G3: Genes|Genomes|Genetics Early Online, published on February 25, 2019 as doi:10.1534/g3.118.200714

Evolution of the highly repetitive PEVK region of titin across mammals

Kathleen Muenzen*, †, Jenna Monroy*, Findley R. Finseth*

*Keck Science Department, Claremont McKenna, Pitzer, and Scripps Colleges, Claremont, CA 91711, USA.

†Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA 98105, USA.

Running title: Titin evolution across mammals

Key words: Titin, PEVK region, comparative genomics, gene prediction, molecular evolution

***Corresponding authors:**

Findley R. Finseth, Keck Science Department, 925 N. Mills Avenue, Claremont McKenna, Pitzer, and Scripps Colleges, Claremont, CA 91711, USA;
ffinseth@kecksci.claremont.edu

Jenna Monroy, Keck Science Department, 925 N. Mills Avenue, Claremont McKenna, Pitzer, and Scripps Colleges, Claremont, CA 91711, USA;
jmonroy@kecksci.claremont.edu

ABSTRACT

The protein titin plays a key role in vertebrate muscle where it acts like a giant molecular spring. Despite its importance and conservation over vertebrate evolution, a lack of high quality annotations in non-model species makes comparative evolutionary studies of titin challenging. The PEVK region of titin—named for its high proportion of Pro-Glu-Val-Lys amino acids—is particularly difficult to annotate due to its abundance of alternatively spliced isoforms and short, highly repetitive exons. To understand PEVK evolution across mammals, we developed a bioinformatics tool, PEVK_Finder, to annotate PEVK exons from genomic sequences of titin and applied it to a diverse set of mammals. PEVK_Finder consistently outperforms standard annotation tools across a broad range of conditions and improves annotations of the PEVK region in non-model mammalian species. We find that the PEVK region can be divided into two subregions (PEVK-N, PEVK-C) with distinct patterns of evolutionary constraint and divergence. The bipartite nature of the PEVK region has implications for titin diversification. In the PEVK-N region, certain exons are conserved and may be essential, but natural selection also acts on particular codons. In the PEVK-C, exons are more homogenous and length variation of the PEVK region may provide the raw material for evolutionary adaptation in titin function. The PEVK-C region can be further divided into a highly repetitive region (PEVK-CA) and one that is more variable (PEVK-CB). Taken together, we find

39 that the very complexity that makes titin a challenge for annotation tools may also
40 promote evolutionary adaptation.
41
42 **INTRODUCTION**
43 One goal of modern biology is to connect the underlying molecular structure of a
44 protein with physiological function. Studies that compare protein sequences in an
45 evolutionary context can illuminate the regions of proteins that are most essential,
46 under functional constraint, or responding to natural selection (e.g., Perutz, 1983;
47 Hughes and Nei, 1989; Kreitman and Akashi, 1995; Galindo et al. 2003; Zhao et al. 2009;
48 Finseth et al. 2015). With the advent of next generation sequencing, the focus of such
49 studies has often moved beyond single gene analyses to comparing entire genomes or
50 transcriptomes (e.g., Finseth et al. 2014; Karlson et al. 2014; Zhang et al. 2014;
51 Pervouchine et al. 2015; Chikina et al. 2016; Partha et al. 2017). Yet, genes that are large,
52 repetitive, and/or poorly annotated can be left behind by this approach. One such
53 example is titin (TTN), a giant filamentous protein expressed in the muscles of all
54 bilaterian metazoans (Steinmetz et al. 2012) that plays a key role in muscle elasticity
55 (Linke, 2018) and is linked to various human muscular diseases (Savarese et al. 2016).
56
57 At nearly 4,000 kDa and greater than 1 micron in length, titin (also known as connectin)
58 is among the largest proteins found in vertebrates (Bang et al. 2001). Titin plays a

fundamental role in vertebrate striated muscle, where it acts as a giant molecular spring responsible for passive and active muscle elasticity (reviewed in Linke, 2018). Titin spans an entire half sarcomere from z-disk to m-line and its I-band region is composed of three domains: the proximal tandem Ig segment; the unique N2A (skeletal muscle), N2B (cardiac muscle), or N2BA (cardiac muscle) sequence; and the PEVK region (Linke et al. 1996; Gregorio et al. 1999). Together, these act as serially linked springs (Labeit and Kolmerer, 1995).

The general role titin plays in sarcomere structure is conserved among vertebrates; however, the elastic and physiological properties of vertebrate muscle vary dramatically across species and tissues. Titin, and the I-band in particular, are at least partially responsible for some of these functional differences (Freiburg et al. 2000, Prado et al. 2005, Trombitas et al. 2000) and likely contribute to variation in muscle physiology among tetrapods (Manteca et al. 2017). For example, cysteine residues in the I-band of titin may be responsible for mechanochemical evolution of vertebrate titin (Manteca et al. 2017). Likewise, variation in the length and structure of alternatively spliced titin isoforms can affect passive and active stiffness of muscles within a single species (Freiburg et al. 2000; Linke et al. 1998, Prado et al. 2005, Powers et al. 2016, Monroy et al. 2017). Moreover, mutations to the TTN gene have been associated with various cardiomyopathies and skeletal muscular dystrophies in humans (Savarese et al. 2016).

Despite its central role in muscle physiology, evolution, and disease, the TTN gene is often poorly or partially annotated in non-model species. One issue is the sheer size of TTN; it is composed of more than 100 kbp with a protein length of > 30,000 amino acids and the gene is more than 60 times the length of the average eukaryotic gene (Bang et al. 2001, Granzier et al. 2007). TTN also contains 364 exons (363 coding exons and a first non-coding exon) with lengths that range from just a few base pairs to greater than 17,000 base pairs, and theoretically can produce more than one million splice variants (Bang et al. 2001, Guo et al. 2010). The abundance of alternatively spliced isoforms means that numerous tissues, samples, and individuals are required for complete cDNA-based or RNA-based annotations.

The PEVK region of TTN, an important determinant of titin and muscle elasticity, also presents a problem for most annotation tools. The PEVK region contains over 100 short, repetitive exons consisting of ~ 70% proline (P), glutamate (E), valine (V), and lysine (K) residues. These features mean that PEVK exons are often missed by automated annotation tools and previous studies have relied upon manual annotation of this region (e.g., Freiburg et al. 2000; Granzier et al. 2007). Nevertheless, the PEVK region is potentially an important target of selection over evolutionary time; it is evolutionarily labile, varies in length, exon structure, and amino acid content across some vertebrates,

and may contribute to evolutionary adaptations in myofibril and whole muscle stiffness (Witt et al. 1998; Greaser et al. 2002; Granzier et al. 2007). The PEVK region also appears to have a hierarchical structure, with greater sequence divergence in the N-terminal among vertebrates than the C-terminal (Witt et al. 1998). With the wealth of genomic data now available, more detailed analyses of the PEVK region across vertebrates are possible and key to understanding its role in muscle evolution.

Here, we characterize the genomic structure of the PEVK region of TTN to gain insight into the structure-function relationships of I-band titin and its evolution across mammals. We first developed a custom tool, PEVK_Finder, to annotate exons within the PEVK region across diverse mammalian species. We then compared exon structure and sequence content both within one individual's PEVK region and across species. Finally, we performed evolutionary analyses to examine the nature of selection acting on the PEVK region of TTN.

MATERIALS AND METHODS

PEVK_Finder development and optimization

We developed a tool called PEVK_Finder to annotate the PEVK region of TTN across vertebrates (Table S1). PEVK_Finder annotates exons according to three criteria: 1) minimum exon length (12), 2) minimum PEVK ratio per exon (0.54), and 3) sliding window length (10; determination of optimal parameter values described below). For a given species, PEVK_Finder translates the complete TTN sequence into three forward reading frames and stores the resulting amino acid sequences in memory. PEVK_Finder then utilizes a sliding window approach to identify windows with a PEVK ratio above a minimum threshold. Note that the PEVK_Finder algorithm also incorporates the amino acid alanine (A) in its PEVK ratio calculations, based on previously observed PPAK motifs in titin (Greaser et al. 2002). However, all ratios reported in the text and figures of this paper will be referred to simply as *PEVK ratio*. All windows that meet the PEVK ratio requirement are stored in memory, and overlapping windows are combined into discrete sequences with unique start and end coordinates. To determine exonic boundaries, PEVK_Finder searches for paired donor and acceptor splice sites within each sequence using canonical mammalian nucleotide splicing patterns (Burset et al. 2000). Although these paired sites are traditionally used to define the 5′ (donor) and 3′ (acceptor) ends of intronic sequences, PEVK_Finder searches for acceptor and donor pairs of adjacent introns. For example, the acceptor site of an intron marks the 5′ end of a likely PEVK exon, while the donor site of the following intron marks the 3′ end of the exon. When no acceptor/donor pairs are found

137 in a given sequence, PEVK_Finder discards the current sequence and moves on to the
138 next likely PEVK exon. When multiple acceptor and/or donor sites are found in a given
139 sequence, the sub-sequence with the minimum distance between any pair of acceptor
140 and donor sites is considered the most likely PEVK exon. We confirmed that each non-
141 overlapping exon was represented by a single reading frame and combined all resultant
142 exons into a single, large exon set. The algorithmic workflow of PEVK_Finder is
143 summarized in Figure S13. The PEVK_Finder program was written in Python using the
144 Biopython bioinformatics tools package (Cock et al, 2009).

145

146 We optimized PEVK_Finder on the well-annotated human and mouse TTN sequences,
147 before applying the tool to a diverse set of mammalian species (Table S1). To determine
148 an optimal set of baseline parameters for PEVK_Finder, we subjected the untranslated
149 DNA reference sequence of human and mouse TTN to a range of parameter settings in
150 PEVK_Finder and compared those results to cDNA-annotated PEVK regions. The
151 complete human and mouse TTN reference DNA were downloaded from the NCBI
152 RefSeq database (O'Leary et al. 2016). The coordinates and DNA sequences for cDNA-
153 annotated human and mouse PEVK exons were gathered from the Ensembl genome
154 browser (Yates et al. 2016; Table S1). Human TTN exons 112-224 were defined as the
155 human PEVK region, as determined by Granzier et al. (2007). Although human TTN
156 exon 225 is also part of the PEVK region, the second ~half of the exon also encodes the

first exon of the distal tandem Ig segment. Exon 224 is the last exclusively-PEVK exon in the PEVK region, and exon 225 was therefore excluded from all further analyses. Note that we refer to human exon numbers according to the NCBI consensus CDS database and provide NCBI exon numbers, location, sequences, and additional numbering schemes in Table S2. Mouse TTN exons 109-207 were defined as the mouse PEVK region, as determined by manual inspection of cDNA-annotated sequences.

All possible combinations of minimum exon length (10-30), minimum PEVK ratio (0.45-0.83), and sliding window length (10-30) were used to generate ~17,000 PEVK exon sets per species. The parameter ranges used for optimization testing were determined by examining the minimum and mean exon lengths and PEVK ratios of all human and mouse PEVK exons. To determine the lower boundary for each parameter range, a number slightly below the minimum was chosen, and a number higher than the mean was chosen to determine the upper boundary. Each resulting exon set was compared with the corresponding cDNA annotated exon set for each species using the Basic Local Alignment Search Tool (BLAST; Altschul et al. 1990). Only exons with 100% identity as determined by BLAST were retained for the next steps of optimization. If multiple hits per exon met these criteria, only the hit with the highest bit score was retained.

A match score was calculated to determine how well parameter sets recover the annotated PEVK regions for the ~17,000 parameter combinations per species. The match score is a weighted score that prioritizes exons that recapitulate annotated exons ("recovered exons"; 70%), rewards identical exons ("perfect exons"; 10%), and minimizes exons identified by PEVK_Finder that are not found in the annotations ("extraneous exons"; 20%). *Recovered exons* were calculated as the proportion of the cDNA annotated exons identified by PEVK_Finder and included exons that generated both partial and full BLAST hits; *perfect exons* were defined as the proportion of recovered exons that were 100% identical for the entire length of the annotated exons; *extraneous exons* were calculated by subtracting the number of PEVK_Finder exons with no matches in the annotated database from 100 and dividing that number by 100. When the number of matchless exons found by PEVK_Finder was >100, the extraneous exons score was 0. The final match score was the sum of the three separate weighted scores and ranged from 0 to 1. Parameter space that encompassed the highest match scores were used to determine an optimal set of parameter ranges that yielded the most accurate sets of PEVK exons for a range of mammalian species. All match score scripts were written in R (R Core Team, 2016).

PEVK Finder validation

To evaluate the utility of PEVK_Finder, we compared the performance of PEVK_Finder on human and mouse TTN with a suite of popular gene annotation tools: Augustus (Stanke and Morgenstern, 2005), FGENESH (Salamov and Solovyev, 2000), geneid (Parra, Blanco and Guigó, 2000) and GENSCAN (Burge and Karlin, 1997). We used each tool with default parameters to predict the exon-intron coordinates of PEVK exons in both the human and mouse TTN sequences. Exon coordinates generated by each tool were manually compared with cDNA-annotated PEVK exon coordinates for the corresponding species. Exons generated by any of the four tools with a partial match (>= 50%) to a corresponding cDNA exon were considered a match, and a tool exon that spanned two or more cDNA exons was considered a single match. cDNA exons with no overlapping coordinates in the tool exon set were considered missing exons. Tool exons with no overlapping coordinates in the cDNA exon set were considered novel exons.

Phylogeny construction and PEVK region characterization across mammals

We downloaded 43 mammalian TTN sequences representing 16 major orders from the NCBI RefSeq database (Table S1). Our taxonomic sampling is a representative subset from Zhao et al. (2009), which generated a phylogeny of 59 mammals that evolved to fill diverse niches including subterranean, aquatic and nocturnal niches. The genomic TTN sequences for all 43 mammals in the species list were run through PEVK_Finder and

used to create novel PEVK exon sets. Forty-one of these exon sets were used for further analysis. To restrict PEVK exons to the PEVK region, terminal exons were manually determined based on the exon spacing and sequence composition patterns observed in the terminal exons of the cDNA-annotated human and mouse PEVK regions. Specifically, the first PEVK exon (human exon 112) has an amino acid sequence identical or nearly identical to "EIPPVVAPPIPLLLPTPEEKKPPPKRI" and is ~3,000 or fewer nucleotides before human TTN exon 114, which has an ortholog in all but one species and can thus be easily recognized. The last PEVK exon has an amino acid sequence identical or nearly identical to "AKAPKEEAAKPKGPI" and is generally ~3,000 or fewer nucleotides after the second to last exon. Exons identified by PEVK_Finder that are located beyond these landmarks may indeed be true PEVK exons, but all exon sets used for further analyses were restricted to this exon range for the sake of remaining consistent with previous studies of the PEVK region.

Using a full time-calibrated phylogeny of extant mammalian species, we created a phylogeny of 39 of the 41 mammalian species in our study to compare PEVK exons across vertebrates and set duck-billed platypus as the outgroup (Kumar et al. 2017). Two bat species, *Myotis lucifugus* and *Myotis davidii*, were not included in the phylogeny because their PEVK regions recapitulated those of other closely related bat species already represented in the figure. Phylogeny importation and editing was performed in

R using the Phytools package (Revell, 2012; R Version 3.3.2, R Core Team, 2016). To facilitate visual comparisons of PEVK regions, we generated exon-intron diagrams for each species with a custom R script that plotted the coordinates of PEVK exons in each exon set, color-coded by percent PEVK per exon.

PEVK_Finder exons were compared with cDNA-based annotations or *in silico* predicted exons for the 41 mammalian exon sets. The NCBI gene prediction program Gnomon (Souvorov et al. 2010) predicts the optimal coding sequence using a combination of partial alignments and *ab initio* modeling. PEVK_Finder annotations were compared with cDNA-based Gnomon predictions when cDNA was available and were compared with *in silico* Gnomon predictions when no cDNA was available. For simplicity, we refer to both sets of predictions as "Gnomon exons." We calculated exons identified by PEVK_Finder or Gnomon, exons that were identified by both ("consensus exons"), and the total number of unique exons identified by both tools. Mean differences between Gnomon and PEVK_Finder were determined using a paired t-test, pairing by species.

Evolutionary analyses

To facilitate comparisons among PEVK_Finder exons, we attempted to identify "orthologous exons". We use the term "orthologous" to distinguish exons that descend

254 from a common DNA ancestral sequence from those that are related by duplication (i.e.
255 paralogous). Orthologous PEVK exons among species were determined using the
256 reciprocal best BLAST method, with the exception that orthologs were determined for
257 each exon rather than the entire TTN gene (Tatusov, 1997; Bork and Koonin, 1998;
258 Koonin, 2005). The nucleotide sequence of the human PEVK exon set was used as the
259 query and compared with exon sets from the other species. Potential orthologs were
260 called when a given human exon's top BLASTn species hit (minimum e-value,
261 maximum bit score among hits) returned the original query exon when performed
262 reciprocally. Only orthologs found in the PEVK exon sets of >10 species (including
263 human) were included in further analyses. The remaining sets of orthologous pairs are
264 referred to as "confident orthologs". Confident orthologs were aligned by codon with
265 the ClustalW (Goujon et al. 2010) algorithm as implemented in MEGA7 (Kumar et al.
266 2016), using all other default parameter settings. Codons missing >15% of data among
267 species were eliminated from the alignment. We estimated average evolutionary
268 divergence over all sequence pairs for each confident ortholog following Tamura et al.
269 (2004) as implemented in MEGA7 (Kumar et al. 2016). Values for each ortholog were
270 calculated in MEGA7 with default parameters and 100 bootstrap replications.

271

272 We also evaluated repetition and duplication within the PEVK region by comparing all
273 PEVK exons for a single species with each other. For a given species, all possible exon

pairs in the PEVK exon set were aligned, and pairwise nucleotide substitutions per exon were calculated using MEGA Proto and MEGA-CC (Kumar et al. 2012). The PEVK region was divided into the subregions PEVK-N and PEVK-C, defined in *Results: PEVK region hierarchical structure*. The pairwise exon alignments were divided into quadrants representing PEVK-C:PEVK-N (Quadrant I), PEVK-C:PEVK-C (Quadrant II), PEVK-N:PEVK-N (Quadrant III), and PEVK-N:PEVK-C (Quadrant IV) comparisons. For each species, we calculated mean substitutions per exon for all quadrants, and results were compared with a one-way ANOVA and a Tukey's HSD post-hoc test. To evaluate repetition at the nucleotide level, we concatenated exonic sequences from the human PEVK-C region and generated a dot plot in the Dotlet web browser (Junier and Pagni, 2000). We additionally made self-dot plots using the R Dotplot package (Delaney, 2017) for all species using the full PEVK-C genomic sequence for each species, including intronic sequences. Finally, we used a BLAST search to find human LINE-1 elements in other mammals. We used the intronic sequence between human TTN exons 180 and 181 as the query and compared with all 41 mammal TTN sequences.

The PEVK-C region was further divided into two subregions PEVK-CA and PEVK-CB for single-species analysis, also defined in *Results: PEVK region hierarchical structure*. Pairwise exon alignments were divided into ninths representing PEVK-N:PEVK-CB (Box i), PEVK-CA:PEVK-CB (Box ii), PEVK-CB:PEVK-CB (Box iii), PEVK-N:PEVK-CA

294 (Box iv), PEVK-CA:PEVK-CA (Box v), PEVK-CB:PEVK-CA (Box vi), PEVK-N:PEVK-N
295 (Box vii), PEVK-CA:PEVK-N (Box viii), and PEVK-CB:PEVK-N (Box ix) comparisons.
296 Mean substitutions per exon were calculated for all species and compared using the
297 same statistical measures used for the quadrant comparisons described above.

298

299 We tested for positive selection acting on individual codons using site models as
300 implemented in the Phylogenetic Analysis by Maximum Likelihood (PAML) software
301 package (Yang, 2007). For each ortholog alignment, a phylogenetic tree was estimated
302 using the M0 model, then models M0 (one ratio), M1a (neutral), and M2a (positive
303 selection), were run. For each model, ω was set first at 0.5, then at 1 and then lastly at 3.
304 We recorded Lnl values for each model and test statistics for two model comparisons:
305 M0 vs. M1a and M1a vs. M2a. The PAML $\chi 2$ calculator was used to determine the p-
306 value of each comparison. All sites under selection, as determined by Bayes Empirical
307 Bayes analysis (Yang et al. 2005), were recorded for estimates with posterior
308 probabilities > 0.5.

309

310 Finally, we tested for charge shifts in codons under positive selection. For each codon,
311 the corresponding amino acid for each species was assigned one of four values based on
312 residue charge: hydrophobic/uncharged, polar, positively charged, or negatively

charged. The ancestral charge status of each node in the corresponding mammal tree was estimated using an equal rates (ER) model. Ancestral state reconstruction was performed in R using the Phytools package (Revell, 2012; R Core Team, 2016).

Data Availability

The PEVK_Finder source code is available for download at https://github.com/kmuenzen/pevk_finder_public. Tables S1 and S2 contain species names, NCBI accession numbers and TTN exon numbers for sequences used in this study. Scripts used in this study have been made available at https://github.com/kmuenzen/pevk_finder. The most up-to-date version of the PEVK_Finder source code can be downloaded from https://github.com/kmuenzen/pevk_finder_public.

RESULTS

PEVK_Finder parameter optimization

PEVK_Finder was developed to annotate the PEVK region of TTN across mammals. We determined optimal parameters (exon length, window size, and PEVK ratio) for downstream applications by comparing PEVK_Finder results with the human and

mouse annotations and calculating a match score. For each human and mouse model, we tested a total of ~17,000 parameter combinations. Optimal match scores for human were achieved when minimum exon length was 10-15 nucleotides, PEVK ratio was 0.54-0.55, and window length was 10 nucleotides (Figure 1a). Optimal match scores for mouse were achieved when minimum exon length was 12-14 nucleotides, PEVK ratio was 0.53-0.54, and window length was 10 nucleotides (Figure S1a). Default consensus parameters were therefore set as follows: window length = 10 nucleotides, PEVK ratio = 0.54, and minimum exon length = 12 nucleotides.

Using the optimal parameter values, PEVK_Finder identified 109 exons in the PEVK region in human TTN (113 annotated) and 93 PEVK exons in mouse TTN (99 annotated). In humans, 106 PEVK_Finder exons recovered annotated exons (94%) and 70 of these were perfect matches. Of the 36 non-perfect exon matches, 7 were overestimations of exon length and 29 were underestimations. When PEVK_Finder exon length was >50% reduced relative to the annotated length, the annotated exons were often long, with low PEVK ratios. In these cases, PEVK_Finder tended to truncate the exons in regions due to low PEVK ratio (e.g., human TTN exon 114). When PEVK_Finder overestimated exon length, all PEVK_Finder exons were <100 nucleotides long and length differences were marginal (3-4 amino acids difference). Patterns were similar with the mouse TTN, with the exception that overestimations were more

353 frequent (12 of 37 non-perfect exon matches), possibly due to non-canonical splice
354 sequences (Figure S1c).

355

356 **PEVK_Finder validation**

357 To evaluate the utility of PEVK_Finder, we compared results generated from
358 PEVK_Finder with four different gene annotation tools. Overall, PEVK_Finder
359 outperformed all other tools when identifying PEVK exons in the TTN gene.
360 PEVK_Finder identified 97% of human TTN exons, while no other tool identified more
361 than 65% of human PEVK exons (Figure 1b). PEVK_Finder also missed far fewer exons
362 than GENSCAN in both human and mouse TTN (grey circles Figures 1c, S1c).
363 PEVK_Finder was also the only tool that identified >80% of both human and mouse
364 exons in the PEVK region (Figures 1b, S1b). FGENESH also identified a high
365 proportion of mouse exons (81%), but was far less effective than PEVK_Finder at
366 annotating human exons (62%). PEVK_Finder performed slightly better on human
367 TTN (94%) than mouse TTN (90%). PEVK_Finder also annotated exons across a range
368 of PEVK ratios and in regions of short, high-density exons (Figures 1c, S1c); these
369 regions were a challenge for GENSCAN. In the few instances when PEVK_Finder did
370 fail to annotate exons, it was usually those exons with low PEVK ratios. GENSCAN
371 also tended to lump distinct exons together, likely due to missed splicing sites.
372 PEVK_Finder, FGENESH, and GENSCAN could be applied to both mouse and human

373 TTN, whereas geneid and Augustus are only suitable for human data. In addition to
374 identifying known exons, PEVK_Finder discovered novel putative PEVK exons in both
375 the human (1; Figure 1c) and mouse (2; Figure S1c) TTN sequences. However, the
376 "novel" exon identified in human TTN is technically outside the bounds of the pre-
377 defined PEVK region (exon 112-224).

378

379 We extended our evaluation of PEVK_Finder by evaluating its performance on other
380 species with annotated TTN sequences. Of the 41 mammalian species tested, only five
381 have cDNA data for TTN in NCBI (Table S3). Human TTN has the most complete
382 annotation, with consensus cDNA from 11 isoforms. The American Pika (*Ochotona*
383 *princeps*) has cDNA data from 5 isoforms, the house mouse (*Mus musculus*) and the
384 Gairdner's Shrewmouse (*Mus pahari*) have cDNA data from 3 isoforms each, and the
385 Chinese Rufous Horseshoe Bat (*Rhinolophus sinicus*) has cDNA data from 1 short
386 isoform. For the remaining 36 species, we compared PEVK_Finder results with the
387 annotations provided by NCBI generated by the Gnomon program (Souvorov et al.
388 2010). Across all species, PEVK_Finder identified significantly more exons (mean: 89.4,
389 SD: 8.33) than the Gnomon-based annotations (mean: 73.9, SD: 10.6; t = 9.37, df = 40, P =
390 $1.23*10^{-11}$; Table S3). The number of exons identified by PEVK_Finder represents a
391 significantly higher percentage of the total exons (mean: 93.6%, SD: 2.35) than Gnomon-
392 based annotations (mean: 76.7%, SD: 9.29; t = 11.10, P = $8.71*10^{-14}$; Table S3).

PEVK_Finder also identified many more putative novel PEVK exons that were not described previously by either Gnomon or cDNA (mean: 22.7; SD: 10.17).

PEVK region hierarchical structure

PEVK_Finder successfully generated exon sets for 41 of the 43 species tested; we were unable to obtain exon sets for two species (*Sorex araneus* and *Leptonychotes weddellii*) due to non-canonical splicing sequences that were not compatible with those used by PEVK_Finder to identify exon-intron boundaries. The PEVK region of the remaining 41 species exhibited structural similarities in both exon-intron spacing and PEVK content per exon (Figure 2). The total number of exons identified by PEVK_Finder ranged from 81 – 116 exons, 74 – 109 of which were within the bounds of the exons defining the start and end of the PEVK region (Table S3). Interestingly, the species with the smallest number of total PEVK exons are either aquatic diving mammals (*Odobenus rosmarus* – walrus (81), *Tursiops truncatus* – dolphin (79), *Neomonachus schaunslandi* – Hawaiian monk seal (83)) or a burrowing mammal, *Condylura cristata* (82). However, two of these species also show gaps in their PEVK regions, which may be real or may be due to assembly artifacts (see below).

412 Exon structure of the upstream ~half of the PEVK region is relatively conserved over
413 mammalian evolution, while the downstream ~half varies considerably. Based on this
414 observation, we divided the PEVK region into two distinct regions: a structurally
415 conserved region ("PEVK-N"; human exons 112– 165) and a structurally variable region
416 ("PEVK-C"; human exons 166 – 224). Because the end of the PEVK-N region and the
417 beginning of the PEVK-C region varies 1-5 exons between species, a visual landmark
418 was used to manually determine the PEVK-N/C boundary for each species. The
419 manually determined PEVK-N/C boundary is preceded by 3-4 very closely spaced, low
420 PEVK-ratio exons, and is followed by 3-4 more low-PEVK ratio exons and a dense, high
421 PEVK-ratio region (Figure 2). Across mammals, intron/exon boundaries and the
422 number of exons are more similar across the PEVK-N (mean: 50.00, SD: 2.51) versus the
423 PEVK-C region (mean: 39.46, SD: 8.12, $f = 0.10$, $P < 1.70*10^{-11}$) (Figure 2, Table S3).
424 Several species (e.g., chimpanzee, dolphin, walrus and alpaca) also show large deletions
425 of PEVK exons in the PEVK-C region. The walrus TTN sequence contains a large string
426 of >9,000 'N' bases, indicating that the gap in the walrus PEVK-C region is likely due to
427 low assembly quality in this region. Alpaca TTN also contains several strings of N's
428 that range from 500-5,000 bases long, which are likely responsible for the large gaps.
429 However, while several small strings of 10-150 'N' bases are found in the PEVK-C
430 region of chimpanzee and dolphin TTN, no large strings of Ns are present. The large
431 deletions in the chimpanzee and dolphin may therefore represent true PEVK exon

"deserts". Alternatively, the gaps could also be an artifact of non-canonical splice sites or small strings of N's interfering with splice site detection or PEVK ratios. Complete genomic sequences of TTN and/or RNA-based annotations could improve resolution of PEVK exons in these species.

PEVK region sequence variability within an individual

Although the PEVK-C region is structurally variable across mammals, its sequence content is more similar within a given individual than the PEVK-N region (Figures 2, 3). Exons in the PEVK-C region tend to be short (mean: 75.76, SE: 0.685) with high PEVK ratios (mean: 0.84, SE: 0.003), whereas those in the PEVK-N region consist of a range of lengths (mean: 93.29, SE: 1.37: $t = 11.44$, $f = 8.68$, $P < 2.2*10^{-16}$) and lower PEVK ratios (mean: 0.79, SE: 0.003: $t = 12.84$, $P < 2.2*10^{-16}$) (Figure 2). Likewise, exon-exon identity comparisons (Figure 3a) and dot plots of nucleotide sequence (Figures 3b, S14-S23) reveal large blocks of highly repetitive sequence in the PEVK-C region. When looking across all species, the mean substitutions per exon varied significantly among the four quadrants (one-way ANOVA: $F_{3,160} = 37.62$; $P = 2 * 10^{-16}$; Figure 3c). Post hoc comparisons using the Tukey HSD test indicated that the mean number of substitutions per exon for PEVK-C:PEVK-C exon comparisons (mean: 1.01; SD: 0.18; Quadrant II) were significantly lower than comparisons for PEVK-C:PEVK-N exons (mean: 1.27; SD:

0.12; $P < 0.0001$; Quadrant I), PEVK-N:PEVK-N exons (mean: 1.22; SD: 0.09; $P < 0.0001$; Quadrant III), and PEVK-N:PEVK-C exons (mean: 1.27; SD: 0.12; $P < 0.0001$; Quadrant IV). No other quadrant: quadrant comparisons were significantly different ($P > 0.05$ in all cases). Together, these analyses suggest that exons within the PEVK-C region of an individual have more similar sequences, while exons in the PEVK-N region are more variable in sequence.

The observed sequence variability within an individual's PEVK-C region (Figure 3a, Quadrant II; self-dot plots of the PEVK-C region for each species, Figures S14-23) suggested that the PEVK-C could be further subdivided into a highly similar PEVK-CA region (exons 166-215 in human TTN) and a more variable PEVK-CB region (~last 9 exons of the PEVK-C). The PEVK-CA/PEVK-CB boundary for each species was manually determined using a visual landmark, where the final exon in the densest part of the PEVK-C region marked the last exon of the PEVK-CA region, and the first exon of the PEVK-CB region directly followed the last PEVK-CA exon. The mean substitutions per exon varied significantly among the nine subregion comparisons (one-way ANOVA: $F_{8,360} = 211.6$; $P = 2 * 10^{-16}$; Figure S25). Post hoc comparisons indicated that the number of substitutions per exon for PEVK-CA:PEVK-CA comparisons (mean: 0.64; SD: 0.13; Box v) were significantly lower than all other PEVK-CA comparisons ($P < 0.0001$ in all cases; Box v vs. Boxes i-iv, vi-ix), as well self-comparisons including the PEVK-

471 CB:PEVK-CB ($P < 0.0001$; Box iii vs Box ix). Additionally, the number of substitutions
472 per exon for PEVK-CB:PEVK-CB exons were significantly greater than the PEVK:N-
473 PEVK-N self-comparisons ($P < 0.0001$; Box iii vs. Box vii). PEVK-CB:PEVK-CB exon
474 comparisons also show similar levels of variability as between region comparisons,
475 such as PEVK-CA and PEVK-CB exon comparisons ($P = 0.93$, Box iii vs ii; $P = 0.9$, Box iii
476 vs vi). These additional analyses suggest that the PEVK-CB region is more variable in
477 sequence than the PEVK-CA region in an individual, and that sequences with high
478 similarity to each other are concentrated in the PEVK-CA region. Likewise, the self-
479 dot plots (Figures S14-S23) show that each species has some highly repetitive regions in
480 the PEVK-CA (appearance of boxes in lower left corner). The distinct patterns of
481 sequence variability with the PEVK-C suggested that the PEVK-CA and PEVK-CB may
482 be evolving differently. Intriguingly, cardiac N2B titin isoforms do not contain exons in
483 the PEVK-CA, and may provide a functional basis for variable selection on the two
484 regions (Greaser et al. 2002; Guo et al. 2010).

485

486 In humans, the movement of long interspersed nuclear elements (LINEs) facilitated the
487 restructuring of the PEVK-C region, resulting in a large triplication event in the PEVK-C
488 (Bang et al. 2001). We used self-dot plots of the PEVK-C region to examine the
489 incidence of large-scale duplications and triplications (off-axis parallel lines) across each
490 species (Figure S24). We find evidence of large duplications in chimpanzee, orangutan,

Norway rat, house mouse and star-nosed mole in the PEVK-CA region. Triplications are only apparent in koala and human (Figures S14b, S16b, S24). In koala, the internal segment of the triplication is flanked by identical sequences, similar to the human triplication event, though it is not known if the sequence is a LINE-1 element as in humans, as LINEs are species-specific. Previous work based on nucleotide divergence placed the insertion of the LINE-1 elements in human PEVK-C after the divergence of humans and other primates (Bang et al. 2001). However, here we find evidence that both orangutan and chimpanzee also have the LINE-1 element in the PEVK-C, though they only show 1 element each (2 in humans; Figures S16a-c, S24). While in humans, the LINE-1 restructuring of the PEVK-C produces a large triplication event, the region is only duplicated in orangutans and the chimpanzee shows two large repeats with a truncated third repeat (Figures S16a, S16c, S24).

Previous work found that exons in the region corresponding to the PEVK-N were proportionally richer in glutamate (Greaser, 2001; Labeit et al. 2003; Forbes et al. 2005). PEVK-Finder replicates that finding here, suggesting again that PEVK_Finder is able to reproduce previously described PEVK exons in humans (Figure 3e).

Evolutionary analysis of the PEVK region across mammals

We performed a reciprocal best BLAST (Tatusov, 1997; Bork and Koonin, 1998; Koonin, 2005) to detect orthologs in the PEVK region. Overall, orthology across the PEVK region is low, with only 22 confident orthologs detected in the PEVK-N region (43% of human PEVK-N exons) and 13 in the PEVK-C region (22% of PEVK-C exons) (Figures 4, S2). More orthologs are detected in the PEVK-N region than the PEVK-C region (red/orange, Figure S2), consistent with a more conserved exon structure in the PEVK-N region. Exons in the PEVK-C region display greater sequence divergence across the mammalian tree (mean: 0.11; SE: 0.01) than the PEVK-N region (mean 0.08; SE: 0.01), though this difference is not significant ($t = 1.67$, $P = 0.11$) and primarily driven by three exons in the PEVK-C region (Figure S3).

The PEVK-N and PEVK-C regions also encode for significantly different amounts of the core P, E, V, and K amino acids. Across all mammal species, PEVK-N exons contain significantly more glutamate (1.7X; $t = 33.51$, $P < 6.84*10^{-31}$), lysine (1.2X; $t = 15.76$, $P < 9.32*10^{-19}$), and valine (1.2X; $t = 13.81$, $P <= 8.30*10^{-17}$) amino acids than PEVK-C exons (Figure 3d). PEVK-C exons contain significantly more proline amino acids (1.1X; $t = 13.83$, $P < 7.79*10^{-17}$). Non-PEVK amino acid residues, such as isoleucine, arginine, threonine, leucine, tyrosine and serine, are also more common in PEVK-N exons (Figure S4).

529

530 We also examined the nature of selection acting on codons in the PEVK region.
531 Specifically, we tested for positive selection using random-sites models in PAML, which
532 allowed ω to vary among sites but not lineages. For six sets of orthologous exons
533 (human exons 114, 135, 137, 138, 145 and 199), models that allowed certain codons to
534 evolve under positive selection ($\omega > 1$; M2a) were significantly better than models that
535 restricted codons to be conserved ($0 < \omega < 1$) or neutral ($\omega = 1$; M1a; Figure 4; Table 1).
536 Model M1a vs. M0 was also significant for six exons investigated (results not shown).
537 Five exons with sites under positive selection were in the PEVK-N region, and one was
538 in the PEVK-C region. In total, 20 codons were under positive selection (Figures S5-S10;
539 Table 1). Results were qualitatively similar for all initial ω values.

540

541 The polarity of amino acids can influence protein-binding interactions and solubility,
542 and therefore may contribute to adaptive evolution of titin. Therefore, we used
543 ancestral state reconstruction to examine charge transitions in TTN exons with
544 confident orthologs. Our analyses show that codons 42 of exon 114 and 6 of exon 135
545 are under positive selection and have substitutions that change the charge of the codon
546 (Figures S11-S12). Codon 42 of exon 114 features shifts from hydrophobic to polar
547 amino acids from a minimum of 3 to ~6 times, with shifts occurring in the branches

548 leading to bats, beaver, cetaceans, lemur and Northern treeshrew. This codon also
549 shifts from a hydrophobic to a positively charged amino acid in *Mus musculus*. Codon 6
550 of exon 135 reveals four independent transitions from positively charged to polar amino
551 acids across mammals. Intriguingly, in cetaceans, this codon shifts from positively
552 charged to hydrophobic. Thus, in both instances of codons with evidence for charge
553 transitions, cetaceans are one of the few groups to experience changes in polarity and
554 sometimes do so in unique ways (Figure S12). Together, these charge changes and low
555 PEVK exon numbers suggest that titin in cetaceans and other aquatic diving mammals
556 may be under different selective pressures than terrestrial mammals.

557

558 **DISCUSSION**

559

560 **PEVK_Finder improves the annotation of the PEVK region across mammals**

561 The PEVK region of TTN possesses many attributes that confound standard gene
562 prediction tools; it contains numerous short, repetitive exons that vary within and
563 across species in sequence content, exon structure, and length. However, it is this very
564 complexity that makes the PEVK region a promising source of raw material for
565 evolutionary adaptation of muscle (Freiburg et al. 2000; Tompa, 2003; Prado et al. 2005).

566 Robust annotations of the PEVK region are therefore needed to link molecular evolution
567 of TTN with physiological performance over evolutionary time.

568

569 While accurate gene prediction may be achieved with RNA-Seq or similar methods,
570 relying on transcriptomic data for gene prediction is not always feasible due to time and
571 cost constraints. For example, only 5 of the 43 species examined in this study had
572 cDNA data supporting their TTN annotations at the time of the study. The plethora of
573 alternatively spliced isoforms of TTN further complicates RNA-based gene predictions,
574 as data from numerous tissues, developmental stages, and/or individuals are required
575 to generate complete annotations (e.g., Savarese et al. 2018). To this end, we developed
576 PEVK_Finder, which combines *ab initio* and signature-based approaches to gene
577 prediction and targets the PEVK region of TTN (Wang et al. 2004; Lerat, 2010; Sleater,
578 2010). Our tool depends solely on nucleotide sequence and searches for exons with
579 motifs and patterns common to well-annotated PEVK regions (i.e., human, mouse).
580 Similar custom tools have been developed for classes of transposable elements, which
581 are also highly repetitive and often missed by standard annotation tools (Lerat, 2010).

582

583 Here, we find that PEVK_Finder consistently improves the annotation of the PEVK
584 region across mammals. One major challenge of annotating repetitive regions with

585 short exons is that exons are commonly missed. PEVK_Finder consistently detects
586 significantly more PEVK exons than standard annotation tools, both in model and non-
587 model species (Figures 1b, 1c; Table S3). Specifically, PEVK_Finder outperforms other
588 tools in regions of short, high-density exons and across broad ranges of PEVK ratios.
589 PEVK_Finder also identifies numerous putative PEVK exons that are not found in
590 cDNA annotations or identified by other bioinformatics tools (Figures 1c, S1c; Table S3).
591 It is possible that these novel exons are not transcribed, but TTN has many alternatively
592 spliced isoforms that may not be completely characterized by cDNA. For example,
593 manual annotation of genomic TTN sequences has discovered novel TTN exons in the
594 past (e.g., Freiburg et al. 2000; Granzier et al. 2007).

595

596 While PEVK_Finder improves available annotations in non-model species, we caution
597 that it does not perfectly replicate the PEVK region in human and mouse TTN.
598 PEVK_Finder occasionally misses, truncates, or joins distinct exons, albeit at a lower
599 frequency than the other tools tested. Exons with low PEVK ratios, internal splice sites
600 and non-canonical splice sites cause most of these errors. Indeed, we were unable to
601 obtain exon sets for two of the 43 species we tested due to the presence of non-canonical
602 splice sites. In the future, incorporation of hidden Markov models into the tool could
603 improve issues with splicing and variable PEVK ratios. Given this, we argue that
604 PEVK_Finder is most useful for determining major trends in PEVK structure and

evolution in non-model species, as exemplified here. When complete, perfectly resolved annotations are required in a non-model species, PEVK_Finder should be used in conjunction with RNA-Seq or cDNA sequencing from numerous tissues, individuals, and developmental times. Likewise, pairing PEVK_Finder with another annotation tool, such as Gnomon, could also provide more comprehensive annotations (Table S3).

The evolution of the PEVK region

Our results reveal contrasting patterns of constraint and divergence across the PEVK region, suggesting two or three subregions with distinct evolutionary dynamics. The PEVK-N region shows relatively conserved length and exon structure over evolutionary time, but evidence of diversifying selection and more variable amino acid content (Figures 2, 4). In contrast, the PEVK-C region varies dramatically in length and exon number across mammals and can be further divided into the highly similar PEVK-CA and more variable PEVK-CB (Figures 2, 3a, 3c, S14-S23). Overall, the documented patterns argue for selection maintaining particular, "essential" PEVK-N and PEVK-CB exons over evolutionary time, with diversifying selection targeting specific codons in the PEVK-N. For example, 17 out of 22 PEVK-N exons with confident orthologs are either constitutively expressed or expressed in >95% of TTN transcripts in skeletal muscle (Figure 4). Likewise, cardiac N2B titin contains exons in the PEVK-N and

PEVK-CB, but not the PEVK-CA (Greaser et al. 2002; Guo et al. 2010). Conservation over both evolutionary time and across isoforms suggest these exons may play key roles in vertebrate muscle function. Additionally, five PEVK-N exons have codons that experienced adaptive evolution and may be implicated in diversification of muscle function over mammalian evolution. Specifically, codons in exons 114 and 135 experienced major shifts in charge over mammalian evolution and may influence titin-protein binding. Conversely, expansion and contraction of the total length of the PEVK-C region, rather than selection on any particular exon, dominates the evolution of the PEVK-C. Such length variation has the potential to contribute to functional differences, as altering the size of the titin "spring" through alternative splicing of more or fewer PEVK repeats can affect titin's compliance, though neutral processes may also be relevant (reviewed in Linke, 2018). Future work can focus on disentangling the effects of natural selection acting on specific codons from PEVK length variation, and how both contribute to evolutionary adaptation of TTN.

What explains the discrepancy between the molecular patterns documented in the PEVK-N and PEVK-C regions? Intraspecific sequence comparisons of PEVK exons offer clues. Within a single species' titin sequence, the sequence content of the PEVK-CA is more repetitive than the PEVK-N or PEVK-CB (Figures 3, S5, S14-S23, S25). This suggests a mechanistic basis for expansion and contraction of the PEVK-CA over the

course of mammalian evolution. Replication slippage and recombinatorial repeat expansions often occur in regions with tandem repeats, and may result in exon duplication and loss (Tomba, 2003). Mobile LINE element insertions have also been implicated in TTN remodeling by causing PEVK exon duplication or differential splicing (Granzier et al. 2007). For example, we find that retrotransposition of LINE-1 elements facilitated tandem duplications of longer PEVK-CA regions in chimpanzee and orangutan, in addition to the previously reported human duplication (Bang et al. 2001; Figures S24, S16). Because we find that the PEVK-CA contains more repeats than the PEVK-N and PEVK-CB, we argue that any process generating exon duplication or loss occurs most frequently in the PEVK-CA.

Implications for titin function

The length and exon structure of the PEVK region vary remarkably over evolutionary time, with implications for titin functionality. The PEVK region of titin has long been known to contribute to passive stiffness of muscle (Gautel and Goulding, 1996; Linke et al. 1998). Through alternative splicing, titin can be expressed as isoforms with varying lengths of the PEVK domain which correlate with the passive properties of different muscle types (Freiburg et al. 2000; Prado et al. 2005). To date, at least 4000 alternative splicing events and twelve distinct isoforms have been identified in human, mouse, and

663 rabbit tissues (Freiburg et al. 2000; Prado et al. 2005; Savarese et al. 2018). Muscles with
664 long PEVK segments have low passive stiffness whereas muscles with shorter segments
665 have higher passive stiffness (Freiburg et al. 2000; Prado et al. 2005). The repetitive
666 nature of PEVK-C exons may allow for variability in the length and stiffness of titin
667 isoforms. Repeats within the PEVK region have been postulated to be functionally
668 equivalent to fulfill entropic chain requirements (Tomba, 2003). While this may be the
669 case for repeats within the PEVK-C region, it is less clear for repeats within the PEVK-N
670 region. Alternatively, variation in exon number across the PEVK region may be a
671 product of neutral processes that result in exon skipping and loss, without affecting
672 TTN function.

673

674 While a mechanism has not yet been elucidated, several researchers have proposed that
675 during muscle activation, titin stiffness increases in the presence of calcium possibly by
676 titin-actin binding (Herzog et al. 2012; Nishikawa et al. 2012; Schappacher-Tilp et al.
677 2015). Evidence suggests that there is a small direct effect of calcium on PEVK stiffness
678 but that this increase cannot fully account for active muscle stiffness (Tatsumi et al.
679 2001; Labeit et al. 2003). Within the PEVK region, exons can be composed of strings of
680 negatively charged glutamate residues (E-rich motifs). Results from single molecule
681 experiments have shown that E-rich PEVK segments bind to actin filaments, which
682 produces a viscous load that could possibly resist the sliding of the thin filament along

683 the thick filament during muscle contraction (Kellermayer and Granzier, 1996; Labeit et
684 al. 2003; Nagy et al. 2004; Bianco et al. 2007). If this interaction is necessary for
685 increasing titin stiffness in activated muscle, splicing of exons that encode for amino
686 acids responsible for changing charge or binding affinity to actin could have dramatic
687 effects on titin function. For example, several studies have implicated titin in the
688 enhancement of active muscle force with stretch (Leonard and Herzog, 2010; Powers et
689 al. 2016; Hessel et al. 2017). If titin stiffness failed to increase as a result of a change in
690 amino acid charge and/or binding to actin or other proteins, muscles could show
691 decreased force during stretch and alter an organism's ability to move. Our data show
692 that PEVK-N exons are conserved and confirm previous work suggesting that PEVK-N
693 exons have a higher percentage of E-rich motifs than PEVK-C exons (Greaser, 2001;
694 Labeit et al. 2003; Forbes et al. 2005; Figure 3e). In addition, PEVK-N exons 114 and 135
695 exhibit charge shifts that could affect PEVK interactions with other proteins (Figures
696 S11, S12). However, no studies to date have shown calcium dependent PEVK-actin
697 binding (Bianco et al. 2007; Nagy et al. 2004; Linke et al. 2002). Thus, it is unlikely this
698 interaction underlies the increase in titin stiffness during muscle activation. It remains
699 unclear how E-rich motifs or change in charge may affect titin function. Further studies
700 on the expression of PEVK-N exons could serve to elucidate the role of the PEVK in
701 active and passive muscle.

702

703 Several cardiac and skeletal muscle diseases have been associated with mutations in the
704 TTN gene including the PEVK region (Chauveau et al. 2014; Savarese et al. 2016; Schafer
705 et al. 2017). Understanding how titin has evolved across mammals may provide insight
706 about the effects that TTN mutations have on muscle physiology and disease. Our
707 analyses have identified evolutionarily conserved PEVK exons in the PEVK-N that have
708 previously been shown to be constitutively expressed (Savarese et al. 2018). These
709 exons may play fundamental roles in titin function and therefore represent sites that
710 could lead to severe deficits in function if modified. Of the 127 TTN sequence variants
711 that have been associated with human muscular disorders, 29 are located in the I-band,
712 5 of which are found in the PEVK region (Chauveau et al. 2014). Cardiomyopathies are
713 often associated with truncating variants in the A-band, although a few variants have
714 been identified in the I-band as well (Schafer et al. 2017). Interestingly, to date, most of
715 the mutations in the PEVK region that have been linked to neuromuscular disease are
716 located in the PEVK-C region (Savarese et al. 2016). It is possible that there are more
717 disease-causing variants in the PEVK-C compared to the PEVK-N region, or it could be
718 that variants in the PEVK-N region have yet to be identified. It is also possible that the
719 highly repetitive nature of the PEVK-C may facilitate errors during replication that
720 increase the frequency of deleterious mutations. Alternatively, it has been suggested
721 that some I-band variants could be circumvented via differential splicing (Schafer et al.

2017). Clearly, further studies on the essentiality of particular exons for muscle function and mechanisms of disease are crucial for better understanding TTN-related disorders.

Conclusions

In summary, PEVK_Finder provides a useful method for determining major trends in PEVK structure and evolution in non-model species. By characterizing the PEVK region of titin in mammals, we identify two potential pathways through which selection could shape titin function: by changes in the amino acid sequences of specific codons, and by variations in the size of the PEVK region. In the PEVK-N, exons with confident orthologs that are constitutively expressed in isoforms are strong candidates for future work testing the essentiality of particular exons that underlie titin stiffness and muscle function. Together, the construction of custom annotation tools for challenging-to-annotate genes can facilitate novel insights into the diversification and nature of selection acting on important proteins like titin.

Acknowledgements

We thank Kiisa Nishikawa, Jocelyn Crawford, Silvia Leblanc, Keon Rabbani, Emma Bekele and two anonymous reviewers for their helpful comments on earlier versions of

this manuscript. This work was supported by the National Science Foundation (IOS-1731917 awarded to J.A.M.) and start-up funds for F.R.F. and J.A.M.

743 **References**

744

745 Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment
746 search tool. J. Mol. Biol. 215: 403–410.

747 Bang, M. L., T. Centner, F. Fornoff, A. J. Geach, M. Gotthardt *et al.* 2001 The complete gene
748 sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with
749 obscurin identify a novel Z-line to I-band linking system. Circ. Res. 89: 1065–1072.

750 Bianco, P., A. Nagy, A. Kengyel, D. Szatmári, Z. Mártonfalvi *et al.* 2007 Interaction forces
751 between F-actin and titin PEVK domain measured with optical tweezers. Biophys. J. 93:
752 2102–2109.

753 Bork, P., and E. V. Koonin, 1998 Predicting functions from protein sequences—where are the
754 bottlenecks? Nat. Genet. 18: 313–318.

755 Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J.
756 Mol. Biol. 268: 78–94.

757 Burset, M., I. A. Seledtsov, and V. V Solovyev, 2000 Analysis of canonical and non-canonical
758 splice sites in mammalian genomes. Nucleic Acids Res. 28: 4364–75.

759 Chauveau, C., J. Rowell, and A. Ferreiro, 2014 A Rising Titan, TTN review and mutation
760 update. Hum Mutat. 35: 1046-59.

761 Chikina, M., J. D. Robinson, and N. L. Clark, 2016 Hundreds of Genes Experienced Convergent
762 Shifts in Selective Pressure in Marine Mammals. Mol. Biol. Evol. 33: 2182–2192.

763 Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox *et al.* 2009 Biopython: freely

764 available Python tools for computational molecular biology and bioinformatics.
765 Bioinformatics 25: 1422–3.

766 Delaney, N., 2017 Dotplot. GitHub Repository, https://github.com/evolvedmicrobe/dotplot.

767 Finseth, F. R., E. Bondra, and R. G. Harrison, 2014 Selective constraint dominates the evolution
768 of genes expressed in a novel reproductive gland. Mol. Biol. Evol. 31(12):3266–3281.

769 Finseth, F. R., Y. Dong, A. Saunders, and L. Fishman, 2015 Duplication and adaptive evolution
770 of a key centromeric protein in mimulus, a genus with female meiotic drive. Mol. Biol.
771 Evol. 32(10):2694–2706.

772 Freiburg, A., K. Trombitas, W. Hell, O. Cazorla, F. Fougerousse *et al.* 2000 Series of exon-
773 skipping events in the elastic spring region of titin as the structural basis for myofibrillar
774 elastic diversity. Circ. Res. 86: 1114–1121.

775 Galindo, B. E., V. D. Vacquier, and W. J. Swanson, 2003a Positive selection in the egg receptor
776 for abalone sperm lysin. Proc. Natl. Acad. Sci. U. S. A. 100: 4639–4643.

777 Gautel, M., and D. Goulding, 1996 A molecular map of titin/connectin elasticity reveals two
778 different mechanisms acting in series. FEBS Lett. 385: 11–4.

779 Goujon, M., H. McWilliam, W. Li, F. Valentin, S. Squizzato *et al.* 2010 A new bioinformatics
780 analysis tools framework at EMBL-EBI. Nucleic Acids Res. 38: W695-9.

781 Granzier, H., M. Radke, J. Royal, Y. Wu, T. C. Irving *et al.* 2007 Functional genomics of chicken,
782 mouse, and human titin supports splice diversity as an important mechanism for
783 regulating biomechanics of striated muscle. Am. J. Physiol. Regul. Integr. Comp. Physiol.
784 293: R557-67.

785 Greaser, M. L., M. Berri, C. M. Warren, and P. E. Mozdziak, 2002 Species variations in cDNA
786 sequence and exon splicing patterns in the extensible I-band region of cardiac titin:
787 Relation to passive tension. J. Muscle Res. Cell Motil. 23: 473–482.

788 Gregorio, C. C., H. Granzier, H. Sorimachi, and S. Labeit, 1999 Muscle assembly: a titanic
789 achievement? Curr. Opin. Cell Biol. 11: 18–25.

790 Guo, W., S. J. Bharmal, K. Esbona, and M. L. Greaser, 2010 Titin diversity: alternative splicing
791 gone wild. J. Biomed. Biotechnol. 2010:1-8.

792 Herzog, W., T. Leonard, V. Joumaa, M. DuVall, and A. Panchangam, 2012 The three filament
793 model of skeletal muscle stability and force production. Mol. Cell. Biomech. 9: 175–91.

794 Hessel A. L., S. L. Lindstedt, K.C. Nishikawa, 2012, Physiological mechanisms of eccentric
795 contraction and its appications: a role for the gian titin protein. Front Physiol. 8:70

796 Hughes, A. L., and M. Nei, 1989 Nucleotide substitution at major histocompatibility complex
797 class II loci: Evidence for overdominant selection. Proc. Nati.Acad. Sci. 86:958-962.

798 Junier, T., and M. Pagni, 2000 Dotlet: diagonal plots in a web browser. Bioinformatics 16: 178–9.

799 Karlsson, E. K., D. P. Kwiatkowski, and P. C. Sabeti, 2014 Natural selection and infectious
800 disease in human populations. Nat. Rev. Genet. 15:379-393.

801 Kellermayer, M. S. Z., and H. L. Granzier, 1996 Calcium-dependent inhibition of in vitro thin-
802 filament motility by native titin. FEBS Lett. 380: 281–286.

803 Koonin, E. V., 2005 Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet. 39:
804 309–338.

805 Kreitman, M., and H. Akashi, 1995 Molecular Evidence for Natural Selection. Annu. Rev. Ecol.

Syst. 26:403-422.

Kumar, S., A. J. Filipski, F. U. Battistuzzi, S. L. Kosakovsky Pond, and K. Tamura, 2012 Statistics and truth in phylogenomics. Mol. Biol. Evol. 29(2):457–472.

Kumar, S., G. Stecher, and K. Tamura, 2016 MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol. 33: 1870–1874.

Kumar, S., G. Stecher, S.B. Hedges, 2017 TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol BIol Evol 34(7):1812-1819

Labeit, S., and B. Kolmerer, 1995 Titins: giant proteins in charge of muscle ultrastructure and elasticity. Science 270: 293–6.

Labeit, D., K. Watanabe, C. Witt, H. Fujita, Y. Wu *et al.* 2003 Calcium-dependent molecular spring elements in the giant protein titin. Proc. Natl. Acad. Sci. 100: 13716–13721.

Leonard, T. R., W. Herzog, 2010 Regulation of muscle force in the absenve of actin-myosin-based cross-bridge interaction. Am J Physiol Cell Physiol. 299(1):C14-20

Lerat, E., 2010 Identifying repeats and transposable elements in sequenced genomes : how to find your way through the dense forest of programs. Heredity. 104: 520–533.

Linke, W. A., 2018 Titin Gene and Protein Functions in Passive and Active Muscle. Annu. Rev. Physiol. 80: 389–411.

Linke, W. A., M. Ivemeyer, P. Mundel, M. R. Stockmeier, and B. Kolmerer, 1998 Nature of PEVK-titin elasticity in skeletal muscle. Proc. Natl. Acad. Sci. 95: 8052–8057.

Linke, W. A., M. Ivemeyer, N. Olivieri, B. Kolmerer, J. C. Rüegg *et al.* 1996 Towards a molecular understanding of the elasticity of titin. J. Mol. Biol. 261: 62–71.

827 Linke, W. A., M. Kulke, H. Li, S. Fujita-Becker, C. Neagoe *et al.* 2002 PEVK domain of titin: An
828 entropic spring with actin-binding properties. J. Struct. Biol. 137: 194–205.

829 Manteca, A., J. Schönfelder, A. Alonso-Caballero, M. J. Fertin, N. Barruetabeña *et al.* 2017
830 Mechanochemical evolution of the giant muscle protein titin as inferred from resurrected
831 proteins. Nat. Struct. Mol. Biol. 24: 652–657.

832 Monroy, J. A., K. L. Powers, C. M. Pace, T. Uyeno, and K. C. Nishikawa, 2017 Effects of
833 activation on the elastic properties of intact soleus muscles with a deletion in titin. J. Exp.
834 Biol. 220:828-836.

835 Nagy, A., P. Cacciafesta, L. Grama, A. Kengyel, A. Málnási-Csizmadia *et al.* 2004 Differential
836 actin binding along the PEVK domain of skeletal muscle titin. J. Cell Sci. 117: 5781–5789.

837 Nishikawa, K. C., J. A. Monroy, T. E. Uyeno, S. H. Yeo, D. K. Pai *et al.* 2012 Is titin a "winding
838 filament"? A new twist on muscle contraction. Proc. R. Soc. B Biol. Sci. 279:981-990.

839 O'Leary, N. A., M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad *et al.* 2016 Reference sequence
840 (RefSeq) database at NCBI: current status, taxonomic expansion, and functional
841 annotation. Nucleic Acids Res. 44: D733-45.

842 Parra, G., E. Blanco, and R. Guigó, 2000 GeneID in Drosophila. Genome Res. 10: 511–5.

843 Partha, R., B. K. Chauhan, Z. Ferreira, J. D. Robinson, K. Lathrop *et al.* 2017 Subterranean
844 mammals show convergent regression in ocular genes and enhancers, along with
845 adaptation to tunneling. Elife. 6:e25884.

846 Perutz, M. F., 1983 Species Adaptation in a Protein Mol. Biol. Evol. 1(1): l-28.

847 Pervouchine, D. D., S. Djebali, A. Breschi, C. A. Davis, P. P. Barja *et al.* 2015 Enhanced

transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression. Nat. Commun. 6:5903.

Powers, K., K. Nishikawa, V. Joumaa, and W. Herzog, 2016 Decreased force enhancement in skeletal muscle sarcomeres with a deletion in titin. J. Exp. Biol. 219: 1311–1316.

Prado, L. G., I. Makarenko, C. Andresen, M. Krüger, C. A. Opitz *et al.* 2005 Isoform Diversity of Giant Proteins in Relation to Passive and Active Contractile Properties of Rabbit Skeletal Muscles. J. Gen. Physiol. 126: 461–480.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna Austria, URL https://www.R-project.org/

Revell, L. J., and R. Graham Reynolds, 2012 A new bayesian method for fitting evolutionary models to comparative data with intraspecific variation. Evolution (N. Y). 66: 2697–2707.

Salamov, A. A., and V. V Solovyev, 2000 Ab initio gene finding in Drosophila genomic DNA. Genome Res. 10: 516–22.

Savarese, M., P. H. Jonson, S. Huovinen, L. Paulin, P. Auvinen *et al.* 2018 The complexity of titin splicing pattern in human adult skeletal muscles. Skelet. Muscle 8: 1–9.

Savarese M., Sarparanta J., Vihola A., Udd B., Hackman P., 2016 Increasing role of titin mutations in neuromuscular disorders. Journal of Neuoromuscular diseases 3: 293-208

Schafer, S. et al. 2017 Titin-truncating variants affect heart function in disease cohorts and the general population. Nature Genetics 49: 46-53.

Schappacher-Tilp, G., T. Leonard, G. Desch, and W. Herzog, 2015 A novel three-filament model of force generation in eccentric contraction of skeletal muscles. PLoS One 10: 1–16.

869 Sleator, R. D., 2010 An overview of the current status of eukaryote gene prediction strategies.
870 Gene 461: 1–4.

871 Souvorov, A., Y. Kaputsin, B. Kiryutin, V. Chetverin, T. Tatusova, D. Lipman. 2010. Gnomon-
872 NCBI eukaryotic gene prediction tool. National Center for Biotechnology Information,
873 Bethesda, MD

874 Stanke, M., and B. Morgenstern, 2005 AUGUSTUS: A web server for gene prediction in
875 eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33: W465-W467.

876 Steinmetz, P. R. H., J. E. M. Kraus, C. Larroux, J. U. Hammel, A. Amon-Hassenzahl *et al.* 2012
877 Independent evolution of striated muscles in cnidarians and bilaterians. Nature 487: 231–
878 234.

879 Tamura K., M. Nei, and S. Kumar, 2004 Prospects for inferring very large phylogenies by using
880 the neighbor-joining method. Proc Nat Sci 101:11030-11035.

881 Tatsumi, R., Maeda K., Hattori. A., Takahashi, K. 2001 Calcium binding to an elastic portion of
882 connectin/titin filaments. J. Muscle Res. Cell Motil. 22:149–62.

883 Tatusov, R. L., E. V Koonin, and D. J. Lipman, 1997 A genomic perspective on protein families.
884 Science 278: 631–7.

885 Tompa, P., 2003 Intrinsically unstructured proteins evolve by repeat expansion. BioEssays 25:
886 847–855.

887 Trombitás, K., A. Redkar, T. Centner, Y. Wu, S. Labeit *et al.* 2000 Extensibility of isoforms of
888 cardiac titin: variation in contour length of molecular subsegments provides a basis for
889 cellular passive stiffness diversity. Biophys. J. 79: 3226–34.

Wang, Z., Y. Chen, and Y. Li, 2004 A brief review of computational gene prediction methods. Genomics. Proteomics Bioinformatics 2: 216–21.

Witt, C. C., N. Olivieri, T. Centner, B. Kolmerer, S. Millevoi *et al.* 1998 A survey of the primary structure and the interspecies conservation of I-band titin's elastic elements in vertebrates. J. Struct. Biol. 122: 206–215.

Yang, Z., 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24: 1586–1591.

Yang, Z., W. S. W. Wong, and R. Nielsen, 2005 Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol. Biol. Evol. 22: 1107–1118.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis *et al.* 2016 Ensembl 2016. Nucleic Acids Res. 44: D710–D716.

Zhang, G., C. Li, Q. Li, B. Li, D. M. Larkin *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation Science. 346 (6215):1311-1320.

Zhao, H., B. Ru, E. C. Teeling, C. G. Faulkes, S. Zhang *et al.* 2009 Rhodopsin molecular evolution in mammals inhabiting low light environments. PLoS One. 4(12): e8326.

Figure 1. PEVK Finder tool optimization and evaluation of human TTN. **a)** Match scores of human PEVK exon sets were generated using different combinations of minimum exon length, PEVK ratio and sliding window length parameter settings. **b)** PEVK_Finder recovered more PEVK exons than other existing gene prediction tools (GENSCAN, Augustus, FGENESH and geneid). **c)** PEVK_Finder outperformed GENSCAN at recovering the exon-intron distribution of human TTN PEVK exons identified by cDNA. Vertical lines indicate exon boundaries, and the thickness of the lines is determined by the exon coordinates. Grey circles indicate exons that were missed by either PEVK Finder or GENSCAN, and black asterisks indicate automatically annotated exons that were not annotated by cDNA. In this figure, the first exon indicated by an asterisk is technically outside the pre-defined bounds of the PEVK region. The PEVK ratio scale is given in the figure.

Figure 2. Phylogenetic comparison of PEVK exon structure. PEVK_Finder exon-intron plots were overlaid onto a time-calibrated phylogeny for 39 of 41 mammalian species. The dotted line indicates the approximate boundary of the PEVK-N and PEVK-C segments of the PEVK region, based on the human boundary between the two segments. PEVK ratio scale is the same as in Fig. 1. In this figure, *Neomonachus schaunslandi* and *Tupaia chinensis* are represented by their alternative names, *Monachus schaunslandi* and *Tupaia belangeri*, respectively.

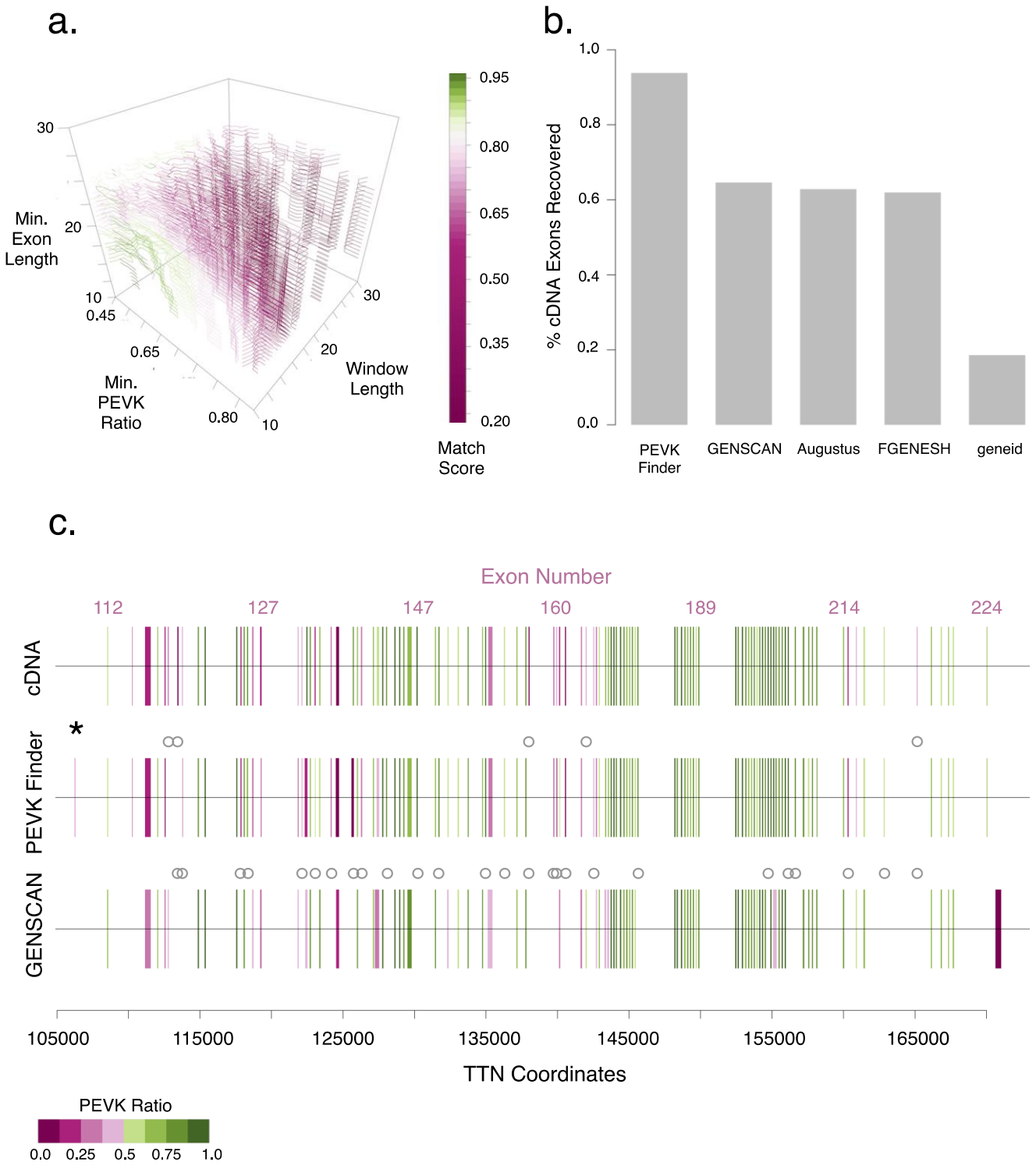
Figure 3. Exon and nucleotide-level comparisons of PEVK-N and PEVK-C regions. **a)** A heat map of substitutions among PEVK exons within an individual: I) PEVK-N vs. PEVK-C, II) PEVK-C vs. PEVK-C, III) PEVK-N vs. PEVK-N, and IV) PEVK-C vs. PEVK-N. Dark pink indicates exon pairs with few substitutions, whereas light pink and green indicate exons pairs with many substitutions. In general, PEVK-C exons are highly repetitive and homogeneous, while PEVK-N exons are more variable**. b)** A reciprocal dot plot of the human PEVK-C nucleotide sequence that shows the repetitive nature of the PEVK-C region in humans (Dotlet, https://dotlet.vital-it.ch). **c)** Mean pairwise substitutions per exon across all 41 species for each quadrant from Figure 3a. Bars represent mean ± s.e. **d)** Mean P,E,V,K amino acids per exon across all 41 mammalian species. There is significantly more glutamate (E), valine (V) and lysine (K) per exon in the PEVK-N region (grey bars) and more proline (P) per exon in the PEVK-C region (black bars) across all 41 species. Bars represent mean ±s.e. Asterisks denote significance at $P < 0.05$. **e)** Ratio of glutamate (E) per exon in human TTN PEVK. PEVK_Finder confirms that there is relatively more glutamate (E) per exon in the PEVK-N region compared to the PEVK-C region.

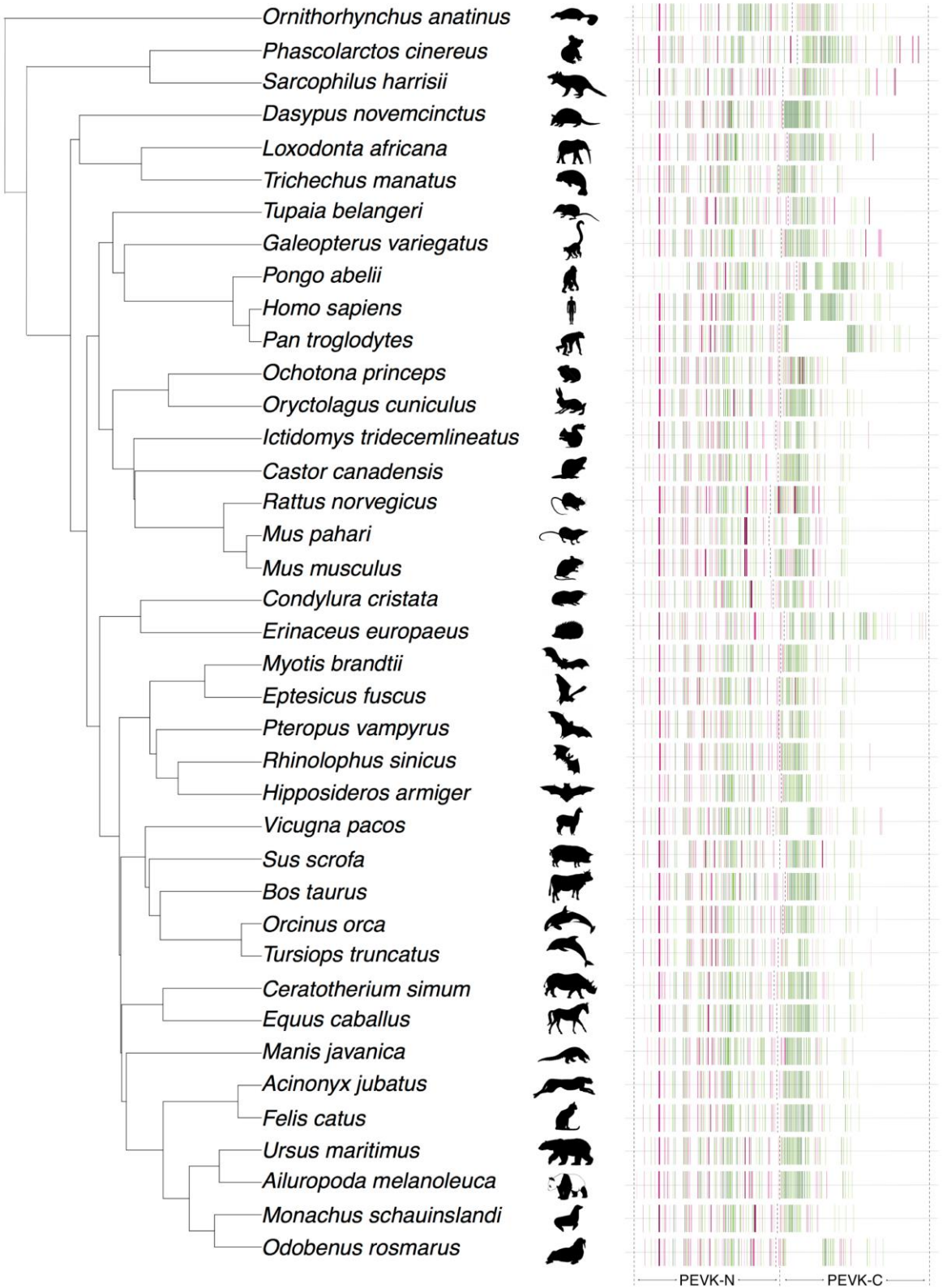
Figure 4. Human PEVK_Finder exon-intron plot depicting orthologous exons and codons under significant positive selection. Dark grey bars (labeled with their respective exon numbers) indicate orthologous exons, grey circles indicate exons missed by PEVK Finder, and asterisks indicate codons under selection at varying levels of significance. Asterisks indicate posterior probabilities of * > 0.5; ** > 0.75; *** > 0.95. The probability for only the most significant codon for a given exon is noted. Exons 114, 116, 122, 125, 130, 135, 136, 137, 138, 141, 142, 143, 144, 146, 152 153, and 161 are either constitutively expressed or expressed in >95% of TTN transcripts in skeletal muscle according to Savarese et al. 2018. Details about codons under selection are in Figures S5-S12. As in figure 1c, the first exon in this figure is technically outside the pre-defined bounds of the human PEVK region.

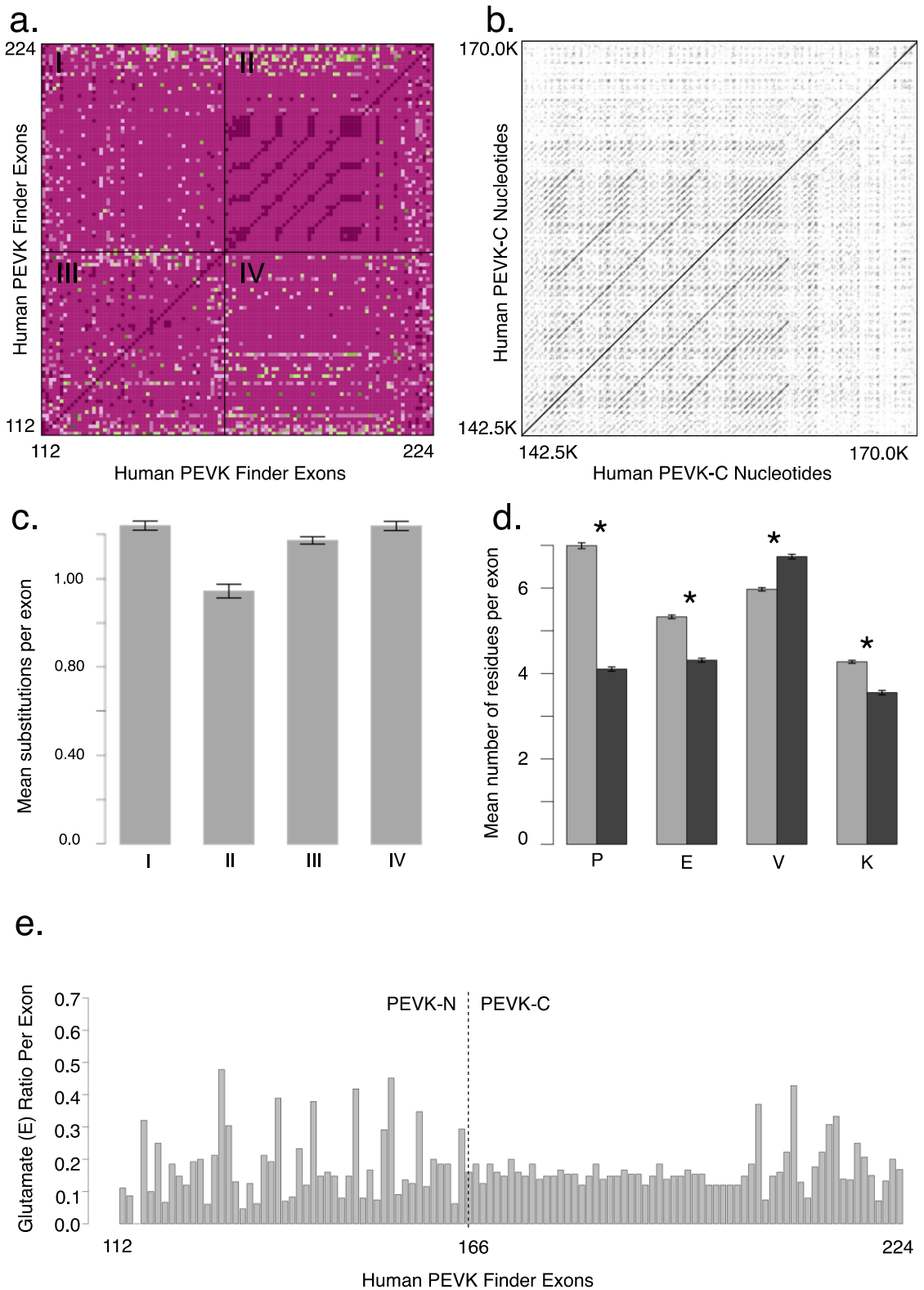
954 **Table 1**

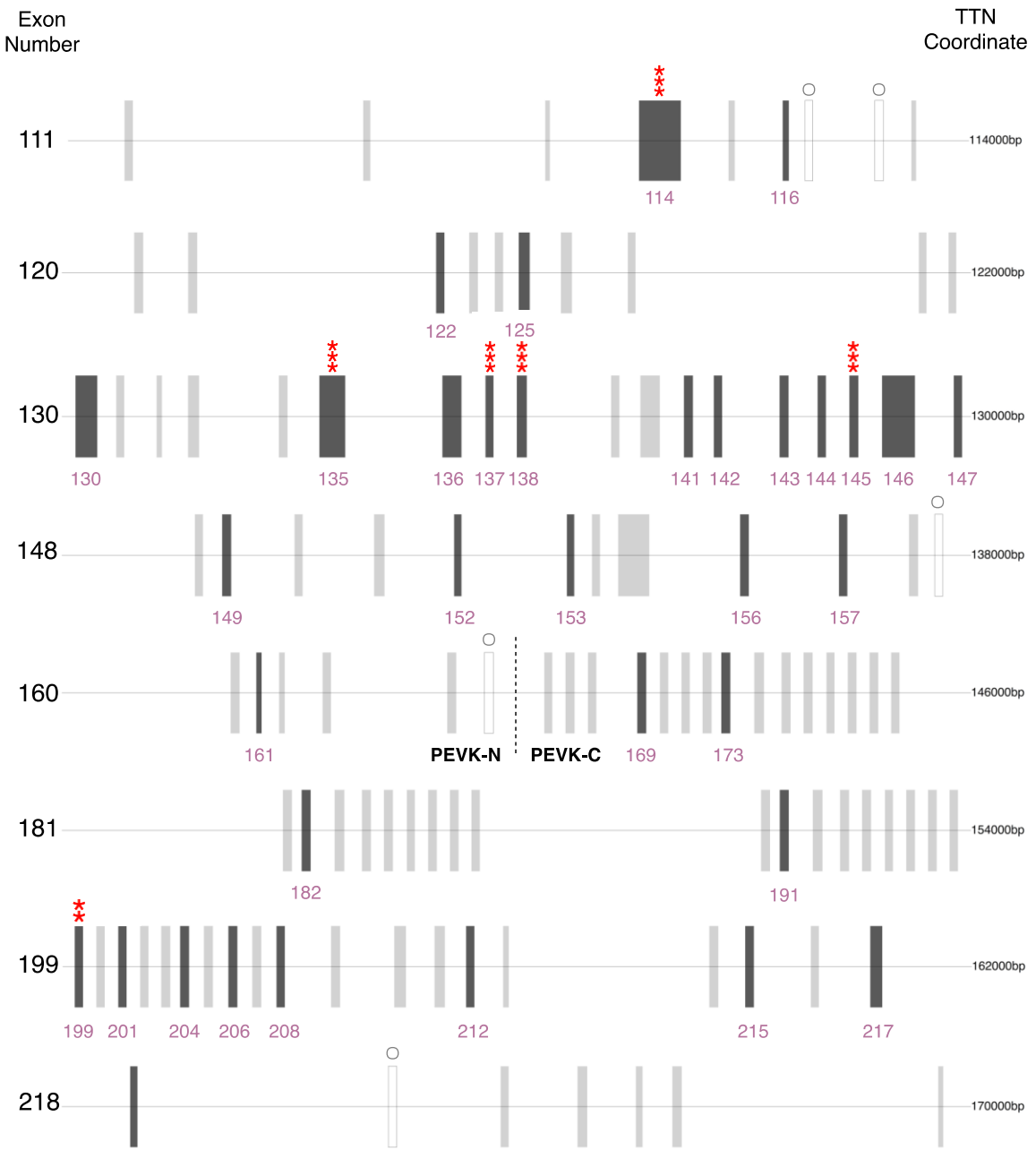
Exon	TTN Start Coordinate	Null LnL (M1)	Alternate LnL (M2)	Test Statistic	*P* Value	Sites > 0.95	Sites > 0.75	Sites > 0.5	Initial ω	Estimated ω	Estimated kappa	Codon Model
114	111156	-2476.26	-2472.43	7.67	0.02160	46F	42F		3	3.04	2.59	3x4
135	124495	-758.84	-749.98	17.71	0.00014	1A, 6R, 51I	41Y		3	7.83	5.59	3x4
137	125975	-334.48	-328.98	11.00	0.00408	1L			3	11.64	4.59	3x4
138	126254	-379.56	-373.27	12.58	0.00186	1P, 4P	13R, 17T	25T	3	4.86	2.64	3x4
145	129217	-437.83	-429.71	16.24	0.00030	7L, 18P		17V, 26P	3	4.59	2.05	3x4
199	154307	-260.39	-257.34	6.10	0.04740		11P, 13T, 14P	20V	3	2.55	4.65	3x4

955


a.
30
Min.
Exon
Length
20
10
0.45
0.65
Min.
PEVK
Ratio
0.80
10
20
30
Window
Length
0.95
0.80
0.65
0.50
0.35
0.20
Match
Score
b.
1.0
0.8
0.6
0.4
0.2
0.0
% cDNA Exons Recovered
PEVK
Finder
GENSCAN
Augustus
FGENESH
geneid
c.
Exon Number
112
127
147
160
189
214
224
cDNA
PEVK Finder
GENSCAN
*
105000
115000
125000
135000
145000
155000
165000
TTN Coordinates
PEVK Ratio
0.0
0.25
0.5
0.75
1.0


Ornithorhynchus anatinus
Phascolarctos cinereus
Sarcophilus harrisii
Dasypus novemcinctus
Loxodonta africana
Trichechus manatus
Tupaia belangeri
Galeopterus variegatus
Pongo abelii
Homo sapiens
Pan troglodytes
Ochotona princeps
Oryctolagus cuniculus
Ictidomys tridecemlineatus
Castor canadensis
Rattus norvegicus
Mus pahari
Mus musculus
Condylura cristata
Erinaceus europaeus
Myotis brandtii
Eptesicus fuscus
Pteropus vampyrus
Rhinolophus sinicus
Hipposideros armiger
Vicugna pacos
Sus scrofa
Bos taurus
Orcinus orca
Tursiops truncatus
Ceratotherium simum
Equus caballus
Manis javanica
Acinonyx jubatus
Felis catus
Ursus maritimus
Ailuropoda melanoleuca
Monachus schauinslandi
Odobenus rosmarus
PEVK-N
PEVK-C


a.
224
I
II
III
IV
Human PEVK Finder Exons
112
112
224
Human PEVK Finder Exons
b.
170.0K
Human PEVK-C Nucleotides
142.5K
142.5K
170.0K
Human PEVK-C Nucleotides
c.
Mean substitutions per exon
1.00
0.80
0.40
0.0
I
II
III
IV
d.
Mean number of residues per exon
6
4
2
0
*
*
*
*
P
E
V
K
e.
Glutamate (E) Ratio Per Exon
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
PEVK-N
PEVK-C
112
166
224
Human PEVK Finder Exons


Exon Number
TTN Coordinate
111
114
116
114000bp
120
122
125
122000bp
130
130
135
136
137
138
141
142
143
144
145
146
147
130000bp
148
149
152
153
156
157
138000bp
160
161
PEVK-N
PEVK-C
169
173
146000bp
181
182
191
154000bp
199
199
201
204
206
208
212
215
217
162000bp
218
170000bp