

1

2 **Phenotyping in the era of genomics: *MaTrics* – a digital character matrix to document**
3 **mammalian phenotypic traits coded numerically**

4

5

6 Clara Stefen¹, Franziska Wagner¹, Marika Aształos¹, Peter Giere², Peter Grobe³, Michael
7 Hiller^{4,5,6,7,8,9}, Rebecca Hofmann⁸, Maria Jähde¹, Ulla Lächele², Thomas Lehmann⁸, Sylvia
8 Ortmann¹⁰, Benjamin Peters¹, Irina Ruf⁸, Christian Schiffmann¹⁰, Nadja Thier¹, Gabi
9 Unterhitzenberger¹⁰, Lars Vogt¹¹, Matthias Rudolf¹², Peggy Wehner¹², Heiko Stuckas¹

10

11

12

13 ¹ Senckenberg Naturhistorische Sammlungen Dresden, Königsbrücker Landstraße 159, 01109
14 Dresden, Germany, clara.stefen@senckenberg.de; heiko.stuckas@senckenberg.de

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

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

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

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

23 ⁶ Center for Systems Biology Dresden, Pfotenhauerstr. 108, 01307 Dresden, Germany

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

26 ⁸ Senckenberg Forschungsinstitut und Naturmuseum Frankfurt, Senckenberganlage 25, 60325
27 Frankfurt, Germany

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

30 ¹⁰ Leibniz Institute for Zoo and Wildlife Research, Alfred-Kowale-Straße 17, 10315 Berlin,
31 Germany

32 ¹¹ TIB Leibniz Information Centre for Science and Technology, Welfengarten 1B, 30167
33 Hannover, Germany

34 ¹² TU Dresden, Institut für Allgemeine Psychologie, Biopsychologie und Methoden der
35 Psychologie, Raum BZW A317, 01062 Dresden

36

37

38

39 Corresponding authors: clara.stefen@senckenberg.de; heiko.stuckas@senckenberg.de

40

41

42

43 ORCID IDs

44 Michael Hiller: 0000-0003-3024-1449

45 Sylvia Ortmann 0000-0003-2520-6251

46 Irina Ruf: 0000-0002-9728-1210

47 Clara Stefen: 0000-0001-79986-110X

48 Heiko Stuckas 0000-0002-5690-0994

49 Lars Vogt: 0000-0002-8280-0487

50 Franziska Wagner 0000-0001-6623-6700

51 Benjamin Peters 0000-0002-2737-7006

52

53

54 **Abstract**

55 A new and uniquely structured matrix of mammalian phenotypes, *MaTrics* (*Mammalian*
56 *Traits for Comparative Genomics*) is presented in a digital form. By focussing on mammalian
57 species for which genome assemblies are available, *MaTrics* provides an interface between
58 mammalogy and comparative genomics.
59 *MaTrics* was developed as part of a project to link phenotypic differences between mammals
60 to differences in their genomes using *Forward Genomics*. Apart from genomes this approach
61 requires information on homologous phenotypes that are numerically encoded (presence-
62 absence; multistate character coding*) in a matrix. *MaTrics* provides these data, links them to
63 at least one reference (e.g., literature, photographs, histological sections, CT-scans, or
64 museum specimens) and makes them available in a machine actionable NEXUS-format. By
65 making the data computer readable, *MaTrics* opens a new way for digitizing collections.
66 Currently, *MaTrics* covers 147 mammalian species and includes 207 characters referring to
67 structure, morphology, physiology, ecology and ethology. Researching these traits revealed
68 substantial knowledge gaps, highlighting the need for substantial phenotyping efforts in the
69 genomic era. Using the trait information documented in *MaTrics*, previous Forward Genomics
70 screens identified changes in genes that are associated with various phenotypes, ranging from
71 fully-aquatic lifestyle to dietary specializations. These results motivate the continuous
72 expansion of phenotype information, both by filling research gaps or by adding additional
73 taxa and traits. *MaTrics* is digitally available online within the data repository Morph·D·Base
74 (www.morphdbase.de).

75
76
77
78
79

80 **Key words** hard tissue, visceral & life history traits, museum specimens, character states,
81 numerical coding

82

83 Expressions indicated with an * are explained in the attached glossary

84

85

86 **Funding** As part of the interdisciplinary research project ‘Identifying genomic loci underlying
87 mammalian phenotypic variability using *Forward Genomics*’ by the Leibnitz Association,
88 grant SAW-2016-SGN-2.

89 **Conflicts of interest/Competing interests** none known for all authors

90 **Ethics approval** Not applicable as no experiments with living animals have been performed
91 **Consent to participate** all authors agreed to the work
92 **Consent for publication** all authors agree to the publication
93 **Availability of data and material:** data are available within Morph·D·Base
94 (www.morphdbase.de)
95 **Code availability** (software application or custom code): Not applicable
96 **Authors' contributions** (optional: please review the submission guidelines from the journal
97 whether statements are mandatory), Cs, HS, MH had the idea for the manuscript, CSt, FW,
98 HS drafted and finalized the manuscript; IR, PGi, MH, Rh, UL, SO, TL, NT commented and
99 participated in writing the manuscript; MA, CSt did the study on cusps in teeth of Carnivora;
100 BP, CSch, CSt, FW, GU, IR, MA, MJ, NT, RH, SO, TL, UL were involved in coding
101 characters in *MaTrics*; PGr, LV provided Morph·D·Base, implemented tools therein and
102 technically finalized *MaTrics*; FW got statistics for *MaTrics*, MR, PW did the statistical
103 calculations on characters

104

105

106

107

108 **Introduction**

109 *Background*

110 Knowing and understanding the organisms around us has always been important for mankind
111 and thus describing and comparing phenotypes has a long tradition that goes beyond the
112 emergence of academic disciplines (e.g., Pruvost et al. 2011). The phenotype of an organism
113 refers to its observable constituents, properties, and relations that can be considered to result
114 from the interaction of the organism's genotype with itself and its environment and include
115 the anatomical organization of an organism, its physical properties, behaviour, ecological
116 features, and lifestyle traits. They characterize an organism and contribute to biodiversity.
117 Morphological* and anatomical* data usually make up the largest part of the phenotypic data
118 available for a species. In mammalogy, specific skeletal, dental as well as body plan, visceral
119 and physiological traits are traditionally used for differentiating species and for describing
120 their inter- and intraspecific variability. Depending on preservation, this can also be applied to
121 fossil species.

122

123 Advances in molecular biology and genetics over the last decades identified many
124 genes and molecular mechanisms that are required for the development of many traits. This
125 work relied primarily on studying model organisms such as the fruit fly (*Drosophila*
126 *melanogaster*), the zebra fish (*Danio rerio*) or the mouse (*Mus musculus*). These models
127 provided decisive insights into the genes behind basic developmental processes, including
128 organ function and morphogenesis (Meunier 2012). Comparing developmental processes
129 from model to non-model organisms opened the field for evolutionary developmental biology
130 and explained the molecular basis of processes such as body plan evolution. However, there
131 are some limitations on what model organisms can tell (Bolker 2012). For instance, insights
132 from experiments on model organisms are restricted to the phenotypes present in that
133 particular species. For example, rodents such as mice do not have canine teeth, making mouse
134 an inappropriate model to study the molecular mechanisms required to make canine teeth.
135 Furthermore, even if model organism research would reveal all genes that are associated with
136 a given phenotype (e.g., the digestive system), it remains unknown which of these genes were
137 altered during evolution and contributed to phenotypic changes between related species (e.g.,
138 mammals that specialized on particular diets).

139

140 With the development of sequencing technologies, sequencing and assembly of whole
141 genomes became possible; the first was published in 1995 (of the bacteria *Haemophilus*
142 *influenzae*, Fleischmann et al. 1995) and the mouse genome “only” was published in 2002
143 (Waterston et al. 2002). Due to advancements in high-throughput DNA sequencing, there are
144 an increasing number of species for which sequenced nuclear genomes are available (e.g.,
145 Genome 10K Community of Scientists 2009; Teeling et al. 2018; Feng et al. 2020; Zoonomia
146 Consortium 2020). This wealth of genomes provides a basis for comparative genomics*
147 (“defined as the comparison of biological information derived from whole-genome
148 sequences” and as discipline / methodology thus only started in 1995 (de Crécy-Lagard and
149 Hanson 2018)). While comparative genomics often aims at identifying genomic elements that
150 are conserved across species and thus likely have an evolutionarily important function
151 (Nobrega and Pennacchio, 2004), comparative genomics can also be used to detect
152 differences in functional genomic elements and associate them with phenotypic differences of
153 species. For example, targeted analyses of genes associated with the formation of dentin
154 (DSPP) and enamel (AMTN, AMBN, ENAM, AMELX, MMP20) across Mammalia and
155 Sauropsida (including Aves, Crocodylia, Testudines, Squamata) showed an association
156 between the loss of these genes and the loss of teeth (Meredith et al. 2009, 2014). Another

157 example are losses of chitinase genes (CHIAs), enzymes that digest chitin, which
158 preferentially occurred in mammalian species that have non-insectivorous diets (Emerling
159 2018).

160

161 Recent advances in comparative genomics follow the idea that convergent phenotypic
162 evolution can be associated with convergent genomic changes e.g., gene loss (Lamichhaney et
163 al., 2019). This assumption is one conceptual foundation of the general *Forward Genomics*
164 approach that performs an unbiased screen for genomic changes that are associated with
165 convergent phenotypic traits (Hiller et al. 2012; Prudent et al. 2016). This approach employs
166 phenotype matrices and genome alignments to search for associations between convergent
167 phenotypic traits and genomic signatures. These genomics signatures (e.g., candidate genes)
168 need to be subjected to functional analyses to explain their association with the phenotype of
169 interest. *Forward Genomics* identified many new links between genomic changes in genes as
170 well as regulatory elements and various phenotypic changes such as adaptations to fully-
171 aquatic lifestyles in cetaceans and manatees (Sharma et al. 2018a), echolocation in bats and
172 toothed whales (Lee et al., 2018), reductions and losses of the mammalian vomeronasal
173 system (Hecker et al., 2019a), the evolution of body armour in pangolins and armadillos
174 (Sharma et al. 2018a), the absence of testicular descent (Sharma et al. 2018b), and the
175 reduction of eye sight in subterranean mammals (Roscito et al., 2018; Langer and Hiller
176 2018).

177

178 Development of Matrics

179 A key requirement Forward Genomics is comprehensive and digitally-available phenotypic
180 knowledge of species considered in the genomic screen. However, in contrast to genomic
181 data, phenotypic data are not readily available in such a digitized form that it can be used by
182 computer programs, not even for well-characterized species such as mammals with sequenced
183 genomes. Research in Zoology and related fields assembled a rich body of phenotypic
184 knowledge. But the information assembled over centuries is usually documented using natural
185 language and thus in the form of texts unstructured for computer-programs and so the
186 information is not machine-actionable* (Vogt et al. 2010). Whereas this is what researchers in
187 Zoology and related fields used, and still use effectively, it limits research in other disciplines
188 where substantial time investments are required to search and extract relevant phenotypic data
189 from published descriptions. As a result, this important cultural and scientific heritage is
190 underutilized in genomics and some disciplines of biomedical research.

191

192 Here, we address the need for digitally-available trait information by creating a phenotypic
193 character matrix. Since the genome “encodes” all traits that have a genetic basis, genomes of
194 many related species (such as mammals) enable Forward Genomic screens for many different
195 traits with convergent changes. Thus, comprehensive information of many traits can be stored
196 in a matrix form, where rows represent species and columns represent traits.

197

198 Constructing a comprehensive phenotypic matrix poses several challenges. While “simple”
199 phenotypes that can be compiled relatively easily across several mammals, more complex
200 phenotypes require experienced researchers in morphology, anatomy, physiology, veterinary
201 science or related fields, since interpreting the collected information on phenotypes requires
202 specialized knowledge on the terminology and taxon of interest. For example, the exact
203 meaning of specialized terms might depend on the described taxon, the author, and the time of
204 publication. Additionally, some terms might refer to spatio-structural properties, others to
205 common function or presumed common evolutionary origin, or to a mixture of both. All this
206 is well understandable to the experts, but, difficult for non-experts and even more so for
207 computer algorithms. Thus, integrating the information on phenotypes in machine actionable
208 form with other sources of data becomes exceedingly challenging and time-consuming
209 (Lamichhaney et al. 2019; Vogt 2019). For integrative research a way is sought to exploit that
210 knowledge without involving experts in each project.

211

212 To enable simpler use (and exploitation) of expert knowledge, more and more
213 information is being digitized, stored, and made accessible online such as current journals or
214 even older and classic books (Biodiversity Heritage Library). There are several databases for
215 storing, editing and publishing information on phenotypes (mainly on morphological ones)
216 covering various taxa (Table 1). All of them have their own purpose and relevance, but none
217 of them so far fulfills the requirements of *Forward Genomics* and other efforts to link
218 phenotypic to genomic differences; most neither provide information on the same character
219 across the selected species nor is the information numerically coded to be directly useful to
220 other computer analysing programs. However, tables on morphological traits for phylogenetic
221 analyses fulfil these requirements, but do not focus on extant mammals with sequenced
222 genomes. Furthermore, while inference of homologies in genomic data (nucleotide sequence
223 alignment) is fully automated, homology analyses (character alignment) of phenotypic data
224 cannot be executed by computer algorithms so far. This is irrespective of the type of basic

225 information available (digitized literature, 2D/3D scans of museum specimens). Matrices
226 usable to link phenotypic differences between species to genomic loci ~~first~~ need to provide
227 homology information.

228

229 To enable the full use of *Forward Genomics*, a trait matrix of mammalian phenotypes
230 was developed that fulfils the above-mentioned criteria. Here, we introduce *MaTrics*
231 (*Mammalian Traits for Comparative genomics*), the first and newly established matrix
232 providing information on phenotypic traits of mammals.

233

234

235 **Design and coding* principles of *MaTrics***

236 *MaTrics*, version 1.0, release January 2021 (in the following still referred to as *MaTrics* only)
237 is implemented in the online data repository* Morph-D-Base (www.morphdbase.de, Grobe
238 and Vogt 2009) and publicly available.

239

240 Principles and data entry

241 *MaTrics* meets all requirements of *Forward Genomics*. We primarily focused on mammalian
242 species for which genome sequences are available. Some basic principles of *MaTrics* are
243 described herein, a detailed user's guide is available online (Wagner et al. 2020). Different
244 types of phenotypic traits were considered (see below) and in each case homology was
245 assured.

246

247 In case a phenotypic trait has several different expressions, it must be coded as a
248 multistate character. According Sereno (2007), the *character* part in a multistate character
249 comprises not only the *locator* but additionally a *variable* (V – the aspect that varies) and a
250 *variable qualifier* (q – the variable modifier). The *character states* of a multistate character in
251 *MaTrics* are numerically coded by 2 to n (Fig. 1B, Table 2). For example, the height of the
252 mandibular canine teeth in relation to the level of the occlusal height (averaged) of the cheek
253 teeth are coded as *short* (2), *occlusal height* (3) or *long* (4) (Fig. 1B).

254

255 According to Sereno (2007), a (phenotypic) trait of an operational taxonomic unit
256 (OUT; here the selected mammalian species) can be represented in a *character statement* that
257 is composed of two parts: *character* and *statement*, and can be divided into four types of
258 logical components (Sereno 2007:Table 4): one or more *locators*, a *variable*, and a *variable*

259 *qualifier* as parts of the character and a *character state* as the *statement*. Not all these
260 components are needed in any case, but a *locator* and a *character state* are the minimum
261 (representing *character* and *statement*). Thus, each *character* consists of at least one *locator*
262 (L – the morphological structure, the structure bearing the trait) and the *statement* of the
263 *character state* (v – mutually exclusive condition of a character) (Fig. 1). Specifying a *locator*
264 and a *character state* is sufficient in case of absent-present *character statements**
265 (numerically coded by 0/1; Fig. 1A, Table 2). Following Sereno's (2007) coding scheme, each
266 character in *MaTrics* is named with a label starting with a single *locator* or a sequence of
267 *locators* starting with L_n to L₁ (the trait bearing structure), which provide all information
268 necessary for unambiguously identifying and locating the trait within the OTU. The sequence
269 of locators (L_n to L₁ as illustrated in Fig. 1) in the character label is hierarchically organized.
270 While Sereno (2007) developed his coding scheme primarily for structural traits, we extended
271 it here and applied it also to ecological or behavioural traits.

272

273 In case a phenotypic trait has several different expressions or patterns, it must be
274 coded as a multistate character. According Sereno (2007), the *character* part in a multistate
275 character comprises not only the *locator* but additionally a *variable* (V – the aspect that
276 varies) and a *variable qualifier* (q – the variable qualifier). The *character states* of a
277 multistate character in *MaTrics* are numerically coded by 2 to n (Fig. 1B, Table 2). For
278 example, the height of the mandibular canine teeth in relation to the level of the occlusal
279 height (averaged) of the cheek teeth are coded as *short* (2), *occlusal height* (3) or *long* (4)
280 (Fig. 1B).

281

282 A key consideration when generating *MaTrics* was to clearly document the source(s)
283 for each phenotypic entry. In *MaTrics*, the *character* part of each *character statement*
284 therefore possesses a short textual definition that is taken from published sources (journals,
285 text books, online references) and includes references to relevant ontology terms from various
286 biomedical ontologies (the following online resources were used for identifying adequate
287 terms: Ontology Lookup Service, OLS, <https://www.ebi.ac.uk/ols/index>, Jupp et al. (2015);
288 Ontobee, <https://www.ontobee.org>, Xiang et al. (2011); Bioportal,
289 <https://bioportal.bioontology.org>, Musen et al. (2012)). If no adequate definition was
290 available, we provided our own definitions and clearly marked them as such.

291

292 The dimensions of *MaTrics* are defined by the number of rows (OTUs) and columns
293 (characters) that result in a specific number of cells (rows x columns). These cells primarily
294 contain the character states. Morph-D-Base enables the addition of further information such as
295 references, photos, illustrations, or museum specimen IDs to each matrix cell. All character
296 states recorded, thus each cell of *MaTrics* is linked to at least one supporting reference. This
297 refers either to citations from the literature (e.g., published journal articles, books, reliable
298 scientific online resources) or to primary data sources. These data sources can cover IDs of
299 museum specimens or direct links to media (e.g., photographs; microscopic and electron
300 microscopic (TEM and SEM) images, magnetic resonance (MRI), computed tomography (,
301 μ CT), or even synchrotron data) which are directly uploaded in MDB. As a result, researchers
302 using *MaTrics* can trace the information to at least one original source.

303 Phenotypic traits coded in *MaTrics* represent by default adult states. Fetal structures or
304 traits that belong to perinatal or not yet fully-grown stages are explicitly indicated as “fetal”
305 (fetal is used as *locator L_n* in the character label). Traits referring to other ontogenetic stages
306 can be coded in a similar way.

307

308 Phenotypic traits included in *MaTrics* represent by default adult stages. Fetal
309 structures, or traits that belong to perinatal or not yet fully-grown stages are explicitly
310 indicated by placing “fetal” in front of the *locator L_n* in the character label. Traits referring to
311 other ontogenetic stages could be considered in a similar way.

312

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

315

316

317 Specificities of MaTrics

318 The primary motivation to generating *MaTrics* was to create a research tool for linking
319 phenotypic differences between species to differences in their genomes. This is the main
320 reason why intraspecific variations of traits such as sexual dimorphism were not considered.
321 Another specificity is that *character states* (presence/absence; multistate) do not encode
322 character polarity. Researchers can decide for each project individually whether to use and
323 determine polarity or not. The characters might be further analysed (e.g., polarity analyses
324 using out-group comparison) if considered for phylogenetic studies or gene loss analyses.
325 Finally, character dependencies were not specifically accounted for during the choice and

326 coding of traits. For each research question, specific characters of interest were added to
327 *MaTrics*. Similarly, for different projects, characters can be selected individually to be
328 retrieved from *MaTrics* for other use. Character dependencies can be avoided or reduced in
329 this way, if needed.

330

331

332 **Current status: *MaTrics***

333 To date, *MaTrics* contains 207 characters for 147 mammalian species to date, resulting in a
334 total of 30,282 documented character states. 153 of the 207 characters (74%) are described as
335 absent-present characters and the remaining 53 (26%) are multistate characters. The
336 mammalian species considered in *MaTrics* include two representatives of Monotremata, five
337 of Marsupialia and 140 of placental mammals (supplementary material Table S1). The
338 number of species from each order neither represents the respective diversity nor the
339 morphological disparity of mammalian orders, as the primary criterion for the inclusion in
340 *MaTrics* was the availability and suitable quality of whole genomes. The characters in
341 *MaTrics* cover structural, ecological, ethological, and physiological phenotypic traits (Table
342 3). All, but one character (*organum vomeronasale*), refer to the adult stage. For one character
343 (*os jugale*), the recording is 100%, so all cells contain coded and referenced character states.
344 Some traits were specifically included for the study in subsets of the listed mammals, and
345 therefore the recording purposely is less complete (for coding status see Table S2).

346

347

348 **Notes on application**

349 The primary motivation for creating *MaTrics* was to provide fully referenced phenotypic
350 information for applications in comparative genomics, especially the *Forward Genomics*
351 approach. The creation and filling of *MaTrics* and studies applying *Forward Genomics* were
352 developed in parallel within the mentioned project. So, phenotypes were coded in *MaTrics*
353 were partially successfully used in earlier studies and simpler shorter tables e.g. by Sharma et
354 al. (2018a) who identified various convergent gene losses associated with some specific
355 convergent mammalian phenotypes. They showed convincingly that tooth and enamel loss are
356 associated with the loss of ACP4 (a gene that is associated with the enamel disorder
357 amelogenesis imperfecta), and that the presence of scales is associated with the loss of the
358 gene DDB2 (which detects substances resulting from UV-light and helps to induce DNA
359 repair). The fully aquatic lifestyle is associated with the loss of MMP12, a gene associated

360 with breathing adaptation. The documented loss of these genes in some mammalian species is
361 functionally explainable either as a consequence of trait loss (the genes ACP4 and DDB2
362 have no function after trait loss) or as putative adaptive genomic alteration, causing novel
363 phenotypes (MMP12-loss is associated with novel lung functions in aquatic mammals)
364 (Sharma et al. 2018a). Such results might help to better understand some related human
365 diseases, as for example in the case of DDB2 whose mutations cause xeroderma pigmentosum
366 which manifests in hypersensitivity to sunlight (Rapic-Otric et al. 2003).

367

368 Another study investigated the gene losses associated with the reduction of the
369 vomeronasal system (VNS) in several mammals. A genomic comparison of 115 mammalian
370 genomes confirmed that *Trpc2* is an indicator for the functionality of the VNS (Hecker et al.
371 2019a). Moreover, it indicated a loss of functionality of the VNS in seals (Phocidae) and
372 otters (Lutrinae). Morphological data is scarce for seals and there is no data for otters (Hecker
373 et al. 2019a; Zhang and Nikaido 2020). A study to test the accuracy of the suggested
374 predictability is under way. This study is an example for testing genotype-phenotype
375 associations in non-model organisms and shows the potential of the combination of
376 comparative morphological and genomic approaches.

377

378 However, the relevance of *MaTrics* is by no means restricted to the *Forward*
379 *Genomics* approach. Characters were also included in *MaTric* for the usage in the
380 contemporary study to explore evolutionary conditions associated with the loss of genes
381 related to convergent evolution of herbivorous and carnivorous diet in mammals (Hecker et
382 al. 2019b). This study included 52 placental species and suggests that the lipase inhibitor gene
383 PNZIPRP1 is preferentially lost in herbivores, whereas the xenobiotic receptor NR1I3 is
384 preferably lost in carnivores. Even though the authors put forward hypotheses, the lack of
385 accessible data on mammalian diet preferences made it difficult to test whether gene losses
386 are associated with dietary fat content and diet-related toxins. Investigating whether
387 convergent gene loss is associated with similar dietary preferences may additionally hold
388 information on whether gene losses might be adaptive (Albalat and Cañestro 2016).
389 Consequently, an ongoing study records dietary categories in *MaTrics* that allow a semi-
390 quantitative of dietary fat content (associated with PNZIPRP1) and diet-related toxins
391 (associated with NR1I3) (Wagner et al. ####). This study provided evidence that the
392 convergent loss of both genes is associated with the convergent evolutionary change of
393 dietary preferences, i.e. the consumption of a diet with reduced fat and toxin contents. The

394 hypotheses of Hecker et al. could be refined and also the evolutionary setting could be
395 reconstructed.

396

397 Future analyses using *MaTrics* have the potential to test how gene losses and dietary
398 composition are related to the presence/absence of structures or organs associated with
399 digestive processes. Even further, it allows investigating whether evolutionary changes in diet
400 composition are not only associated with the loss/presence of single molecules (e.g., lipase
401 inhibitor, xenobiotic receptor), but also with changes in complex structures and their
402 associated gene. For instance, it is interesting to note that first statistical investigations
403 (methods given in document S3) have not yet proven a significant association between the
404 presence of a gall bladder and the diet ($p=0.74$) as well as the lipase inhibitor gene PNLRP1
405 ($p=0.49$). This observation motivates the further development of *MaTrics*, i.e. by adding
406 further traits and species.

407

408 These two studies show how genomic and morphological studies are entangled:
409 current knowledge of morphology serves as basis for creating phenotypic trait matrices like
410 *MaTrics* which – on the other hand – forms the basis of genomic research, especially the
411 *Forward Genomics* approach. Hypotheses associated with findings of candidate loci, may in
412 turn inspire further morphological research.

413

414 The most obvious application are morphological studies. Although mammal dentitions
415 are well studied and a lot is known about teeth number, form, and shape in particular in
416 relation to dietary specialization (see Thenius 1989; Hillson 2005; Ungar 2010), we still have
417 many gaps of knowledge, e.g., concerning functional adaptations and evolutionary
418 transformations. Thus, Sole and Ladevèze (2017) aimed to put forward new ideas on how the
419 basic mammalian tribosphenic molar was transformed to sectorial teeth in hypercarnivorous
420 mammals. They (Sole and Ladevèze 2017) included only carnivores as defined by flesh-
421 eating and the presence of carnassial teeth, representatives of the living Carnivoramrpha
422 (including the extinct Nimravidae) and Dasyuromorphia, as well as from the extinct
423 Sparassoodonta, Oxyaenodonta, and Hyaenodontida in their study. Comparing the cusp
424 pattern/morphology of the upper and lower molars of these species Sole and Ladevèze
425 (2017:fig. 4) derived a scheme for the morphological evolution of the sectorial teeth in
426 hypercarnivorous mammals. They also aimed at providing new arguments to discuss the
427 developmental aspects of the evolution of hypercarnivory by associating their morphological

428 observations with ontogenetic studies. The latter highlighted the importance of the expression
429 of ectodysplasin A (Eda): increased levels are able to modify the number, shape, and position
430 of cusps in mice during tooth development (Kangas et al. 2004). Further, Häärä et al.
431 (2012:3189) showed – again in mice – that “Fgf20 is a major downstream effector of Eda and
432 affects Eda-regulated characteristics of tooth morphogenesis, including the number, size, and
433 shape of teeth. Fgf20 function is compensated for by other Fgfs”. Inspired by the observations
434 and the model of Solé and Ladevèze (2017), we started a study with a subsample of Carnivora
435 (Table S3) collected in *MaTrics* with two aims: firstly, to test the suitability of *MaTrics* in
436 comparative morphological studies and, secondly to set the basis to proceed with genome
437 wide searches for genomic causes correlated with the loss of cusps. This seems to be
438 promising with the development of new methods to include searches for regulatory elements
439 (see below).

440

441 For the selected Carnivora (Table S4) the absence and presence of individual tooth
442 cusps for the fourth upper premolar (P^4) and all molar teeth were recorded in *MaTrics*. The
443 nomenclature of the cusps followed Thenius (1989, exemplified in Fig. 2). The detailed
444 descriptions of cusp patterns for the species are given in the supplementary document S5 and
445 examples are illustrated in Fig. 3 and detailed in Table S6. Some of our results confirmed the
446 observations of Solé and Ladevèze (2017), who focused on carnivores as defined by the
447 presence of carnassials. We confirm that parastyle and protocone of the P^4 are generally
448 reduced in hypercarnivorous carnivorans. Interestingly, both structures are more reduced in
449 the Canidae and the polar bear (*Ursus maritimus*) than in the members of the Felidae and
450 Hyaenidae. Solé and Ladevèze (2017) reported that in the upper molars protocone, paraconule
451 and metaconule are reduced in hypercarnivorous mammals which is also in line with our
452 findings.

453 These structures are reduced in the Canidae, and totally absent in the Felidae and
454 Hyaenidae. Solé and Ladevèze (2017) also found, that metaconid and talonid are generally
455 lost in hypercarnivorous mammals, especially felid-like and hyaenid-like hypercarnivores.
456 Based on our study, we found that metaconid and talonid are completely reduced only in the
457 Felidae (except the cheetah, *Acinonyx jubatus*) and the spotted hyena (*Crocuta crocuta*). Like
458 in the Canidae and the striped hyena (*Hyaena hyaena*), both structures are also present in
459 *Ursus maritimus*. The specialized hypercarnivorous diet of several Feliformia lead to an
460 extreme reduction of the tribosphenic molar, whereas the Canidae and *Ursus maritimus* also
461 eat fruits and vegetables and therefore need crushing structures. The presence of protocone

462 and talonid seems to be necessary for an omnivorous diet (Solé and Ladevèze 2017), but
463 based on our study we can confirm that this is also true for herbivorous species (e.g., red
464 panda, *Ailurus fulgens*; giant panda, *Ailuropoda melanoleuca*).

465

466 Except for the Pacific walrus (*Odobenus rosmarus*) at least 10 specimens per species
467 were analysed (Table S3); and for several species, exceptions of the common pattern in the
468 presence of cusps were observed (Table 4). *MaTrics* was not designed to take intraspecific
469 variability into account, therefore only the most common cusp patterns for each species were
470 recorded. Deviations from the cusp patterns are present in several cusps in domestic dog,
471 brown bear (*Ursus arctos*) and for one cusp in the red fox (*Vulpes vulpes*). Such exceptions
472 are important as they might indicate evolutionary trends. However, variations within a species
473 cannot be reflected in *MaTrics* as maximally one character state is given for each character
474 representatively for a species here. Only in this way the (common) absence or presence of a
475 trait can be compared with the genome of again one representative of a species. Studies on
476 intraspecific variability of certain characters would need additional matrices with different
477 intentions.

478

479

480 **Conclusion and Future Perspectives**

481 Recent advances in molecular techniques lead to a rapid increase in the assembly and
482 publication of genomes from various organisms. However, knowledge of the genome
483 sequences is only a first step to understand the relationships between genomic changes, the
484 phenotype of an organisms and phenotypic differences between different organisms (Hardison
485 2003). The systematic description of phenotypic information in matrix form like in *MaTrics* is
486 necessary to understand the genome information and to deal with questions related to
487 evolutionary biology and biomedicine. This is not restricted to mammals as the coding
488 principles of *MaTrics*, which comply with the requirements of molecular research, can serve
489 as a template for matrices comprising trait knowledge of other vertebrate and non-vertebrate
490 groups. The establishment of trait matrices for various taxa could lead to a broad
491 documentation of phenotypes for applications in comparative genomics, and, hence, enable a
492 systematic exploration of genotype-phenotype associations.

493

494 However, trait collections such as *MaTrics* also revealed a tremendous research gap on
495 phenotypic data. In fact, filling *MaTrics* with information on different phenotypic traits across

496 mammals showed that detailed information on structural, physiological, or life history traits
497 was often not available for many species, even with intensive literature research. For example,
498 reductions of the vomeronasal system (VNS) are clearly documented in several mammals and
499 our previous genomic comparison of 115 mammalian genomes uncovered several genes
500 whose loss is associated with a reduced or non-functional VNS (Hecker et al. 2019a). This
501 genomic screen also revealed that seals (Phocidae) and otters (Lutrinae) have lost some of
502 these genes, indicating a reduced VNS. However, to the best of our knowledge, information
503 concerning the vomeronasal organ of Phocidae and Lutrinae is not available. Indeed, the
504 recording status in *MaTrics* for the character “vomeronasal organ” with the states
505 absent/present is only 37%. Another example of a character, that would be assumed to be
506 well-known, is the absence/presence of the gall bladder (“*Vesica bilaris*”), with a recording
507 status of 70%. In other words, the recording status of the characters in the *MaTrics*
508 demonstrate the lack of information on phenotypic traits in several species. These research
509 gaps can only be filled by specimen-based research (e.g. Thier and Stefen 2020). Although
510 individual studies are valuable scientific contributions, they may not suffice to close the
511 substantial research gaps in short time. The authors see the need for more basic zoological
512 research complementing the systematic exploration of the genomic basis of biodiversity, i.e.
513 research activities on biodiversity genomics could be assisted by research initiatives on
514 biodiversity phenomics (= systematically phenotyping animals in matrices like *MaTrics*).
515

516 Most of the genomic studies mentioned above identified protein coding genes
517 associated with complex body plan changes (e.g., aquatic and aerial lifestyle of cetaceans and
518 bats, respectively). However, evolutionary theory predicts that changes in cis-regulatory
519 genetic elements are probably more important for morphological changes than protein-coding
520 genes. For instance, Roscito et al. (2018) stated that the loss of morphological traits is (often)
521 associated with the decay of the cis-regulatory elements. Consequently, the Forward Genomic
522 approach has been further developed to include methodologies that can be successfully
523 associate phenotypes with the loss or presence of regulatory elements (e.g., Langer et al.
524 2018; Langer and Hiller 2019). In awareness of these developments, the phenotype matrix
525 presented here already provides a whole bunch of morphological characters that will be
526 subject to further exploration in the near future. Thus, the phenotypic information compiled in
527 *MaTrics* will be of increasing importance. This applies for instance to those referring to tooth
528 morphology and tooth cusps discussed above. In fact, tooth characters are known to be the
529 result of a complex signalling network involving timely graded activation and deactivation of

530 genes controlled by regulatory elements (e.g., Jernvall and Thesleff 2000; Thesleff et al.
531 2001).

532

533 A last aspect to be mentioned refers to the way how phenotypic information is
534 documented. So far, filling *MaTrics* with information is still mostly conducted by hand;
535 experienced scientists have to control the content and to check for homology. However, some
536 recent developments may open the door to the partial automation of this work. First, the
537 implementation of ontologies and semantic phenotypes in the platform Morph-D-Base. The
538 development of a respective semantic description module is already initiated (Vogt and Baum
539 2019; Vogt 2019). This is expected to allow the development of computer algorithms to mine
540 data on homologous structures to establish matrices more automatically (Vogt 2018).

541

542 *MaTrics* is a new and unique data collection of phenotypic traits of mammalian
543 species. By including homologous phenotypic traits across (an increasing number of) species,
544 *MaTrics* and similar matrices can serve as basis for a variety of research fields as illustrated
545 herein. The recorded phenotypic traits are well defined and fully referenced (*characters* as
546 well the *character state* for each species). Not only literature data are accepted for the latter,
547 but also references to specimens in collections, which contributes in a specific way to the
548 digitalization of collection material. *MaTrics* data are directly useful in genomic studies since
549 the *character states* are numerically coded and hence can be extracted as NEXUS file to be
550 machine-actionable. The scientific potential of digitized phenotype matrices is apparent and
551 motivates thinking about future development.

552

553

554

555 **Acknowledgement**

556 We want to thank members of the German Society of Mammalogy (DGS) for stimulating
557 discussions at the annual DGS meeting 2019 which were useful to shape the manuscript.
558 Also, the helpful comments of the reviewers (....) are thankfully acknowledged.

559 **Funding**

560 This work was funded by the interdisciplinary research project ‘Identifying genomic loci
561 underlying mammalian phenotypic variability’ by the Leibnitz Association, grant SAW-2016-
562 SGN-2. MH was supported by the Max Planck Society and the LOEWE-Centre for

563 Translational Biodiversity Genomics (TBG) funded by the Hessen State Ministry of Higher
564 Education, Research and the Arts (HMWK).

565

566

567 **References**

568 Albalat R, Cañestro C (2016) Evolution by gene loss. *Nature Rev Genet* 17:379-391

569 Asher RJ (2007) A web-database of mammalian morphology and a reanalysis of placental
570 phylogeny. *BMC Evol Biol* 7:108. <https://doi.org/10.1186/1471-2148-7-108>

571 Bolker J (2012) Model organisms: There's more to life than rats and flies. *Nature*
572 491(7422):31

573 De Crécy-Lagard V, Hanson AD (2018) Comparative Genomics. Reference Module in
574 Biomedical Sciences.

575 <https://www.sciencedirect.com/topics/neuroscience/comparative-genomics>

576 Edmunds RC, Su B, Balhoff JP, Dahdul WM, Lapp H, Lundberg JG, Vision TJ, Dunham RA,
577 Mabee PM, Westerfield M (2016) Phenoscape: Identifying candidate genes for
578 species-specific phenotypes. *Molec Biol Evol* 33:13–24. [doi:10.1093/molbev/msv223](https://doi.org/10.1093/molbev/msv223)

579 Emerling CA, Delsuc F, Nachman MW (2018) Chitinase genes (CHIAs) provide genomic
580 footprints of a post-Cretaceous dietary radiation in placental mammals. *Science*
581 Advances 4(5):eaar6478

582 Feng S, Stiller J, Deng Y, Armstrong J, et al Zhang G. (2020) Dense sampling of bird
583 diversity increases power of comparative genomics. *Nature* 587:252-257.
584 <https://doi.org/10.1038/s41586-020-2873-9>

585 Fleischmann R, Adams M, White O, Clayton R, Kirkness E, Kerlavage A, Bult C, Tomb J,
586 Dougherty B, Merrick J, et al. (1995) Whole-genome random sequencing and
587 assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496–512.
588 [doi:10.1126/science.7542800](https://doi.org/10.1126/science.7542800)

589 Genome 10K Community of Scientists (2009) Genome 10K: a proposal to obtain whole
590 genome sequence for 10 000 vertebrate species. *J Hered* 100.6: 659–674

591 Grobe P, Vogt L (2009) Documenting Morphology: Morph·D·Base. In: Wägele JW,
592 Bartolomaeus T (eds) Deep Metazoan Phylogeny: The Backbone of the Tree of Life –
593 New Insights from Analyses of Molecules, Morphology, and Theory of Data Analysis.
594 De Gruyter, Berlin, pp 475-503. <http://www.morphdbase.de>

595 Häärä O, Harjunmaa E, Lindfors PH, Huh SH, Fliniaux I, Åberg T, Jernvall J, Ornitz DM,
596 Mikkola ML, Thesleff I (2012) Ectodysplasin regulates activator-inhibitor balance in
597 murine tooth development through Fgf20 signaling. *Development* 139(17):3189–3199

598 Hardison RC (2003) Comparative genomics. *PLoS Biol*, 1(2):e58

599 Harrow JL, Steward CA, Frankish A, Gilbert JG, Gonzalez JM, Loveland JE, et al, Wilming
600 LG (2014) The vertebrate genome annotation browser 10 years on. *Nuc Acid Res*
601 42(D1):D771–D779

602 Hecker N, Lächele U, Stuckas H, Giere P, Hiller M (2019a) Convergent vomeronasal system
603 reduction in mammals coincides with convergent losses of calcium signalling and
604 odorant degrading genes. *Mol Ecol* 28(16):3656–3668

605 Hecker N, Sharma V, Hiller M (2019b) Convergent gene losses illuminate metabolic and
606 physiological changes in herbivores and carnivores. *Proc Natl Acad Sci USA*
607 116(8):3036–3041

608 Hiller M, Schaar BT, Indjeian VB, Kingsley DM, Hagey LR, Bejerano G (2012) A “*Forward*
609 *genomics*” approach links genotype to phenotype using independent phenotypic losses
610 among related species. *Cell reports* 2(4):817–823

611 Hillson S (2005) Teeth. Cambridge university press, Cambridge

612 Huelsmann M, Hecker N, Springer MS, Gatesy J, Sharma V, Hiller M (2019) Genes lost
613 during the transition from land to water in cetaceans highlight genomic changes
614 associated with aquatic adaptations. *Science advances* 5(9):eaaw6671

615 Jernvall J (2000). Linking development with generation of novelty in mammalian teeth. *Proc*
616 *Nat Acad Sci* 97(6):2641-2645

617 Jernvall J, Thesleff I (2000) Reiterative signalling and patterning during mammalian tooth
618 morphogenesis. *Mechanisms dev* 92(1):19–29

619 Jupp S, Burdett T, Leroy C, Parkinson HE (2015) A new Ontology Lookup Service at EMBL-
620 EBI. In: Malone J et al. (eds.) *Proceedings of SWAT4LS International Conference*
621 2015, pp 118–119

622 Kangas AT, Evans AR, Thesleff I, Jernvall J (2004) Nonindependence of mammalian dental
623 characters. *Nature* 432(7014):211-214

624 Lamichhaney S, Card DC, Grayson P, Tonini JF, Bravo GA, Nämpflin K, et al, Sackton TB
625 (2019) Integrating natural history collections and comparative genomics to study the
626 genetic architecture of convergent evolution. *Phil Trans Royal Soc B*
627 374(1777):20180248

628 Langer BE, Hiller M (2019) TFforge utilizes large-scale binding site divergence to identify
629 transcriptional regulators involved in phenotypic differences. *Nuc acids res* 47(4):e19-
630 e19

631 Langer BE, Roscito JG, Hiller M (2018) REforge associates transcription factor binding site
632 divergence in regulatory elements with phenotypic differences between species. *Mol*
633 *Biol Evol* 35(12):3027–3040

634 Lecocq T, Benard A, Pasquet A, Nahon S, Ducret A, Dupont-Marin K, Lang I, Thomas M
635 (2019) TOFF, a database of traits of fish to promote advances in fish aquaculture.
636 *Scientific Data* 6(1):1–5

637 Lee JH, Lewis KM, Moural TW, Kirilenko B, Bogdanova B, Prange G, Koessl M,
638 Huggenberger S, Kang C, Hiller M (2018) Molecular parallelism in fast-twitch muscle
639 proteins in echolocating mammals. *Science Adv* 4(9): eaat9660

640 Meredith RW, Gatesy J, Murphy WJ, Ryder OA, Springer MS (2009) Molecular decay of the
641 tooth gene enamelin (ENAM) mirrors the loss of enamel in the fossil record of
642 placental mammals. *PLoS Genet* 5(9):e1000634

643 Meredith RW, Gatesy J, Springer MS (2013) Molecular decay of enamel matrix protein genes
644 in turtles and other edentulous amniotes. *BMC evol biol* 13(1):20

645 Meunier R (2012) Stages in the development of a model organism as a platform for
646 mechanistic models in developmental biology: Zebrafish, 1970–2000. *Studies in*
647 *History and Philosophy of Science Part C: Studies in History and Philosophy of*
648 *Biological and Biomedical Sciences* 43:522–531

649 Musen MA, Noy NF, Shah NH, Whetzel PL, Chute CG, Story MA, Smith B, NCBO team
650 (2012) The National Center for Biomedical Ontology. *J Am Med Inform*
651 *Assoc* 19:190–5. Epub 2011

652 Nobrega MA, Pennacchio LA (2004) Comparative genomic analysis as a tool for biological
653 discovery. *J physiol* 554(1):31–39

654 O'Leary MA, Kaufman S (2011) MorphoBank: phylophenomics in the “cloud”. *Cladistics*
655 27:1–9

656 Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR (2012) What is needed for next-
657 generation ecological and evolutionary genomics?. *TREE* 27(12):673–678

658 Porter IH (1973) From gene to phene. *J Invest Dermatol* 60(6):360–368

659 Pruvost M, Bellone R, Benecke N, Sandoval-Castellanos E, Cieslak M, Kuznetsova T,
660 Morales-Muñiz A, O'Connor T, Reissmann M, Hofreiter M, Ludwig A (2011)
661 Genotypes of predomestic horses match phenotypes painted in Paleolithic works of

662 cave art. Proc Natl Acad Sci USA 108(46):18626-18630.
663 doi:10.1073/pnas.1108982108

664 Prieto-Marquez A, Erickson GM, Seltmann K, Ronquist F, Riccardi GA, Maneva-Jakimoska
665 C, et al, Deans A (2007) Morphbank, an avenue to document and disseminate
666 anatomical data: phylogenetic and paleohistological test cases. J Morph 268:1120–
667 1120

668 Prudent X, Parra G, Schwede P, Roscito JG, Hiller M (2016) Controlling for phylogenetic
669 relatedness and evolutionary rates improves the discovery of associations between
670 species' phenotypic and genomic differences. Molec biol evol 33(8):2135-2150

671 Roscito JG, Sameith K, Parra G, Langer BE, Petzold A, Moebius C, Bickle M, Rodrigues
672 MT, Hiller M (2018) Phenotype loss is associated with widespread divergence of the
673 gene regulatory landscape in evolution. Nat Commun 9:737.
674 <https://doi.org/10.1038/s41467-018-0712>

675 Rosenthal N., Brown S. (2007) The mouse ascending: perspectives for human-disease models.
676 Nature cell biol 9:993–999

677 Ruzicka L, Bradford YM, Frazer K, Howe DG, Paddock H, Ramachandran S, Singer A, Toro
678 S, Van Slyke CE, Eagle AE, Fashena D, Kalita P, Knight J, Mani P, Martin R, Moxon
679 SA, Pich C, Schaper K, Shao X, Westerfield M (2015) ZFIN, the Zebrafish Model
680 Organism Database: Updates and new directions. Genesis 53(8):498–509

681 Schulz S, Jansen L. (2013) Formal ontologies in biomedical knowledge representation. IMIA
682 Yearb Med Inform 8(1):132–46

683 Schulz S, Stenzhorn H, Boekers M, Smith B (2007) Strengths and limitations of formal
684 ontologies in the biomedical domain. Electron J Commun Inf Innov Health 3(1):31–45

685 Sereno PC (2007) Logical basis for morphological characters in phylogenetics. Cladistics
686 23:565–587

687 Sharma V, Hecker N, Roscito JG, Foerster L, Langer BE, Hiller M (2018a) A genomics
688 approach reveals insights into the importance of gene losses for mammalian
689 adaptations. Nat Commun 9:1215. <https://doi.org/10.1038/s41467-018-03667-1>

690 Sharma V, Lehmann T, Stuckas H, Funke L, Hiller M (2018b) Loss of RXFP2 and INSL3
691 genes in Afrotheria shows that testicular descent is the ancestral condition in placental
692 mammals. PLoS Biology 16e2005293

693 Smith B (2003) Ontology. In: Floridi L (ed) Blackwell Guide to the Philosophy of Computing
694 and Information. Blackwell Publishing, Oxford, pp 155–166

695 Solé F, Ladevèze S (2017) Evolution of the hypercarnivorous dentition in mammals
696 (Metatheria, Eutheria) and its bearing on the development of tribosphenic molars. Ev
697 Dev 19(2):56–68

698 Teeling EC, Vernes SC, Dávalos LM, Ray DA, Gilbert MTP, Myers E, Bat1K Consortium
699 (2018) Bat biology, genomes, and the Bat1K project: to generate chromosome-level
700 genomes for all living bat species. Annu Rev Anim Biosci 6:23–46

701 Thenius E. (1989) Zähne und Gebiss der Säugetiere. Handbuch der Zoologie. volume 8,
702 Mammalia, part 56, Walter de Gruyter, Berlin

703 Thesleff I, Keranen S, Jernvall J (2001) Enamel knots as signaling centers linking tooth
704 morphogenesis and odontoblast differentiation. Advances Dent Res 5(1):14–18

705 Thier N, Stefen C (2020) Morphological and radiographic studies on the skull of the straw-
706 coloured fruit-bat *Eidolon helvum* (Chiroptera: Pteropodidae). Vertebrate Zoology
707 70(4). <https://doi.org/10.26049/VZ70-4-2020-05>

708 Vaughan TA, Ryan JM, Czaplewski NJ (2015) ["Chapter 4: Classification of Mammals"](#)
709 (PDF). Mammalogy (Sixth ed.)

710 Vogt L (2018) The logical basis for coding ontologically dependent characters. Cladistics
711 34(4):438–458

712 Vogt L (2019) Organizing phenotypic data—a semantic data model for anatomy. J. Biomed.
713 Semant. 10 (2019) 12. <https://doi:10.1186/s13326-019-0204-6>

714 Vogt L, Bartolomaeus T, Giribet G (2010) The linguistic problem of morphology: structure
715 versus homology and the standardization of morphological data. Cladistics 26:301–
716 325

717 Vogt L, Baum R (2019) Using named graphs and knowledge graph template patterns for
718 efficiently organizing FAIR anatomy data and metadata. Biodiversity Information
719 Science and Standards 2019. doi:10.3897/biss.3.37205

720 Vogt L, Baum R, Bhatty P, Köhler C, Meid S, Quast B, et al. (2019) SOCCOMAS: a FAIR
721 web content management system that uses *knowledge* graphs and that is based on
722 semantic programming. Database. 2019 (baz067):1–22

723

724 Wagner F, Peters B, Giere P, Grobe P, Hofmann R, Jähde M, Lächele U, Lehmann, T,
725 Ortmann S, Ruf I, Schiffmann C, Stefen C, Stuckas H, Thier N, Unterhitzenberger G,
726 Vogt L (2020) How to use [*Mammalian Traits for Comparative Genomics \(MaTrics*](#)
727 Design Principles of a project trait matrix in Morph·D·Base. URL will follow

728 Wagner F, Ruf I, Hofmann R, Lehmann T, Ortmann S, Schiffmann C, Hiller M, Stefen C,
729 Stukas H (####) Convergent evolutionary changes in mammalian composition are
730 associated to convergent gene loss: a case study for the lipase inhibitor PNLLPRP1 and
731 the xenobiotic receptor NR1I3

732 Waterston RH, Lindblad-Toh K, Birney E, et al. (2002) Initial sequencing and comparative
733 analysis of the mouse genome. *Nature* 420(6915):520–562. [doi:10.1038/nature01262](https://doi.org/10.1038/nature01262).
734 [PMID 12466850](https://pubmed.ncbi.nlm.nih.gov/12466850/)

735 Wilson DE, Reeder DM (2005) Mammal species of the world: a taxonomic and geographic
736 reference. 3rd Ed, John Hopkins University Press, Baltimore.

737 Xiang Z, Mungall C, Ruttenberg A, He Y (2011) [Ontobee: A Linked Data Server and](#)
738 [Browser for Ontology Terms](#). Proceedings of the 2nd International Conference on
739 Biomedical Ontologies (ICBO), July 28-30, 2011, Buffalo, NY, USA. pp 279-281.
740 <http://ceur-ws.org/Vol-833/paper48.pdf>

741 Zhang Z, Nikaido M (2020) Inactivation of ancV1R as a predictive signature for the loss of
742 vomeronasal system in mammals. *Genome Biol Evol* 12(6):766-778

743 Zoonomia Consortium: Genereux DP, Serres A, Armstrong J, Johnson J, Marinescu V, Murén
744 E, et al, Damas J (2020) A comparative genomics multitool for scientific discovery
745 and conservation. *Nature* 587(7833):240-245.
746 <https://www.nature.com/articles/s41586-020-2876-6>

747

748

749 **Glossary**

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

755

756 Character Coding – The parameterized description of a quality or relation of an operational
757 taxonomic unit.

758

759 Character Statement – see Sereno 2007

760

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

764

765 Machine actionable – Data and metadata that are structured in a formalized and consistent
766 way so that machines (i.e. computers) can read and use them with algorithms that were
767 programmed against this structure. Machine-actionability of data and metadata includes for
768 instance the use of persistent identifiers for data creators (e.g. ORCIDs), organizations and
769 funding agencies, but also open accessibility of data for machines through a corresponding
770 application programming interface (API), and basic semantics that allow algorithms to
771 distinguish different categories of information and apply rules to them. Machine-actionability
772 in this sense goes beyond machine-readability which only requires data and metadata to be
773 readable by a machine, i.e. data and metadata must be provided in a machine-readable format.

774 Machine-readability does not necessarily require data and metadata to provide basic semantics
775 for allowing algorithms to distinguish different categories of information contained in them.

776

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

782

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

786

787 Phenotypic trait – A particular part of the phenotype of an organism. The Phenotype of an
788 organism refers to its observable constituents, properties, and relations that can be considered
789 to result from the interaction of the organism’s genotype with itself and its environment.

790 Anatomy is the part of the phenotype that refers to the physical and structural properties of the
791 organism.

792

793 Ontology – Ontologies are dictionaries that can be used for describing a certain reality. They
794 consist of labeled classes and relations between classes, both with clear definitions that are
795 ideally created by experts through consensus and that are formulated in a highly formalized
796 canonical syntax and standardized format with the goal to yield a lexical or taxonomic
797 framework for knowledge representation (Smith 2003). Each ontology class and relation (also
798 called property) possesses its own Uniform Resource Identifier (URI*) through which it can

799 be identified and individually referenced. Ontologies contain expert-curated domain
800 knowledge about specific kinds of entities together with their properties and relations in the
801 form of classes defined through universal statements (Schulz et al. 2009, Schulz and Jansen
802 2013). Ontologies in this sense do not include statements about particular entities (i.e.,
803 empirical data). (Vogt et al. 2019)

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

807 http://purl.obolibrary.org/obo/CL_0000255 (for eukaryotic cell)

808

809

810 **Tables**

811 Table 1 Examples of data repositories in which phenotypic data of different vertebrate taxa
812 are collected. The table lists the projects with their URL and aim and/or type of information
813 that is stored and, if available, references in which the project is introduced.

814

Project	Link	Aim, type of information	Reference
Morphobank	http://www.morphobank.org	Homology of phenotypes over the web; building the Tree of Life with phenotypes, 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	
Morphological Image database	http://people.pw.cam.ac.uk/rja58/database/morphsite_bmc07.html		Asher 2007
Phenoscope	http://kb.phenoscope.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. 2015
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	Lecocq et al. 2019

		measurements	
--	--	--------------	--

815

816

817

818 Table 2 The numerical coding options for (A) absent/present and (B) multistate characters in
819 *MaTrics*. The numerals 0 and 1 refer to the *character states* ‘absent’ and ‘present’, thus, the
820 coding for multistate character states starts at 2 and continues to
821

State	State name	Description
<i>(A) Absent/present characters</i>		
?	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
<i>(B) Multistate characters</i>		
?	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

822

823

824

825

826 Table 3 Gross categories of 206 characters included in *MaTrics* and number of characters in
827 these categories

828

Gross category	Subcategory	n
Anatomy/Morphology		164
	Body plan	1
	Cranial skeleton	30
	Dentition	94
	Gastrointestinal tract	5
	Head	4
	Integument	3
	Postcranial skeleton	26
	Sense organs	1
Ecology		30
Ethology		5
Physiology		6
Embryonic		1
Total		206

829

830

831 Table 4 Deviations in cusp patterns in the studied Carnivora. Abbreviations: M1–3, upper
832 (indicated by number in superscript)/lower molar tooth (indicated by subscript); P⁴ – upper 4th
833 premolar

834

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 palatinal at P ⁴ Third lingual cusp at M ₂ Three metaconid-cusps at M ₂ Third palatinal cusp at M ¹

835

836 **Figure legends**

837

838 Fig. 1 Schematic illustration showing how phenotypic traits are reflected in *character*
839 *statements* and in the character labels in *MaTrics*. The basic nomenclature is based on Sereno
840 (2007: table 4, Scheme 3), see A1 and B1. A) Illustrates the structure for characters which can
841 be described with only two *character states*: absent and present. B) Illustrates the structure for
842 characters which require more than two *character states* (multistate characters). A2 and B2
843 give the terminology for the examples from *MaTrics* named in A3/B3. Sereno's (2007)
844 terminology recognizes *character statements* (CS) consisting of *characters* (C) and
845 *statements* (S). The *character* is represented by a (list of) *locators* (L_n, ...L₁; hierarchically
846 organized and forming the structure tree) and optionally the *variable* (V) and the *variable*
847 *qualifier* (q). The different expressions of the *variable* are given as *character states* (v₀, ... to
848 v_n) representing the *statement*. A4/B4 are examples how this nomenclature is given in the
849 character label in *MaTrics*. The *character states* are defined in the “states” field and assigned
850 to each cell of *MaTrics*. Whether a character can be described by the two states absent/present
851 or several states is indicated in the character label by the addition [a/p] and [m], respectively.

852

853

854 Fig. 2 Some examples for the presence of cusps in the studied Carnivora. A) the spotted hyaen
855 *Crocuta Crocuta* MTD B4936, B) the red panda *Ailurus fulgens* MTD B17478, C) the panda
856 *Ailuropda melanoleuca* ZMB_Mam_17246 and D) the Weddell seal *Leptonychotes weddellii*
857 MTD B5029. For each species the upper P4 and molars (1, 2) and lower molars (3, 4) are
858 illustrated as present and the cusps labelled. The teeth are photographed in lateral (1, 3) and
859 occlusal (2, 4) view. Abbreviations alphabetically: **En^d** – entoconid, **Enl^d** – entoconulid, **Hy** –
860 hypocone, **Hy^d** – hypoconid, **Hyl^d** – hypoconulid, **Me** – metacone, **Mec** – metaconule, **Me^d** –
861 metaconid, **Mes** – mesostyle, **Ms** – metastyle, **Pa** – paracone, **Pac** – paraconule, **Pa^d** –
862 paraconid, **Pr** – protocone, **Pr^d** – protoconid and **Ps** – parastyle

863

864

865 Fig. 3 The presence and absence of cusps in P⁴ and M¹ exemplified in A) the spotted hyaena
866 *Crocuta Crocuta*, B) the red panda *Ailurus fulgens*, C) the panda *Ailuropda melanoleuca* and
867 D) the Weddell seal *Leptonychotes weddellii* (teeth illustrated in Fig. 3). Abbreviations as
868 they appear in table: **Ps** – parastyle, **Pa** – paracone, **Pr** – protocone, **Ms** – metastyle, **Hy** –
869 hypocone, **Mes** – mesostyle, **Me** – metacone, **Mec** – metaconule, and **Pac** – paraconule

870

871 **Supplementary Material**

872 Table S1 List of the mammal species allocated to order, sorted alphabetically, included in
873 *MaTrics* so far (as of December 2020)

874

875

876 Table S2 Recording status of *MaTrics* as of December 2020.

877 A) recording progress of the 30,282 cells for the specific character traits (absent/present, a/p;
878 or multistate, m) as well as *missing* and *inapplicable*. Missing is the default setting and can
879 mean a) the cell has not been treated, the information on the character state for the taxon is not
880 known, or the information is known, but currently not retrievable (for example a specimen is
881 known in a distant collection). The number of relevant cells as well as the percentage is given.
882 B) Recording progress of the 206 characters for the 147 mammalian species included in
883 *MaTrics*. The table lists the number of cells which are recorded to 100% (i.e., for all species),
884 and to at least 75% and 50% of the species, respectively.

885

886 A

Character states	n	%
Recorded data (a/p and m)	18,389	60.7
<i>missing</i>	8,982	29.7
<i>inapplicable</i>	2,911	9.6

887

888 B

Recording progress (147 species)	n (traits)	% (of total number of traits)
100% (all species)	1	0.5
≥ 75% (≥ 111 species)	71	34
≥ 50% (≥ 74 species)	129	63
< 50% (<74 species)	77	37

889

890

891 Supplementary Material document S3 Brief description of statistical methods, samples and
892 observed *p*-values mentioned in the text

893

894

895 Supplementary Material Table S4 Species and the assigned material studied in the different
896 collections (SNSD – Senckenberg Naturhistorische Sammlungen Dresden, MfN – Museum
897 für Naturkunde Berlin) for the test study on Carnivora.

898

899

900

901 Supplementary Material document S5 Description of the tooth cusp patterns in 20 selected
902 Carnivora.

903

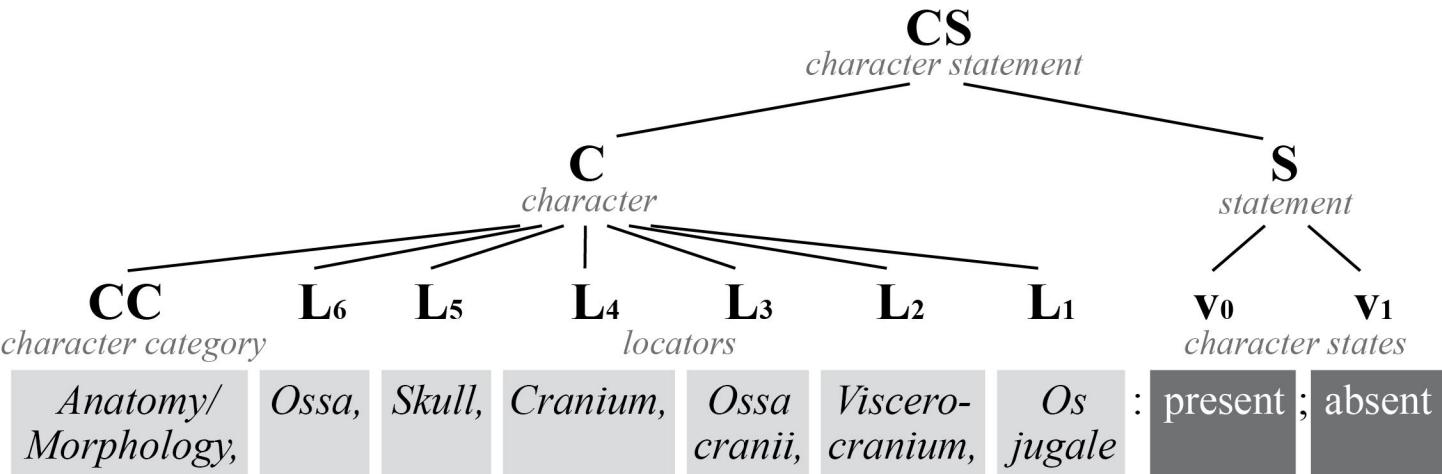
904

905 Supplementary Material Table S6 Absence (0) and presence (1) of the analyzed cusps in the
906 studied teeth of the carnivoran species

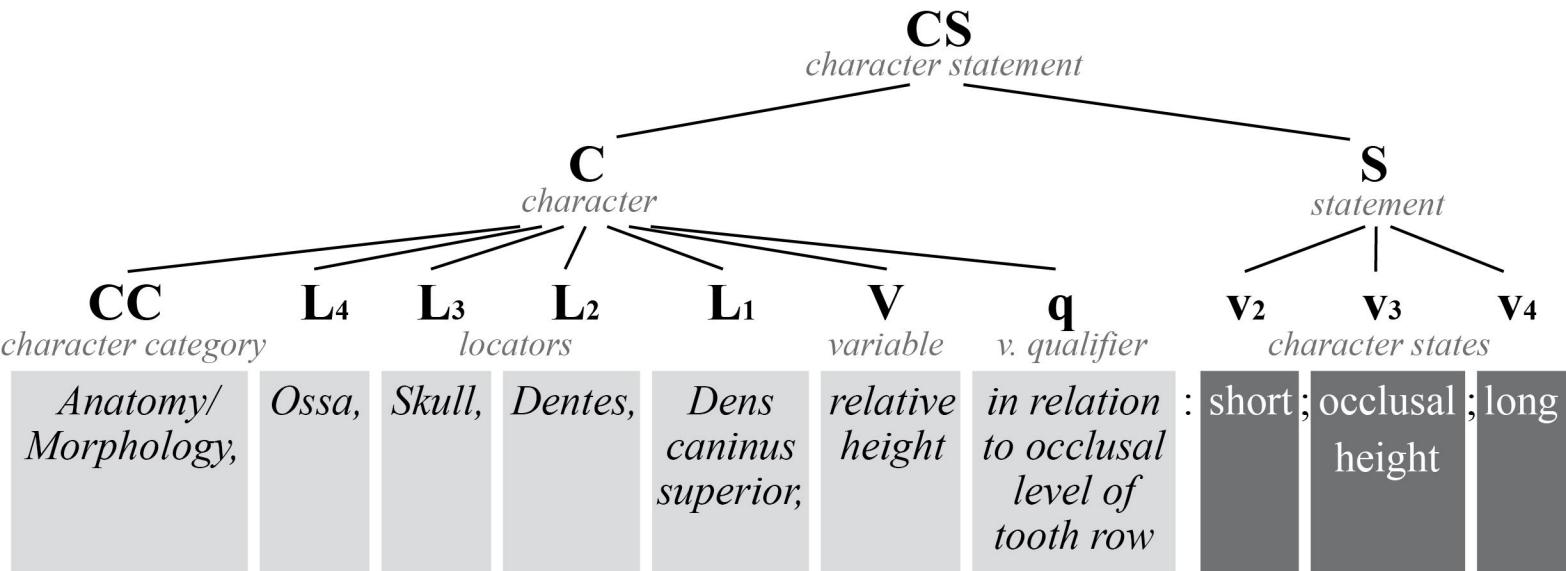
907

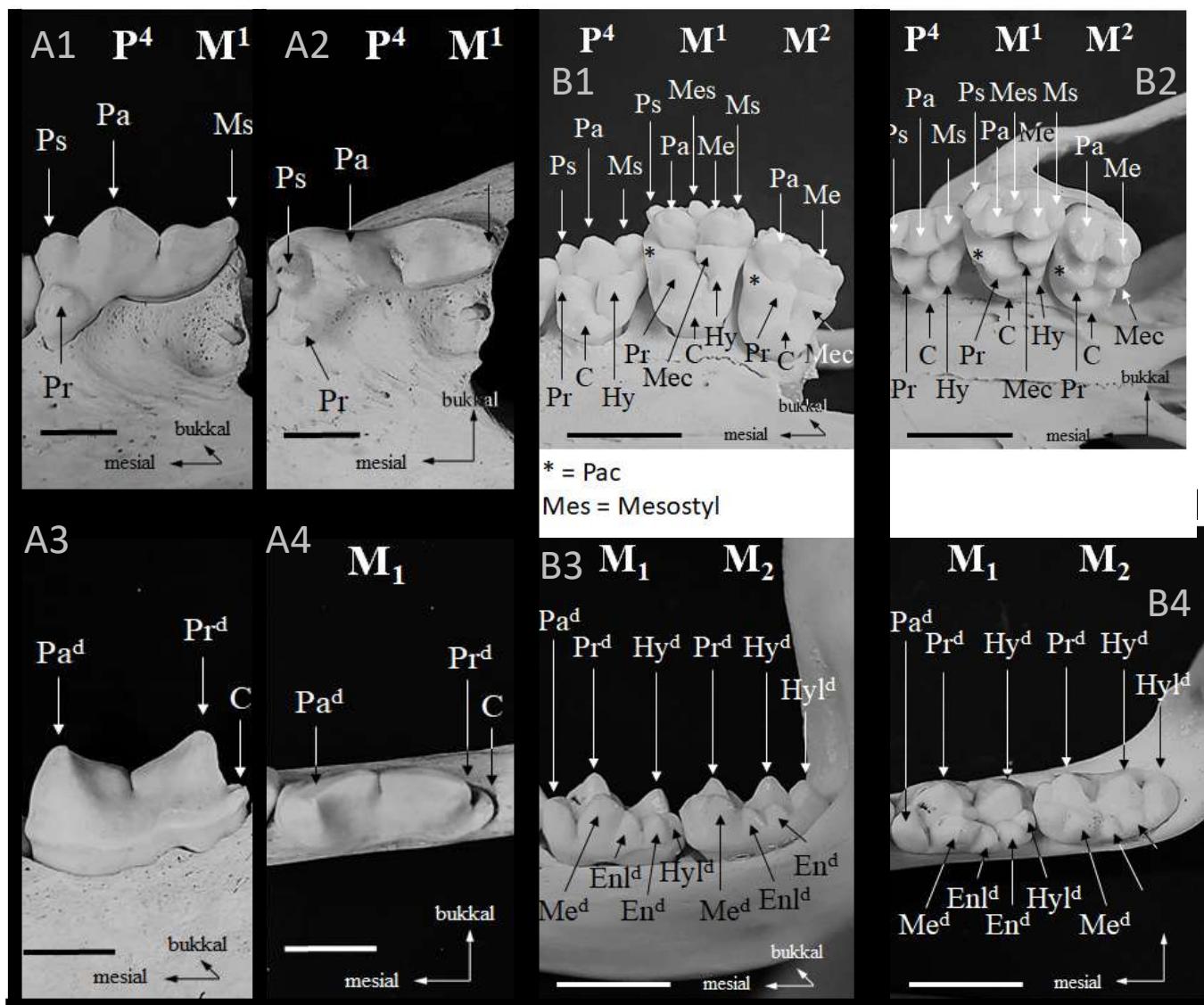
908

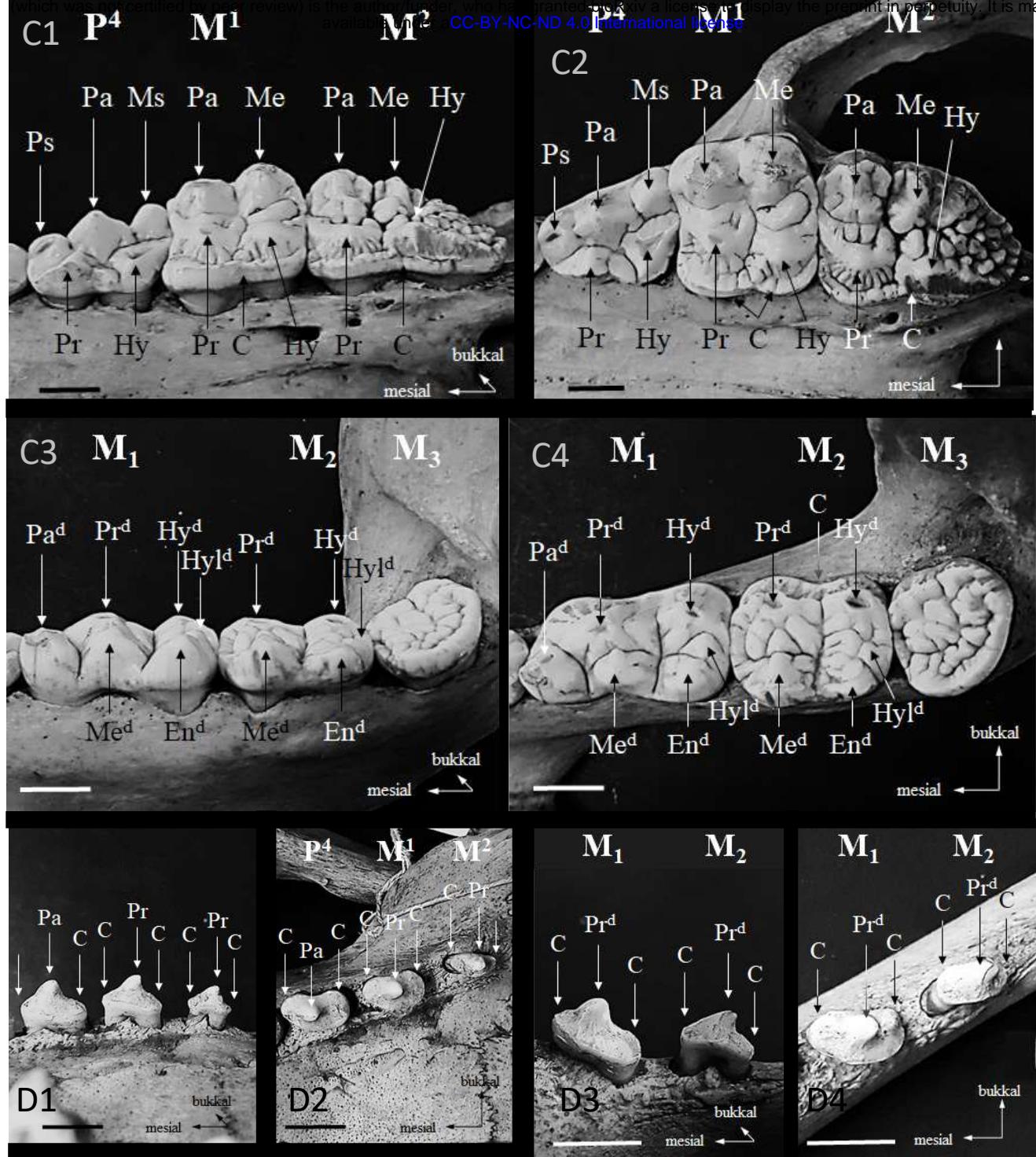
absent-present character statement



multistate character statement







Family	Genus	Species	Diet	P ⁴					M ¹								
									Trigon				Talon				
				Ps	Pa	Pr	Ms	Hy	Ps	Mes	Ms	Pa	Pr	Me	Pac	Mec	Hy
Canoidae	Ailuridae	<i>Ailurus fulgens</i>	herbivorous	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Canidae	<i>Canis lupus</i>	carnivorous	0	1	1	1	0	0	0	0	1	1	1	1	1	0
	Canidae	<i>Canis familiaris</i>	carnivorous	0	1	1	1	0	0	0	0	1	1	1	1	1	0
	Canidae	<i>Vulpes vulpes</i>	carnivorous	0	1	1	1	0	0	0	0	1	1	1	1	1	1
	Mustelidae	<i>Mustela putorius</i>	omnivorous	1	1	1	1	0	0	0	0	1	1	1	0	0	0
	Phocidae	<i>Leptonychotes weddellii</i>	piscivorous	0	1	0	0	0	0	0	0	0	1	0	0	0	0
	Procyonidae	<i>Bassariscus astutus</i>	omnivorous	1	1	1	1	1	0	0	0	1	1	1	1	1	1
	Procyonidae	<i>Nasua nasua</i>	omnivorous	1	1	1	1	1	0	0	0	1	1	1	1	1	1
	Procyonidae	<i>Procyon lotor</i>	omnivorous	1	1	1	1	1	0	0	0	1	1	1	1	1	1
	Odobenidae	<i>Odobenus rosmarus</i>			X	X	X	X	X	X	X	X	X	X	X	X	
	Ursidae	<i>Ailuropoda melanoleuca</i>	herbivorous	1	1	1	1	1	0	0	0	1	1	1	0	0	1
	Ursidae	<i>Ursus arctos</i>	omnivorous	0	1	1	1	0	0	0	0	1	1	1	0	0	1
	Ursidae	<i>Ursus maritimus</i>	hypercarnivorous	0	1	1	1	0	0	0	0	1	1	1	0	0	1
Feloidae	Felidae	<i>Acinonyx jubatus</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	1	0	0	0	0
	Felidae	<i>Felis catus</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	1	0	0	0	0
	Felidae	<i>Panthera leo</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	1	0	0	0	0
	Felidae	<i>Panthera tigris</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	1	0	0	0	0
	Felidae	<i>Puma concolor</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	1	0	0	0	0
	Hyaenidae	<i>Crocuta crocuta</i>	hypercarnivorous	1	1	1	1	0	0	0	0	1	0	0	0	0	0
	Hyaenidae	<i>Hyaena hyaena</i>	hypercarnivorous	1	1	1	1	0	1	0	0	1	1	1	0	0	0