SCIENTIFIC REPORTS

OPEN

Aminode: Identification of Evolutionary Constraints in the Human Proteome

Received: 5 July 2017
Accepted: 5 January 2018
Published online: 22 January 2018

Kevin T. Chang[1], Junyan Guo[1,2], Alberto di Ronza[1] & Marco Sardiello[1]

Evolutionarily constrained regions (ECRs) are a hallmark for sites of critical importance for a protein's structure or function. ECRs can be inferred by comparing the amino acid sequences from multiple protein homologs in the context of the evolutionary relationships that link the analyzed proteins. The compilation and analysis of the datasets required to infer ECRs, however, are time consuming and require skills in coding and bioinformatics, which can limit the use of ECR analysis in the biomedical community. Here, we developed Aminode, a user-friendly webtool for the routine and rapid inference of ECRs. Aminode is pre-loaded with the results of the analysis of the whole human proteome compared with proteomes from 62 additional vertebrate species. Profiles of the relative rates of amino acid substitution and ECR maps of human proteins are available for immediate search and download on the Aminode website. Aminode can also be used for custom analyses of protein families of interest. Interestingly, mapping of known missense variants shows great enrichment of pathogenic variants and depletion of non-pathogenic variants in Aminode-generated ECRs, suggesting that ECR analysis may help evaluate the potential pathogenicity of variants of unknown significance. Aminode is freely available at http://www.aminode.org.

Evolutionary changes along a protein sequence occur at rates that are inversely correlated with the strength of specific constraints at each site. Constrained regions are considered to be under functional constraint owing to a role in protein stability, post-translational modifications, subcellular localization, interaction with other molecules, or enzymatic function[1–4]. Because constraint can vary widely along a given protein sequence, profiling the rates of evolutionary changes can provide information useful to identify the key residues or domains of the protein.

Several studies have shown that evolutionarily constrained regions (ECRs) can pinpoint the position of residues that are relevant for the function of enzymes or other protein types and can even provide significant information to predict the effects of specific mutations[5–11]. Therefore, the identification of ECRs may help inform investigation and experimental design of protein studies. For example, profiling evolutionary constraint can indicate regions to avoid or to target for protein tagging when the function or interactions of the protein must be preserved. Conversely, highly constrained regions might be an excellent choice for functional studies based on mutagenesis analysis[7,8,12]. In the absence of prior experimental data, the identification of ECRs may indeed point towards candidate positions in a protein that, if mutated, may have a deleterious effect on the protein function. The underlying reasoning is that if a site has been refractory to changes over long periods of evolutionary time—as inferred from a comparison of numerous and distantly related taxa—any change at that site is likely deleterious[13,14].

Effective methods of profiling a set of homologous proteins to determine ECRs require the simultaneous analysis of amino acid sequences and phylogenetic relationships of the proteins under examination[15,16]. A general approach to identify ECRs consists of a multi-step procedure[15]: First, orthologs of the protein of interest are selected and a multiple alignment is generated to allow the measurement of the relative rate of substitution at each protein position. Depending on the analysis to be performed, paralogs may also be included—closely related paralogs if the analysis is focused on specific structural features of the protein under examination, or both close and distant paralogs if the analysis is aimed at identifying general constraints of the protein family[5,15]. Next, the number of substitutions that have occurred at each protein position is computed based on the phylogenetic

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, 77030, USA. [2]Present address: Microsoft Corporation, 1 Microsoft Way, Redmond, WA, 98052, USA. Correspondence and requests for materials should be addressed to M.S. (email: sardiell@bcm.edu)

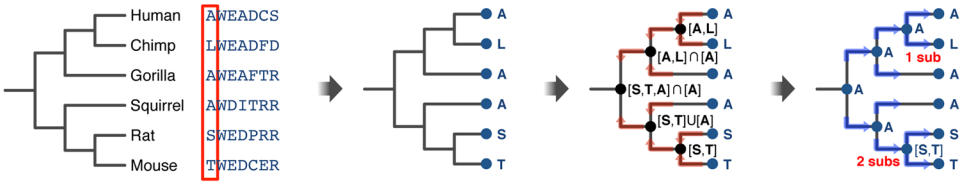

Figure 1. An example of the procedure used to compute ancestor nodes based on the Hartigan algorithm.

relationships among the proteins under examination; the information is then used to calculate the relative rate of substitutions in a sliding window of a fixed length over the entire protein multiple alignment, where each window's relative rate is obtained by dividing the substitution rate in that window by the average of all windows. Relative rates are finally plotted as a function of their position along the protein alignment, and ECRs are identified as corresponding to the "valleys" in the plot[15]. These procedures are time consuming and require skills in coding and bioinformatics. As a service to the biomedical community, we have developed a web tool, Aminode, which automatically profiles protein evolutionary constraints with a minimal amount of information from the user. Aminode is freely available, includes a pre-computed analysis of the human proteome, and allows download of high-resolution graphs and computed data for immediate use.

Results

Aminode Scope. Aminode calculates the relative amino acid substitution rates of the protein(s) of interest and identifies evolutionarily constrained regions (ECRs) via a comparative analysis of multiple protein homologs in the context of their evolutionary relationships. The Aminode pipeline performs analyses based on two inputs: (i) The amino acid sequences of the protein homologs, and (ii) a phylogenetic tree that describes the evolutionary relationships of the inputted protein homologs. Aminode implements a user-friendly, web-based interface that allows two modalities of analysis:

Pre-computed analysis of the human proteome. Users can retrieve the results from the pre-computed analysis of the human proteome cross-analyzed against 62 vertebrate proteomes available in Ensembl genome browser[17]. For this analysis, the Aminode pipeline was executed using annotated vertebrate orthologs of human proteins, which resulted in the determination of the relative amino acid substitution rates and the identification of evolutionary constrained regions for a total of 18,713 human proteins.

Custom analysis. Users can analyze their proteins of interest by submitting the proteins' amino acid sequences and, optionally, their phylogenetic tree to profile the rate of evolutionary changes and identify ECRs using customizable parameters.

The Aminode Pipeline. Protein multiple alignments are obtained by using Multalin[18] (http://multalin.toulouse.inra.fr/multalin/) with default parameters. Columns containing gaps in more than 50% of aligned proteins are eliminated from the multiple sequence alignment. The phylogenetic tree is converted to a noded tree where the end nodes are the current species used in the analysis, and the ancestor nodes represent the last common ancestor of each branch. The Hartigan algorithm provides a framework for calculating best fits of a given tree according to a maximum parsimony approach[19] and is here used for calculating the minimum mutation fits at all aligned amino acid positions. Briefly, according to the parsimony criterion, the algorithm seeks a phylogenetic history that explains tree topology and/or amino acid changes with the fewest number of evolutionary events. In the Aminode pipeline, the tree topology is either fixed (the pre-computed analysis of the human proteome is based on comparison with species with known phylogenetic relationships) or calculated based on the input sequences in custom analyses (see below). Thus, in Aminode the Hartigan algorithm was used to infer amino acid identities in the ancestral nodes of the given evolutionary tree. In particular, for each amino acid position, a bottom-up procedure compares the amino acids from the child nodes to their immediate ancestral node and establishes that each ancestral node is equal to the intersection of its child nodes if the intersection is not empty (that is, if the child nodes share the same amino acid); otherwise, it is equal to their union (see example in Fig. 1). The subsequent top-down refinement retains, at each node, the amino acid that gives the minimum node substitution score[19] (NSS) (Fig. 1), which is assigned based on a modified BLOSUM62 Target Frequencies matrix[20] available at the NIH Repository (ftp://ftp.ncbi.nih.gov/repository/blocks/unix/blosum). We obtained a scoring system in which node substitution scores can vary from 0 to 1, with 0 denoting no changes (amino acid self-substitution), scores increasing with the rarity of BLOSUM62 substitution occurrence, and 1 as a theoretical maximum (amino acid substitution observed zero times). This was obtained by normalizing to 1 the sum of the frequencies of substitution (F_S) for each amino acid (including self-substitution, F_{SS}) to take into account differences in amino acid abundances, and then by calculating each node substitution score as $NSS = 1 - (F_S/F_{SS})$. A graphical representation of the matrix of amino acid substitution scores is reported in Fig. 2. For each position of the multiple alignment, a substitution score (SS) is calculated as the sum of all node substitution scores at that position. A relative substitution score is then obtained by diving the SS by the number of informative sequences (no. of sequences with an amino acid in that position); these values are finally normalized by the mean relative substitution score, and then averaged by using an 11-amino acid-long sliding window across the whole protein length, with two consecutive smoothing steps using a 7-amino acid-long sliding window[5]. The resulting profile

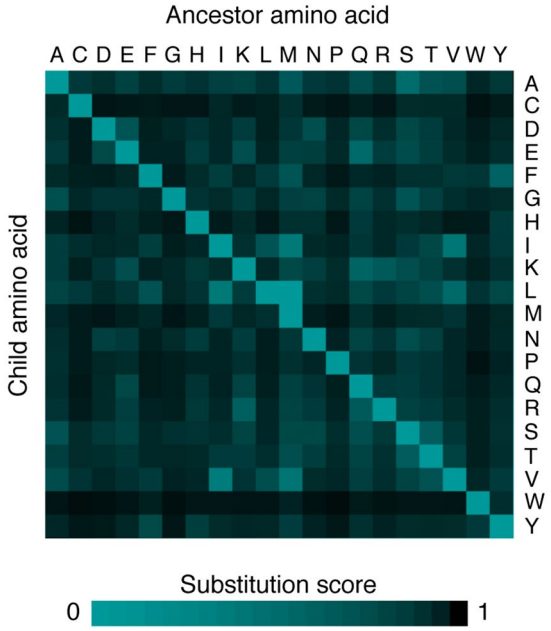

Figure 2. Heatmap of the ancestor-to-child amino acid substitution scores used in the Aminode pipeline.

describing the weighted relative rate of amino acid substitution is plotted as a function of alignment position in a two-dimensional array using the JFreeChart Java library (http://www.jfree.org/jfreechart). Scanning the array from the bottom (minimum) to the top (maximum) leads to the identification of local minima or evolutionarily constrained regions (ECRs), whose extent is defined by the closest proximal and distal positions where the second derivative of the plot is zero. Computed data are transferred to Excel files using the Apache POI Java library (https://poi.apache.org/) and are available for download.

To execute the analysis of the human proteome, the protein sequences and phylogenetic tree of 63 species (human plus 62 additional vertebrate species) were downloaded from the Ensembl genome browser[17] (release 84), and the Aminode pipeline was executed on each ortholog series. Proteins containing long stretches of incomplete sequences or long out-of-frame regions possibly derived from annotation errors were excluded from the analysis.

Backend Information. The Aminode website (www.aminode.org) is hosted on Heroku and uses the Spark Framework for the web server. Precomputed files are hosted on Microsoft Azure Storage and Github using Large File Storage (LFS), and bulk data are hosted on Google Drive. All files are retrieved via Aminode Search through link generation. The frontend uses the AngularJS framework, and the backend uses Java to process data and generate various output files.

Aminode Content. Aminode contains results from evolutionary constrained region analyses for human proteins that have at least two vertebrate orthologs annotated in Ensembl, Release 84 (18,713 proteins). Figure 3 shows the pipeline for the generation of Aminode graphs. Aminode pre-generated outputs provide a visual representation of the relative rate of amino acid substitution as a line plotted over the multiple sequence alignment (one example is reported below). Local minima indicate regions with low rates of substitution relative to the surrounding protein regions, while maxima indicate relative high rates. The valleys in the graphical output therefore indicate protein regions that are evolutionarily more constrained than the regions identified by the peaks. The positions of the predicted ECRs are marked by yellow bars placed above the multiple alignment. As a reference, the human protein index is reported on the top of the multiple alignment.

Aminode is searchable by the HGNC designated gene name (standard gene symbol). Full-resolution images and tabulated data can be downloaded from the links provided as result of the search. Computed data is downloaded as an Excel file that contains the information processed to execute the Aminode ECR analysis starting from the protein sequences. This file is provided for maximum ease in further processing of Aminode data. The Excel file includes the following tabs: (1) The "Substitution Scores" tab, which contains the human protein index, the filtered alignment index, the human amino acid sequence, the substitution scores, and the information relative to the relative substitution scores used to generate the Aminode graph (plot of relative substitution scores and ECR indexes). (2) The "Raw Substitution Scores" tab contains the raw aligned index, the human protein index, the human amino acid sequence, and the substitution scores. (3) The "Aligned Sequences" tab contains the full multiple alignment of the ortholog series. At the top of the alignment there are three indexes that track the sequences according to the Human Protein Index, Filtered Aligned Index, and Raw Aligned Index, respectively. The Filtered Aligned Index keeps track of the protein after filtering the data for gaps in the sequences. The Raw Aligned Index keeps track of the protein after the multiple alignment. The Human Protein Index keeps track of the original index of the human protein sequence.

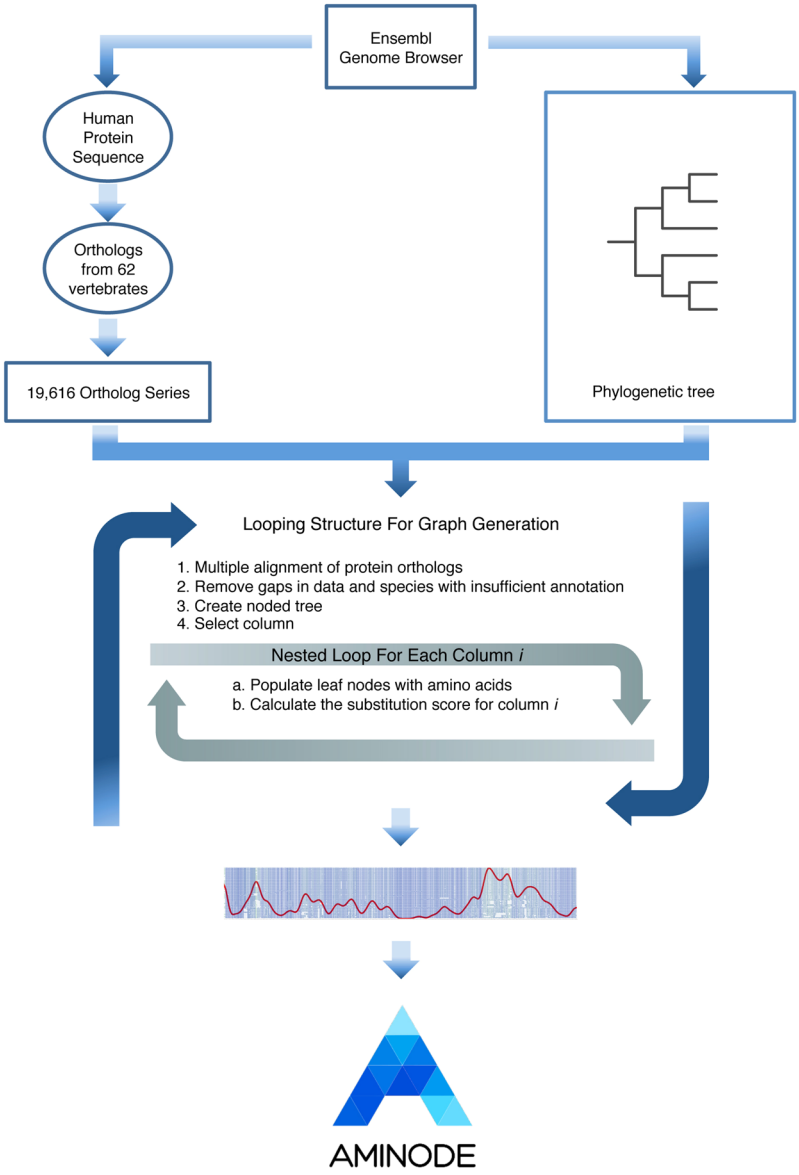

Figure 3. Schematic of the pipeline for the generation of Aminode graphs.

Links to a Gene Summary page (containing information automatically extracted from MyGene.info[21]) and to the queried gene's entry in other gene information sites, including NCBI[22], UniProt[23], and GeneCards[24] are provided. UniProt ID and NCBI ID are obtained from MyGene.info.

Additional links provide access to the amino acid sequences of all the orthologs used in the queried gene's Aminode-generated analysis, and to a Github Repository that contains the information generated and used in the Aminode analysis, available for download. In summary, each entry in Aminode provides access to a graph with the protein evolutionary profile plotted over the multiple protein alignment, raw data (original FASTA files), processed files (multiple alignments), list of rates of substitutions, scraped data, and excel files with the processed data formatted and graphed. Text files with bulk data (aligned and non-aligned sequences and relative substitution scores) are also available for download.

Custom Protein Analysis. The Aminode pipeline is also available to perform analyses with either a different species focus or a custom set of protein sequences. The user is prompted to (i) submit a set of protein sequences in standard FASTA format, and (ii) either submit a phylogenetic tree describing the protein evolutionary relationships in Newick format[25] or, alternatively, generate the tree via the option offered by Aminode, which uses the Multalin algorithm[18]. The names used to label proteins (or species) in the submitted protein sequence file must match the names of the leaf nodes in the submitted phylogenetic tree. The sequences do not need to be in any specific order. The user can adjust parameters such as filter threshold, font size and graph colors for the generation of the graphical output.

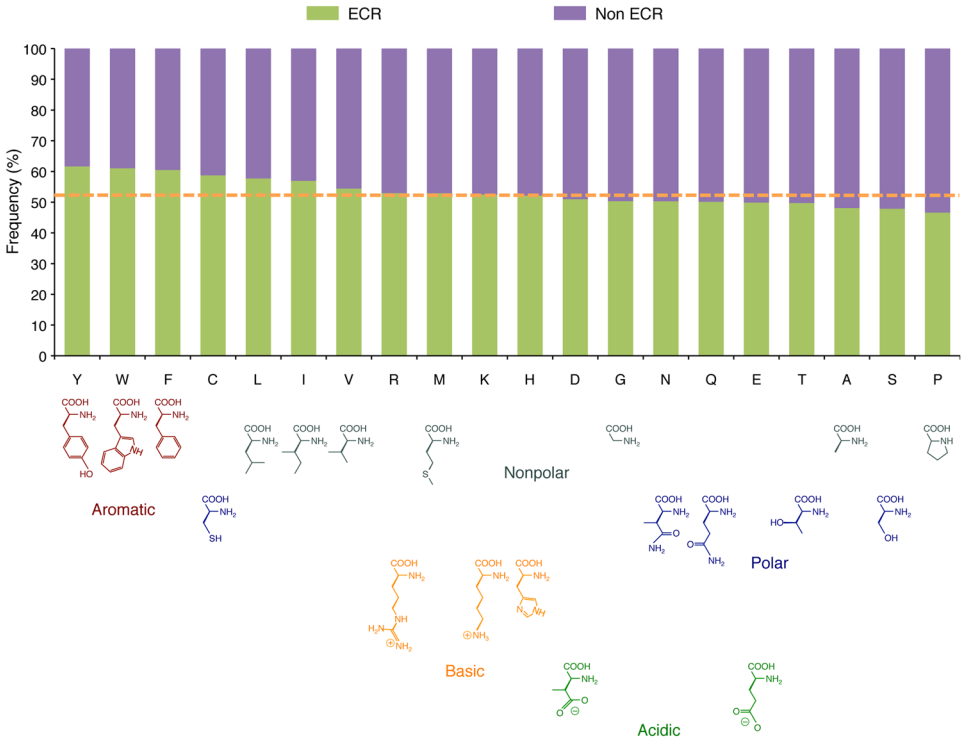

Figure 4. Frequencies of amino acids in evolutionarily constrained regions (ECRs) and non-ECRs across the human proteome. Amino acids are arranged left to right accordingly to their decreasing ECR/non-ECR frequency. The orange dotted line indicates the frequency expected by random distribution.

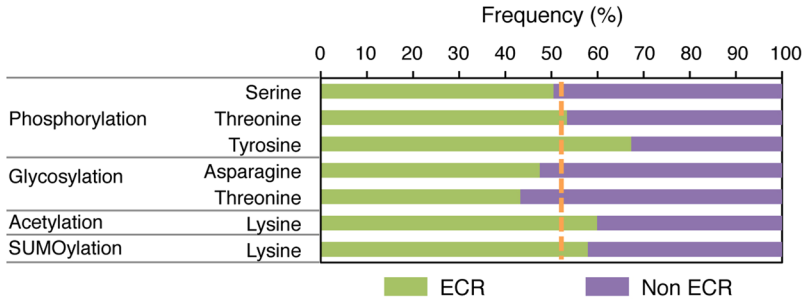

Figure 5. Frequencies of phosphorylated, glycosylated, acetylated and SUMOylated amino acids in evolutionarily constrained regions (ECRs) and non-ECRs across the human proteome. The orange dotted line indicates the frequency expected by random distribution.

Mapping amino acid frequencies, post-translational modifications and human missense variants. We investigated whether ECRs have different amino acid composition from non-ECRs by examining amino acid frequencies in ECRs and non-ECRs. The analysis showed that the aromatic amino acids (tyrosine, tryptophan and phenylalanine) have the most skewed distribution, showing a significant enrichment in ECRs (Bonferroni-adjusted Fisher's $P < 10^{-4}$ for all) (Fig. 4). Conversely, in the group of polar amino acids, serine, threonine, glutamine and asparagine are depleted, while cysteine is enriched, in ECRs ($P < 10^{-4}$ for all). The distribution of basic amino acid does not differ between ECRs and non-ECRs, while both glutamic acid and aspartic acid are depleted in ECRs ($P < 10^{-4}$ for both). Finally, among non-polar amino acids, leucine, isoleucine and valine are enriched, while alanine and proline are depleted, in ECRs ($P < 10^{-4}$ for all) (Fig. 4). We also examined the distribution of annotated sites[23] of the most common types of post-translational modification. While phosphoserine and phosphothreonine do not show skewed distribution, phosphotyrosine is enriched in ECRs ($P < 10^{-4}$) (Fig. 5), and all three have slightly increased frequencies in ECRs compared to their non-phosphorylated counterparts ($P < 10^{-4}$ for all), which may indicate structural or functional relevance for some of these sites. Acetylated or SUMOylated lysine also shows enrichment in ECRs ($P < 10^{-4}$ for both). Conversely, both glycosylated asparagine and threonine show significant depletion from ECRs, indicating relatively smaller structural conservation despite glycosylation relies on a target motif that extends shortly beyond the target amino acid[26]. This effect may depend

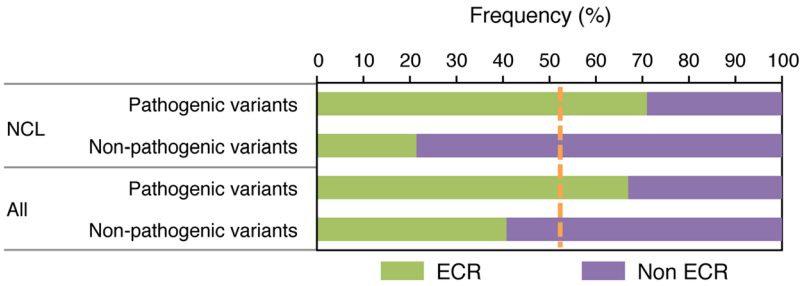

Figure 6. Distribution of pathogenic and non-pathogenic variants described in neuronal ceroid lipofuscinosis (NCL) proteins and in the human annotated proteome in evolutionarily constrained regions (ECRs) and non-ECRs. The orange dotted line indicates the frequency expected by random distribution.

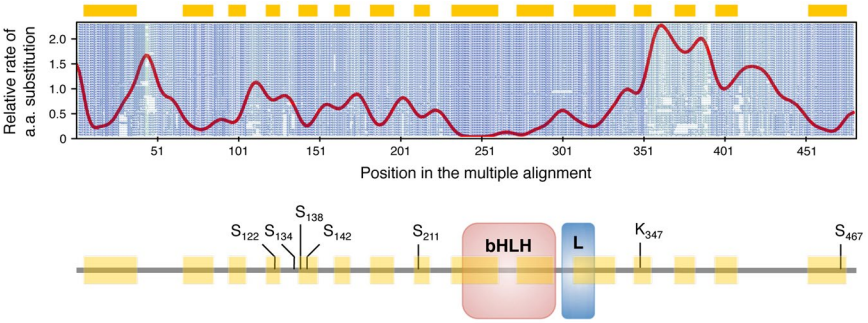

Figure 7. Aminode example: Analysis of transcription factor EB (TFEB). *Upper panel*, Aminode graphical output. The red line represents the relative rate of amino acid substitution calculated at each protein position. Local minima (highlighted by yellow bars above the graph) are constrained regions with relatively low substitution rates. Peaks (local maxima) indicate regions with relatively high substitution rates. The amino acid sequences for each ortholog are shown in shades of blue to green (more conserved to less conserved) in the background of the graph. *Lower panel*, Schematic of TFEB protein structure showing the position of the DNA-binding bHLH domain, the leucine zipper domain (L) and various known sites of post-translational modification that regulate TFEB function (references in the text).

on the fact that glycans, rather than amino acids, may direct protein interaction or function at the modified sites, or on a masking effect that bulky glycans may exert on the nearby amino acids, thus making them less available to selection-driving interactions.

We also investigated the distribution of known human missense variants in ECRs by examining the lists of pathogenic and nonpathogenic variants reported in UniProt[23]. We first focused on a group of neurodegenerative diseases named neuronal ceroid lipofuscinoses or Batten disease, for which high-quality annotations of pathogenic mutations are available[27]. Interestingly, 71% of annotated pathogenic missense mutations in Batten disease proteins map in ECRs, compared to 21% of nonpathogenic variants ($P < 10^{-4}$) (Fig. 6). Similarly, 67% of reported pathogenic variants across the entire proteome were found to fall within ECRs, compared to 41% of non-pathogenic variants ($P < 10^{-4}$).

Discussion

ECR analysis may help pinpoint protein sites that are under purifying selection over a certain evolutionary time scale. The selection maintains conservation at sites crucial to structure and function of the protein. Thus, from such observed evolutionary constraints one may deduce and predict the relative importance of specific protein sites[1–4]. Aminode enables the execution of complex sequence analyses in order to identify protein regions that are either evolutionarily constrained or unconstrained. The Aminode webtool allows researchers the swift identification of ECRs in proteins of interest and specifically provides the results of evolutionarily constrained region analyses for vertebrate proteome data available from Ensembl with a focus on the human proteome. In addition, Aminode enables user-customized analyses for proteins of interest.

The potential importance of *in silico* support for ECRs is multifold. First, ECRs can predict functional importance, providing researchers with key information to design their bench experiments. ECRs may indeed contain residues that are part of the active site in enzymes, map sites that are essential to the protein structure or function, and help identify post-translational modification sites[5–11]. The example reported in Fig. 7 represents the analysis of the transcription factor EB (TFEB), a master transcriptional regulator of lysosomal degradative pathways[28–30] that is being studied in our laboratory. The example reports a schematic of the structure of TFEB and shows that the DNA-binding bHLH domain, the leucine zipper domain, and six out of seven experimentally validated

post-translational modification sites of TFEB that regulate TFEB function[31–38] fall within Aminode-identified ECRs (Fig. 7), underlying the overlap between ECRs and functionally relevant sites.

Second, researchers executing experiments of protein manipulation could benefit from Aminode use. Terminal or internal protein tagging can be designed on the basis of Aminode analyses to select unconstrained regions to minimize the potential impact of the tag to the protein's function or interactions; conversely, targeted disruption of constrained regions may be used to experimentally identify essential protein sites. The identification of ECRs could also be useful to evaluate the potential impact of vector insertions in large-scale mutagenesis projects[39–44].

Third, due to the observed tendency of pathogenic variants to fall within ECRs, Aminode can serve as a tool to help evaluate which variants of unknown significance are more likely to be pathogenic and/or require further investigation. The integration of Aminode analysis with that of tools such as PhastCons[45] and PhyloP[46], which investigate evolutionary conservation at the nucleotide level, may provide a wider perspective on the potential impact of variants that cause changes in the amino acid sequence of a protein.

Aminode will be continuously updated as genome assemblies are updated and newly sequenced genomes become available and curated in Ensembl.

Data Availability. All data generated or analyzed during this study are available at the Aminode website: http://www.aminode.org.

References

1. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* **257**, 342–358, https://doi.org/10.1006/jmbi.1996.0167 (1996).
2. Lichtarge, O., Bourne, H. R. & Cohen, F. E. Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. *Proc Natl Acad Sci USA* **93**, 7507–7511 (1996).
3. Dean, A. M. & Golding, G. B. Enzyme evolution explained (sort of). *Pac Symp Biocomput*, 6–17 (2000).
4. Karchin, R., Cline, M. & Karplus, K. Evaluation of local structure alphabets based on residue burial. *Proteins* **55**, 508–518, https://doi.org/10.1002/prot.20008 (2004).
5. Sardiello, M., Annunziata, I., Roma, G. & Ballabio, A. Sulfatases and sulfatase modifying factors: an exclusive and promiscuous relationship. *Hum Mol Genet* **14**, 3203–3217, https://doi.org/10.1093/hmg/ddi351 (2005).
6. Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* **6**, e1001162, https://doi.org/10.1371/journal.pgen.1001162 (2010).
7. Ko, D. C., Binkley, J., Sidow, A. & Scott, M. P. The integrity of a cholesterol-binding pocket in Niemann-Pick C2 protein is necessary to control lysosome cholesterol levels. *Proc Natl Acad Sci USA* **100**, 2518–2525, https://doi.org/10.1073/pnas.0530027100 (2003).
8. Ota, M., Kinoshita, K. & Nishikawa, K. Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J Mol Biol* **327**, 1053–1064 (2003).
9. Kashuk, C. S. *et al*. Phenotype-genotype correlation in Hirschsprung disease is illuminated by comparative analysis of the RET protein sequence. *Proc Natl Acad Sci USA* **102**, 8949–8954, https://doi.org/10.1073/pnas.0503259102 (2005).
10. Jackson, P. J. *et al*. Structural and molecular evolutionary analysis of Agouti and Agouti-related proteins. *Chem Biol* **13**, 1297–1305, https://doi.org/10.1016/j.chembiol.2006.10.006 (2006).
11. Lin, R. J., Blumenkranz, M. S., Binkley, J., Wu, K. & Vollrath, D. A novel His158Arg mutation in TIMP3 causes a late-onset form of Sorsby fundus dystrophy. *Am J Ophthalmol* **142**, 839–848, https://doi.org/10.1016/j.ajo.2006.06.003 (2006).
12. Spatuzza, C. *et al*. Physical and functional characterization of the genetic locus of IBtk, an inhibitor of Bruton's tyrosine kinase: evidence for three protein isoforms of IBtk. *Nucleic Acids Res* **36**, 4402–4416, https://doi.org/10.1093/nar/gkn413 (2008).
13. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* **15**, 978–986, https://doi.org/10.1101/gr.3804205 (2005).
14. Binkley, J. *et al*. ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. *Genome Res* **20**, 142–154, https://doi.org/10.1101/gr.097121.109 (2010).
15. Simon, A. L., Stone, E. A. & Sidow, A. Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc Natl Acad Sci USA* **99**, 2912–2917, https://doi.org/10.1073/pnas.042692299 (2002).
16. Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449 (2000).
17. Yates, A. *et al*. Ensembl 2016. *Nucleic Acids Res* **44**, D710–716, https://doi.org/10.1093/nar/gkv1157 (2016).
18. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16**, 10881–10890 (1988).
19. Hartigan, J. A. Minimum Mutation Fits to a Given Tree. *Biometrics* **29**, 53–65, https://doi.org/10.2307/2529676 (1973).
20. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**, 10915–10919 (1992).
21. Xin, J. *et al*. High-performance web services for querying gene and variant annotation. *Genome Biol* **17**, 91, https://doi.org/10.1186/s13059-016-0953-9 (2016).
22. Coordinators, N. R. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, doi:gkw1071 (2016).
23. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43**, D204–212, https://doi.org/10.1093/nar/gku989 (2015).
24. Stelzer, G. *et al*. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr Protoc Bioinformatics* **54**, 1 30 31–31 30 33, https://doi.org/10.1002/cpbi.5 (2016).
25. Cardona, G., Rossello, F. & Valiente, G. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics* **9**, 532, https://doi.org/10.1186/1471-2105-9-532 (2008).
26. Ohtsubo, K. & Marth, J. D. Glycosylation in cellular mechanisms of health and disease. *Cell* **126**, 855–867, https://doi.org/10.1016/j.cell.2006.08.019 (2006).
27. Mole, S. E. & Cotman, S. L. Genetics of the neuronal ceroid lipofuscinoses (Batten disease). *Biochim Biophys Acta* **1852**, 2237–2241, https://doi.org/10.1016/j.bbadis.2015.05.011 (2015).
28. Sardiello, M. *et al*. A gene network regulating lysosomal biogenesis and function. *Science* **325**, 473–477, https://doi.org/10.1126/science.1174447 (2009).
29. Sardiello, M. Transcription factor EB: from master coordinator of lysosomal pathways to candidate therapeutic target in degenerative storage diseases. *Ann N Y Acad Sci* **1371**, 3–14, https://doi.org/10.1111/nyas.13131 (2016).
30. Sardiello, M. & Ballabio, A. Lysosomal enhancement: a CLEAR answer to cellular degradative needs. *Cell Cycle* **8**, 4021–4022, https://doi.org/10.4161/cc.8.24.10263 (2009).
31. Palmieri, M. *et al*. mTORC1-independent TFEB activation via Akt inhibition promotes cellular clearance in neurodegenerative storage diseases. *Nat Commun* **8**, 14338, https://doi.org/10.1038/ncomms14338 (2017).

32. Vega-Rubin-de-Celis, S., Pena-Llopis, S., Konda, M. & Brugarolas, J. Multistep regulation of TFEB by MTORC1. *Autophagy* **13**, 464–472, https://doi.org/10.1080/15548627.2016.1271514 (2017).
33. Martina, J. A., Chen, Y., Gucek, M. & Puertollano, R. MTORC1 functions as a transcriptional regulator of autophagy by preventing nuclear transport of TFEB. *Autophagy* **8**, 903–914, https://doi.org/10.4161/auto.19653 (2012).
34. Roczniak-Ferguson, A. *et al*. The transcription factor TFEB links mTORC1 signaling to transcriptional control of lysosome homeostasis. *Sci Signal* **5**, ra42, https://doi.org/10.1126/scisignal.2002790 (2012).
35. Settembre, C. *et al*. A lysosome-to-nucleus signalling mechanism senses and regulates the lysosome via mTOR and TFEB. *EMBO J* **31**, 1095–1108, https://doi.org/10.1038/emboj.2012.32 (2012).
36. Li, Y. *et al*. Protein kinase C controls lysosome biogenesis independently of mTORC1. *Nat Cell Biol* **18**, 1065–1077, https://doi.org/10.1038/ncb3407 (2016).
37. Miller, A. J., Levy, C., Davis, I. J., Razin, E. & Fisher, D. E. Sumoylation of MITF and its related family members TFE3 and TFEB. *J Biol Chem* **280**, 146–155, https://doi.org/10.1074/jbc.M411757200 (2005).
38. Palmieri, M., Pal, R. & Sardiello, M. AKT modulates the autophagy-lysosome pathway via TFEB. *Cell Cycle*, 1–2, https://doi.org/10.1080/15384101.2017.1337968 (2017).
39. Nord, A. S. *et al*. The International Gene Trap Consortium Website: a portal to all publicly available gene trap cell lines in mouse. *Nucleic Acids Res* **34**, D642–648, https://doi.org/10.1093/nar/gkj097 (2006).
40. Hansen, J. *et al*. A large-scale, gene-driven mutagenesis approach for the functional analysis of the mouse genome. *Proc Natl Acad Sci USA* **100**, 9918–9922, https://doi.org/10.1073/pnas.1633296100 (2003).
41. Austin, C. P. *et al*. The knockout mouse project. *Nat Genet* **36**, 921–924, https://doi.org/10.1038/ng0904-921 (2004).
42. Cobellis, G. *et al*. Tagging genes with cassette-exchange sites. *Nucleic Acids Res* **33**, e44, https://doi.org/10.1093/nar/gni045 (2005).
43. Venken, K. J. *et al*. MiMIC: a highly versatile transposon insertion resource for engineering Drosophila melanogaster genes. *Nat Methods* **8**, 737–743 (2011).
44. Schnutgen, F. *et al*. Genomewide production of multipurpose alleles for the functional analysis of the mouse genome. *Proc Natl Acad Sci USA* **102**, 7221–7226, https://doi.org/10.1073/pnas.0502273102 (2005).
45. Siepel, A. *et al*. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050, https://doi.org/10.1101/gr.3715005 (2005).
46. Cooper, G. M. *et al*. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–913, https://doi.org/10.1101/gr.3577405 (2005).

Acknowledgements

We thank Dr. Michael Kohn for helpful suggestions and critical reading of the manuscript. This work was supported by NIH grant NS079618 to M.S. and by a grant from the Beyond Batten Disease Foundation to M.S.

Author Contributions

M.S. conceived and supervised the study. K.T.C. and J.G. developed Aminode and performed all analyses. A.D.R. supervised the mutational analyses. K.T.C., J.G. and M.S. wrote the manuscript. All authors read and approved the final manuscript.

Additional Information

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.