working with information on phenotypes of mammals across all orders is still difficult for non-experts and even more so for computer algorithms. Thus, integrating the information on phenotypes in a machine actionable form with other sources of data becomes exceedingly challenging and time-consuming (Lamichhaney et al. 2019; Vogt 2019). For integrative research, a way is sought to allow exploiting this knowledge without involving experts in each project.

To improve the accessibility and usability (and to take full advantage) of expert knowledge, more and more information is being digitized, stored, and made accessible online such as current journals or even older and classic books (e.g., Biodiversity Heritage Library). There are currently several online databases that allow for storing, editing or publishing information on phenotypes (mainly on morphological ones) covering various taxa. Some examples are given in Table 1, but it does not represent an exhaustive view of all current efforts in that direction. Each of these databases have their own research purpose and relevance. Many gather state-coded morphological traits on a selection of taxa, in order to perform phylogenetical analyses (e.g., Morphobank; see Table 1). Even though in these cases encoded characters are available they are matrices from individual projects and not combined in one extendable matrix with cross-linked references, specimens, pictures or other information. Some databases provide illustrations of anatomical structures, but do not give detailed description, nor encode characters (e.g., Digimorph; see Table 1). Thus, most of existing matrices with information on phenotypic traits do not fulfil all requirements of *Forward Genomics* (or other comparative genomics methods). On the other hand, Washington et al. (2009) and Haendel et al. (2015), however) proposed ways to create matrices that function as an interface between phenotype and genotype, but they focused on human diseases only. With *MaTrics*, we created a machine actionable dataset about phenotypic traits of mammals specifically tailored for comparative genomics research. The focus of

Table 1 Examples of data repositories in which phenotypic data of different vertebrate taxa are collected.

Project	Link	Aim, type of information	Reference
Morphobank	http://www.morphobank.org	<i>Homology of phenotypes over the web; building the Tree of Life with phenotypes</i> , publicly accessible containing images and matrices	O'Leary and Kaufmann (2011)
Digimorph	http://www.digimorph.org	A National Science Foundation Digital Library at The University of Texas Austin, a dynamic archive that holds high-resolution X-ray computed tomography of biological specimens	
Morphbank	http://www.morphbank.net	A continuously growing database of Biological Imaging and stores images that scientists use for international collaboration, research and education	
Morphomuseum	https://morphomuseum.com	A special case: an online journal with associated data repository; <i>MorphoMuseum (M3)</i> is a publication of the Department of Paleontology of the Institut des Sciences de l'Évolution from Montpellier, France	Lebrun & Orliac (2016)
Morphological Image database	http://people.pwf.cam.ac.uk/rja58/database/morphsite_bmc07.html	A database of morphological characters and a combined-data reanalysis of placental mammal phylogeny.	Asher (2007)
Phenoscape	http://kb.phenoscape.org	Data resource that is ontology-driven and contains information about mutant zebrafish (<i>Danio rerio</i>) phenotypes curated by the zebrafish model organism database, ZFIN at http://zfin.org	Ruzicka et al. (2015), Edmunds et al. (2016)
TOFF	http://toff-project.univ-lorraine.fr	An open source repository focusing on fish functional traits. It aims to combine behavioural, morphological, phenological, and physiological traits with environmental measurements	Lecocq et al. (2019)

The table lists the projects with their URL and aim and/or type of information that is stored and, if available, references in which the project is introduced.

phenotypic traits recorded in *MaTrics* is on convergent traits, coded without a distinction between apomorphies or plesiomorphies.

Our goal is to create a knowledge pool on all “phenomes*” that represents the actual mammalian diversity. As a first step, the current *MaTrics* version intends to gather phenotypic information for taxa with well-aligned sequenced genomes in a machine actionable way to simplify and enhance the use of *Forward Genomics*.

While inference of putative homologies in genomic data, resulting in nucleotide sequence alignments, is fully automated, analyses of homology or more precisely 'comparative homology' (Vogt 2017) of phenotypic data cannot be executed by computer algorithms so far. This is irrespective of the type of basic information available (e.g., digitized literature, 2D/3D scans of museum specimens). Establishing comparative homology requires the identification of units of comparison across different OTUs, resulting in a phylogenetic character matrix. When creating matrices usable to link phenotypic differences between species to genomic loci one must first identify the phenotypic units that can be compared across the OTUs (identification of comparative homologies) prior to coding the *MaTrics*.

Design and coding* principles of *MaTrics*

MaTrics (version 1.0, released in January 2021; <https://www.morphdbase.de/?MaTrics-Mx-v1>) is implemented in the online data repository* Morph-D-Base (MDB, www.morphdbase.de, Grobe and Vogt 2009) and publicly available to anybody who registers. MDB is a state-of-the-art data repository to document phenotypic data and its matrix module is best suited to host *MaTrics*.

Principles and data entry

MaTrics meets all requirements of *Forward Genomics*. We primarily focused on mammalian species for which genome sequences are available. Some basic principles of *MaTrics* are described herein; a detailed user's guide is available online (Wagner et al. 2021).

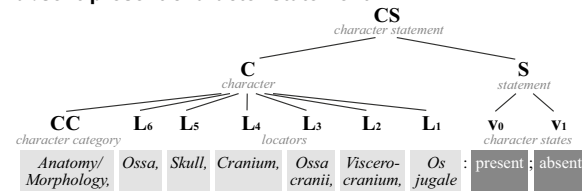
According to Sereno (2007), a (phenotypic) trait of an operational taxonomic unit (OTU; here the respective mammalian species) can be represented in a *character statement*, that is composed of two parts: *character* and *statement*, and can be divided into four types of logical

components (Sereno 2007: Table 4): one or more *locators*, a *variable*, and a *variable qualifier* as parts of the character and a *character state* as the *statement*. Not all these components are needed in any case, but a *locator* and a

character state are the minimum (representing *character* and *statement*). Thus, each *character* consists of at least one *locator* (L – the morphological structure, the structure bearing the trait) and the *statement* of the *character state* (v – mutually exclusive condition of a character) (Fig. 1). Specifying a *locator* and a *character state* is sufficient in case of absent-present *character statements*. Such kind of *character statements* can easily be encoded in a discretized form, i.e., coded as 0 or 1 (Fig. 1A, Table 2). Following Sereno's (2007) coding scheme, each character in *MaTrics* is named with a label starting with a single *locator* or a sequence of *locators* starting with L_n to L_1 (the trait-bearing structure), which provide all information necessary for unambiguously identifying and locating the trait within the OTU. The sequence of locators (L_n to L_1 as illustrated in Fig. 1) in the character label is hierarchically organized. And for clearer organization and orientation of characters a character category was added at the beginning of the 'character label' in MDB. While Sereno (2007) developed his coding scheme primarily for structural traits, we extended it here and applied it also to ecological or behavioural traits.

In case a phenotypic trait may have several different expressions or patterns, it must be coded as a 'multistate' character. Such a character needs a *variable qualifier* (q – the variable qualifier) (Fig. 1, Table 2). The *character states* of a multistate character in *MaTrics* are as discrete states coded using integers 2 to n. For example, the height of the mandibular canine teeth in relation to the level of the occlusal height (averaged) of the cheek teeth are coded as *low* (2), *occlusal height* (3) or *high* (4) (Fig. 1). Absence is recorded alongside other states, which is a shortcoming discussed by Sereno (2007). However, the state 'absent' is needed for the application of *Forward Genomics*.

absent-present character statement



multistate character statement

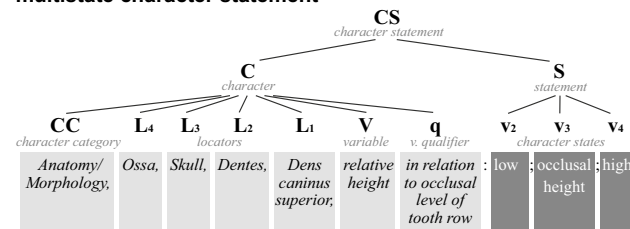


Fig. 1 Schematic illustration showing how phenotypic traits are reflected in *character statements* and in the character labels (shaded in grey) in *MaTrics*. The basic nomenclature is based on Sereno (2007: table 4, scheme 3). Top: structure for characters which can be described with only two *character states* (absent and present) exemplified by the jugal bone. Bottom: structure for characters which require more than two *character states* (multistate characters) exemplified by the length of the canine tooth in relation to the tooth row. Sereno's (2007) terminology recognizes *character statements* (CS) consisting of *characters* (C) and *statements* (S). The *character* is represented by *locators* (L_n, \dots, L_1 ; hierarchically organized) and optionally the *variable* (V) and the *variable qualifier* (q). The different expressions of the *variable* are given as *character states* (v_0, \dots, v_n) representing the *statement*.

Table 2 The options for numeric code options for (A) absent/present and (B) multistate characters in *MaTrics*.

Cell filling and state	State name	Description
(A) Absent/present characters		
?	Missing	Information is missing
–	Inapplicable	Refers to traits which are part of a structural complex which is absent in a species (e.g., a/p recording of roots in a toothless species)
0	Absent	Absence of the trait
1	Present	Presence of the trait
(B) Multistate characters		
?	Missing	Information is missing
–	Inapplicable	Refers to traits which are part of a structural complex which is absent in a species (e.g., trait "prehensile tail" in a tailless species)
2	State 2	Lowest expression (or absence) of the character variable
3, 4, 5, ..., n	state 3, 4, 5, ... n	Each different state of increasing expression of the character variable, either nominal or scaled, is given with a number starting with 3

The numerals 0 and 1 refer to the *character states* 'absent' and 'present' in absent/present characters, thus, the coding for multistate character states starts with 2. The default setting for a matrix cell in MDB is "missing"

Multistate characters in *MaTrics* can be described by different states or specifications. These states may be ordered based on their nature (like for example for the character clavícula pattern which might be absent, reduced or fully developed), or might even be metric (i.e. number of incisors inferior or superior), or nominal (e.g., the shape of the anterior nasals). The definition of each state is given by the author who included the character in *MaTrics* for a certain purpose. Another person for another purpose might define and use different states. Up to ten states per character states can be entered in *MaTrics*, so that the state definitions can be adjusted to fit other purposes.

A key consideration when generating *MaTrics* was to clearly document the source(s) of evidence for each phenotypic entry. The *character* part of each *character statement* possesses a short textual definition that is extracted from published sources (e.g. journals, text books, online references); it includes references to relevant ontology terms from various biomedical ontologies. The following online resources were used for the identification of adequate terms: Ontology Lookup Service, OLS, <https://www.ebi.ac.uk/ols/index>, Jupp et al. (2015); Ontobee, <https://www.ontobee.org>, Xiang et al. (2011); Bioportal, <https://bioportal.bioontology.org>, Musen et al. (2012). If no adequate definition was available, we provided a definition and clearly marked it as such.

Phenotypic traits coded in *MaTrics* represent by default adult states.

The dimensions of *MaTrics* are defined by the number of rows (OTUs) and columns (characters) that result in a specific number of cells (rows x columns). These cells primarily contain the character states. Morph-D-Base enables the addition of further information such as references, photos, illustrations, or museum specimen IDs to each matrix cell. All recorded character states and thus each cell of *MaTrics* is linked to at least one supporting reference. This refers either to citations from the literature (e.g., published journal articles, books, reliable scientific online resources) or to primary data sources. These data sources can cover IDs of museum specimens or media (e.g., photographs, images taken by microscopy, electron microscopy (TEM and SEM), magnetic resonance tomography (MRT), micro computed tomography (μ CT), or synchrotron) which can be uploaded in MDB or larger datasets might be linked to MDB. As a result, researchers using *MaTrics* can trace the information to at least one original source. This makes data entries not only revisable but offers the opportunity to *post hoc* re-analyse phenotypes for instance based on user-defined categories or even is raw data sets, e.g., continuous data sets.

The *MaTrics* or individual characters can be exported as a NEXUS file that provides data in a structured way and can be used as input in various software analysis tools.

Specificities of *MaTrics*

The primary motivation generating *MaTrics* was to create a research tool to link phenotypic differences between species to differences in their genomes. With this aim we follow the “from genome to phenome” approach initiated with the Human Phenome Project (Freimer and Sabatti 2003) and also discussed by others (see Edwards and Batley 2004; Scriver 2004). This is the reason why intraspecific variation of traits such as sexual dimorphism was not considered. *Character states* (presence/absence; multistate) do not take character polarity into account and character dependencies were not specifically considered. Specific characters of interest were added to *MaTrics* for some each research question, under certain considerations. Similarly, for different projects, characters can be selected individually to be retrieved from *MaTrics* for other use. Character dependencies can be avoided or reduced in this way, if needed.

Current status: *MaTrics* (version 1.0, release January 2021)

To date, *MaTrics* contains 231 characters for 147 mammalian species, resulting in a total of 33,957 matrix cells. The mammalian species considered in *MaTrics* include two representatives of Monotremata, five of Marsupialia and 140 of placental mammals (Supplementary Material Table S1). The number of species from each major clade of mammals neither represents the respective diversity nor morphological disparity of the respective trait. This is due to that the primary criterion for the inclusion in *MaTrics* was the availability and suitable quality of whole genomes when taxa were selected in 2016. A majority of the characters, 186 out of 231 (=80.52%), are coded as absent-present characters and the remaining 45 (19.48%) are multistate characters. The characters in *MaTrics* cover structural, ecological,

Table 3 Gross categories of 231 characters included in *MaTrics* and number of characters in these categories

Gross category	Subcategory	<i>N</i>
Anatomy/Morphology	Body plan	1
	Cranial skeleton	25
	Dentition	126
	Gastrointestinal tract	5
	Head	3
	Integument	3
	Postcranial skeleton	25
	Sense organs	1
	Ecology	31
Ethology	5	
Physiology	6	
Total		231

ethological, and physiological phenotypic traits (Table 3). All refer to the adult stage. For three characters (*os jugale*; *fully aquatic*; *body armor in the form of scales*), the recording is 100%, so all cells for these characters contain coded and referenced character states. Some traits were specifically included for the study in subsets of the listed mammals, and, therefore, the recording purposely is less complete and these characters include more cells still filled with “missing”. For overall coding status see Supplementary Material Table S2.

Notes on application

The primary motivation for creating *MaTrics* was to provide fully referenced phenotypic information for applications in comparative genomics, especially the *Forward Genomics* approach. The creation and filling of *MaTrics* and studies applying *Forward Genomics* were developed in parallel within the mentioned project. So, some phenotypes recorded in *MaTrics* were successfully used in earlier studies and simpler shorter tables, e.g., by Sharma et al. (2018a) who identified various convergent gene losses associated with some specific convergent mammalian phenotypes. They showed convincingly that tooth and enamel loss are associated with the loss of ACP4 (a gene that is associated with the enamel disorder *amelogenesis imperfecta*) and that the presence of scales is associated with the loss of the gene DDB2 (which detects substances resulting from UV light and helps to induce DNA repair). The fully aquatic lifestyle is associated with the loss of MMP12, a gene associated with breathing adaptation. The documented loss of these genes in some mammalian species is functionally explainable either as a consequence of trait loss (the genes ACP4 and DDB2 have no function after trait loss) or as putative adaptive genomic alteration, causing novel phenotypes (MMP12-loss is associated with novel lung functions in aquatic mammals) (Sharma et al. 2018a). Such results might help to better understand some related human diseases, as for example in the case of DDB2 whose mutations cause *xeroderma pigmentosum* which manifests in hypersensitivity to sunlight (Rapić-Otrin et al. 2003).

Another study investigated the gene losses associated with the reduction of the vomeronasal system (VNS) in several mammals. A genomic comparison of 115 mammalian genomes confirmed that *Trpc2* is an indicator for the functionality of the VNS (Hecker et al. 2019a). Moreover, it indicated a loss of functionality of the VNS in seals (Phocidae) and otters (Lutrinae). Morphological data are scarce for seals and there are no data for otters (Hecker et al. 2019a; Zhang and Nikaido 2020) and, therefore, we will proceed to test the accuracy of the suggested predictability. This study on the VNS is an example for testing genotype–phenotype associations in non-model organisms and shows the potential of

the combination of comparative morphological and genomic approaches.

Nevertheless, the relevance of *MaTrics* is by no means restricted to the *Forward Genomics* approach. Characters were also included in *MaTrics* for the usage in the contemporary study to explore evolutionary conditions associated with the loss of genes related to convergent evolution of herbivorous and carnivorous diet in mammals (Hecker et al. 2019b). This study included 52 placental species and suggests that the lipase inhibitor gene PNLIPRP1 is preferentially lost in herbivores, whereas the xenobiotic receptor NR1I3 is preferably lost in carnivores. Even though the authors put forward hypotheses, the lack of accessible data on mammalian diet preferences made it difficult to test whether gene losses are associated with dietary fat content and diet-related toxins. Investigating whether convergent gene loss is associated with similar dietary preferences may additionally hold information on whether gene losses might be adaptive (Albalat and Cañestro 2016). Consequently, an ongoing study records dietary categories in *MaTrics* that allow a semi-quantitative encoding of dietary fat content (associated with PNLIPRP1) and diet-related toxins (associated with NR1I3) (Wagner et al. accepted). This study will test whether the convergent loss of both genes is associated with the convergent evolutionary change of dietary preferences, i.e., the consumption of a diet with reduced fat and toxin contents.

Future analyses using *MaTrics* have the potential to test how gene losses and dietary composition are related to the presence/absence of structures or organs associated with digestive processes. Even further, it allows investigating whether evolutionary changes in diet composition are not only associated with the loss / presence of single molecules (e.g., lipase inhibitor, xenobiotic receptor), but also with changes in complex structures and their associated genes.

The two studies by Hecker et al. (2019b) and Wagner et al. (accepted) mentioned above show how genomic and morphological studies are entangled: current knowledge of morphology serves as basis for creating phenotypic trait matrices like *MaTrics* which — on the other hand — forms the basis of genomic research, especially the *Forward Genomics* approach. Hypotheses associated with findings of candidate loci, may in turn inspire further morphological research.

The most obvious applications are morphological studies. Although mammal dentitions are well studied and a lot is known about teeth number, form, and shape in particular in relation to dietary specialization (see Thenius 1989; Hillson 2005; Ungar 2010), we still have many knowledge gaps, e.g., concerning functional adaptations and evolutionary transformations. Thus, Sole and Ladevèze (2017) aimed to put forward new ideas on how the basic mammalian tribosphenic molar was transformed to sectorial teeth in hypercarnivorous mammals. The study only included carnivores as defined

by flesh-eating and the presence of carnassial teeth, representatives of the living Carnivoramorpha (including the extinct Nimravidae) and Dasyuromorpha, as well as from the extinct Sparassodonta, Oxyaenodonta, and Hyaenodonta. Comparing the cusp pattern/morphology of the upper and lower molars of these species Solé and Ladevèze (2017: fig. 4) derived a scheme for the morphological evolution of the sectorial teeth in hypercarnivorous mammals. They also aimed at providing new arguments to discuss the developmental aspects of the evolution of hypercarnivory by associating their morphological observations with ontogenetic studies. The latter highlighted the importance of the expression of ectodysplasin A (Eda): increased levels are able to modify the number, shape, and position of cusps in mice during tooth development (Kangas et al. 2004). Further, Häärä et al. (2012:3189) showed — again in mice — that “Fgf20 is a major downstream effector of Eda and affects Eda-regulated characteristics of tooth morphogenesis, including the number, size, and shape of teeth. Fgf20 function is compensated for by other Fgfs”. A study of hairless dog phenotypes in primary teeth (deciduous premolars and permanent molars) indicated that “the haploinsufficiency of *FOXI3* leads to an incomplete development of the lingually positioned cusps in the trigon(id) and talon(id) parts of both upper and lower molars and deciduous fourth premolars, respectively” (Kupczik et al. 2017:5). The ectodermal development regulator gene *Foxi3* is known to be a target of Eda and to be involved in tooth cusp development (Drögemüller et al. 2008); it suppresses epithelial differentiation (Jussila et al. 2015). Inspired by the observations and the model of Solé and Ladevèze (2017), we started a study with teeth and cusps in a subsample of Carnivora collected in *MaTrics* with two aims: first, to test the suitability of *MaTrics* in comparative morphological studies, and second, to set the basis to proceed with genome wide searches for genomic causes correlated with the loss of cusps. This seems to be promising with the development of new methods to include searches for regulatory elements (see below).

For the selected Carnivora (Supplementary Material Table S3) the absence and presence of individual tooth cusps for the fourth upper premolar (P^4) and all molar teeth were recorded in *MaTrics*. The nomenclature of the cusps followed Thenius (1989). The detailed descriptions of cusp patterns for the species are given in the Supplementary Material document S4 and examples are illustrated in Fig. 2 and detailed in Supplementary Material Table S5. Some of our results confirmed the observations of Solé and Ladevèze (2017). So, we confirm that parastyle and protocone of the P^4 are generally reduced in hypercarnivorous carnivorans. Interestingly, both structures are more reduced in the Canidae and the polar bear (*Ursus maritimus*) than in the members of the Felidae and Hyaenidae. Solé and Ladevèze (2017) reported that in the upper molars

protocone, paraconule and metaconule are reduced in hypercarnivorous mammals which is also in line with our findings (Fig. 3D–F; Supplementary Material Table S5).

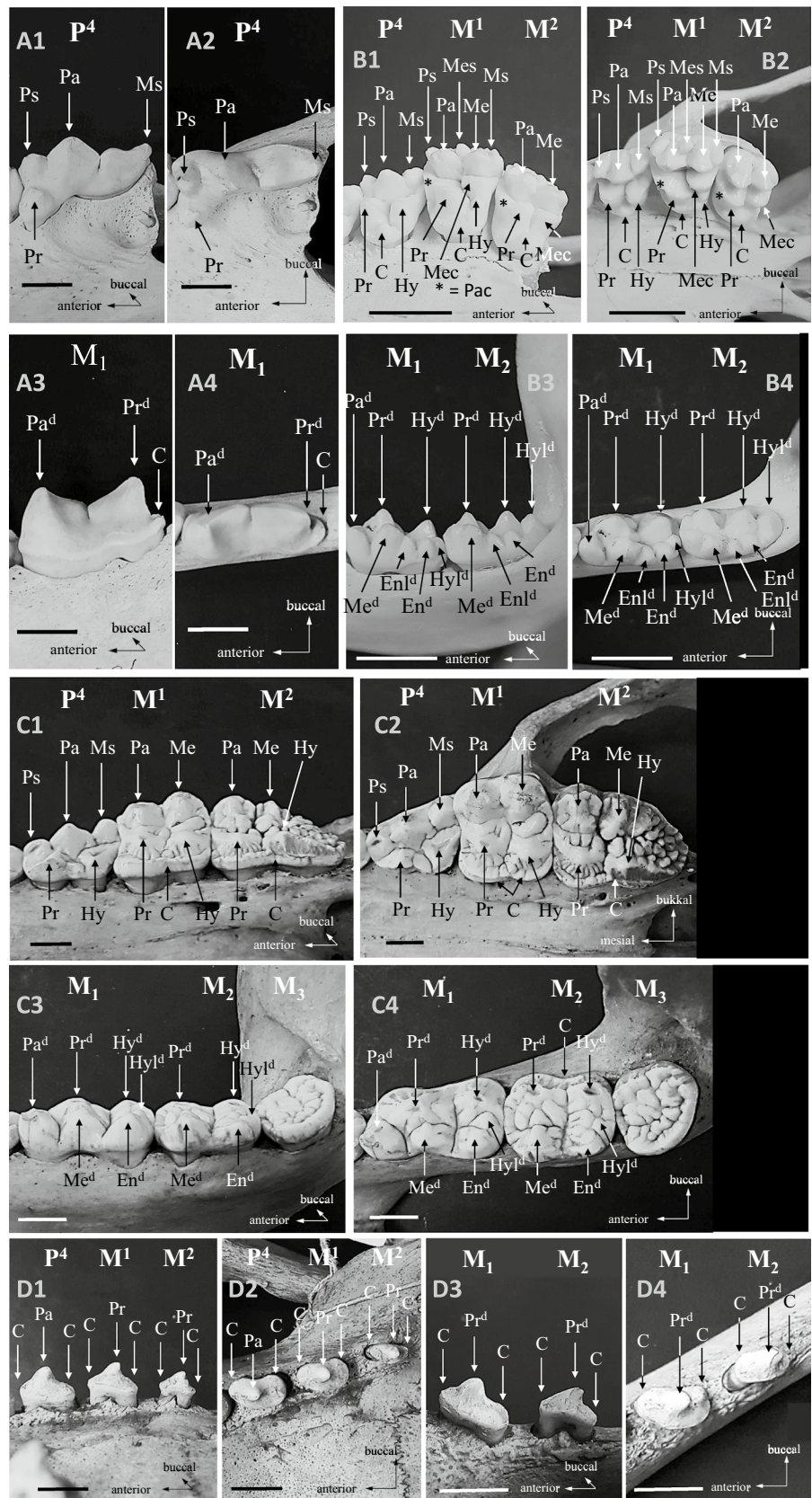
Solé and Ladevèze (2017) also observed that metaconid and talonid are generally lost in hypercarnivorous mammals, especially felid-like and hyaenid-like hypercarnivores. Based on our study, we found that metaconid and talonid are completely reduced only in the Felidae (except the cheetah, *Acinonyx jubatus*) and the spotted hyena (*Crocuta crocuta*). Like in the Canidae and the striped hyena (*Hyaena hyaena*), both structures are also present in *Ursus maritimus*. The specialized hypercarnivorous diet of several Feliformia lead to an extreme reduction of the tribosphenic molar, whereas the Canidae and *Ursus maritimus* also eat fruits and vegetables and, therefore, need crushing structures. The presence of protocone and talonid seems to be necessary for an omnivorous diet (Solé and Ladevèze 2017), but, based on our study, we can confirm that this is also true for herbivorous species (e.g., red panda, *Ailurus fulgens*; giant panda, *Ailuropoda melanoleuca*).

Except for the Pacific walrus (*Odobenus rosmarus*) at least ten specimen per species were analysed and for several species' individual deviations from the common cusp pattern were observed (Table 4). *MaTrics* was not designed to take intraspecific variability into account. Therefore, only the most common cusp patterns for each species were recorded. Variations of the cusp patterns can affect several cusps in domestic dog and brown bear (*Ursus arctos*), but only one cusp in the red fox (*Vulpes vulpes*). Such exceptions are important as they might indicate evolutionary trends. However, variations within a species cannot be reflected in *MaTrics* as only one-character state is attributed to a given species for each character here. Only in this way the (common) absence or presence of a trait can be compared with the genome of again one representative of a species. Studies on intraspecific variability of certain characters would need matrices redesigned for this purpose. There are two options for this: first, record characters in Morph-D-Base for specimens instead of species which would result in several rows of specimen for one taxon. These could be pooled in a phylogenetic analysis by restricting the tree-space for searching the best tree to trees that include clades that comprise all pooled specimens of a species. Or, second, one could enable the recording of several character states for the same character in the matrix, thus representing the variability found across the various specimens of a given species.

Conclusion and future perspectives

Recent advances in molecular techniques lead to a rapid increase in the assembly and publication of genomes from various organisms. However, knowledge of the genome

Fig. 2 Some examples for the presence of cusps in the studied Carnivora in P^4 as well as upper and lower molars. **A** The spotted hyaena *Crocuta crocuta* MTD B4936, **B** the red panda *Ailurus fulgens* MTD B17478, **C** the giant panda *Ailuropoda melanoleuca* ZMB_Mam_17246 and **D** the Weddell seal *Leptonychotes weddellii* MTD B5029. For each species P^4 and upper molars (insets 1, 2) as well as the lower molars (insets 3, 4) are illustrated. The teeth are photographed in lateral (insets 1, 3) and occlusal (insets 2, 4) view. Abbreviations alphabetically: En^d – entoconid, Enl^d – entoconulid, Hy – hypocone, Hy^d – hypoconid, Hyl^d – hypoconulid, Me – metacone, Mec – metaconule, Me^d – metaconid, Mes – mesostyle, Ms – metastyle, Pa – paracone, Pac – paraconule, Pa^d – paraconid, Pr – protocone, Pr^d – protoconid and Ps – parastyle.



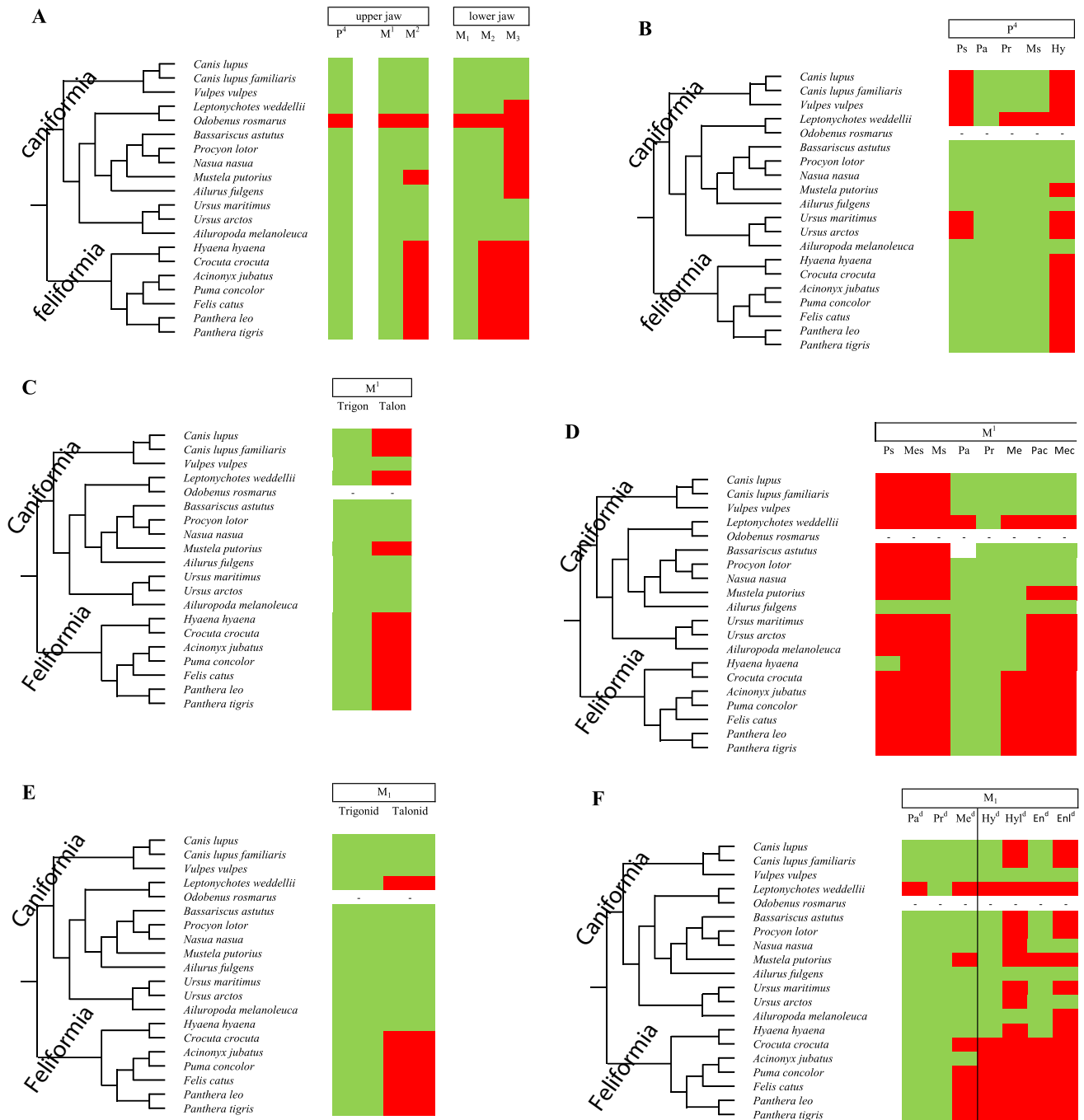


Fig. 3 Comparison of the presence or absence of individual teeth (P⁴-M¹-M² and M₁-M₃), of trigonid and talonid (M¹, M₁), and of their cusps (P⁴, M¹, M₁ in some species of Carnivora plotted on a phylogenetic tree based on Nyakatura and Bininda-Emonds (2012) and Agnarsson et al. (2010) **A** Presence/Absence of individual teeth, **B** presence/absence of cusps on the P⁴, **C** presence/absence of the trigonid and talonid on the M¹, **D** presence/absence of cusps on the M¹,

E presence/absence of the trigonid and talonid on the M₁, **F** presence/absence of cusps on the M₁. En^d – entoconid, En^l – entoconulid, Hy – hypocone, Hy^d – hypoconid, Hy^l – hypoconulid, Me – metacone, Me^d – metaconid, Mec – metaconule, Mes – mesostyle, Ms – meta-style, Pa – paracone, Pac – paraconule, Pa^d – paraconid, Pr – protocone, Pr^d – protoconid and Ps – parastyle

sequences is only a first step to understand the relationships between genomic changes, the phenotype of organism and phenotypic differences between different organisms

(Hardison 2003). The systematic description of phenotypic information in matrix form like in *MaTrics* is necessary to understand the genome information and to deal with

Table 4 Deviations in cusp patterns in the studied Carnivora.

Species	Deviation from common cusp pattern for species
<i>Canis familiaris</i>	Metaconid and hypoconid at M ³ Small cusp mesial of paracone at P ⁴ Entoconulid (mesial of entoconid) at M ₁ Additional fourth lower molar
<i>Vulpes vulpes</i>	Small cusp mesial of paracone at P ⁴
<i>Ursus arctos</i>	Second cusp palatal at P ⁴ Third lingual cusp at M ₂ Three metaconid-cusps at M ₂ Third palatal cusp at M ¹

M¹⁻³, upper (indicated by number in superscript)/lower molar tooth (indicated by subscript); P⁴ – upper 4th premolar

questions related to evolutionary biology and biomedicine. This is not restricted to mammals as the coding principles of *MaTrics*, which comply with the requirements of molecular research, can serve as a template for matrices comprising trait knowledge of other vertebrate and non-vertebrate groups. The establishment of trait matrices for various taxa could lead to a broad documentation of phenotypes for applications in comparative genomics, and hence, enable a systematic exploration of genotype-phenotype associations.

However, trait collections such as *MaTrics* also revealed a tremendous research gap in phenotypic data. In fact, filling *MaTrics* with information on different phenotypic traits across mammals showed that detailed information on structural, physiological, or life history traits was often not available for many species, even with intensive literature research. For example, reductions of the vomeronasal system (VNS) are documented in several mammals and our previous genomic comparison of 115 mammalian genomes uncovered several genes whose loss is associated with a reduced or non-functional VNS (Hecker et al. 2019a). This genomic screen also revealed that seals (Phocidae) and otters (Lutrinae) have lost some of these genes, indicating a reduced VNS. However, to the best of our knowledge, information concerning the vomeronasal organ of Phocidae and Lutrinae is not available. Indeed, the recording status in *MaTrics* for the character “vomeronasal organ” with the states absent/present is only 37%. Another example of a character, that would be assumed to be well-known, is the absence/presence of the gall bladder (“*Vesica biliaris*”), with a recording status of 70%. In other words, the recording status of the characters in *MaTrics* demonstrate the lack of information on phenotypic traits in several species. These research gaps can only be filled by specimen-based research (e.g., Thier and Stefen 2020). Although individual studies are valuable scientific contributions, they may not suffice to close the substantial research gaps in short time. The authors see the need for more basic zoological research complementing the

systematic exploration of the genomic basis of biodiversity, i.e., research activities on biodiversity genomics could be assisted by research initiatives on biodiversity phenomics (= systematically phenotyping animals in matrices like *MaTrics*).

Most of the genomic studies mentioned above identified protein coding genes associated with complex body plan changes (e.g., aquatic and aerial lifestyle of cetaceans and bats, respectively). However, evolutionary theory predicts that changes in cis-regulatory genetic elements are probably more important for morphological changes than protein-coding genes. For instance, Roscito et al. (2018) stated that the loss of morphological traits is (often) associated with the decay of the cis-regulatory elements. Consequently, the *Forward Genomics* approach has been further developed to include methodologies that can successfully associate phenotypes with the loss or presence of regulatory elements (e.g., Langer et al. 2018; Langer and Hiller 2019). In awareness of these developments, the phenotype matrix presented here already provides a number of morphological characters that will be subject to further exploration in the near future. Thus, the phenotypic information compiled in *MaTrics* will be of increasing importance. This applies for instance to those referring to tooth morphology and tooth cusps discussed above. In fact, tooth characters are known to be the result of a complex signalling network involving timely graded activation and deactivation of genes controlled by regulatory elements (e.g., Jernvall and Thesleff 2000; Thesleff et al. 2001).

A last aspect to be mentioned refers to the way how phenotypic information is documented. So far, filling *MaTrics* with information is still mostly conducted by hand; experienced scientists have to control the content and to check for homology. However, some recent developments may open the door to the partial automation of this work. First, the implementation of ontologies and semantic phenotypes in the platform Morph-D-Base. The development of a respective semantic description module is already initiated (Vogt and Baum 2019; Vogt 2019). This is expected to allow the development of computer algorithms to mine data on homologous structures to establish matrices more automatically (Vogt 2018).

MaTrics is a new and unique data collection of phenotypic traits of mammal species. By including homologous phenotypic traits across (an increasing number of) species, *MaTrics* and similar matrices can serve as basis for a variety of research fields as illustrated herein. The recorded phenotypic traits are well defined and fully referenced (*characters* as well the *character state* for each species). Not only literature data are accepted for the latter, but also references to specimens in collections, which contribute in a specific way to the digitalization of collection material. *MaTrics* data are directly useful in genomic studies since the *character*

states are numerically coded and hence can be extracted as NEXUS file to be machine actionable. The scientific potential of digitized phenotype matrices is apparent and motivates thinking about future development.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42991-021-00192-5>.

Acknowledgement We want to thank members of the German Society of Mammalogy (DGS) for stimulating discussions at the annual DGS meeting 2019 which were useful to shape the manuscript. Also, the helpful comments of the reviewers (one anonymous and Martín Ramírez) are thankfully acknowledged.

Authors' contributions (optional: please review the submission guidelines from the journal whether statements are mandatory), CS, HS, MH had the idea for the manuscript, CSt, FW, HS drafted and finalized the manuscript; IR, PGI, MH, RH, UL, SO, TL, NT commented and participated in writing the manuscript; MA, CSt did the study on cusps in teeth of Carnivora; BP, CSch, CSt, FW, GU, IR, MA, MJ, NT, RH, SO, TL, UL were involved in coding characters in *MaTrics*; PGr, LV provided Morph-D-Base, implemented tools therein and technically finalized *MaTrics*; FW computed statistics for *MaTrics*, MR, PW did the statistical calculations on characters.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was funded as part of the interdisciplinary research project 'Identifying genomic loci underlying mammalian phenotypic variability using *Forward Genomics*' by the Leibnitz Association, grant SAW-2016-SGN-2.

Availability of data and material Data are available within Morph-D-Base (www.morphdbase.de).

Code availability (software application or custom code): Not applicable.

Declarations

Conflicts of interest None known for all authors.

Ethics approval Not applicable as no experiments with living animals were performed.

Consent to participate All authors agreed to the work.

Consent for publication All authors agree to the publication.

Glossary

Anatomy – "The demonstrable facts of animal structure, or also, by transference to the object, the structure or even the tissue of the animal itself." (Snodgrass 1951:173). In other words, anatomy is the part of the phenotype of an organism that refers to its physical and structural properties. At the same time, it refers to the science of anatomy, with anatomical data being facts about the anatomy of organisms.

Data repository – A large database infrastructure that collects, manages, and stores data sets for data analysis, sharing and reporting. A data repository is also known as a data library or data archive. *NCBI GenBank* is an example of a data repository for a sequence database.

Machine actionable – Data and metadata that are structured in a formalized and consistent way so that machines (i.e., computers) can read and use them with algorithms that were programmed against this structure. Machine actionability of data and metadata includes for instance the

use of persistent identifiers for data creators (e.g., ORCID), organizations and funding agencies, but also open accessibility of data for machines through a corresponding application programming interface (API), and basic semantics that allow algorithms to distinguish different categories of information and apply rules to them. Machine actionability in this sense goes beyond machine readability which only requires data and metadata to be readable by a machine, i.e., data and metadata must be provided in a machine readable format. Machine readability does not necessarily require data and metadata to provide basic semantics for allowing algorithms to distinguish different categories of information contained in them.

Morphology – "Our philosophy or science of animal form, a mental concept derived from evidence based on anatomy and embryology, usually incapable of proof, attempting to discover structural homologies and to explain how animal organization has come to be as it is." (Snodgrass 1951:173). In other words, morphology refers to the interpretations of anatomical facts within theories and hypotheses such as homology.

NEXUS file – A file format widely used in bioinformatics. It stores information about taxa, phenotypic characters, trees, and other information relevant for phylogenetics. Several phylogenetic programs such as PAUP, MrBayes, and Mac Clade use this format.

Phenome – the entirety of all observable physical or physiological traits or characteristics of an organism

Phenotypic trait – A specific part of the phenotype of an organism. The phenotype of an organism refers to its observable constituents, properties, and relations that can be considered to result from the interaction of the organism's genotype with itself and its environment. Anatomy is the part of the phenotype that refers to the physical and structural properties of the organism.

Ontology – Ontologies are dictionaries that can be used for describing a certain reality. They consist of labeled classes and relations between classes, both with clear definitions that are ideally created by experts through consensus and that are formulated in a highly formalized canonical syntax and standardized format with the goal to yield a lexical or taxonomic framework for knowledge representation (Smith 2003). Each ontology class and relation (also called property) possesses its own Uniform Resource Identifier (URI*) through which it can be identified and individually referenced. Ontologies contain expert-curated domain knowledge about specific kinds of entities together with their properties and relations in the form of classes defined through universal statements (Schulz et al. 2009, Schulz and Jansen 2013). Ontologies in this sense do not include statements about particular entities (i.e., empirical data) (Vogt et al. 2019).

URI – A Uniform Resource Identifier (URI) is a string of characters that follows a specific structure and unambiguously identifies a particular resource. The URI can also serve as a URL (web address), and can be resolved to an IP address.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.













References

- Agnarsson I, Kuntner M, May-Collado LJ (2010) Dogs, cats, and kin: a molecular species-level phylogeny of Carnivora. *Molec Phy Evol* 54:726–745. <https://doi.org/10.1016/j.ympev.2009.10.033>
- Albalat R, Cañestro C (2016) Evolution by gene loss. *Nature Rev Genet* 17:379–391
- Asher RJ (2007) A web-database of mammalian morphology and a reanalysis of placental phylogeny. *BMC Evol Biol* 7:108. <https://doi.org/10.1186/1471-2148-7-108>
- Bolker J (2012) Model organisms: There's more to life than rats and flies. *Nature* 491(7422):31
- De Crécy-Lagard V, Hanson AD (2018) Comparative Genomics. Reference Module in Biomedical Sciences. <https://www.sciencedirect.com/topics/neuroscience/comparative-genomics>
- Drögemüller C, Karlsson EK, Hytönen MK, Perloski M, Dolf G, Sainio K, Leeb T (2008) A mutation in hairless dogs implicates *FOXI3* in ectodermal development. *Science* 321(5895):1462–1462
- Edmunds RC, Su B, Balhoff JP, Dahdul WM, Lapp H, Lundberg JG, Vision TJ, Dunham RA, Mabee PM, Westerfield M (2016) Phenoscope: Identifying candidate genes for species-specific phenotypes. *Molec Biol Evol* 33:13–24. <https://doi.org/10.1093/molbev/msv223>
- Edwards D, Batley J (2004) Plant bioinformatics: from genome to phenotype. *TRENDS in Biotechnology* 22(5):232–237
- Emerling CA, Delsuc F, Nachman MW (2018) Chitinase genes (CHIAs) provide genomic footprints of a post-Cretaceous dietary radiation in placental mammals. *Science Advances* 4(5):eaar6478
- Feng S, Stiller J, Deng Y, Armstrong J, Zhang G et al (2020) Dense sampling of bird diversity increases power of comparative genomics. *Nature* 587:252–257. <https://doi.org/10.1038/s41586-020-2873-9>
- Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J, Dougherty B, Merrick J et al (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–512. <https://doi.org/10.1126/science.7542800>
- Freimer N, Sabatti C (2003) The human genome t. *Nature genetics* 34(1):15–21
- Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole genome sequence for 10 000 vertebrate species. *J Hered* 100(6):659–674
- Grobe P, Vogt L (2009) Documenting Morphology: Morph-D-Base. In: Wägele JW, Bartolomeaus T (eds) *Deep Metazoan Phylogeny: The Backbone of the Tree of Life –New Insights from Analyses of Molecules, Morphology, and Theory of Data Analysis*. De Gruyter, Berlin, pp 475–503. <http://www.morphdbase.de>
- Häärä O, Harjunmaa E, Lindfors PH, Huh SH, Fliniaux I, Åberg T, Jernvall J, Ornitz DM, Mikkola ML, Thesleff I (2012) Ectodysplasin regulates activator-inhibitor balance in murine tooth development through Fgf20 signaling. *Development* 139(17):3189–3199
- Haendel MA, Vasilevsky N, Brush M, Smedley D (2015) Disease insights through cross-species phenotype comparisons. *Mammalian Genome* 26(9):548–555
- Hardison RC (2003) Comparative genomics. *PLoS Biol* 1(2):e58
- Harrow JL, Steward CA, Frankish A, Gilbert JG, Gonzalez JM, Loveland JE, Wilming LG et al (2014) The vertebrate genome annotation browser 10 years on. *Nuc Acid Res* 42(D1):D771–D779
- Hecker N, Lächele U, Stuckas H, Giere P, Hiller M (2019) Convergent vomeronasal system reduction in mammals coincides with convergent losses of calcium signalling and odorant degrading genes. *Mol Ecol* 28(16):3656–3668
- Hecker N, Sharma V, Hiller M (2019) Convergent gene losses illuminate metabolic and physiological changes in herbivores and carnivores. *Proc Natl Acad Sci USA* 116(8):3036–3041
- Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G (2012) A “*Forward genomics*” approach links genotype to phenotype using independent phenotypic losses among related species. *Cell reports* 2(4):817–823
- Hillson S (2005) *Teeth*. Cambridge University Press, Cambridge
- Horovitz I, Sánchez-Villagra MR (2003) A morphological analysis of marsupial mammal higher-level phylogenetic relationships. *Cladistics* 19(3):181–212
- Huelsmann M, Hecker N, Springer MS, Gatesy J, Sharma V, Hiller M (2019) Genes lost during the transition from land to water in cetaceans highlight genomic changes associated with aquatic adaptations. *Sci Adv* 5(9):eaaw6671
- Jernvall J (2000) Linking development with generation of novelty in mammalian teeth. *Proc Nat Acad Sci* 97(6):2641–2645
- Jernvall J, Thesleff I (2000) Reiterative signalling and patterning during mammalian tooth morphogenesis. *Mechanisms dev* 92(1):19–29
- Jupp S, Burdett T, Leroy C, Parkinson HE (2015) A new Ontology Lookup Service at EMBL–EBI. In: Malone J et al. (eds.) *Proceedings of SWAT4LS International Conference 2015*, pp 118–119
- Jussila M, Aalto AJ, Navarro MS, Shirokova V, Balic A, Kallonen A, Thesleff I (2015) Suppression of epithelial differentiation by *Foxi3* is essential for molar crown patterning. *Development* 142(22):3954–3963
- Kangas AT, Evans AR, Thesleff I, Jernvall J (2004) Nonindependence of mammalian dental characters. *Nature* 432(7014):211–214
- Kupczik K, Cagan A, Brauer S, Fischer MS (2017) The dental phenotype of hairless dogs with *FOXI3* haploinsufficiency. *Sci Rep* 7(1):1–8
- Lamichhaney S, Card DC, Grayson P, Tonini JF, Bravo GA, Näpflin K, Sackton TB et al (2019) Integrating natural history collections and comparative genomics to study the genetic architecture of convergent evolution. *Phil Trans Royal Soc B* 374(1777):20180248
- Langer BE, Hiller M (2019) TFforge utilizes large-scale binding site divergence to identify transcriptional regulators involved in phenotypic differences. *Nuc acids res* 47(4):e19–e19
- Langer BE, Roscito JG, Hiller M (2018) REforge associates transcription factor binding site divergence in regulatory elements with phenotypic differences between species. *Mol Biol Evol* 35(12):3027–3040
- Lebrun R, Orliac MJ (2016) MorphoMuseuM: an online platform for publication and storage of virtual specimens. *Paleontol Soc Papers* 22:183–195. <https://doi.org/10.1017/scs.2017.14>
- Lecocq T, Benard A, Pasquet A, Nahon S, Ducret A, Dupont-Marín K, Lang I, Thomas M (2019) TOFF, a database of traits of fish to promote advances in fish aquaculture. *Scientific Data* 6(1):1–5
- Lee JH, Lewis KM, Moural TW, Kirilenko B, Bogdanova B, Prange G, Koessl M, Huggenberger S, Kang C, Hiller M (2018) Molecular parallelism in fast-twitch muscle proteins in echolocating mammals. *Science Adv* 4(9):eaat9660
- Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS (2009) Molecular decay of the tooth gene enamel (ENAM) mirrors the loss of enamel in the fossil record of placental mammals. *PLoS Genet* 5(9):e1000634
- Meredith RW, Gatesy J, Springer MS (2013) Molecular decay of enamel matrix protein genes in turtles and other edentulous amniotes. *BMC evl biol* 13(1):20
- Meunier R (2012) Stages in the development of a model organism as a platform for mechanistic models in developmental biology: Zebrafish, 1970–2000. *Studies History Philosophy Sci Part C: Stud History Philosophy Biological Biomedical Sci* 43:522–531

- Milinkovitch MC, Tzika A (2007) Escaping the mouse trap: the selection of new Evo-Devo model species. *J Exper Zool B Mol Dev Evol* 308(4):337–346
- Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, SmithNCBO team B (2012) The national center for biomedical ontology. *J Am Med Inform Assoc* 19:190–5 (Epub 2011)
- Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological discovery. *J physiol* 554(1):31–39
- Nyakatura K, Bininda-Emonds ORPp, (2012) Updating the evolutionary history of Carnivora (Mammalia): a new species-level super-tree complete with divergence time estimates. *BMC Biology* 10:12. <https://doi.org/10.1186/1741-7007-10-12>
- O’Leary MA, Kaufman S (2011) MorphoBank: phylophenomics in the “cloud.” *Cladistics* 27:1–9
- Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-generation ecological and evolutionary genomics? *TREE* 27(12):673–678
- Porter IH (1973) From gene to phene. *J Invest Dermatol* 60(6):360–368
- Prieto-Marquez A, Erickson GM, Seltmann K, Ronquist F, Riccardi GA, Maneva-Jakimoska C, Deans A et al (2007) Morphbank, an avenue to document and disseminate anatomical data: phylogenetic and paleohistological test cases. *J Morph* 268:1120–1120
- Prudent X, Parra G, Schwede P, Roscito JG, Hiller M (2016) Controlling for phylogenetic relatedness and evolutionary rates improves the discovery of associations between species’ phenotypic and genomic differences. *Molec Biol Evol* 33(8):2135–2150
- Pruvost M, Bellone R, Benecke N, Sandoval-Castellanos E, Cieslak M, Kuznetsova T, Morales-Muñiz A, O’Connor T, Reissmann M, Hofreiter M, Ludwig A (2011) Genotypes of predomestic horses match phenotypes painted in Paleolithic works of cave art. *Proc Natl Acad Sci USA* 108(46):18626–18630. <https://doi.org/10.1073/pnas.1108982108>
- Rapić-Otrin V, Navazza V, Nardo T, Botta E, McLenigan M, Bisi DC, Levine AS, Stefanini M (2003) True XP group Epitopes have a defective UV-damaged DNA binding protein complex and mutations in DDB2 which reveal the functional domains of its p48 product. *Hum Mol Genet* 12(13):1507–1522
- Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues MT, Hiller M (2018) Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution. *Nat Commun* 9:737. <https://doi.org/10.1038/s41467-018-0712>
- Rosenthal N, Brown S (2007) The mouse ascending: perspectives for human-disease models. *Nature Cell Biol* 9:993–999
- Ruzicka L, Bradford YM, Frazer K, Howe DG, Paddock H, Ramachandran S, Singer A, Toro S, Van Slyke CE, Eagle AE, Fashena D, Kalita P, Knight J, Mani P, Martin R, Moxon SA, Pich C, Schaper K, Shao X, Westerfield M (2015) ZFIN, the Zebrafish Model Organism Database: Updates and new directions. *Genesis* 53(8):498–509
- Schulz S, Jansen L (2013) Formal ontologies in biomedical knowledge representation. *IMIA Yearb Med Inform* 8(1):132–46
- Schulz S, Stenzhorn H, Boekers M, Smith B (2007) Strengths and limitations of formal ontologies in the biomedical domain. *Electron J Commun Inf Innov Health* 3(1):31–45
- Scriven CR (2004) After the genome—the phenome? *J Inherit Metab Dis* 27(3):305–317
- Sereno PC (2007) Logical basis for morphological characters in phylogenetics. *Cladistics* 23:565–587
- Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M (2018) A genomics approach reveals insights into the importance of gene losses for mammalian adaptations. *Nat Commun* 9:1215. <https://doi.org/10.1038/s41467-018-03667-1>
- Sharma V, Lehmann T, Stuckas H, Funke L, Hiller M (2018b) Loss of RXFP2 and INSL3 genes in Afrotheria shows that testicular descent is the ancestral condition in placental mammals. *PLoS Biology*. 16e2005293
- Smith B (2003) Ontology. In: Floridi L (ed) *Blackwell guide to the philosophy of computing and information*. Blackwell Publishing, Oxford, pp 155–166
- Smith CL, Goldsmith CAW, Eppig JT (2005) The Mammalian Phenotype Ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol* 6(1):1–9
- Snodgrass RE (1951) Anatomy and morphology. *J New York Entomol S* 59(2):71–73
- Solé F, Ladevèze S (2017) Evolution of the hypercarnivorous dentition in mammals (Metatheria, Eutheria) and its bearing on the development of tribosphenic molars. *Ev Dev* 19(2):56–68
- Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Bat1K Consortium (2018) Bat biology, genomes, and the Bat1K project: to generate chromosome-level genomes for all living bat species. *Annu Rev Anim Biosci* 6:23–46
- Thenius E. (1989) Zähne und Gebiss der Säugetiere. *Handbuch der Zoologie*. volume 8, Mammalia, part 56, Walter de Gruyter, Berlin
- Thesleff I, Keranen S, Jernvall J (2001) Enamel knots as signaling centers linking tooth morphogenesis and odontoblast differentiation. *Advances Dent Res* 5(1):14–18
- Thier N, Stefen C (2020) Morphological and radiographic studies on the skull of the straw-coloured fruit-bat *Eidolon helvum* (Chiroptera: Pteropodidae). *Vertebrate Zoology*. 70(4). <https://doi.org/10.26049/VZ70-4-2020-05>
- Ungar PS (2010) Mammal teeth: origin, evolution, and diversity. JHU Press.
- Vaughan TA, Ryan JM, Czaplewski NJ (2015) Chapter 4: Classification of Mammals. *Mammalogy* (Sixth ed.). http://samples.jbpub.com/9781284032093/9781284032093_CH04_Secure.pdf
- Vogt L (2017) Assessing similarity: on homology, characters and the need for a semantic approach to non-evolutionary comparative homology. *Cladistics* 33(5):513–539. <https://doi.org/10.1111/clad.12179>
- Vogt L (2018) The logical basis for coding ontologically dependent characters. *Cladistics* 34(4):438–458
- Vogt L, Baum R (2019) Using named graphs and knowledge graph template patterns for efficiently organizing FAIR anatomy data and metadata. *Biodiv Info Sci Standards*. <https://doi.org/10.3897/biss.3.37205>
- Vogt L, Bartolomeaus T, Giribet G (2010) The linguistic problem of morphology: structure versus homology and the standardization of morphological data. *Cladistics* 26:301–325
- Vogt L, Baum R, Bhatti P, Köhler C, Meid S, Quast B et al (2019) SOCCOMAS: a FAIR web content management system that uses *knowledge* graphs and that is based on semantic programming. *Database* 2019(baz067):1–22
- Vogt L (2019) Organizing phenotypic data—a semantic data model for anatomy. *J Biomed Semant*. 10 (1). <https://doi.org/10.1186/s13326-019-0204-6>
- Wagner F, Peters B, Giere P, Grobe P, Hofmann R, Jähde M, Lächele U, Lehmann T, Ortman S, Ruf I, Schiffmann C, Stefen C, Stuckas H, Thier N, Unterhitzberger G, Vogt L (2021) How to use *Mammalian Traits for Comparative Genomics (MaTrics)* Design Principles of a project trait matrix in Morph•D•Base. <https://doi.org/10.20363/mbd.ref-5293>
- Wagner F, Ruf I, Hofmann R, Lehmann T, Ortman S, Schiffmann C, Hiller M, Stefen C, Stukas H (in revision) Reconstruction of evolutionary changes in fat and toxin consumption reveals associations with gene losses in mammals: a case study for the lipase inhibitor PNLIPRP1 and the xenobiotic receptor NR1I3. *JEB*
- Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE (2009) Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol* 7(11):e1000247

- Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562. <https://doi.org/10.1038/nature01262>
- Wilson DE, Reeder DM (2005) *Mammal species of the world: a taxonomic and geographic reference*, 3rd edn. John Hopkins University Press, Baltimore
- Xiang Z, Mungall C, Rutenber A, He Y (2011) *Ontobee: A Linked Data Server and Browser for Ontology Terms*. Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO), July 28–30, 2011, Buffalo, NY, USA. pp 279–281. <http://ceur-ws.org/Vol-833/paper48.pdf>
- Zhang Z, Nikaido M (2020) Inactivation of ancVIR as a predictive signature for the loss of vomeronasal system in mammals. *Genome Biol Evol* 12(6):766–778
- Zoonomia Consortium: Genereux DP, Serres A, Armstrong J, Johnson J, Marinescu V, Murén E, et al, Damas J (2020) A comparative genomics multitool for scientific discovery and conservation. *Nature* 587(7833):240–245. <https://www.nature.com/articles/s41586-020-2876-6>

Authors and Affiliations

Clara Stefen¹  · Franziska Wagner¹  · Marika Asztalos¹  · Peter Giere²  · Peter Grobe³ · Michael Hiller^{4,5,6,7,8,9}  ·
 Rebecca Hofmann^{8,10}  · Maria Jähde¹ · Ulla Lächele^{2,11} · Thomas Lehmann⁸  · Sylvia Ortman¹²  ·
 Benjamin Peters¹  · Irina Ruf^{8,10}  · Christian Schiffmann¹² · Nadja Thier¹ · Gabriele Unterhitzberger¹² ·
 Lars Vogt¹³  · Matthias Rudolf¹⁴ · Peggy Wehner¹⁴ · Heiko Stuckas¹ 

¹ Senckenberg Naturhistorische Sammlungen Dresden, Königsbrücker Landstraße 159, 01109 Dresden, Germany

² Museum für Naturkunde, Leibniz Institute for Evolution and Biodiversity Science, Invalidenstr. 43, 10115 Berlin, Germany

³ Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, 53113 Bonn, Germany

⁴ Max Planck Institute of Molecular Cell Biology and Genetics, Pfötenhauerstr. 108, 01307 Dresden, Germany

⁵ Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38, 01187 Dresden, Germany

⁶ Center for Systems Biology Dresden, Pfötenhauerstr. 108, 01307 Dresden, Germany

⁷ LOEWE Center for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

⁸ Abteilung Messelforschung und Mammalogie, Senckenberg Forschungsinstitut und Naturmuseum, Senckenberganlage 25, 60325 Frankfurt am Main, Germany

⁹ Faculty of Biosciences, Goethe-University, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany

¹⁰ Institut für Geowissenschaften, Goethe-Universität, Altenhöferallee 1, 60438 Frankfurt am Main, Germany

¹¹ Institut für Biologie, Humboldt-Universität zu Berlin, Invalidenstr. 42, 10115 Berlin, Germany

¹² Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowalke-Straße 17, 10315 Berlin, Germany

¹³ TIB Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167 Hannover, Germany

¹⁴ TU Dresden, Institut für Allgemeine Psychologie, Biopsychologie und Methoden der Psychologie, Raum BZW A317, 01062 Dresden, Germany