Vole genomics links determinate and indeterminate growth of teeth

AUTHORS:

Zachary T. Calamari[1,2,3,4,*], zachary.calamari@baruch.cuny.edu

Andrew Song[1,5], ajs557@cornell.edu

Emily Cohen[1,6], ec4744@nyu.edu

Muspika Akter[1], muspika.akter@baruchmail.cuny.edu

Rishi Das Roy[7], rishi.dasroy@helsinki.fi

Outi Hallikas[7], outi.hallikas@helsinki.fi

Mona M. Christensen[7], mona.christensen@helsinki.fi

Pengyang Li[3,8], pengyang.li@cshs.org

Pauline Marangoni[3,8], pauline.marangoni@cshs.org

Jukka Jernvall[7,9], jernvall@fastmail.fm

Ophir D. Klein[3,8,*] ophir.klein@cshs.org

[1]Baruch College, City University of New York, One Bernard Baruch Way, New York, NY 10010, USA

[2]The Graduate Center, City University of New York, 365 Fifth Ave, New York, NY 10016, USA

[3]Program in Craniofacial Biology and Department of Orofacial Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[4]Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, NY, 10024, USA

[5]Cornell University, 616 Thurston Ave, Ithaca, NY 14853, USA

24 [6]New York University College of Dentistry, 345 E 34th St, New York, NY 10010

25 [7]Institute of Biotechnology, University of Helsinki, FI-00014 Helsinki, Finland

26 [8]Department of Pediatrics, Cedars-Sinai Guerin Children's, 8700 Beverly Blvd., Suite 2416, Los

27 Angeles, CA 90048, USA

28 [9]Department of Geosciences and Geography, University of Helsinki, FI-00014 Helsinki, Finland

29 *Corresponding authors

30

31 ABSTRACT

32 Continuously growing teeth are an important innovation in mammalian evolution, yet genetic

33 regulation of continuous growth by stem cells remains incompletely understood. Dental stem

34 cells responsible for tooth crown growth are lost at the onset of tooth root formation. Genetic

35 signaling that initiates this loss is difficult to study with the ever-growing incisor and rooted

36 molars of mice, the most common mammalian dental model species, because signals for root

37 formation overlap with signals that pattern tooth size and shape (i.e., cusp patterns). Different

38 species of voles (Cricetidae, Rodentia, Glires) have evolved rooted and unrooted molars that

39 have similar size and shape, providing alternative models for studying roots. We assembled a *de*

40 *novo* genome of *Myodes glareolus*, a vole with high-crowned, rooted molars, and performed

41 genomic and transcriptomic analyses in a broad phylogenetic context of Glires (rodents and

42 lagomorphs) to assess differential selection and evolution in tooth forming genes. We identified

43 15 dental genes with changing synteny relationships and six dental genes undergoing positive

44 selection across Glires, two of which were undergoing positive selection in species with unrooted

45 molars, *Dspp* and *Aqp1*. Decreased expression of both genes in prairie voles with unrooted

46 molars compared to bank voles supports the presence of positive selection and may underlie

differences in root formation. Bulk transcriptomics analyses of embryonic molar development in bank voles also demonstrated conserved patterns of dental gene expression compared to mice, with species-specific variation likely related to developmental timing and morphological differences between mouse and vole molars. Our results support ongoing evolution of dental genes across Glires, revealing the complex evolutionary background of convergent evolution for ever-growing molars.

Keywords: Evolution, selection, Glires, molar, root, dental, development, genome, rodent, tooth

DECLARATIONS

Ethics approval: The University of California, San Francisco (UCSF) Institutional Animal Care and Use Program and the Finnish national animal experimentation board approved protocols for humane euthanasia and collection of tissues for animals used in this study under protocols AN189916 (UCSF) and KEK16-021, KEK19-019, and KEK17-030 (University of Helsinki).

Availability of data and materials: The datasets supporting the conclusions of this article are available in the GenBank repository under the BioProject PRJNA1050237 (genome accession number JBBHLL000000000) and in the article's additional files.

Competing interests: The authors declare that they have no competing interests.

Funding: This research was supported by National Science Foundation grants CNS-0958379, CNS-0855217, OAC-1126113, and OAC-2215760 through the City University of New York

High Performance Computing Center at the College of Staten Island; OAC-1925590 through the MENDEL high performance computing cluster at the American Museum of Natural History; Academy of Finland to JJ; Doctoral Programme in Biomedicine, University of Helsinki to MMC; and National Institutes of Health NIDCR R01-DE027620 and R35-DE026602 to ODK.

Authors' contributions: ZTC and ODK designed the study. ZTC and PM performed animal husbandry. ZTC performed and oversaw tissue sampling, sequencing, genome assembly and annotation for *Myodes glareolus*. ZTC, AS, EC, and MA performed genome computational analyses. PL performed qPCR analyses. OH, MMC, RDR, and JJ designed and implemented RNA sequencing experiments. ZTC wrote and all authors contributed to and approved the manuscript.

Acknowledgements: The authors thank A. Joo, N. Ahituv, G. Amato, A. Narechania, S. Singh, A. Scott, and A. Paasch for advice on methods and access to cluster computing resources.

INTRODUCTION

Hypselodonty, or the presence of unrooted and thus ever-growing teeth, has evolved multiple times in mammals. Glires—the clade containing rodents, rabbits, and their relatives—have hypselodont incisors (1), and multiple Glires have also evolved hypselodont molars (Fig. 1). At least in rodents, molar hypselodonty evolved considerably later than hypsodont molars, which are high crowned but rooted, which in turn evolved later than hypselodont incisors. In Glires, molars appear to increase in crown height from brachydonty (low-crowned, rooted), through hypsodonty (high-crowned, rooted), toward hypselodonty (high-crowned, unrooted) (2).

93 Mice (*Mus musculus*), the primary mammalian model species of dental research, have
94 hypselodont incisors but retain brachydont molars. Because of this, mice cannot provide
95 information about the hypsodont teeth that likely preceded hypselodonty.

96 Mammalian teeth sit in bony sockets, held in place by soft tissue (periodontal ligament)
97 attached to cementum-covered tooth roots (3). Ligamentous tooth attachment may have arisen
98 along with a reduction in the rate of tooth replacements, providing greater flexibility for
99 repositioning the teeth as the dentary grows (4,3). Consequently, the limited replacement of
100 mammalian teeth (two sets of teeth in most mammals and one in Glires) may have spurred the
101 evolution of hypsodont and hypselodont teeth, both with high crowns that compensate for tooth
102 wear from gritty or phytolith-heavy diets (5,6), and resulted in further modification of the
103 anchoring roots. The convergent evolution of unrooted molars in Glires presents an opportunity
104 to identify whether consistent developmental and genomic changes underlie the formation of
105 hypselodont teeth in different species, in turn revealing the conserved mechanisms that produce
106 tooth roots. Furthermore, the relatively recent evolution of molar hypselodonty, starting in the
107 Middle Miocene (approximately 16-12 Ma) (2), should provide molecular evidence for the steps
108 required to make a continuously growing organ.

109 Dental development proceeds from the tooth germ, composed of epithelium and
110 mesenchyme, through phases known as the bud, cap, and bell (7). Multipotent enamel epithelium
111 differentiates into the cells that form the tooth crown (8–11). As development progresses in
112 rooted teeth, the epithelium at the tooth apex transitions first to a tissue called Hertwig's
113 epithelial root sheath, and eventually cementum-covered roots (9,10). Studies have identified
114 numerous candidate genes and pathways with various roles during root development, such as
115 *Fgf10*, which decreases in expression at the beginning of root formation (12–18). Although

116 research on mouse molars has identified genetic signals related to root formation, a number of
117 the key genes studied have broad developmental roles, such as *Wnt* family members (14), or
118 overlap considerably with genes also involved in patterning the size and shape of the tooth
119 (17,19–22). This overlap between shape and root expression patterns confounds our ability to
120 identify a clear signal initiating root formation.

121 Evolutionary novelties such as high-crowned hypsodont and hypselodont molars can
122 arise from differences in gene expression and regulation (23–26). Evolutionarily conserved gene
123 expression levels produce conserved phenotypes, and changes in gene regulatory networks have
124 long been linked to morphological evolution (27,28). The order of genes along a chromosome
125 (synteny) can affect gene expression and regulation, as regulatory sequences are often located
126 near their target genes (cis-regulatory elements) (29–31). Genome rearrangements that place
127 genes near new regulatory elements may change the expression and selective environment of
128 those genes; these small-scale rearrangements of genes may be common in mammals (32–34).
129 Likewise, regions of chromosomes that form topologically associated domains may experience
130 similar selective pressures, including selection against rearrangement (35,36). Genes involved in
131 molar development are not syntenic in the mouse genome nor are genes with organ-specific
132 expression (37), and thus the regulatory or selection effects of co-localization need not apply to
133 all dental genes at once. Changes in genome architecture between Glires species thus may result
134 in different selective and expression environments for dental genes that could result in the
135 evolution of hypselodont molars.

136 To establish a model rodent species with hypsodont molars for close comparison to
137 hypselodont molars, we sequenced and annotated a highly-complete *de novo* genome of *Myodes*
138 *glareolus*, the bank vole. The bank vole is increasingly used in medical and environmental

139 research, ranging from studying zoonotic diseases (38) to immune responses (39,40), and even
140 assessing environmental remediation efforts through heavy metals that accumulate in vole teeth
141 (41,42), thus our efforts may be of use beyond dental research. The bank vole's hypsodont
142 molars bridge the gap between low-crowned mouse and hypselodont prairie vole (*Microtus*
143 *ochrogaster*) molars, reducing the effects of morphological differences on root formation
144 signaling. We performed a suite of genomic and transcriptomic tests of our new bank vole
145 genome in a broad phylogenetic context to test the hypothesis that dental genes are undergoing
146 positive selection and exhibit different expression patterns in species with unrooted, hypselodont
147 molars. We predicted that genes without conserved syntenic relationships in these species would
148 be more likely to have sites under positive selection or significantly different expression. Our
149 analyses revealed loss of synteny and positive selection for dental genes in Glires with unrooted
150 molars compared to those with rooted molars. We also demonstrated strong conservation of
151 dental gene expression patterns between bank voles and mice, with key differences related to the
152 timing and patterning of tooth morphology.
153
154 RESULTS
155 *Orthology assessment and loss of synteny*
156 To identify which sequences in our bank vole (*Myodes glareolus*) genome and annotation
157 had the same evolutionary history as dental genes identified in other Glires and assess genome
158 rearrangements, we performed orthology and synteny analyses in a broad phylogenetic context.
159 OrthoFinder identified 20,547 orthogroups representing 97.9% of the genes across all 24
160 analyzed genomes (including the human outgroup). Of the orthogroups, 6,158 had all species
161 present. In our *de novo* bank vole genome, there were 27,824 annotated genes, of which 84.2%
162 were assigned to an orthogroup. Bank vole genes were present in 16,250 orthogroups. On

163 average, the genomes included in the OrthoFinder analysis had 19,814 genes, with 98.2% of
164 those assigned to orthogroups.

165 The completeness and large scaffold N50 (4.6 Megabases) of our bank vole assembly
166 supported its inclusion in generating a Glires synteny network. Using the infomap clustering
167 algorithm, we produced 19,694 microsynteny clusters from this overall synteny network. We did
168 not expect dental genes to share the same microsynteny cluster, and instead examined whether
169 each gene was in the same microsynteny cluster in species with rooted or unrooted molars. We
170 identified 15 hierarchical orthogroups in which synteny was not conserved for at least half of the
171 Glires with unrooted molars (Fig. 2). The genes form two groups (Fig. 2A), group 1, lacking
172 synteny across Glires, and group 2, lacking synteny mainly in species with unrooted molars.
173 Most of these genes also are missing from the orthogroups; only *Mmp20*, *Irx6*, *Aqp3*, *Sema3b*,
174 and *Col4a1* were well represented in their orthogroups but not in their synteny networks (full
175 comparisons of orthology and synteny are in Additional information 1). Overall, these genes
176 represent multiple categories of "keystone" dental genes. Null mutations in keystone dental
177 genes affect embryonic dental development (43): "shape" genes cause morphological errors;
178 "eruption" genes prevent tooth eruption; "progression" genes stop the developmental sequence;
179 "tissue" genes cause defects in tissues; "developmental process" genes are annotated with the
180 "GO:0032502" gene ontology term; "dispensable" genes, while dynamically expressed in
181 developing teeth, have no documented effect on phenotype; and "double" genes function
182 redundantly with a paralog and only produce a phenotype when both genes are mutated. The
183 group "other" is composed of the remaining protein coding genes (43). Most genes lacking
184 conserved synteny in species with unrooted molars are in the "dispensable" category (Fig. 2D),

185 thus the relationship between differences in these genes and tooth phenotypes is unclear, at least
186 during embryonic development.
187
188 *Multiple dental genes under positive selection*
189 We hypothesized that dental genes are undergoing positive selection in species with
190 unrooted molars. Our positive selection analyses in PAML (phylogenetic analysis by maximum
191 likelihood (44)) identified 6 dental gene orthogroups undergoing site-specific positive selection
192 across Glires (Table 1). Four orthogroups with site-specific positive selection lacked synteny
193 among Glires with unrooted molars: *Col4a1*, *Dspp*, *Runx3,* and the four-gene orthogroup with
194 sequences similar to *Runx3* (Fig. 2A). We then assessed genes for site-specific positive selection
195 in species with unrooted molars compared to species with rooted molars (branch-and-site-
196 specific positive selection (45)), focusing on those genes with site-specific positive selection or
197 evidence for loss of synteny. Two genes, *Dspp* and *Aqp1* were undergoing this branch-and-site
198 specific positive selection. Both genes had a single highly supported site (posterior probability >
199 0.95) under positive selection in species with unrooted molars based on the Bayes Empirical
200 Bayes method for identifying sites under selection implemented in PAML (46). *Dspp* also had
201 multiple sites with moderate support (posterior probability > 0.75). The overall selection patterns
202 on each gene differed. Maximum likelihood estimates of selection for *Dspp* showed the
203 percentage of sites under purifying and neutral selection on all branches were nearly equal (47%
204 and 44%, respectively). Percentages of sites under positive selection in the species with unrooted
205 molars (foreground branches) were nearly evenly divided as well, with 5% of sites from
206 branches where the species with rooted molars (background branches) were undergoing
207 purifying selection and 4% of sites from branches where the species with rooted molars were

208 under neutral selection. For *Aqp1*, nearly all sites were under purifying selection on all branches
209 (91%), and few sites were under neutral selection on all branches (7%). Few sites were
210 undergoing positive selection in the foreground branches and their distribution also was unevenly
211 split between sites under purifying and neutral selection on background branches (0.6% and
212 0.04%, respectively). The complete list of dental genes with hierarchical orthogroups,
213 microsynteny clusters, and positive selection test results are available in Additional file 1.
214 Because genes under positive selection are often expressed at lower levels than genes
215 under purifying selection (47–50), we also compared expression levels of *Dspp* and *Aqp1* in first
216 molars (M1) at postnatal days 1, 15, and 21 (P1, P15, and P21) in bank voles (rooted molars) and
217 prairie voles (unrooted molars) using quantitative PCR. Prairie vole molars expressed *Aqp1* at
218 significantly lower levels than bank vole molars across all three ages (Fig. 3). Prairie vole P1
219 molars expressed significantly lower levels of *Dspp* than bank vole molars; at P15 and P21, their
220 molars expressed *Dspp* at lower, but not statistically significantly different, levels than their bank
221 vole equivalent. For both genes, the prairie vole had consistent expression levels across three
222 biological replicates, while the bank vole had greater variation in expression levels across
223 replicates.
224
225 *Few changes of secondary structure at positively selected sites*
226 To detect whether substitutions at sites under positive selection influenced protein
227 structure and evolution, we analyzed ancestral states and secondary structure across Glires. We
228 first reconstructed ancestral sequences along the internal nodes of the Glires phylogeny for the
229 genes undergoing branch-and-site specific positive selection to assess potential secondary
230 structural changes in their protein sequences. At the best-supported site in *Dspp* (position 209 in

231 the gapped alignment, Additional file 2), there were three major amino acid changes. The
232 ancestral Glires sequence started with an asparagine (N) in this position. Two of the three species
233 with unrooted molars represented in the *Dspp* dataset had amino acid substitutions at this
234 position, with *Oryctolagus cuniculus* substituting a leucine (L) and *Dipodomys ordii* substituting
235 an aspartic acid (D) at this position (Fig. 4A). All muroids (the clade including the voles in
236 family Cricetidae and mice and rats in family Muridae) in our phylogeny substituted histidine
237 (H) for the asparagine at this position. The secondary structure predicted at this position was a
238 coil for most sequences but a helix for the *D. ordii* sequence (Fig. 5). *Aqp1* sequences varied
239 greatly at the position under putative positive selection in species with unrooted molars (position
240 294 in the gapped alignment, Additional file 3). The ancestral state reconstruction showed twelve
241 changes of the amino acid at this position across Glires (Fig. 4B), yet these changes did not
242 affect the predicted secondary structure of the protein near this residue, which was a coil for all
243 sequences tested. All secondary structure predictions are available in Additional file 4.
244
245 *Bank vole molar gene expression is similar to that of other Glires*
246 We also assessed differential gene expression between mouse and bank vole molars
247 across early development to study the effects of morphology on expression levels of dental
248 genes. Our gene expression analysis focused on keystone dental gene categories. Our bank vole
249 genome was like the mouse and rat genomes in terms of the numbers and expression patterns of
250 genes annotated from these keystone categories (Table 2). Ordination of gene expression results
251 from the bank vole and mouse data at embryonic day 13, 14, and 16 (E13, E14, E16) (43) by
252 principal components analysis showed a distinct separation between the mouse and bank vole
253 along the first principal component (PC1) of the 500 most variable genes (Fig. 6A). PC1

254 explained 82.81% of the variance in these genes; there are distinct, species-specific expression
255 patterns in these tissues. Along PC2 (7.47% of variance explained), E13 and E14 samples differ
256 from the E16 samples, although the difference in time points is much greater in bank voles.
257 Ordination of just the keystone dental genes showed clear separations between tissues based on
258 species and age (Fig. 6B). Within this focused set of genes, however, PC1 and PC2 explain less
259 variance (44.8% and 28.84% respectively), and have a less clear relationship to species and age.
260 There are two distinct, parallel trajectories for the mouse and bank vole. Although within each
261 species there is separation by age along PC1 and PC2, mouse E16 and bank vole E13 occupy a
262 similar position along PC1, and mouse E13 and bank vole E16 occupy a similar position along
263 PC2.

264 Examining individual genes underlying the differences between mouse and vole molars,
265 we note several upregulated genes in our vole molars are broadly expressed in developing molars
266 of other vole species (51,52). Relative to the mouse molars, vole molars overexpressed genes
267 related to forming tooth cusps, including *Bmp2*, *Shh*, *p21* (also known as *Cdkn1a*), and *Msx2*, a
268 difference explained by the faster patterning and larger number of cusps in the vole molar
269 compared to the mouse molar (52). Another gene upregulated in the patterning stage vole molar
270 is *Fgf10*, which is associated with delayed root formation later in vole molar development (9).

271 Nevertheless, developing bank vole molars at E13, E14, and E16 expressed keystone
272 dental genes in overall proportions like those observed at analogous stages of mouse and rat
273 molar development (Fig. 7). Permutation tests within each bank vole sample showed that log
274 counts for the set of genes related to the progression of dental development were significantly
275 higher than those in the tissue, dispensable, developmental process, and "other" categories at E14
276 and E16. The progression gene counts in E13 molars were higher for all of these except the

277 dispensable category. Shape category genes also were significantly higher than "other" category
278 genes in the E14 tissue. Overall, even though we observed conserved expression patterns of
279 dental genes at the system level, individual genes involved in cusp patterning and morphology
280 differed between the mouse and the vole.
281
282 DISCUSSION
283 Our two goals in sequencing the genome of *Myodes glareolus* were to support the
284 development of a comparative system for studying tooth root development and to investigate the
285 evolution of dental genes in Glires, a clade in which ever-growing molars have evolved multiple
286 times (1). Our new *M. glareolus* assembly and annotation captured nearly all of the single-copy
287 orthologs for Euarchontoglires and provided scaffolds with sufficient length for synteny
288 analyses. It was well represented in ortholog groups and microsynteny clusters across Glires. We
289 tested the hypothesis that dental genes are undergoing site-specific positive selection in species
290 with unrooted molars (branch-and-site specific positive selection (45)). We predicted that lack of
291 conserved syntenic relationships in species with unrooted molars could place dental genes in
292 regulatory and selective environments that promote changes among genes relevant to tooth root
293 formation. Our analyses identified 15 dental genes without conserved syntenic relationships
294 across Glires and two dental genes, *Dspp* and *Aqp1*, under positive selection in species with
295 unrooted molars. We also demonstrated conserved patterns of gene expression among dental
296 keystone genes between bank voles and mice during early embryonic development, and
297 deviations from these conserved patterns likely related to differences in molar morphology
298 between the two species.

299 We identified 15 genes which were not syntenic in at least half of the species with
300 unrooted molars, and six genes undergoing site-specific positive selection across all Glires.
301 Although four of the orthogroups with site-specific positive selection lacked synteny in species
302 with unrooted molars, only *Col4a1* was well represented among these species in its orthogroup.
303 The two genes undergoing branch-and-site-specific positive selection in species with unrooted
304 molars, *Dspp* and *Aqp1*, both maintained their synteny relationships across the Glires studied.
305 Although we predicted loss of synteny for dental genes in Glires with unrooted molars could
306 result in sequence evolution by placing genes in new selective contexts, our analyses did not
307 support a strong relationship between non-syntenic genes and branch-and-site-specific positive
308 selection. Maximum likelihood estimates of selection on each site for the genes with branch-
309 specific positive selection revealed different overall selective pressures on *Dspp* and *Aqp1*; *Dspp*
310 sites on background branches (i.e., branches with species that have rooted molars) were under a
311 mix of purifying and neutral selection, while nearly all *Aqp1* background branch sites were under
312 purifying selection. These selection regimes suggest there is greater conservation for *Aqp1*
313 function across Glires than for *Dspp* function. Gene duplication can result in functional
314 redundancy and evolution toward a novel function in some genes (53–56), which may explain
315 positive selection in *Aqp1*, as there are other aquaporin family genes present. Although *Dspp* has
316 no paralogs, it overlaps functionally with other SIBLING family proteins (e.g., *Opn*, *Dmp1*)
317 (57,58).
318 *Aqp1* and *Dspp* play different functional roles during dental development. Under the
319 keystone dental development gene framework, *Aqp1* is a "dispensable" gene: developing teeth
320 express it, but tooth phenotypes do not change in its absence. *Aqp1* is expressed in endothelia of
321 microvessels in the developing tooth (59,60). *Dspp* may be particularly relevant for the

322 formation of an unrooted phenotype if its expression domain or function have been modified in
323 species with unrooted molars. *Dspp* is a "tissue" category keystone dental gene, meaning the
324 main effects of a null mutation occur during the tissue differentiation stage of dental
325 development (43). Null mutations of *Dspp* cause dentin defects in a condition called
326 dentinogenesis imperfecta (61,62); in some patients, teeth form short, brittle roots (62,63). *Dspp*
327 knockout mice also exhibit the shortened root phenotype, among a variety of other defects in
328 both endochondral and intramembranous bone, due to the disruption of collagen and bone
329 mineralization (64–66).

330 Our ancestral sequence reconstructions and estimated secondary protein structures
331 allowed us to assess whether nonsynonymous substitutions at sites under positive selection
332 resulted in structural differences, thus potentially affecting protein function. Although unrooted
333 molars are a convergent phenotype across Glires, the sites under positive selection did not
334 converge on the same amino acid substitution in species with unrooted molars, and *Aqp1*
335 appeared particularly labile at this residue. The non-synonymous substitutions at these sites often
336 resulted in changes of properties of the amino acid in the sequence, for example in *Dspp*, polar
337 asparagine was replaced with non-polar leucine in *O. cuniculus*. Only one of these substitutions
338 changed the predicted secondary structure. Nevertheless, single amino acid substitutions do
339 produce dental phenotypes for both *Dspp* (67) and *Aqp1* (68), thus we cannot rule out functional
340 changes in these genes in species with unrooted molars.

341 Although the exact relationship between gene expression and sequence divergence
342 remains unclear (69), studies of genome evolution across small numbers of mammal species
343 show correlations between gene sequence divergence and levels of expression (70). In particular,
344 highly-expressed genes are more likely to experience purifying selection (47–50), while lowly-

345 expressed genes and tissue-specific genes may experience positive selection (48). The decreased
346 expression of *Dspp* and *Aqp1* in prairie vole M1 compared to that of the bank vole M1 thus
347 supports our finding of positive selection in these genes in species with unrooted molars. If all
348 species with unrooted molars also exhibit decreased expression levels of *Dspp* and *Aqp1*, it could
349 suggest a strong link between lower levels of the genes and the unrooted phenotype.

350 Without analyses of functional variation caused by positive selection at these coding
351 sites, or spatial sampling to determine where these genes may be expressed during development,
352 we are limited from exploring the specific effects of *Dspp* and *Aqp1* on root formation.
353 Nevertheless, we found evidence for evolution of these genes in Glires with unrooted molars,
354 and *Dspp* especially has clinical relevance for tooth root formation. Future studies should explore
355 the spatial distribution of *Dspp* expression, which could be relevant to functional changes in
356 Glires with unrooted molars. If positive selection and corresponding amino acid changes
357 identified in *Dspp* here modify its expression domain or its interaction with yet-unidentified root
358 formation co-factors, it may serially reproduce the unrooted incisor phenotype in molars.

359 Our RNA sequencing results supported the bank vole as a suitable system for studying
360 dental development. Although molar morphology differs considerably across mammals,
361 candidate-gene approaches have identified numerous conserved genes involved in tooth
362 development and morphological patterning (71). Studies of single genes or gene families have
363 identified shape-specifying roles common to multiple species (52,72–74), and high-throughput
364 sequencing of mouse and rat molars demonstrate that both species express sets of dental
365 development genes in similar proportions during early stages of tooth development (43). The
366 similarity of our high-throughput RNA sequencing results (Fig. 7) to the mouse and rat results in
367 previous studies suggest overall expression patterns of keystone dental development genes

368 within each stage are conserved across Glires. Our principal component analyses and differential
369 expression analyses measuring changes between mouse and bank vole molars, however, showed
370 that several dental genes' expression levels differed significantly by species and age. Previous
371 research has documented organ expression patterns that are conserved across species early in
372 development and diverge over time, with some major organs displaying heterochronic shifts in
373 some species (75). If the major source of variation in keystone dental gene expression patterns
374 between mice and bank vole molars were solely attributable to species, we might expect to see
375 clear separation between the species along the first or second principal component (PC1 or PC2),
376 like that observed in PC1 of the 500 most variable genes (Fig. 6). If molar development follows
377 the diverging expression patterns observed in other organs, we might expect just the earliest age
378 classes to align on one, or multiple, PCs. Instead, we found two trajectories that were nearly
379 parallel across PC1 and PC2 and multiple keystone dental genes that were significantly
380 differentially expressed with respect to species and age. This variation between species is likely
381 driven by the larger number of cusps in the vole molar, and corresponding upregulation of genes
382 regulating cusp formation. The overall acceleration of patterning in vole molars likely explains
383 the significance of the age variable in our expression results, causing a heterochronic shift in the
384 expression patterns.

385 Our analyses were limited by the small number of rodent species with sufficiently
386 annotated genomes to be included in synteny and positive selection analyses. This limitation left
387 us with a small phylogeny for our ancestral state reconstructions, which thus did not encompass
388 the full diversity of Glires tooth roots, and potentially weakened model-based genomic analyses.
389 Although positive selection analyses using the Bayes Empirical Bayes criterion are robust to
390 smaller sample sizes (46), incomplete sampling can affect estimations of ancestral characteristics

391 (76). Innovations in paleoproteomics also offer the opportunity to compare fossil species' dental
392 gene sequences directly to living and estimated ancestral sequences (77,78). By incorporating
393 data for extinct Glires in both morphological and molecular analyses, we can further elucidate
394 links between dental gene evolution and unrooted teeth.
395
396 CONCLUSIONS
397 Our genomics and transcriptomics analyses, based on our newly sequenced, high-quality
398 draft bank vole genome assembly and annotation, showed that bank vole early tooth
399 development is comparable to other commonly used rodent models in dental development
400 research. We identified 6 dental gene orthogroups that were undergoing site-specific positive
401 selection across Glires and two genes, *Dspp* and *Aqp1*, that were undergoing site-specific
402 positive selection in Glires with unrooted molars. *Dspp* appears particularly relevant to root
403 formation, as loss-of-function mutations cause a dentin production defect that can result in
404 shortened tooth roots. Future research must explore the functional role that *Dspp* plays in tooth
405 root formation in Glires and other clades. The rodent dentary is an exciting system for
406 understanding tooth development; it provides an easily manipulated set of tissues that can be
407 produced quickly and features a lifelong population of stem cells in the incisor with genomic
408 mechanisms that are potentially replicated across other teeth in species with unrooted molars.
409 Our results identify candidate genes for future analyses, and our draft bank vole genome and
410 annotation improve the utility of this species for comparative dental research that can uncover
411 the genetic mechanisms of tooth root formation.
412
413 METHODS

414 *Tissue collection and sequencing*

415 To assemble the bank vole genome, we sequenced tissues from a single adult male
416 specimen housed in a colony at the UCSF Mission Center Animal Facility. We euthanized the
417 animal according to UCSF IACUC protocol AN189916 and harvested muscle, kidney, heart, and
418 liver tissue, which were immediately frozen at -80°C. Tissues were sent to a third-party
419 sequencing service, where they were combined and homogenized to achieve appropriate mass
420 for high molecular weight DNA extraction. We targeted 60x coverage with 150 base pair (bp)
421 reads using 10X Chromium linked-read chemistry (79,80) sequenced on the Illumina platform.
422 We also targeted 10x coverage with Pacific Biosciences SMRT long-read chemistry. For genome
423 annotation and gene expression analyses, we collected seven biological replicates each of first
424 molars at embryonic days 13-16 (E13, E14, E15, E16), second molars at E16, and jaw tissues at
425 E14 under University of Helsinki protocols KEK16-021, KEK19-019, and KEK17-030 and
426 stored them in RNAlater at -80°C for RNA sequencing, following a tissue harvesting protocol
427 established for mice and rats (43). We extracted RNA from these tissues using a guanidium
428 thiocyanate and phenol-chloroform protocol combined with an RNeasy column purification kit
429 (Qiagen) based on the keystone dental gene protocol (43). Single-end 84 bp RNA sequencing
430 was performed using the Illumina NextSeq 500 platform.

431

432 *Genome assembly and quality control*

433 We first assembled only the 10X Chromium linked reads using the default settings in
434 Supernova 2.1.1. (79,80). We selected the "pseudohaplotype" (pseudohap) output format, which
435 randomly selects between potential alleles when there are two possible contigs assembled for the
436 same region. This option produces two assemblies, each with a single resolved length of the

437 genome sequence (79–81). We used our lower-coverage, long-read data for gap filling and
438 additional scaffolding. First, we estimated the genome's length using the raw sequence data in
439 GenomeScope (82), which predicted a length of 2.6 gigabases. We then performed error
440 correction of the long reads using Canu (83), removing reads shorter than 500 base pairs (bp) and
441 disregarding overlaps between reads shorter than 350 bp. We kept only those reads with
442 minimum coverage of 3x for scaffolding. Following long read error correction, we used Cobbler
443 and RAILS (84) with a minimum alignment length of 200 bases to accept matches for gap filling
444 and scaffolding of both pseudohap assemblies.

445 For quality control, we assessed both unscaffolded and long-read scaffolded pseudohap
446 assemblies by standard assembly length statistics with QUAST (85) and presence of single-copy
447 orthologs with BUSCO v3 (86). Both scaffolded assemblies were approximately 2.44 Gigabases
448 long, with an N50 (the length of the shortest scaffold at 50% of the total assembly length) of 4.6
449 Megabases; we refer to them as Pseudohap1+LR and Pseudohap2+LR. The Pseudohap1+LR
450 assembly had 17,528 scaffolds over 1000 bp long, and the Pseudohap2+LR assembly had 17,518
451 scaffolds over 1000 bp long (Table 3). BUSCO searched for universal single-copy orthologs
452 shared by Euarchontoglires, recovering 89.4% of these genes in the scaffolded Pseudohap1+LR
453 assembly and 92.8% of the single-copy orthologs in the scaffolded Pseudohap2+LR assembly
454 (Fig. 8). The two assemblies were similar length and contiguity, but we based annotation and
455 downstream analyses on Pseudohap2+LR because it recovered more single-copy orthologs.
456

457 *Genome annotation*

458 We annotated the genome using multiple lines of evidence in three rounds of the
459 MAKER pipeline (87–89). For evidence from gene transcripts, we assembled a *de novo*

460 transcriptome assembly of the single-end RNA sequences pooled from all molar and jaw tissues
461 using Trinity (90). We also included cDNA sequences from the *Mus musculus* assembly
462 GRCm38 to provide additional transcript evidence from a close relative with a deeply annotated
463 genome. We used SwissProt's curated protein database to identify protein homology in the
464 genome. Two libraries of repeats provided information for repeat masking: the Dfam Rodentia
465 repeat library (91–93) and a custom library specific to the bank vole estimated with a protocol
466 modified from Campbell et al. (88). The custom library features miniature inverted-repeat
467 transposable elements identified with default settings in MiteFinder (94), long terminal repeat
468 retrotransposons extracted with the GenomeTools LTRharvest and LTRdigest functions (95)
469 based on the eukaryotic genomic tRNA database, and *de novo* repeats identified with
470 RepeatModeler (96). We combined elements identified by these programs into a single repeat
471 library, then removed any elements that matched to a custom SwissProt curated protein database
472 excluding known transposons. The custom repeat library is available in Additional file 5. We
473 trained a custom gene prediction model for MAKER as well. The first iteration of the model
474 came from BUSCO's implementation of augustus (97). Between each round of MAKER
475 annotation, we further updated the gene prediction model with augustus.
476 MAKER considered only contigs between 10,000-300,000 bp long during annotation.
477 Our second and third iterations of MAKER used the same settings but excluded the
478 "Est2genome" and "protein2genome" functions, as recommended in the MAKER tutorial. We
479 included a SNAP (98) gene prediction model based on the output of the first round of annotation
480 during the second and third iterations of MAKER annotation. Annotation quality (i.e., agreement
481 between different lines of evidence and the MAKER annotation) was assessed visually in
482 JBrowse after each iteration and using *compare_annotations_3.2.pl* (99), which calculates the

483 number of coding and non-coding sequences in the annotation in addition to basic statistics about
484 sequence lengths. Our MAKER annotation covered 2.41 Gb of the scaffolded Pseudohap2
485 assembly in 4,125 scaffolds. These scaffolds contained 27,824 coding genes (mRNA) and 15,320
486 non-coding RNA sequences. The average gene length was 12,705 bp. Most annotations (91.4%)
487 had an annotation edit distance (AED) of 0.5 or better. AED is a measure of congruency between
488 the different types of evidence for an annotation, where scores closer to zero represent better-
489 annotated genes (100).
490
491 *Orthology and synteny analyses*
492 We analyzed orthology and synteny of the bank vole genome to understand gene and
493 genome evolution related to dental development across Glires with rooted and unrooted molars.
494 We obtained genomes from Ensembl for 23 Glires species and one phylogenetic outgoup, *Homo*
495 *sapiens* (Table 4). These genomes all had an N50 over 1 Mb, which improves synteny
496 assessment (101). We first analyzed all 24 genomes for groups of orthologous genes
497 (orthogroups) in OrthoFinder (102), providing a tree topology based on the Ensembl Compara
498 reference tree (Fig. 1) to guide orthology detection. Because we would not analyze the human
499 outgroup in downstream analyses, we implemented the OrthoFinder option that splits
500 orthogroups at the root of Glires (hierarchical orthogroups), thus any group of orthologs studied
501 here represents only genes with shared, orthologous evolutionary history within Glires. We
502 selected MAFFT (103) for multiple sequence alignment and fastme (104) for phylogenetic tree
503 searches within OrthoFinder. We retained the gene trees estimated for each orthogroup for
504 downstream analyses.

505 Although dental development genes are spread throughout the genome, we were
506 interested in whether each gene remained in the same local arrangement across species of Glires.
507 We prepared each genome annotation and sequence file for synteny analysis using the
508 reformatting functions of Synima (105) to extract each peptide sequence associated with a gene
509 coding sequence in the Ensembl annotation. Collinear synteny blocks estimated by MCScanX
510 (106) formed the basis for synteny network analyses using the SynNet pipeline (107–109). We
511 inferred networks from the top five hits for each gene, requiring any network to have a minimum
512 of 5 collinear genes and no more than 15 genes between a collinear block, settings that perform
513 well for analyzing mammal genomes (109). Using the infomap algorithm, we clustered the
514 synteny blocks into microsynteny networks, from which we extracted network clusters
515 corresponding to the list of keystone dental genes (43). For each dental gene hierarchical
516 orthogroup, we assessed whether genes of species with unrooted molars were missing from the
517 synteny networks that contained other Glires species' sequences, representing loss of synteny for
518 those species.
519
520 *Positive selection analysis*
521 We aligned protein sequences for each dental gene orthogroup with clustal omega (110)
522 using default settings. Based on universal translation tables, we obtained codon-based nucleotide
523 alignments with pal2nal (111), removing sites in which any species had an indel (i.e., ungapped)
524 and formatting the output for analysis in PAML (44). We pruned and unrooted the orthogroup
525 gene trees from OrthoFinder to contain only tips representing the genes in each synteny network
526 or orthogroup under analysis in PAML. We tested whether any of the genes were undergoing
527 positive selection using a likelihood ratio test comparing site-specific models of "nearly neutral"

528 and positive selection. In these models, ω, the ratio of nonsynonymous to synonymous
529 nucleotide substitutions (also known as dN/dS), can vary at each codon site. In the "nearly
530 neutral" model, ω can take values between 0 and 1, while the positive selection model allows
531 sites to assume ω values greater than 1 (46,112). We estimated κ (the ratio of transitions to
532 transversions) and ω from initial values of 1 and 0.5, respectively, for both tests.

533 Dental genes with significant site-specific positive selection or those lacking synteny in
534 species with unrooted molars formed the basis for our second set of positive selection tests using
535 a branch-and-site model of positive selection. This model allows ω to vary not only among
536 codon sites, but also between "foreground" and "background" lineages (46). We marked the
537 species with unrooted molars as foreground lineages, then ran the model twice: once with ω
538 unconstrained to detect sites undergoing positive selection only on foreground branches, and a
539 second time and with ω fixed to 1, or neutral selection. A likelihood ratio test of the two models
540 determined whether the lineage-specific positive selection model was more likely than a neutral
541 model, and Bayes Empirical Bayes analyses (46) produced posterior probabilities to identify
542 sites under positive selection.

543 Genes under positive selection also tend to have lower expression levels (48), thus we
544 compared expression of the genes with branch-and-site specific positive selection between the
545 prairie (unrooted molars) and the bank vole (rooted molars) to provide further support for
546 selective differences. We collected three biological replicates of first molars from both species at
547 three postnatal stages (P1, P15, and P21) and immediately preserved them at -80°C in lysis
548 buffer (Buffer RLT; Qiagen) supplemented with 40 μM dithiothreitol. RNA was extracted from
549 homogenized tissues using a RNeasy column purification kit (Qiagen). We assessed
550 concentration and purity of extracted RNA using a NanoDrop 2000 spectrophotometer

551 (ThermoFisher Scientific). Using 1 μg of RNA, we synthesized cDNA using a high-capacity
552 cDNA reverse transcription kit (ThermoFisher Scientific). We used 1 μL diluted cDNA (1:3 in
553 ddH_2O) and iTaq Universal SYBR Green Supermix (Bio-rad) in the Bio-rad CFX96 real-time
554 PCR detection system for qPCR experiments, producing three technical replicates for each
555 biological replicate. We normalized cycle threshold (CT) values of genes of interest to GAPDH
556 expression levels and calculated relative expression levels as $2^{-\Delta\Delta CT}$. A two-tailed unpaired t-test
557 calculated in Prism 9 measured whether expression of these genes significantly differed between
558 bank voles and prairie voles. The oligonucleotide primers for each species and gene are in
559 Additional file 6.
560
561 *Sequence and secondary structure evolution*
562 We performed ancestral sequence reconstruction on the codon sequences of the genes
563 that had evidence of branch-and-site specific positive selection to understand how the sequence
564 has changed through time. The gapped clustal omega alignments were the basis for ancestral
565 sequence reconstruction on the Glires species tree (Fig. 1) using pagan2 (113). For each gene, we
566 plotted amino acid substitutions at the site with potential positive selection. Finally, we predicted
567 secondary structures (i.e., helices, beta sheets, and coils) for each unrooted species' protein
568 sequence and the reconstructed ancestral sequence prior to the change at the site under positive
569 selection using the PSIPRED 4.0 protein analysis workbench (114,115). Comparing these
570 predictions across the phylogeny, we assessed how these substitutions at the site under selection
571 may affect the structure of each protein.
572
573 *Developmental gene expression*

574 We performed quality control and filtering of the short reads for the seven replicates of
575 first molar tissues at E13, E14, and E16 using the nf-core/rnaseq v. 3.11.2 workflow (116) for
576 comparability to previous mouse and rat analyses (43). RNAseq reads were evaluated and
577 adapter sequences were filtered using FastQC v. 0.11.9 (117) and Cutadapt v. 3.4 (118), and
578 ribosomal RNA was removed using SortMeRNA v. 4.3.4 (119). We then aligned trimmed
579 sequences to our bank vole annotation using Salmon v. 1.10.1 (120). Counts were then
580 normalized by gene length. We categorized gene count data into functional groups based on their
581 established roles in tooth bud development (43) using the one-to-one orthology list between our
582 bank vole genome and the mouse GRCm39.103 genome annotation generated from our
583 OrthoFinder output. Using the rlog function of DESeq2 (121), we normalized gene counts within
584 each functional group on a log2 scale. A permutation test assessed whether the mean counts of
585 the progression, shape, and double functional groups were significantly different from genes in
586 the tissue, dispensable, and "other" groups (which are potentially relevant later in development)
587 based on 10,000 resampling replicates of the dataset (43).
588 We also assessed differential expression between the bank vole first molar and published
589 mouse M1 data at the same three time points (GEO accession GSE142199 (43)), combining the
590 data based on the one-to-one orthology relationships used in the functional permutation analysis.
591 Using the mouse E13 molar as the reference level, we modeled expression as a response to
592 species (mouse or vole), embryonic day (E13, E14, or E16), and the interaction between species
593 and day. We considered as significant any gene with a log fold change greater than 1, log fold
594 change standard error less than 0.5, and false discovery rate adjusted p value less than 0.05.
595
596 TABLES

597 **Table 1 – Genes undergoing site-specific and branch-and-site-specific positive selection**

Gene	*Mus* transcript	*Myodes* transcript	Site	Branch-and-site
Aqp1	ENSMUST00000004774	Mglareolus_00011822	Yes	Yes
Col4a1	ENSMUST00000033898	Mglareolus_00032740	Yes	No
Dspp	ENSMUST00000112771	Mglareolus_00014030	Yes	Yes
Fgf20	ENSMUST00000034014	Mglareolus_00013079	Yes	No
Runx3	ENSMUST00000056977	Mglareolus_00033992	Yes	No
similar to *Runx3*	–	–	Yes	–*

598 Table 1 Legend: *HOG only contained four genes with one unrooted species' sequence, could

599 not be tested for branch-and-site specific selection.

600

601 **Table 2 – P-values of permutation tests between keystone gene categories in bank vole M1**

602 **at embryonic days 13, 14, and 16**

		Tissue	Dispensable	Dev. Process	Other
E13	**Progression**	*0.0310*	0.0942	*0.0436*	*0.0402*
	Shape	0.6431	0.9041	0.2289	0.0995
	Double	0.1292	0.1521	0.0716	0.0655
E14	**Progression**	*0.0136*	*0.0383*	*0.0437*	*0.0401*
	Shape	0.3115	0.4725	0.0922	*0.0454*
	Double	0.1288	0.0945	0.0709	0.0630
E16	**Progression**	*0.0140*	*0.0401*	*0.0303*	*0.0274*
	Shape	0.3770	1	0.1831	0.0662
	Double	0.1343	0.1099	0.0638	0.0596

603 Table 2 Legend: Italicized values are statistically significant ($p < 0.05$)

604

605 **Table 3 – QUAST assembly statistics for *de novo* bank vole (*Myodes glareolus*) genome**

606 **assemblies**

	Pseudohap1	Pseudohap1+LR	Pseudohap2	Pseudohap2+LR*
Largest contig	27939478	32658832	27937749	32657565
Total length	2434151515	2441426554	2434099357	2441472313
GC (%)	41.88	41.89	41.88	41.89
N50	4187179	4579815	4187179	4558134
N75	1689669	1818134	1687188	1810460
L50	170	153	170	154
L75	388	357	388	358
Ns per 100 kbp	1151.99	1030.75	1151.96	1030.48

607 Table 3 Legend: *assembly used for annotation and downstream analyses in this paper.

608

609 **Table 4 – Genomes used in orthology, synteny, and positive selection analyses**

Species	Assembly	Citation
Myodes glareolus	CUNY_Mgla_1.0	This paper
*Cavia porcellus**	Cavpor3.0	(122)
*Cavia aperea**	CavAp1.0	(123)
Marmota marmota	marMar2.1	(124)
*Microtus ochrogaster**	MicOch1.0	(125)
Mus musculus	GRCm39	(126)
*Oryctolagus cuniculus**	OryCun2.0	(122)
*Dipodomys ordii**	Dord_2.0	(122)
Jaculus jaculus	JacJac1.0	(127)

Rattus norvegicus	Rnor_6.0	(128)
Mus pahari	PAHARI_EIJ_v1.1	(129)
Mus caroli	CAROLI_EIJ_v1.1	(129)
Mus spretus	SPRET_EiJ_v1	(130)
Mus spicilegus	MUSP714	(131)
Cricetulus griseus	CHOK1GS	(132)
Mesocricetus auratus	MesAur1.0	(133)
Peromyscus maniculatus	HU_Pman_2.1	(134)
Nannospalax galili	S.galili_v1.0	(135)
*Octodon degus**	OctDeg1.0	(136)
Heterocephalus glaber (F)	HetGla_female_1.0	(137)
*Chinchilla lanigera**	ChiLan1.0	(138)
Urocitellus parryi	ASM342692v1	(139)
Ictidomys tridecemlineatus	SpeTri2.0	(140)
*Homo sapiens***	GRCh38	(141)

610 Table 4 Legend: *Species with unrooted molars; **Peptide annotation used as outgroup only in
611 OrthoFinder analysis.
612
613 FIGURES

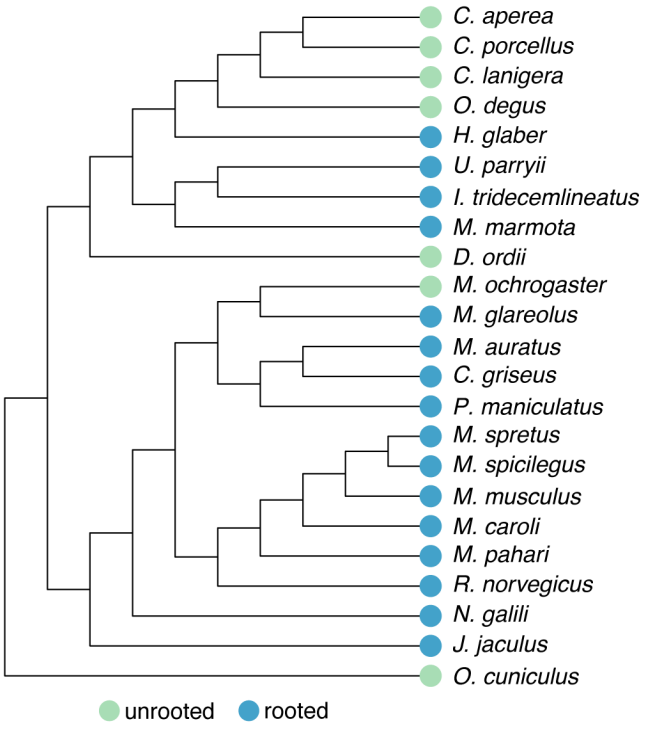


Figure 1 – Species tree of Glires based on the Ensembl Compara species tree. Whether each species has rooted or unrooted molars is indicated by colored circles at the tip of each branch. Note that unrooted, or hypselodont, molars have evolved multiple times across Glires. This topology was the basis for our orthology analysis.

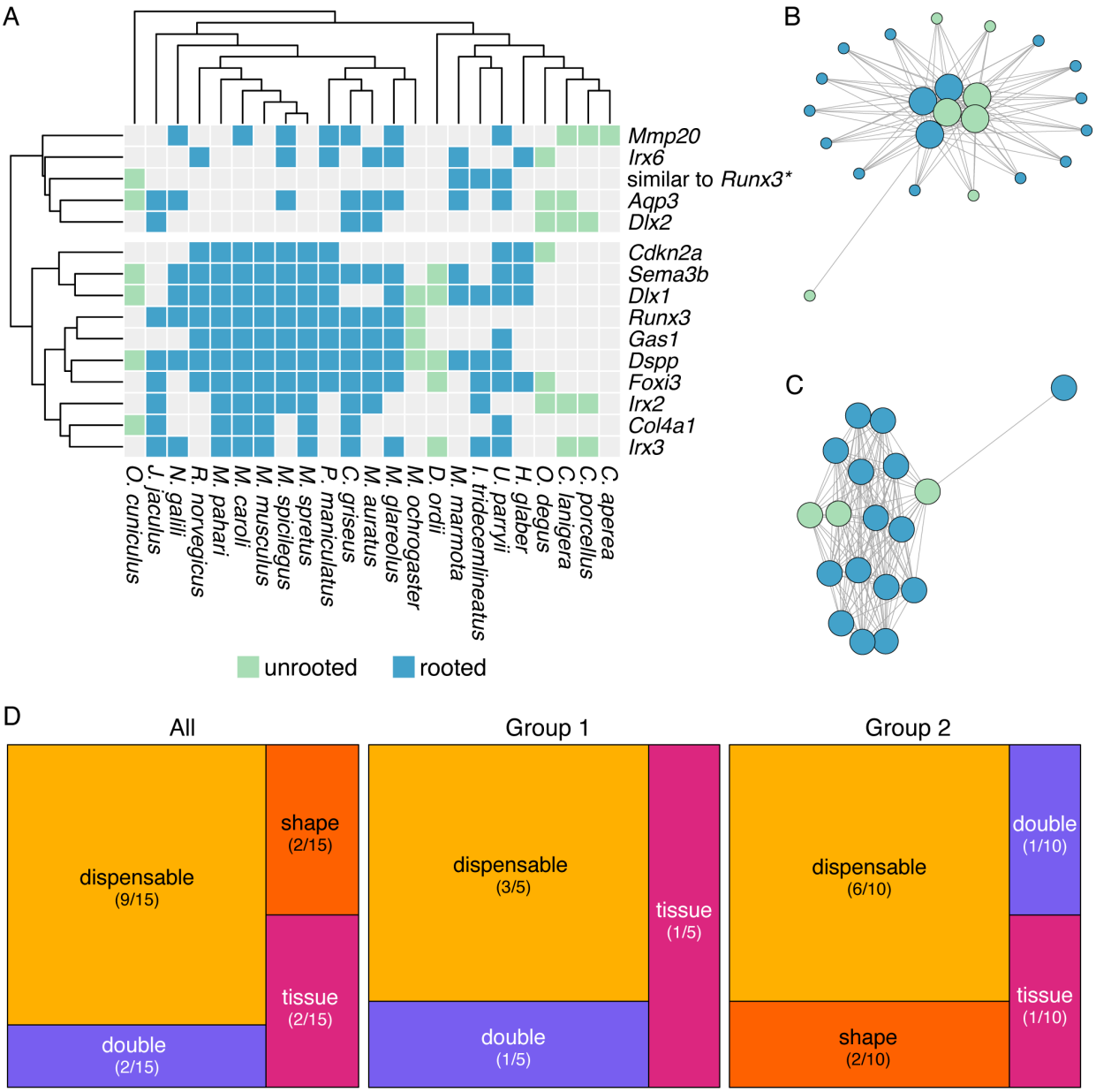


Figure 2 – A Presence (colored boxes) or absence (gray boxes) of gene sequences for each species in hierarchical orthogroups where fewer than half of the species with unrooted molars had conserved synteny. Columns are ordered according to phylogenetic positions (top) and rows are ordered by Euclidean distance clustering. Rows are split into two major groups: group 1, in which synteny is not conserved across Glires, and group 2, in which synteny is not conserved

625 mainly in species with unrooted molars. * = One hierarchical orthogroup represented only four
626 gene sequences annotated based on similarity to *Runx3*. **B** An example of a synteny network for
627 genes in Group 1, displayed using the Fruchterman-Reingold layout algorithm in the R package
628 *iGraph* (142). Small circles represent genes in the synteny network that are not part of the
629 hierarchical orthogroup, large circles represent genes in the hierarchical orthogroup, and lines
630 between circles represent a syntenic relationship between two species. Circle color represents
631 whether species has rooted or unrooted molars following the same key in A. **C** An example
632 synteny network for genes in Group 2, displayed using the Fruchterman-Reingold layout
633 algorithm in the R package *iGraph* (142). Circles represent genes in the hierarchical orthogroup,
634 and lines between circles represent a syntenic relationship between two species. Circle color
635 represents whether species has rooted or unrooted molars following the same key in A. **D**
636 Treemaps representing the keystone gene categories for all hierarchical orthogroups, the Group 1
637 hierarchical orthogroups, and the Group 2 hierarchical orthogroups. Most genes in each group
638 are in the "dispensable" keystone gene category, which includes genes that are dynamically
639 expressed during dental development but have no documented effect on phenotypes.

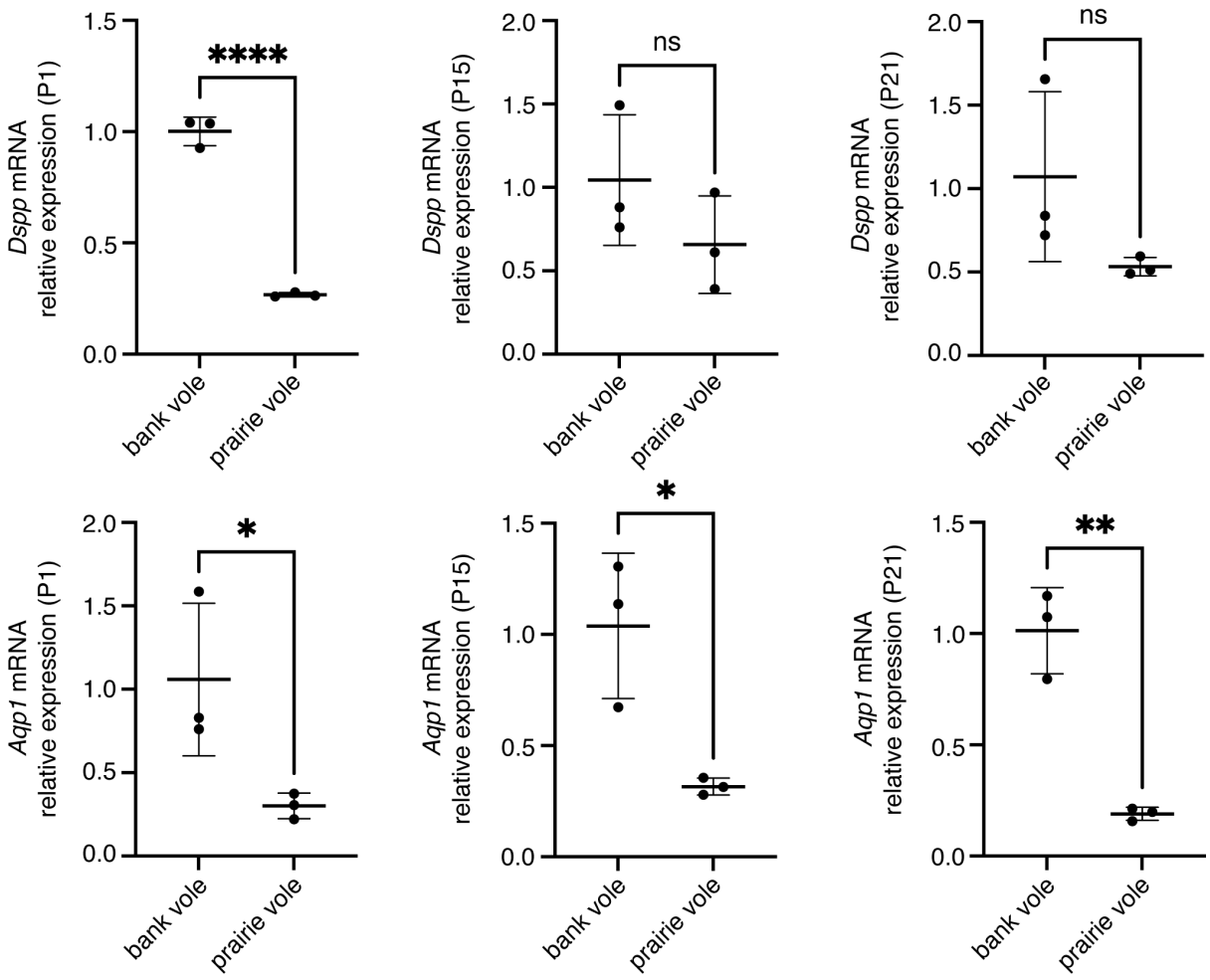
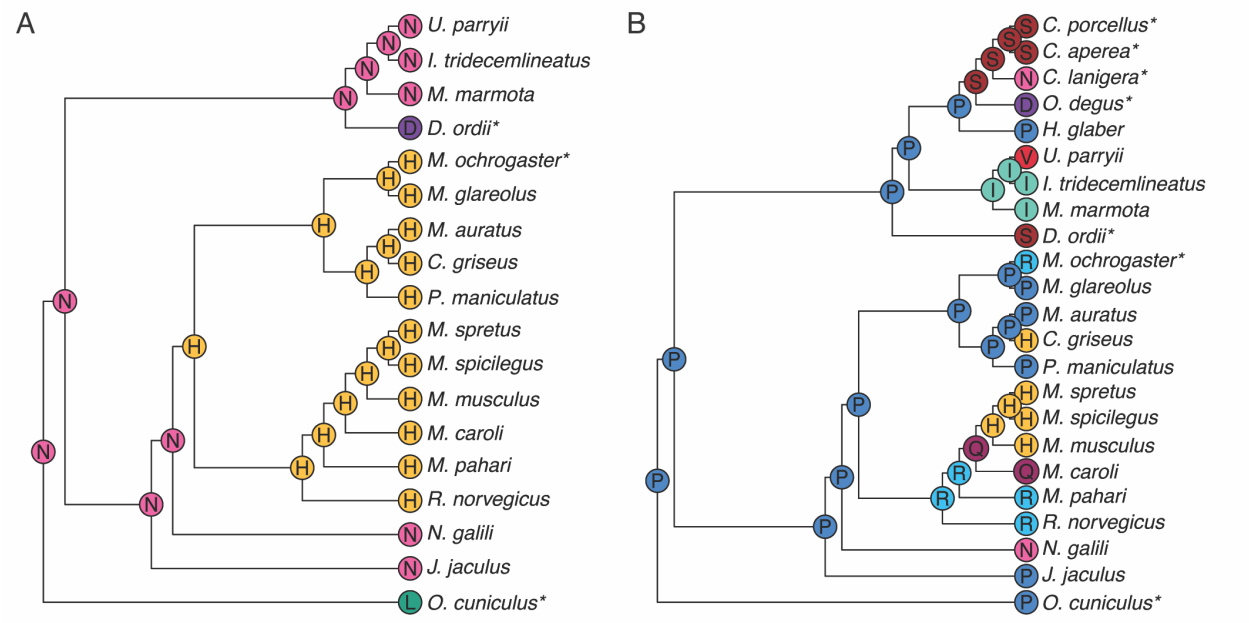


Figure 3 – Quantitative PCR comparisons of *Dspp* and *Aqp1* expression between bank vole and prairie vole M1 at postnatal days 1, 15, and 21 (P1, P15, P21). Expression levels for both genes are lower in the prairie vole (unrooted molars), which supports the positive selection detected for these genes in species with unrooted molars.

645



646 **Figure 4** – Ancestral state reconstructions of the residue under positive selection in PAML tests.

647 Letters at tips and internal nodes represent IUPAC codes for amino acids and * denotes species

648 with unrooted molars. **A** *Dspp*; **B** *Aqp1*.

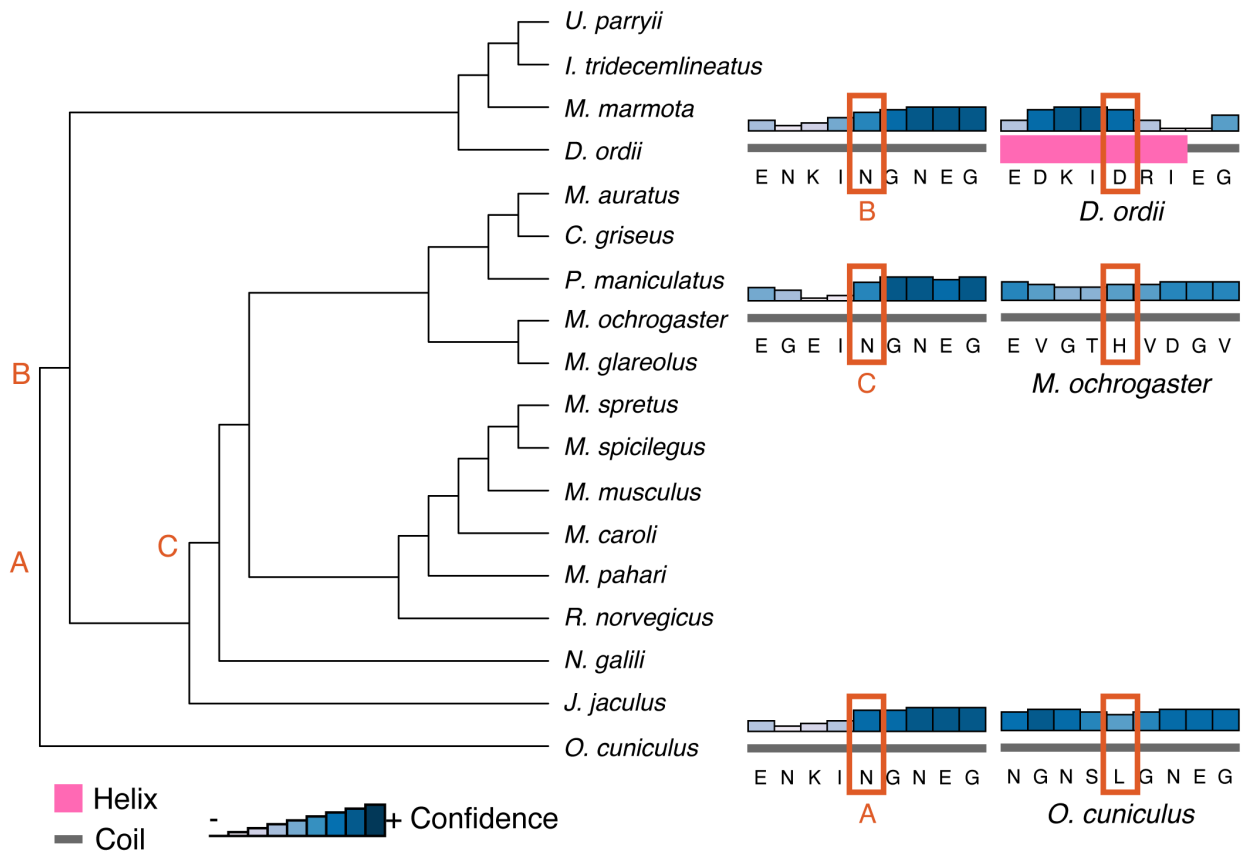


Figure 5 – PSIPRED secondary structure predictions for the three species with unrooted molars represented in the *Dspp* sequences. Letters correspond to the most recent ancestor of each tip species where the amino acid at the site under positive selection differed: A, the predicted ancestor of *O. cuniculus*; B, the predicted ancestor of *D. ordii*; and C, the predicted ancestor of *M. ochrogaster*. Structure predictions, the relative confidence of the prediction, and the amino acid sequence for each pair of extant species and ancestor are on the right.

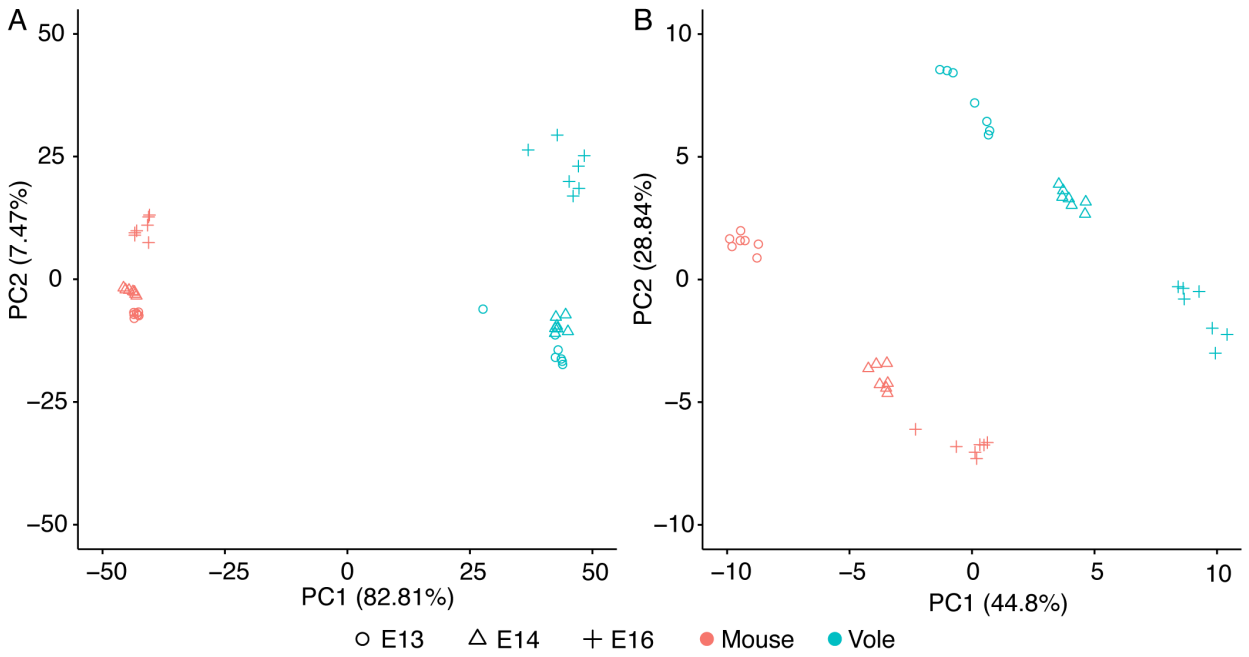


Figure 6 – Principal component (PC) analyses of differentially expressed genes in mouse and bank vole M1. **A** PC1 and PC2 of the 500 most variable genes, showing a clear differentiation between species along PC1 and differentiation between age classes along PC2. **B** PC1 and PC2 of the keystone dental genes. Both PC1 and PC2 separate age classes within, but not between, the species, likely due to differences in developmental timing and molar morphology between mice and voles.

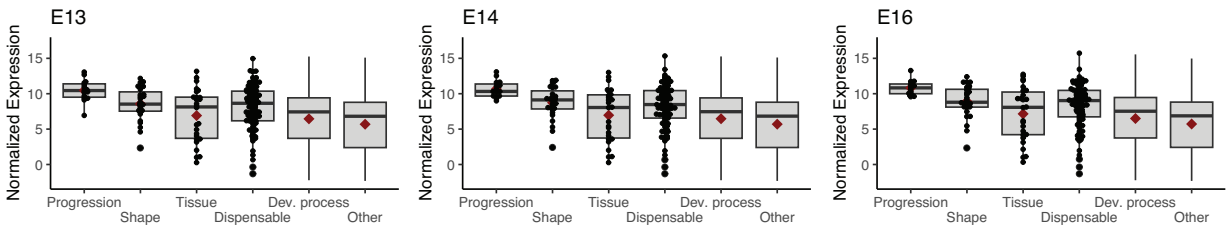


Figure 7 – Box and whisker plots showing normalized log base 2 expression levels for each keystone gene category in bank vole M1 at embryonic days 13, 14, and 16. Horizontal bar and

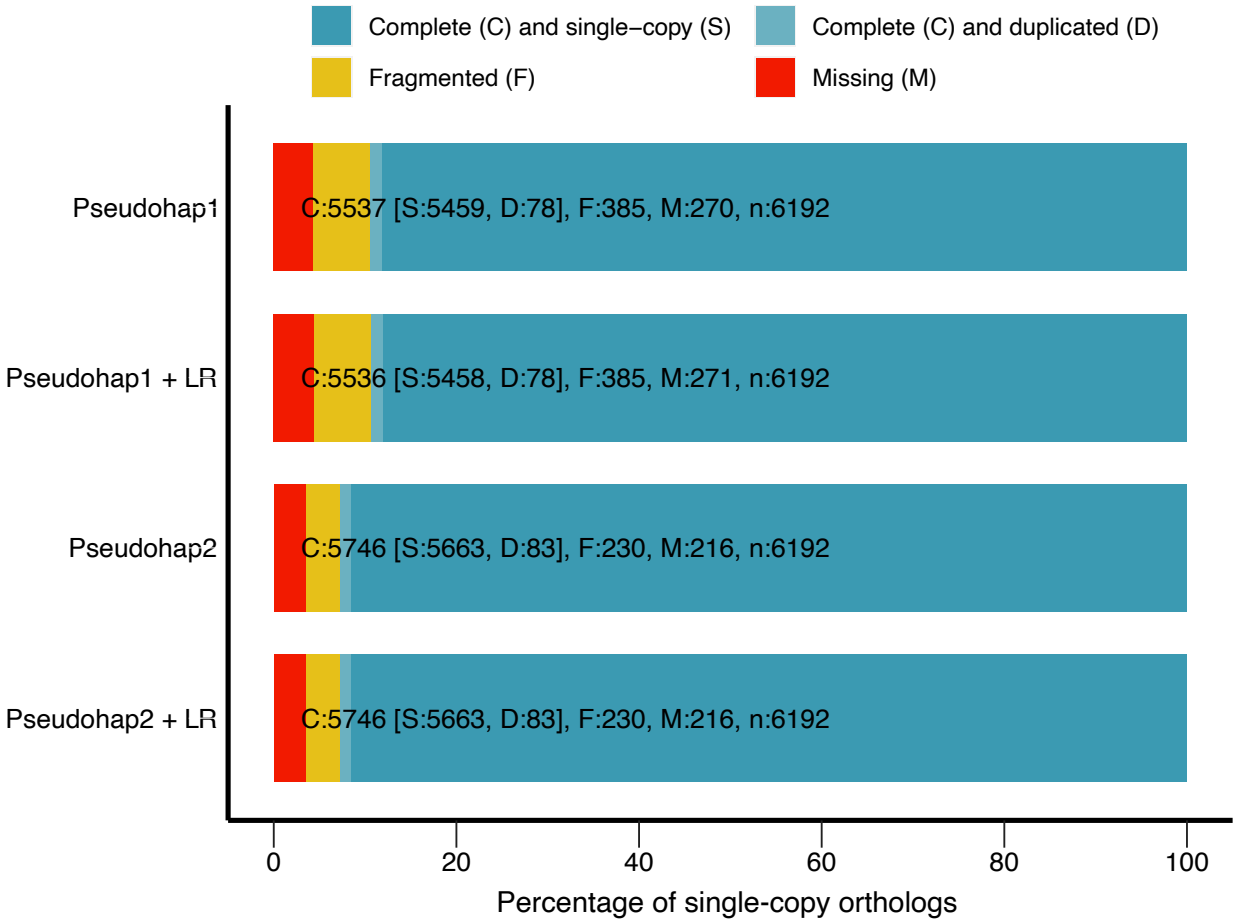
667 diamond within each box represent the median and mean values. Individual datapoints are

668 displayed for smaller keystone gene categories. Gene expression profiles at these stages are

669 comparable to mouse and rat molars at analogous developmental stages, as seen in Hallikas et al.

670 2021.

671



672

673 **Figure 8** – BUSCO single-copy ortholog recovery for each "pseudohaploid" version of our draft

674 bank vole genome assembly and these version after long-read scaffolding (denoted by "+ LR").

675 Each bar represents the cumulative proportion of the 6,192 single-copy orthologs for

676 Euarchontoglires identified by BUSCO represented by complete single-copy, complete-

duplicated, fragmented, and missing orthologs. The Pseudohap2 and Pseudohap2 + LR assemblies had the best single-copy ortholog recovery.

ADDITIONAL FILES

Additional file 1 [.xlsx] Dental gene results – Full table of orthology, synteny, and positive selection test results for all dental genes assessed.

Additional file 2 [.txt] *Dspp* gapped alignment – Gapped codon-based alignment for *Dspp* in fasta formatted sequences.

Additional file 3 [.txt] *Aqp1* gapped alignment – Gapped codon-based alignment for *Aqp1* in fasta formatted sequences.

Additional file 4 [.pdf] Structure predictions – PSIPRED Secondary structure predictions for each ancestral node and unrooted molar tip species for *Dspp* and *Aqp1*.

Additional file 5 [.txt] Custom repeat library – Custom repeat library of fasta formatted sequences used in annotation of the draft *Myodes glareolus* genome. See Methods for description of the process used to generate the library.

Additional file 6 [.pdf] Oligonucleotide primers – List of oligonucleotide primers for *Dspp*, *Aqp1*, and *GAPDH* used in bank vole and prairie vole qPCR experiments.

REFERENCES

1. Renvoisé E, Michon F. An Evo-Devo perspective on ever-growing teeth in mammals and dental stem cell maintenance. Front Physiol. 2014;5(324):1–12.

698 2. Tapaltsyan V, Eronen JT, Lawing AM, Sharir A, Janis C, Jernvall J, et al. Continuously
699 growing rodent molars result from a predictable quantitative evolutionary change over 50
700 million years. Cell Rep. 2015;11(5):673–80.
701 3. LeBlanc ARH, Brink KS, Whitney MR, Abdala F, Reisz RR. Dental ontogeny in extinct
702 synapsids reveals a complex evolutionary history of the mammalian tooth attachment
703 system. Proc R Soc B Biol Sci. 2018 Nov 7;285(1890):20181792.
704 4. Saffar JL, Lasfargues JJ, Cherruau M. Alveolar bone and the alveolar process: the socket that
705 is never stable. Periodontol 2000. 1997;13(1):76–90.
706 5. Davit-Béal T, Tucker AS, Sire JY. Loss of teeth and enamel in tetrapods: Fossil record,
707 genetic data and morphological adaptations. J Anat. 2009;214(4):477–501.
708 6. Damuth J, Janis CM. On the relationship between hypsodonty and feeding ecology in
709 ungulate mammals, and its utility in palaeoecology. Biol Rev. 2011;86(3):733–58.
710 7. Miletich I, Sharpe PT. Normal and abnormal dental development. Hum Mol Genet. 2003 Apr
711 2;12(suppl_1):R69–73.
712 8. Harada H, Kettunen P, Jung HS, Mustonen T, Wang YA, Thesleff I. Localization of putative
713 stem cells in dental epithelium and their association with Notch and FGF signaling. J Cell
714 Biol. 1999;147(1):105–20.
715 9. Tummers M, Thesleff I. Root or crown: a developmental choice orchestrated by the
716 differential regulation of the epithelial stem cell niche in the tooth of two rodent species.
717 Development. 2003;130(6):1049–57.
718 10. Thesleff I, Tummers M. Tooth organogenesis and regeneration. In: StemBook. Cambridge,
719 MA: Harvard Stem Cell Institute; 2008.

720 11. Krivanek J, Buchtova M, Fried K, Adameyko I. Plasticity of dental cell types in
721 development, regeneration, and evolution. J Dent Res. 2023 Jun 1;102(6):589–98.
722 12. Luan X, Ito Y, Diekwisch TGH. Evolution and development of Hertwig's epithelial root
723 sheath. Dev Dyn. 2006;235(5):1167–80.
724 13. Kumakami-Sakano M, Otsu K, Fujiwara N, Harada H. Regulatory mechanisms of Hertwig's
725 epithelial root sheath formation and anomaly correlated with root length. Exp Cell Res.
726 2014;325(2):78–82.
727 14. Wen Q, Jing J, Han X, Feng J, Yuan Y, Ma Y, et al. *Runx2* regulates mouse tooth root
728 development via activation of WNT inhibitor *NOTUM*. J Bone Miner Res.
729 2020;35(11):2252–64.
730 15. Yang S, Choi H, Kim TH, Jeong JK, Liu Y, Harada H, et al. Cell dynamics in Hertwig's
731 epithelial root sheath are regulated by β-catenin activity during tooth root development. J
732 Cell Physiol. 2021;236(7):5387–98.
733 16. Yamashiro T, Tummers M, Thesleff I. Expression of bone morphogenetic proteins and Msx
734 genes during root formation. J Dent Res. 2003;82(3):172–6.
735 17. Yokohama-Tamaki T, Ohshima H, Fujiwara N, Takada Y, Ichimori Y, Wakisaka S, et al.
736 Cessation of Fgf10 signaling, resulting in a defective dental epithelial stem cell
737 compartment, leads to the transition from crown to root formation. Development.
738 2006;133(7):1359–66.
739 18. Ota MS, Vivatbutsin P, Nakahara T, Eto K. Tooth root development and the cell-based
740 regenerative therapy. J Oral Tissue Eng. 2007;4(3):137–42.
741 19. Jernvall J, Thesleff I. Reiterative signaling and patterning during mammalian tooth
742 morphogenesis. Mech Dev. 2000;92:19–29.

743 20. Harada H, Toyono T, Toyoshima K, Yamasaki M, Itoh N, Kato S, et al. FGF10 maintains
744 stem cell compartment in developing mouse incisors. Dev Camb Engl. 2002;129(6):1533–
745 41.
746 21. Tapaltsyan V, Charles C, Hu J, Mindell D, Ahituv N, Wilson GM, et al. Identification of
747 novel *Fgf* enhancers and their role in dental evolution. Evol Dev. 2016;18(1):31–40.
748 22. Christensen MM, Hallikas O, Das Roy R, Väänänen V, Stenberg OE, Häkkinen TJ, et al.
749 The developmental basis for scaling of mammalian tooth size. Proc Natl Acad Sci. 2023
750 Jun 20;120(25):e2300374120.
751 23. Chen ZJ. Genetic and epigenetic mechanisms for gene expression and phenotypic cariation
752 in plant polyploids. Annu Rev Plant Biol. 2007;58(1):377–406.
753 24. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, et al. Relative impact
754 of nucleotide and copy number variation on gene expression phenotypes. Science. 2007 Feb
755 9;315(5813):848–53.
756 25. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution
757 of gene regulation. Nat Rev Genet. 2012 Jul;13(7):505–16.
758 26. de Montaigu A, Giakountis A, Rubin M, Tóth R, Cremer F, Sokolova V, et al. Natural
759 diversity in daily rhythms of gene expression contributes to phenotypic variation. Proc Natl
760 Acad Sci. 2015 Jan 20;112(3):905–10.
761 27. Erwin DH, Davidson EH. The last common bilaterian ancestor. Development. 2002 Jul
762 1;129(13):3021–32.
763 28. Irie N, Kuratani S. Comparative transcriptome analysis reveals vertebrate phylotypic period
764 during organogenesis. Nat Commun. 2011;2:248.

765 29. Koonin EV. Evolution of genome architecture. Int J Biochem Cell Biol. 2009 Feb
766 1;41(2):298–306.
767 30. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007
768 Mar;8(3):206–16.
769 31. Acemel RD, Maeso I, Gómez-Skarmeta JL. Topologically associated domains: a successful
770 scaffold for the evolution of gene regulation in animals. WIREs Dev Biol. 2017;6(3):e265.
771 32. Coghlan A, Eichler EE, Oliver SG, Paterson AH, Stein L. Chromosome evolution in
772 eukaryotes: a multi-kingdom perspective. Trends Genet. 2005 Dec 1;21(12):673–82.
773 33. Swenson KM, Blanchette M. Large-scale mammalian genome rearrangements coincide with
774 chromatin interactions. Bioinformatics. 2019 Jul 15;35(14):i117–26.
775 34. Long HS, Greenaway S, Powell G, Mallon AM, Lindgren CM, Simon MM. Making sense of
776 the linear genome, gene function and TADs. Epigenetics Chromatin. 2022 Jan 29;15(1):4.
777 35. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B.
778 Topologically associating domains are ancient features that coincide with Metazoan clusters
779 of extreme noncoding conservation. Nat Commun. 2017 Sep 5;8(1):441.
780 36. Szabo Q, Bantignies F, Cavalli G. Principles of genome folding into topologically
781 associating domains. Sci Adv. 2019 Apr 10;5(4):eaaw1668.
782 37. Das Roy R, Hallikas O, Christensen MM, Renvoisé E, Jernvall J. Chromosomal
783 neighbourhoods allow identification of organ specific changes in gene expression. PLOS
784 Comput Biol. 2021 Sep 10;17(9):e1008947.
785 38. Torelli F, Zander S, Ellerbrok H, Kochs G, Ulrich RG, Klotz C, et al. Recombinant IFN-γ
786 from the bank vole *Myodes glareolus*: a novel tool for research on rodent reservoirs of
787 zoonotic pathogens. Sci Rep. 2018;8(1):1–11.

788 39. Kloch A, Babik W, Bajer A, Siński E, Radwan J. Effects of an MHC-DRB genotype and
789 allele number on the load of gut parasites in the bank vole *Myodes glareolus*. Mol Ecol.
790 2010;19(SUPPL. 1):255–65.
791 40. Migalska M, Sebastian A, Konczal M, Kotlík P, Radwan J. *De novo* transcriptome assembly
792 facilitates characterisation of fast-evolving gene families, MHC class I in the bank vole
793 (*Myodes glareolus)*. Heredity. 2017;118(4):348–57.
794 41. Appleton J, Lee KM, Sawicka Kapusta K, Damek M, Cooke M. The heavy metal content of
795 the teeth of the bank vole (*Clethrionomys glareolus*) as an exposure marker of
796 environmental pollution in Poland. Environ Pollut. 2000;110:441–9.
797 42. Gdula-Argasińska J, Appleton J, Sawicka-Kapusta K, Spence B. Further investigation of the
798 heavy metal content of the teeth of the bank vole as an exposure indicator of environmental
799 pollution in Poland. Environ Pollut. 2004;131(1):71–9.
800 43. Hallikas O, Das Roy R, Christensen MM, Renvoisé E, Sulic AM, Jernvall J. System-level
801 analyses of keystone genes required for mammalian tooth development. J Exp Zoolog B
802 Mol Dev Evol. 2021;336(1):7–17.
803 44. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug
804 1;24(8):1586–91.
805 45. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for
806 detecting positive selection at the molecular level. Mol Biol Evol. 2005 Dec;22(12):2472–9.
807 46. Yang Z, Wong WSW, Nielsen R. Bayes Empirical Bayes inference of amino acid sites under
808 positive selection. Mol Biol Evol. 2005 Apr 1;22(4):1107–18.
809 47. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. Why highly expressed proteins
810 evolve slowly. Proc Natl Acad Sci. 2005 Oct 4;102(40):14338–43.

811 48. Kosiol C, Vinař T, Fonseca RR da, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of
812 positive selection in six mammalian genomes. PLOS Genet. 2008 Aug 1;4(8):e1000144.
813 49. Martincorena I, Luscombe NM. Non-random mutation: The evolution of targeted
814 hypermutation and hypomutation. BioEssays. 2013;35(2):123–30.
815 50. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. High burden
816 and pervasive positive selection of somatic mutations in normal human skin. Science. 2015
817 May 22;348(6237):880–6.
818 51. Keränen SVE, Åberg T, Kettunen P, Thesleff I, Jernvall J. Association of developmental
819 regulatory genes with the development of different molar tooth shapes in two species of
820 rodents. Dev Genes Evol. 1998;208(9):477–86.
821 52. Jernvall J, Keränen SVE, Thesleff I. Evolutionary modification of development in
822 mammalian teeth: Quantifying gene expression patterns and topography. Proc Natl Acad
823 Sci. 2000;97(26):14444–8.
824 53. Hughes AL. The evolution of functionally novel proteins after gene duplication. Proc R Soc
825 Lond B Biol Sci. 1997 Jan;256(1346):119–24.
826 54. Wagner A. Selection and gene duplication: a view from the genome. Genome Biol. 2002 Apr
827 15;3(5):reviews1012.1.
828 55. David KT, Oaks JR, Halanych KM. Patterns of gene evolution following duplications and
829 speciations in vertebrates. PeerJ. 2020 Mar 31;8:e8813.
830 56. Copley SD. Evolution of new enzymes by gene duplication and divergence. FEBS J.
831 2020;287(7):1262–83.
832 57. Fisher LW. DMP1 and DSPP: Evidence for duplication and convergent evolution of two
833 SIBLING proteins. Cells Tissues Organs. 2011 Aug;194(2–4):113–8.

834 58. Bouleftour W, Juignet L, Bouet G, Granito RN, Vanden-Bossche A, Laroche N, et al. The
835 role of the SIBLING, Bone Sialoprotein in skeletal biology — Contribution of mouse
836 experimental genetics. Matrix Biol. 2016 May 1;52–54:60–77.
837 59. Felszeghy S, Módis L, Németh P, Nagy G, Zelles T, Agre P, et al. Expression of aquaporin
838 isoforms during human and mouse tooth development. Arch Oral Biol. 2004 Apr
839 1;49(4):247–57.
840 60. Yoshii T, Harada F, Saito I, Nozawa-Inoue K, Kawano Y, Maeda T. Immunoexpression of
841 aquaporin-1 in the rat periodontal ligament during experimental tooth movement. Biomed
842 Res. 2012;33(4):225–33.
843 61. Zhang X, Zhao J, Li C, Gao S, Qiu C, Liu P, et al. *DSPP* mutation in dentinogenesis
844 imperfecta Shields type II. Nat Genet. 2001 Feb;27(2):151–2.
845 62. de La Dure-Molla M, Philippe Fournier B, Berdal A. Isolated dentinogenesis imperfecta and
846 dentin dysplasia: revision of the classification. Eur J Hum Genet. 2015 Apr;23(4):445–51.
847 63. Shields ED, Bixler D, El-Kafrawy AM. A proposed classification for heritable human
848 dentine defects with a description of a new entity. Arch Oral Biol. 1973 Apr 1;18(4):543-
849 IN7.
850 64. Sreenath T, Thyagarajan T, Hall B, Longenecker G, D'Souza R, Hong S, et al. Dentin
851 Sialophosphoprotein knockout mouse teeth display widened predentin zone and develop
852 defective dentin mineralization similar to human dentinogenesis imperfecta type III. J Biol
853 Chem. 2003 Jul 4;278(27):24874–80.
854 65. Verdelis K, Ling Y, Sreenath T, Haruyama N, MacDougall M, van der Meulen MCH, et al.
855 DSPP effects on *in vivo* bone mineralization. Bone. 2008 Dec 1;43(6):983–90.

856 66. Chen Y, Zhang Y, Ramachandran A, George A. DSPP is essential for normal development
857 of the dental-craniofacial complex. J Dent Res. 2016 Mar 1;95(3):302–10.
858 67. von Marschall Z, Mok S, Phillips MD, McKnight DA, Fisher LW. Rough endoplasmic
859 reticulum trafficking errors by different classes of mutant dentin sialophosphoprotein
860 (DSPP) cause dominant negative effects in both dentinogenesis imperfecta and dentin
861 dysplasia by entrapping normal DSPP. J Bone Miner Res. 2012;27(6):1309–21.
862 68. Smith BL, Preston GM, Spring FA, Anstee DJ, Agre P. Human red cell aquaporin CHIP. I.
863 Molecular characterization of ABH and Colton blood group antigens. J Clin Invest. 1994
864 Sep 1;94(3):1043–9.
865 69. Jordan IK, Mariño-Ramírez L, Koonin EV. Evolutionary significance of gene expression
866 divergence. Gene. 2005 Jan 17;345(1):119–26.
867 70. Warnefors M, Kaessmann H. Evolution of the correlation between expression divergence
868 and protein divergence in mammals. Genome Biol Evol. 2013;5(7):1324–35.
869 71. Jernvall J, Thesleff I. Tooth shape formation and tooth renewal: evolving with the same
870 signals. Development. 2012;139(19):3487–97.
871 72. Mitsiadis TA. Role of Islet1 in the patterning of murine dentition. Development.
872 2003;130(18):4451–60.
873 73. Charles C, Pantalacci S, Peterkova R, Tafforeau P, Laudet V, Viriot L. Effect of *eda* loss of
874 function on upper jugal tooth morphology. Anat Rec. 2009;292(2):299–308.
875 74. Zurowski C, Jamniczky H, Graf D, Theodor J. Deletion/loss of bone morphogenetic protein
876 7 changes tooth morphology and function in *Mus musculus*: implications for dental
877 evolution in mammals. R Soc Open Sci. 2018 Jan 3;5(1):170761.

878 75. Cardoso-Moreira M, Halbert J, Valloton D, Velten B, Chen C, Shao Y, et al. Gene
879 expression across mammalian organ development. Nature. 2019 Jul;571(7766):505–9.
880 76. Finarelli JA, Flynn JJ. Ancestral state reconstruction of body size in the Caniformia
881 (Carnivora, Mammalia): the effects of incorporating data from the fossil record. Syst Biol.
882 2006;55(2):301–13.
883 77. Welker F, Collins MJ, Thomas JA, Wadsley M, Brace S, Cappellini E, et al. Ancient proteins
884 resolve the evolutionary history of Darwin's South American ungulates. Nature. 2015
885 Jun;522(7554):81–4.
886 78. Warinner C, Korzow Richter K, Collins MJ. Paleoproteomics. Chem Rev. 2022 Aug
887 24;122(16):13401–46.
888 79. Zheng GXY, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping
889 germline and cancer genomes with high-throughput linked-read sequencing. Nat
890 Biotechnol. 2016 Feb;34:303.
891 80. Marks P, Garcia S, Martinez A, Belhocine K. Resolving the full spectrum of human genome
892 variation using linked-reads. 2017;
893 81. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid
894 genome sequences. Genome Res. 2017;27(5):757–67.
895 82. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
896 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017
897 Jul 15;33(14):2202–4.
898 83. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and
899 accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome
900 Res. 2017 May 1;27(5):722–36.

901 84. Warren RL. RAILS and Cobbler: Scaffolding and automated finishing of draft genomes
902 using long DNA sequences. J Open Source Softw. 2016 Nov 17;1(7):116.
903 85. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
904 assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.
905 86. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing
906 genome assembly and annotation completeness with single-copy orthologs. Bioinformatics.
907 2015 Oct 1;31(19):3210–2.
908 87. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use
909 annotation pipeline designed for emerging model organism genomes. Genome Res.
910 2008;18:188–96.
911 88. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: A tool
912 kit for the rapid creation, management, and quality control of plant genome annotations.
913 Plant Physiol. 2014 Feb 1;164(2):513–24.
914 89. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using
915 MAKER and MAKER-P. Curr Protoc Bioinforma. 2014 Dec 12;48:4.11.1-4.11.39.
916 90. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity:
917 reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat
918 Biotechnol. 2011 May 15;29(7):644–52.
919 91. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, et al. Dfam: a database of
920 repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013
921 Jan;41(Database issue):D70-82.
922 92. Caballero J, Smit AFA, Hood L, Glusman G. Realistic artificial DNA sequences as negative
923 controls for computational genomics. Nucleic Acids Res. 2014 Jul;42(12):e99.

924 93. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of
925 repetitive DNA families. Nucleic Acids Res. 2016 Jan 4;44(D1):D81–9.
926 94. Hu J, Zheng Y, Shang X. MiteFinder: A fast approach to identify miniature inverted-repeat
927 transposable elements on a genome-wide scale. In: 2017 IEEE International Conference on
928 Bioinformatics and Biomedicine (BIBM). 2017. p. 164–8.
929 95. Gremme G, Steinbiss S, Kurtz S. GenomeTools: A comprehensive software library for
930 efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol
931 Bioinform. 2013 May 1;10(03):645–56.
932 96. Smit A, Hubley R. RepeatModeler Open-1.0. 2008.
933 97. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method
934 employing protein multiple sequence alignments. Bioinformatics. 2011 Mar 15;27(6):757–
935 63.
936 98. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004 May 14;5(1):59.
937 99. Campbell MS. compare_annotations_3.2.pl [Internet]. 2015. Available from:
938 https://github.com/mscampbell/Genome_annotation/blob/master/compare_annotations_3.2.
939 pl
940 100. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and
941 comparison of annotated genomes. BMC Bioinformatics. 2009 Feb 23;10(1):67.
942 101. Liu D, Hunt M, Tsai IJ. Inferring synteny between genome assemblies: a systematic
943 evaluation. BMC Bioinformatics. 2018 Jan;19(1):26.
944 102. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome
945 comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015
946 Aug 6;16(1):157.

947 103. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple
948 sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002
949 Jul;30(14):3059–66.
950 104. Lefort V, Desper R, Gascuel O. FastME 2.0: A comprehensive, accurate, and fast distance-
951 based phylogeny inference program. Mol Biol Evol. 2015 Oct 1;32(10):2798–800.
952 105. Farrer RA. Synima: a Synteny imaging tool for annotated genome assemblies. BMC
953 Bioinformatics. 2017 Nov 21;18(1):507.
954 106. Wang Y, Tang H, DeBarry JD, Tan X, Li J, Wang X, et al. MCScanX: a toolkit for
955 detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res.
956 2012 Apr;40(7):e49.
957 107. Zhao T, Schranz ME. Network approaches for plant phylogenomic synteny analysis. Curr
958 Opin Plant Biol. 2017 Apr 1;36:129–34.
959 108. Zhao T, Holmer R, de Bruijn S, Angenent GC, van den Burg HA, Schranz ME.
960 Phylogenomic synteny network analysis of MADS-Box transcription factor genes reveals
961 lineage-specific transpositions, ancient tandem duplications, and deep positional
962 conservation. Plant Cell. 2017 Jun 1;29(6):1278–92.
963 109. Zhao T, Schranz ME. Network-based microsynteny analysis identifies major differences
964 and genomic outliers in mammalian and angiosperm genomes. Proc Natl Acad Sci. 2019
965 Feb 5;116(6):2165–74.
966 110. Sievers F, Higgins DG. Clustal Omega. Curr Protoc Bioinforma. 2014;48(1):3.13.1-
967 3.13.16.

968 111. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence
969 alignments into the corresponding codon alignments. Nucleic Acids Res. 2006 Jul
970 1;34(suppl_2):W609–12.
971 112. Wong WSW, Yang Z, Goldman N, Nielsen R. Accuracy and power of statistical methods
972 for detecting adaptive evolution in protein coding sequences and for identifying positively
973 selected sites. Genetics. 2004 Oct 1;168(2):1041–51.
974 113. Löytynoja A, Vilella AJ, Goldman N. Accurate extension of multiple sequence alignments
975 using a phylogeny-aware graph algorithm. Bioinformatics. 2012 Jul 1;28(13):1684–91.
976 114. Jones DT. Protein secondary structure prediction based on position-specific scoring
977 matrices. J Mol Biol. 1999 Sep 17;292(2):195–202.
978 115. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. Nucleic
979 Acids Res. 2019 Jul 2;47(W1):W402–7.
980 116. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core
981 framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020
982 Mar;38(3):276–8.
983 117. Andrews S. FastQC: a quality control tool for high throughput sequence data. [Internet].
984 2010. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
985 118. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.
986 EMBnet.journal. 2011 May 2;17(1):10–2.
987 119. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs
988 in metatranscriptomic data. Bioinformatics. 2012 Dec 1;28(24):3211–7.
989 120. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-
990 aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417–9.

991 121. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
992 RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550.
993 122. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution
994 map of human evolutionary constraint using 29 mammals. Nature. 2011
995 Oct;478(7370):476–82.
996 123. Weyrich A, Schüllermann T, Heeger F, Jeschek M, Mazzoni CJ, Chen W, et al. Whole
997 genome sequencing and methylome analysis of the wild guinea pig. BMC Genomics. 2014
998 Nov 28;15(1):1036.
999 124. Gossmann TI, Ralser M. *Marmota marmota*. Trends Genet. 2020 May;36(5):383–4.
1000 125. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1001 *Microtus ochrogaster*. Broad Inst [Internet]. 2012; Available from:
1002 https://www.ncbi.nlm.nih.gov/bioproject/72443
1003 126. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E,
1004 Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome.
1005 Nature. 2002 Dec 5;420(6915):520–62.
1006 127. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1007 *Jaculus jaculus*. Broad Inst [Internet]. 2012; Available from:
1008 https://www.ncbi.nlm.nih.gov/bioproject/72445
1009 128. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al.
1010 Genome sequence of the Brown Norway rat yields insights into mammalian evolution.
1011 Nature. 2004 Apr;428(6982):493–521.

1012 129. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome
1013 assembly of large and complex genomes using multiple references. Genome Res. 2018 Nov
1014 1;28(11):1720–32.
1015 130. Lilue J, Doran AG, Fiddes IT, Abrudan M, Armstrong J, Bennett R, et al. Sixteen diverse
1016 laboratory mouse reference genomes define strain-specific haplotypes and novel functional
1017 loci. Nat Genet. 2018 Nov;50(11):1574–83.
1018 131. Couger MB, Arévalo L, Campbell P. A high quality genome for *Mus spicilegus*, a close
1019 relative of house mice with unique social and ecological adaptations. G3
1020 GenesGenomesGenetics. 2018 May 24;8(7):2145–52.
1021 132. Chinese hamster CHOK1GS assembly and gene annotation. Horiz Eagle [Internet]. 2017;
1022 Available from: https://www.ensembl.org/Cricetulus_griseus_chok1gshd/Info/Annotation
1023 133. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1024 *Mesocricetus auratus*. Broad Inst [Internet]. 2012; Available from:
1025 https://www.ncbi.nlm.nih.gov/bioproject/77669
1026 134. Lassance JM, Hopi Hoekstra. Improved assembly of the deer mouse *Peromyscus*
1027 *maniculatus* genome. Harv Univ Hughes Med Inst [Internet]. 2018; Available from:
1028 https://www.ncbi.nlm.nih.gov/bioproject/494228
1029 135. Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, et al. Genome-wide adaptive
1030 complexes to underground stresses in blind mole rats *Spalax*. Nat Commun. 2014 Jun
1031 3;5(1):3966.
1032 136. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1033 *Octodon degu*. Broad Inst [Internet]. 2012; Available from:
1034 https://www.ncbi.nlm.nih.gov/bioproject/74595

1035 137. Keane M, Craig T, Alföldi J, Berlin AM, Johnson J, Seluanov A, et al. The Naked Mole Rat
1036 Genome Resource: facilitating analyses of cancer and longevity-related adaptations.
1037 Bioinforma Oxf Engl. 2014 Dec 15;30(24):3558–60.
1038 138. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1039 *Chinchilla lanigera*. Broad Inst [Internet]. 2012; Available from:
1040 https://www.ncbi.nlm.nih.gov/bioproject/68239
1041 139. V. Federov, Dalen L, Olsen RA, Goropashnaya AV, Barnes BM. The genome of the Arctic
1042 ground squirrel *Urocitellus parryii.* Inst Arct Biol [Internet]. 2018; Available from:
1043 https://www.ncbi.nlm.nih.gov/bioproject/477386
1044 140. Di Palma F, Alföldi J, Johnson J, Berlin A, Gnerre S, Jaffe D, et al. The draft genome of
1045 *Ictidomys tridecemlineatus.* Broad Inst [Internet]. 2012; Available from:
1046 https://www.ncbi.nlm.nih.gov/bioproject/61725
1047 141. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen HC, Kitts PA, et al. Evaluation
1048 of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of
1049 the reference assembly. Genome Res. 2017 May 1;27(5):849–64.
1050 142. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, et al. igraph: Network analysis
1051 and visualization in R [Internet]. 2024. Available from: https://CRAN.R-
1052 project.org/package=igraph

1053