Alternative Splice Variants in TIM Barrel Proteins from Human Genome Correlate with the Structural and Evolutionary Modularity of this Versatile Protein Fold

Adrián Ochoa-Leyva1,*, Gabriela Montero-Morán2, Gloria Saab-Rincón3, Luis G. Brieba2, Xavier Soberoñ1,3*

1 Instituto Nacional de Medicina Genómica (INMEGEN), México City, México, 2 Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Irapuato, Guanajuato, México, 3 Departamento de Ingeniería Celular y BiocatalISIS, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

Abstract

After the surprisingly low number of genes identified in the human genome, alternative splicing emerged as a major mechanism to generate protein diversity in higher eukaryotes. However, it is still not known if its prevalence along the genome evolution has contributed to the overall functional protein diversity or if it simply reflects splicing noise. The (ββ) barrel or TIM barrel is one of the most frequent, versatile, and ancient fold encountered among enzymes. Here, we analyze the structural modifications present in TIM barrel proteins from the human genome product of alternative splicing events. We found that 87% of all splicing events involved deletions; most of these events resulted in protein fragments that corresponded to the (ββ) (ββ) (ββ) (ββ) (ββ) subdomains of TIM barrels. Because approximately 7% of all the splicing events involved internal β-strand substitutions, we decided, based on the genomic data, to design β-strand and α-helix substitutions in a well-studied TIM barrel enzyme. The biochemical characterization of one of the chimeric variants suggests that some of the splice variants in the human genome may be evolving novel functions via either the oligomeric state or substrate specificity. We provide results of how the splice variants represent subdomains that correlate with the independently folding and evolving structural units previously reported. This work is the first to observe a link between the structural features of the barrel and a recurrent genetic mechanism. Our results suggest that it is reasonable to expect that a sizeable fraction of splice variants found in the human genome represent structurally viable functional proteins. Our data provide additional support for the hypothesis of the origin of the TIM barrel fold through the assembly of smaller subdomains. We suggest a model of how nature explores new proteins through alternative splicing as a mechanism to diversify the proteins encoded in the human genome.

Citation: Ochoa-Leyva A, Montero-Morán G, Saab-Rincón G, Brieba LG, Soberoñ X (2013) Alternative Splice Variants in TIM Barrel Proteins from Human Genome Correlate with the Structural and Evolutionary Modularity of this Versatile Protein Fold. PLoS ONE 8(8): e70582. doi:10.1371/journal.pone.0070582

Editor: Stefan Maas, NIGMS, NIH, United States of America

Received February 13, 2013; Accepted June 20, 2013; Published August 12, 2013

Copyright: © 2013 Ochoa-Leyva et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by National Institute of Genomic Medicine grants 08/2012/I and 07/2012/E. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: aochoa@inmegen.gob.mx (AOL); xsoberon@inmegen.gob.mx (XS)
¤ Current address: Unidad de Genómica de Poblaciones Aplicada a la Salud, Facultad de Química, UNAM-Instituto Nacional de Medicina Genómica (INMEGEN), México City, México

Introduction

Protein functions and folds evolve through changes in the amino acid sequence such as insertions, deletions and substitutions [1]. Furthermore, structural and functional diversity can be generated through rearrangements of protein modules taken from pre-existing archetypical domain repertoires. Sequence diversity can be generated by both homologous and nonhomologous recombination processes. Nonhomologous recombination has been theorized to be the most effective mechanism at enabling new structures and functions to emerge during protein evolution [1], suggesting that combinatorial assembly and modular substitutions between segments of unrelated proteins play an important role in evolution [2]. In addition, alternative splicing, which combines small gene segments and occurs as a normal process in eukaryotes, greatly increases the sequence diversity of proteins encoded in a genome [3]. Through alternative splicing, a single gene may encode multiple proteins. For instance, it is estimated that up to 95% of multi-exon genes are alternatively spliced in humans [4]. Alternative splicing is a regulatory process working independently of transcriptional regulation and provides an additional control of tissue-specific gene expression [5–8]. Furthermore, alternative splicing can also influence the mechanisms of enzyme activity regulation, protein oligomerization and protein-protein interactions [9]. However, there is still limited knowledge derived from experimental data regarding the structural and functional consequences of alterations in proteins resulting from alternative splicing events and their role in expanding the functionality of a eukaryotic proteome [9–11].

The large increase of available RNA sequences for the human genome, which were exponentially expanded by the incursion of Next-Generation Sequencing technologies, has opened the possibility for analyses of the role that insertion, deletion, and
subdomains were reassembled noncovalently to form the catalytic sequences of previous information about the functional and structural consequences of the expansion of human proteome diversity. In the absence of provide insights regarding the evolution of this fold and its role in Furthermore, we discuss how the alternative splicing process can the speculation that it evolved through duplication and fusion of ancient fold encountered among the known enzymes [17–19]. Together, these studies suggest that many (\(
\beta\alpha\)8) barrel proteins derived from the human genome as a model enzyme. The experiments consisted of the substitution of one \(\beta\)-strand and one \(\alpha\)-helix of the E. coli TrpF by gene segments derived from a different enzyme. Our analyses reinforces the notion that the (\(\beta\alpha\)8) barrel proteins can be made up of structural subdomains. We suggests a model of how the barrel subdomains can be rearranged to diversify the (\(\beta\alpha\)8) barrel proteins encoded in the human genome as opposed to just noise. Our model offers an explanation to the large diversity of proteins generated through alternative splicing.

Results
Identification of the (\(\beta\alpha\)8) barrel proteins encoded in the human genome
First we retrieved all of the predicted proteins containing structural superfamily assignments from the human genome using Gene3D (Figure 1) [30]. From the 14,708 protein-coding genes of the human genome that have at least 1 CATH structural superfamily assignment, we retrieved 172 protein-coding genes containing a predicted (\(\beta\alpha\)8) barrel domain (Table S1). Genes were selected for further analysis if, according to the UniProtKB database, there is experimental evidence for proteins derived from alternative splicing or at least one mRNA with correct intron/exon boundaries [31]. Out of the 172 putative (\(\beta\alpha\)8) barrel proteins, 70 protein coding genes have experimentally confirmed

22,111 protein coding genes of the human genome

14,708 proteins have a structural superfamily assignment

172 proteins are predicted as containing the (\(\beta\alpha\)8) barrel domain (Table S1)

70 proteins have experimental confirmed splice variants (Table S2)

135 splice variants (Table S3)

Mapping the 135 splice variants onto the structures of the 70 full-length proteins (Table S4)

Image 1. Pipeline of bioinformatics analysis to identify the (\(\beta\alpha\)8) barrel proteins with splice variants in the human genome. Summary of our bioinformatics data flow to extract the 135 experimentally confirmed (\(\beta\alpha\)8) barrel splice variants in the human genome.
doi:10.1371/journal.pone.0070582.g001
splice variants (Table S2), resulting in a total of 135 splice variants (Table S3).

Mapping of the splice variants onto the structures of the full-length proteins

Twenty-six of the 70 full-length proteins have a structure reported in the Protein Data Bank, and high-quality structural models were obtained for 44 additional proteins using ModBase (see Materials and Methods). To identify the structural changes as a consequence of the alternative splicing process, we mapped all of the 135 splice variants onto the structures of the 70 full-length proteins (Table S4). We inferred that the alternative splicing events directly affected the structure of the (β2)8 barrel for 67 splice variants (Table S5). This set of 67 splice variants represent a small subset of all possible for the (β2)8 barrels predicted for the human genome. However, it contains curated information on the experimental existence of the mRNA from the splice variants. In addition, the corresponding full-length proteins of this set contain protein members that represent 11 of the 17 CATH superfamilies found for all of the (β2)8 barrel proteins from the human genome (Table S6). The majority of the alternative splicing events fall into the loops or in the first residue of the β-strand or α-helix elements of the (β2)8 barrel structure. Interestingly, alternative splicing events fall in the middle of a β-strand or α-helix element only in 13% of cases. The fact that splicing events fall into non-structured regions of the (β2)8 barrel structure, suggest the presence of well-structured complete subdomains as exchanging units.

Analysis of structural modifications of the (β2)8 barrel as a consequence of alternative splicing

We found that 54/67 splicing events involve sequence deletions in which a part of the full-length protein sequence was removed (Figure 2a). These results are in agreement with previous observations that demonstrated that splice variants with deletions are the most abundant of all splicing events [11]. The majority of splicing events resulted in protein fragments corresponding to the (β2)4, (β2)5, (β2)6, and (β2)7 subdomains of the barrel (Figure 2a). The remaining 13 alternative splicing events resulted in (β2)8 barrel proteins with a loop insertion/deletion, β-strand, and (β2) substitution and α-helix deletion. Figure 2b shows the subset of the 13 alternative splice variants that was predicted to cause modifications to (β2)8 barrel proteins without altering their overall structure. Interestingly, β-strand substitutions that occur in the inner part of the barrel are the majority of the splicing events from this subgroup (Figure 2b) and constitute 7.4% of all splice variants (Figure 2a).

Protein engineering experiments comparable to the structural changes of splice variants

The large data set of experiments with the (β2)8 barrel allows us to analyze if the structural changes found in the splice variants have been previously observed in protein engineering experiments and to explore its consequences on the structure-function of the (β2)8 barrel proteins (Table 1). To that end, we also examined the 67 splice variants for experimental functional annotations in the UniProtKB and we found them (either at protein level, protein solubility, and differential gene expression) for 17 splice variants (Table 2). Many of these splice variants containing barrel subdomains have been previously confirmed at the protein level and were found also to be differentially expressed in different human tissues (Table 2). Interestingly, the subdomain configurations found in the splice variants (Figure 2a) have been previously observed by protein engineering experiments as soluble and stable protein fragments or as structured folding intermediates in different (β2)8 barrel proteins (Table 1) [23,28,29,36]. Proteins with a (β2)7 composition were the most abundant splice variants. Three of the four splice variants that contain the (β2)7 subdomain were previously reported as soluble proteins (Table 2) and this subdomain, which contains only seven β-strands, has been reported as a natural deviation from the canonical topology occurring in certain cellulases, flavoproteins, and in different members of the nicotinate/quinolate PRTase C-terminal domain-like superfamily (Table 1) [38-40]. We found that splice variant Q00722–2 has a deletion of four residues in a β-loop and has been reported as a soluble and active enzyme (Table 2). Sequence insertion/deletion and substitution at β-loops have been shown to have no significant disruption of the structure and function in some members of the fold (Table 1) [24-26].

From the sequence substitutions observed in splice variants, the β-strand substitution was the most common genetic event constituting 7.4% of all splice variants (Figure 2a). Splice variant CSMRZ0-2 has the substitution of the β1 of the barrel and this protein variant has been previously reported as a soluble and catalytically active enzyme (Table 2). However, functional and structural consequences of β-strand or α-helix substitutions have not been reported on protein members of the fold.

β-strand and an α-helix substitutions in a model enzyme of the (β2)8 barrel fold

The Phosphorybysyl anthranylate isomerase (TrpF, enzyme, EC 5.3.1.24) from E. coli, a well-studied (β2)8 barrel involved in tryptophan biosynthesis, was selected as a scaffold for the substitution of the β-strand 7 and α-helix 3 (Figure S1). The β-strand and α-helix elements used as substitutions were selected from different structural positions of an enzyme not functionally related to the PRA isomerise activity of TrpF. These structural positions were chosen to avoid functional, structural, and evolutionary relationships, as previously observed in the substitution of β2-loops [25]. Thus, the α-helix 7 and β-strand 3 from the MetR enzyme (methyltransferase), were selected (Figure S1). In order to obtain folded and functional variants for further experimental analysis, we included the introduction of sequence diversity at the initial and terminal positions of each substitution as was previously suggested for β2-loops substitutions. Two libraries were constructed on the TrpF scaffold: a α-helix library that has the α-helix-7 of MetR replacing the original α-helix-3, and a β-strand library that has the β-strand-3 of MetR replacing the original β-strand-7 (Figure S1).

Because sequence diversity was introduced in both libraries, we analyzed the resulting abundance of folded and functional proteins after the application of corresponding selection pressures. The folded sequence space for the β-strand 7 library was higher than that of the α-helix 3 library: 80% and 39%, respectively (Figure S2). Despite the high number of folded variants observed in the β-strand library, only 4% of the total variants were functional, and these data are in agreement with the 6% observed in the α-helix library (Figure S2). Our results show that both types of secondary-structure substitutions can be tolerated by a (β2)8 barrel structure, although the original β-strand and α-helix elements that were replaced are located in conserved regions with different amino acid lengths compared to TrpF (Figure S1).
Sequence analysis of functional variants and biochemical characterization of a chimeric variant

To investigate which amino acids were selected under functional constraints, we performed a sequence analysis of the variants complementing the PRA isomerase activity and compared those with the residue frequency observed without this selective pressure in each library. After normalizing the observed frequencies of each amino acid under selective versus nonselective pressure (Tables S7, S8, S9, and S10), we found a statistically significant overrepresentation of certain amino acids (Figure 3a and 3b). To compare the selected residues of the chimeric variants with the residues naturally found in the wild-type enzymes, we performed a structural alignment of the TrpF and MetR enzymes, and the sequence distribution observed at the corresponding randomized positions is shown in Figure S3. From this analysis, high sequence conservation for the positions preceding the swapped elements was observed (Figure S3). The amino acids found as over-represented in the functional chimeric TrpF variants (Figure 3a and 3b) are similar to the amino acids at this position in wild-type TrpF enzymes (Figure S3). In contrast, these amino acids are markedly different from the most conserved amino acid in the donor or host protein for the corresponding position. The selected residues in the variability positions of both \( \beta \)-strand and \( \alpha \)-helix elements have a preference for the original residue in TrpF.

To analyze the functional and structural effects of \( \beta \)-strand or \( \alpha \)-helix substitutions on the TrpF scaffold at the molecular level, we performed steady-state enzyme kinetics and structural studies in a...
chimeric variant that retained function. We selected a variant from the β-strand library because the replacement of an internal strand from a β-sheet can result in the loss of hydrogen bonds on both sides of the strand and requires the formation of several new ones to retain the native-like structure of the (βββ) barrel. The analyzed variant, Beta_1, showed a 5-fold decrease in its kcat/Km with respect to the PRA isomerase activity of the wild-type TrpF, largely due to a reduction in kcat (Table 3). Hence, the modification of the original thermal melting temperature (Tm, app) of 55°C suggests only slight structural changes at the secondary structure level as a consequence of the β-strand substitution relative to the wild-type enzyme (Figure 4a). A thermal denaturation curve shows that the Beta_1 variant forms a stable structure with an apparent melting temperature of 55°C, which is 5 degrees higher than the Tm of the wild-type enzyme (Figure 4b and Table 3). This change in thermostability can be related to the increase of the dimer population of this variant (Figure 4c), as previously suggested for other mutants of the TrpF enzyme [25].

All the experimental analyses of the β-strand and α-helix substitution can improve our understanding of the biological role that these structural changes may be playing in protein evolution and in the functional expansion of the human proteome.

### Discussion

On the functional level, splicing has been shown to contribute the structural and functional diversity of proteins [9]. The effects of alternative splicing on the novel functions of proteins range from changes in substrate or interacting partner specificity to the regulation of DNA-binding properties, the subcellular localization and the allosteric regulation sites in a target protein [9,53,54]. Alternative splicing can also influence the functional diversity at the level of protein expression [9,54]. After extensive search of the biological literature for our splice variants, we found 15 splice variants with evidence for functional tissue-specific expression in the cell and 14 splice variants with evidence at protein level (Table 2). In addition, we identified clear evidence of stable protein products originating from such splice variants for the (βββ) barrel (Table 2). It has been previously suggested that variants Q13231-3 of gene CHIT1_HUMAN and Q8TDX5-2 of gene ACMSD_HUMAN might still be able to bind their substrates [48,49]. The authors suggest that if this is the case, the metabolic relevance of the splice variant Q8TDX5-2 would be related to its capacity for binding and sequestering a reactive intermediate [49]. Variant P16278-2 of gene BGAL_HUMAN also termed EBP (elastin binding protein) in which half of the barrel was removed is an enzymatically inactive splice variant of lysosomal beta-galactosidase, but plays functional roles in the formation of extracellular elastic fibers (elastogenesis) and in the development of connective tissue [45]. Together, these observations suggest the existence of novel functional diversity for several splice variants of (βββ) barrel proteins. It has been previously suggested that species-specific splice variants, relative to conserved splice variants, may less frequently play important functional roles [55]. In this regard, we found that the majority of splice variants have been conserved in more than two species (data not shown). We propose that these splice variants are more likely to maintain critical gene activities and they are not reflecting splicing noise or simply the result of selective pressure against insoluble proteins.

The results of our protein engineering experiments of β-strand substitution are consistent with the notion that alternatively spliced variants with β-strand substitutions, which was the main type of all

### Table 2. Functional annotations of human splice variants.

| Barrel subdomain | Gene name      | Splice variant | Detected at protein level | Solubility | Catalytic activity | Tissue-specific gene expression | Refs. |
|------------------|----------------|----------------|--------------------------|-----------|-------------------|---------------------------------|-------|
| (βαβ)_3          | HYAL3_HUMAN    | Q43820-4       | Yes*                     | ND        | Inactive          | Yes                             | [43]  |
| (βαβ)_2          | HYAL3_HUMAN    | Q43820-3       | Yes*                     | ND        | Inactive          | Yes                             | [43]  |
| (βαβ)_1          | HYAL1_HUMAN    | Q12794-5       | Yes*                     | ND        | Inactive          | Yes                             | [43]  |
| (βαβ)_4          | ADC_HUMAN      | Q96A70-5       | Yes*                     | ND        | Inactive          | Yes                             | [44]  |
| (βαβ)_3          | HYAL1_HUMAN    | Q12794-3       | Yes*                     | ND        | Active            | Yes                             | [43]  |
|                  | BGAL_HUMAN     | P16278-2       | Yes                      | Soluble   | Inactive          | Yes                             | [45]  |
|                  | CHIA_HUMAN     | Q98ZP6-3       | ND                       | ND        | ND                | Yes                             | [46]  |
| (βαβ)_3          | HYAL1_HUMAN    | Q12794-4       | Yes*                     | ND        | Inactive          | Yes                             | [43]  |
| (βαβ)_3          | HMGCL_HUMAN    | P35914-2       | ND                       | ND        | ND                | Yes                             | [47]  |
| (βαβ)_7          | HYAL3_HUMAN    | O43820-2       | Yes*                     | ND        | Inactive          | Yes                             | [43]  |
|                  | HYAL1_HUMAN    | Q12794-2       | Yes                      | Soluble   | Inactive          | Yes                             | [43]  |
|                  | CHIT1_HUMAN    | Q13231-3       | Yes                      | Soluble   | Inactive          | ND                              | [48]  |
|                  | ACMSD_HUMAN    | Q8TDX5-2       | Yes                      | Soluble   | Inactive          | Yes                             | [49]  |
| (βαβ)_6 with loop insertion or deletion | PLCB2_HUMAN | Q00722-2       | Yes                      | Soluble   | Active            | Yes                             | [50]  |
| (βαβ)_6 with βα substitution | ADC_HUMAN | Q96A70-4       | Yes*                     | ND        | Inactive          | Yes                             | [44]  |
| (βαβ)_5 with βα substitution | NPL_HUMAN | Q98XDS-2       | ND                       | ND        | ND                | Yes                             | [51]  |
| (βαβ)_4 with ββ strand substitution | AK1BF_HUMAN | C9JZK2-2       | Yes                      | Soluble   | Active            | ND                              | [52]  |

*Reported by in vitro translation. ND: Not determined.

doi:10.1371/journal.pone.0070582.t002
the observed substitutions, can represent $\beta_5$ barrels that retain their original function. One example is the splice variant C9JRZ8-2 from AK1BF_HUMAN, where the original $\beta_1$ of the barrel is substituted by another sequence, which was reported as an expressed, soluble, and active enzyme [52]. Moreover, alternative splicing allows a single gene to produce several splice variants which may affect several properties, such as structure, function, binding properties and stability of the encoded proteins [9]. For instance, we found that in the Beta_1 variant, the single $\beta$-strand substitution changes its stability, its catalytic activity and its capacity to oligomerize. Furthermore, the Beta_1 variant showed a 5-fold decrease in its $k_{\text{cat}}/K_m$ with respect to the PRA isomerase activity of the wild-type TrpF, largely due to a reduction in $k_{\text{cat}}$ (Table 3). This work also reinforces the notion that DNA swapping of unrelated sequences can be relevant for the generation of molecular diversity [1,2]. Although introducing diversity at both ends of the secondary-structure substitutions cannot be explained by alternative splicing events, one has to assume that extant proteins with functional splice variants were subject to some prior sequence variation during evolution. For instance, Tawfik and co-workers reported that compensatory substitutions may follow short insertions and deletions (InDels) accumulation [56]. The authors observed increased substitutions rates in the sequential vicinity of InDels. The primary assumption is that these correlated substitutions compensate for the deleterious effects of InDels and are therefore fixed by positive selection [56].

We found that the majority of splice variants are formed by $\beta_5$ and $\beta_7$ subdomains. Interestingly, the $\beta_7$ subdomain has been reported as an independently folding substructure in different barrel enzymes (Table 1). This subdomain can also maintain a significant amount of secondary structure and native-like tertiary conformation with a propensity to dimerize [28,29]. Deviations from the canonical topology of the barrel structure have been reported. For example, quinolinic acid phosphoribosyltransferase, certain cellulases and flavoprotein contain only seven $\beta$-strands [38–40]. The existence of natural proteins containing the $\beta_7$ subdomain suggests that the splice variants resulting in this subdomain can retain a soluble and stable structure. The experimental existence of smaller substructures in the folding

| Table 3. Biochemical characterization of a chimeric variant. |
|---------------------------------|
| Variant | $k_{\text{cat}}$ (min$^{-1}$) | $K_m$ (uM) | $k_{\text{cat}}/K_m$ (M$^{-1}$ s$^{-1}$) | $T_m, \text{app}$ (°C) $^a$ |
|---------------------------------|
| Beta_1  | 0.015 | 124.3 | 8.9 | 54.5 |
| TrpF$^b$ | 2028 | 28.9 | 1.2×10$^9$ | 49.3 |

$^a$The overall standard errors of enzyme kinetic parameters are less than 20%.
$^b$The apparent thermal melting temperature (°C).

$^\text{Data obtained from Ochoa-Leyva et al., 2011 [25].}$

doi:10.1371/journal.pone.0070582.t003
process and soluble and stable subdomains formed by $\langle \beta \alpha \rangle_2$, $\langle \beta \alpha \rangle_4$, $\langle \beta \alpha \rangle_5$ or $\langle \beta \alpha \rangle_6$ have been demonstrated in different $\langle \beta \alpha \rangle_8$ barrel proteins [20–22,28,29]. Structural and functional resilience towards substitutions of $\beta \alpha$ subdomains and $\beta \alpha$-loops have also been demonstrated in this fold [24–26,32] which is in agreement with our observations of resilience to $\beta$-strand substitutions. Based on all of the experimental work surrounding diverse protein members of the $\langle \beta \alpha \rangle_n$ barrel fold, we suggest that it is reasonable to hypothesize that a sizeable fraction of splice variants found in the human genome may represent structurally viable functional proteins.

Our results suggest that parts of the genes that encode different subdomains, such as the $\langle \beta \alpha \rangle_2$, $\langle \beta \alpha \rangle_4$ and $\langle \beta \alpha \rangle_6$ part-barrels, can be fused, mixed, and matched through genomic evolution to yield new $\langle \beta \alpha \rangle_n$ barrel proteins (Figure 5). This is in agreement with hypotheses on the origin of the $\langle \beta \alpha \rangle_n$ barrel fold [22]. One hypothesis suggests that they have evolved by tandem duplication and fusion of an ancestral half-barrel [20]. The half-barrel can be formed by two rounds of gene duplication and fusion from a quarter-barrel ancestor [21]. An alternative evolutionary hypothesis suggests that three-quarter-barrels were viable evolutionary intermediates and that the combinatorial assembly of diverse pools of part-barrel subdomains gave rise to the many distinct lineages of $\langle \beta \alpha \rangle_n$ barrel proteins [22]. Our analyses are consistent with both hypotheses, as they reveal many splice variants where a quarter-barrel, a half-barrel, and a three-quarter-barrel were created by alternative splicing events. In addition, we suggest that some of the splice variants function as homo-dimers or hetero-dimers to reconstruct the complete $\langle \beta \alpha \rangle_n$ barrel in vivo (Figure 5). This notion is in agreement with the in vivo and in vitro subdomain reassembly previously observed for different barrel enzymes [23,34,57], including the reassembly of two half-barrels derived from different enzymes [50]. The expression of various alternative splice variants might offer an increased functional expansion of the proteome through combination of quarter-barrels, half-barrels, and three-quarter-barrels containing different functional sites in homodimeric or heterodimeric complexes (Figure 5). For instance, splice variants P16278-2, Q12794-3 and Q9BZP6-3, where one half of the barrel is removed, have been described to be expressed in a tissue-specific manner, and experimental validation of the existence of a stable protein product demonstrated for variant P16278-2 (Table 2).

Sequence diversity resulting from alternative splicing events is maintained in eukaryotic genomes throughout evolution. The human genome revealed a much smaller number of genes than anticipated, strengthening the notion that alternative splicing is a major factor in expanding protein diversity. The decade after the completion of the human genome sequencing has seen dramatic developments in sequencing technologies. These developments have generated a plethora of new sequences originating from alternative splicing events and have demonstrated that these events are not only involved in normal molecular processes but are also associated with human diseases such as cancer [59–61]. This work highlights the necessity of studying the folded and functional sequence space of alternatively spliced protein products in the laboratory and the importance of such studies in answering questions about the development and pathogenesis of many diseases where alternative splicing is involved. Our work suggests the possibility of crossing borders in sequence space without changing the natural robustness of a $\langle \beta \alpha \rangle_n$ barrel and reinforces the hypothesis that $\langle \beta \alpha \rangle_n$ barrel proteins appear to be considerably more tolerant to deletions, insertions, and replacements than previously stated.

Figure 4. Structural analysis of variant Beta_1. The variant Beta_1 has a leucine residue in the N-terminal variable position and a glycine residue in the C-terminal variable position. A. Far-UV CD spectra. B. Thermal unfolding curves. C. Analytical gel-filtration chromatograms. A. The variant Beta1 showed only slight structural changes at the secondary structure level with respect to the wild-type enzyme. B. The variant Beta_1 (brown squares) has a $T_m$ that is 5 degrees higher than the wt-TrpF enzyme (green triangles). C. The variant Beta_1 has an increase in the dimer population relative to the wild-type TrpF.

doi:10.1371/journal.pone.0070582.g004
Materials and Methods

Annotation of (βα)₈ barrel proteins from human genome

The predicted (βα)₈ barrel domain proteins encoded in *Homo sapiens* (taxon ID:9606) genome (Ensembl, GRCh37) were obtained using the fold recognition algorithms of the Gene3D program (v:11.0.0, CATH: v3.5.0) [30]. From the 22,111 protein-coding genes annotated for the human genome, 14,708 have at least one CATH Superfamily assignment. From these, 172 protein-coding genes were predicted with the CATH topology: 3.20.20, which corresponds to the (βα)₈ barrel domain (Table S1). The splice variants for each gene were obtained from UniProtKB, and only the protein coding genes with experimental evidence that their splice variants exist at the mRNA level were selected (Table S2), resulting in 135 splice variants (Table S3). The existence at mRNA level of all the isoforms was also confirmed by searching in VEGA and ENSEMBL databases. The identification of the full-length protein sequence (canonical sequence) was extracted from the UniProtKB [31] database for each gene. The so-called “canonical sequence” is the most prevalent sequence and/or the most similar sequence among orthologous species [31]. The three types of splicing events (insertion, deletion and substitution) were identified after the sequence alignment of all splice variants for each gene, and all positional information annotated in the Table S4 and Tables S5 refers to the canonical sequence.

Structures and models of the Full-Length (βα)₈ barrels

The three-dimensional structures of the 70 full-length proteins were obtained from two sources. We first obtained the structures for the full-length proteins through mapping their amino acid sequences to the Protein Data Bank [62]. This searching returned PDB structures for 26 proteins. We then obtained structural models for the rest of the full-length proteins in our data set by searching their sequences in the MODBASE database [63], resulting in 44 structural models of high quality. All structural models were also optimized through energy minimization functions using the Chimera Package from UCSF [64] and through the Repair Object function of the FOLDX program [65,66].

Mapping spliced variants to the (βα)₈ barrel structures

To identify the structural changes produced by the alternative splicing events in the original full-length proteins, each alternative splice variant was mapped to its corresponding protein structure, and the features of the regions being affected were analyzed and annotated manually using the Swiss-PDB viewer [67].

Library design and construction of β-strand and α-helix substitutions

The two libraries were constructed using a previously reported strategy [25]. The diversity was generated by replacing the normal codon with an NNS codon at the respective positions and the replacements were performed using an overlap PCR strategy. Briefly, two oligonucleotides, which are partially complementary (12 bp), were designed for each secondary-structure replacement, one oligonucleotide corresponds to the noncoding DNA strand and the other oligonucleotide to the coding DNA strand. The coding oligonucleotides for α-helix and β-strand substitutions were 5’-CGGACCTTCCCTGCGATGGCGCCNNSTC-
GCTGGGGGGAGTAAA3’ and 5’-TCGCTGGGCTATATCGGTATACTAACCGGNSAT- GATCNNSGGGGGTCTAGGGCGAGAT3’, respectively. The noncoding oligonucleotides were 5’-AAAG- GAAGTGGGTTTAATCGGGNSATCGTTGAAG- GAACAC-3’ and 5’-TAAGCCGCNNNGATCTATNNSGGC- CAAGCGATTGACCATT-3’, respectively. In all oligonucleotides, an NNS codon (underlined) that replaced the variability positions was introduced. Libraries were independently constructed by PCR and final products were ligated into the pDAN3 plasmid. The resulting libraries contained 103 different variants that broadly covered the theoretical sequence diversity of 1024 variants for each library. Approximately 10 plasmids for each library were sequenced to confirm the corresponding β-strand or α-helix substitution and to analyze the statistical distribution of the sequence diversity introduced at both variable positions. From the sequence analysis of the pool of variants from each library, we can conclude that approximately 99% of the generated variants were correctly constructed and the sequence diversity introduced at both ends of the β-strand and α-helix elements is according to the sequence distribution for a NNS codon (data not shown).

Assignment of folded and functional sequence spaces

Folded and functional sequence spaces were estimated using the strategy previously reported to analyze the Structure-Function Loop Adaptability in βαβ8 proteins [25]. Functional sequence space was calculated as the ratio of the number of variants that maintain the functional proficiency for PRA isomerase activity. To this end, the number of CFU (colony forming units) complementing the tryptophan auxotroph E. coli JM101ΔtrpF strain in M9 minimal media was measured and compared with the CFU grown without this functional selective pressure in LB medium. The libraries were previously fused to the CAT gene as an in vivo folding reporter. Folded sequence space was estimated according to the capacity of the variants to grow in the presence of chloramphenicol. Thus, the folding competence was calculated as the ratio of the number of CFU grown in LB amp plus chloramphenicol media (under folding selective pressure) to the number of CFU grown in LB amp media (without folding selective pressure).

Sequence analysis of the functional sequence space

The sequences found in the variability positions from the variants growing under selective pressure for PRA isomerase activity (M9 minimal media) were compared with the sequences from the variants growing without this selective pressure (LB ampicillin). Sequences found in both conditions are illustrated in Tables S7, S8, S9, and S10. The amino acids observed in these positions were converted to frequencies and the difference between these indicates the discrepancy of occurrence for each residue between these two conditions (Figure 3). The average and standard deviation of the frequencies were used to determine, with 95% confidence, the negative or positive selection (red colored histograms in Figure 3) for specific amino acids in the corresponding variability positions.

Biochemical and biophysical characterization of one variant

Expression and purification of variant Beta_1 was performed using the expression vector pET-28b (Stratagene), as previously described for wild-type TrpF [23]. Michaelis-Menten enzyme kinetics of PRA isomerase activity was determined using the protocol previously reported [25]. The kinetic parameters were obtained by fitting initial rates to the Michaelis-Menten model using nonlinear fit analysis with the public available program MicroCal Origin 5.0. The kinetic data shown in Table 3 represent the average of at least three independent experiments using freshly purified enzyme. The CD measurements were carried out using a J-715 spectropolarimeter (Jasco) equipped with a Peltier temperature control supplied by Jasco. The far-UV CD spectra were collected from 190 to 260 nm at 25°C in a 0.1-cm path length cell. Proteins were measured at a concentration of 0.3 mg/ml in 10 mM potassium phosphate buffer at pH 7.6, 1 mM EDTA and 1 mM beta-mercaptoethanol. Eight replicate spectra were collected from each sample to improve the signal-to-noise ratio. The thermal denaturation process was analyzed by measuring the change in ellipticity at 220 nm as a function of temperature, which was increased at a rate of 0.3°C min⁻¹. The thermal denaturation curves were normalized assuming a linear temperature dependence of the baselines for native and denatured states. The apparent thermal melting temperature (Tm, app) was determined by identifying a midpoint temperature between the native form (linear interpolation of the native region) and the denatured form (the lowest point or linear interpolation of the unfolded region) on the thermal unfolding curves. The intermolecular associations were analyzed by size exclusion chromatography in an AKTA FPLC with a superose HR12 column (GE Healthcare). Purified protein in an initial volume of 0.15 ml was eluted at a flow rate of 0.4 ml min⁻¹ on a Superdex 200 column (GE Healthcare) that was previously equilibrated in 50 mM HEPES buffer (pH 7.6) and 100 mM NaCl.

Supporting Information

Figure S1 Design of β-strand and α-helix substitutions in a model (βαβ8) barrel enzyme.
(DOCX)

Figure S2 Theoretical, folded and functional sequence spaces for libraries secondary-structure substitutions.
(DOCX)

Figure S3 Sequence analysis found at variable positions in natural TrpF and MetR enzymes.
(DOCX)

Table S1 Number of protein coding genes from the human genome predicted with a (βαβ8) barrel domain.
(XLSX)

Table S2 Full-length protein coding genes of (βαβ8) barrel domains and number of experimentally confirmed splice variants for each gene.
(XLSX)

Table S3 Experimentally reported splice variants of the 70 full length proteins.
(XLSX)

Table S4 Mapping of splicing variants into the structures of the full-length (βαβ8) barrel proteins.
(XLSX)

Table S5 Sequence and structural details of the spliced variants with predicted structural changes affecting the (βαβ8) barrel domain.
(DOCX)

Table S6 Superfamily assignment for the splice variants in which the structure of the (βαβ8) barrel is affected.
(DOCX)
Table S7 Sequences found without selective pressure for β-strand library.

Table S8 Sequences found under selective pressure for β-strand library.

Table S9 Sequences found without selective pressure for α-helix library.

Table S10 Sequences found under selective pressure for α-helix library.

References

1. Bogard AD, Deem MW (1999) A hierarchical approach to protein molecular evolution. Proc Natl Acad Sci U S A 96: 2391–2395.

2. Bharat TA, Eisenbeis S, Zeki K, Hocker B (2008) A beta alpha-barrel built by the combination of fragments from different folds. Proc Natl Acad Sci U S A 105: 9942–9947.

3. Chen L, Towar-Corona JM, Urrutia AO (2012) Alternative splicing: a potential source of functional innovation in the eukaryotic genome. Int J Evol Biol 2012: 596274.

4. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40: 1413–1415.

5. Groso AR, Gomes AQ, Barbosa-Morais NL, Caldeira S, Thorne NP, et al. (2008) Tissue-specific splicing factor gene expression signatures. Nucleic Acids Res 36: 4823–4832.

6. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, et al. (2007) Discovery of tissue-specific exons using comprehensive exon microarrays. Genome Biol 8: R54.

7. Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30: 3574–3586.

8. Yeo G, Holste D, Krizman G, Burge CB (2004) Variation in alternative splicing across human genomes. Genome Biol 5: R74.

9. Steefeld J, Rueng MA (2003) Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci 30: 513–521.

10. Biczke F, Csaba G, Zimmer R (2008) Alternative splicing and protein structure evolution. Nucleic Acids Res 36: 5537–5548.

11. Wang Y, Berlow RB, Loria JP (2009) Role of loop-loop interactions in stabilizing an α-subunit of trimeric phosphoribosylanthranilate isomerase. Protein Struct Function 68: 1030–1035.

12. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321: 956–960.

13. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456: 470–476.

14. Gilbert W (1987) The exon theory of genes. Cold Spring Harb Symp Quant Biol 52: 901–905.

15. Liu M, Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of comprehensive exon microarrays. Nature 436: 470–476.

16. Gilbert W, de Souza SJ, Long M (1997) Origin of genes. Proc Natl Acad Sci U S A 94: 7689–7703.

17. Wu Z, Hui S, Zhang X, Weiblen K, Petersen J (2008) Protein structure: protein structure prediction through an ensemble of multiple methods. Bioinformatics 24: 1022–1028.

18. Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30: 3574–3586.

19. Yeo G, Holste D, Krizman G, Burge CB (2004) Variation in alternative splicing across human genomes. Genome Biol 5: R74.

20. Clark TA, Schweitzer AC, Chen TX, Staples MK, Lu G, et al. (2007) Discovery of tissue-specific exons using comprehensive exon microarrays. Genome Biol 8: R54.

21. Richter M, Bosnali M, Carstensen L, Seitz T, Durchschlag H, et al. (2010) Structural genomics analysis of a Thousand genomes. Nucleic Acids Res 40: D71–75.

22. Setiyaputra S, Mackay JP, Patrick WM (2011) The structure of a truncated phosphoribosylanthranilate isomerase suggests a unified model for evolution of the (beta/alpha)8-barrel fold. J Mol Biol 416: 145–157.

23. Ochoa-Leyva A, Soberon X, Sanchez F, Arguello M, Montero-Moran G, et al. (2009) Protein design through systematic catalytic loop exchange in the (beta/alpha)8 fold. J Mol Biol 387: 949–964.

24. Afriat-Jurnou L, Jackson CJ, Tawfik DS (2012) Reconstructing a Missing Link in the Evolution of a Recently Diverged Phosphotriesterase by Active-Site Loop Remodeling. Biochemistry 51: 6047–6055.

25. Zitewitz JA, Gualdetti PJ, Perkins IA, Wasta SA, Matthews CR (1999) Identifying the structural boundaries of independent folding domains in the alpha subunit of tryphophan synthase, a beta/alpha barrel protein. Protein Sci 8: 1280–1299.

26. Ochoa-Leyva A, Saa B, Weihrauch D, Nagase T, Enyedi A, et al. (2009) Alpha subunit of trimeric phosphoribosylanthranilate isomerase: a beta/alpha barrel protein. J Biol Chem 284: 3751–3757.

27. Afriat-Jurnou L, Jackson CJ, Tawfik DS (2012) Reconstructing a Missing Link in the Evolution of a Recently Diverged Phosphotriesterase by Active-Site Loop Remodeling. Biochemistry 51: 6047–6055.

28. Asmolov A, Grishin NV, Vakser IA, Wolf I, Poch O, et al. (2006) Comparative protein structure and function prediction server. Nucleic Acids Res 34: D251–D256.

29. Xue F, Guo Y, Seo H, Jiang X, Bai L, et al. (2008) Analysis of alternative splicing in the human transcriptome. Nucleic Acids Res 36: 1772–1781.

30. Enright AJ, Ouzounis CA (1998) Protein function prediction through comparative genome analysis. In: Methods in Enzymology. Academic Press: New York.

31. Consortium TU (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71–75.

32. Shukla A, GuptaSarma P (2004) Folding of beta/alpha-unit scrambled forms of S. cerevisiae triosephosphate isomerase: evidence for autonomy of substructure formation and plasticity of hydrophobic and hydrogen bonding interactions in core of beta/alpha@-barrel. Proteins 55: 548–557.

33. Silverman JA, Harbury PB (2002) The equilibrium unfolding pathway of a (beta/alpha)8 barrel protein: generation of variability by recombination. FEBS Lett 560: 167–172.

34. Soberon X, Fuentes-Gallego P, Saab-Rincon G (2004) In vivo fragment complementation of a (beta/alpha)8 barrel protein: generation of variability by recombination. FEBS Lett 560: 167–172.

35. Patrick WM, Blackburn JM (2005) In vitro selection and characterization of a stable subdomain of phosphoribosylanthranilate isomerase. FEBS J 272: 3604–3607.

36. Eder J, Kirschner K (1992) Stable substructures of eightfold beta alpha-barrel proteins: fragment complementation of phosphoribosylanthranilate isomerase. Biochemistry 31: 3617–3623.

37. Zitewitz JA, Matthewes CR (1999) Molecular dissection of the folding mechanism of the alpha subunit of tryptophan synthase: an amino-terminal autonomous folding unit controls several rate-limiting steps in the folding of a single domain protein. Biochemistry 38: 10205–10214.

38. Toumazou J, Bergfors T, Terri T, Knowles JK, Jones TA (1999) Three-dimensional structure of cellulohydrolase II from Trichoderma reesi. Science 249: 380–386.

39. Eadk JC, Ozturk D, Wedler TB, Grubmeyer C, Sacchettini JC (1997) A new function for a common fold: the crystal structure of quinolinic acid phosphoribosyltransferase. Structure 5: 47–58.

40. Moore SA, James MN, O’Kane DJ, Lee J (1995) Crystal structure of a flavoprotein related to the subunits of bacterial luciferase. Embryo 12: 1767–1774.

41. Wang Y, Berlow RB, Loria JP (2009) Role of loop-loop interactions in coordinating motions and enzymatic function in triosephosphate isomerase. Biochemistry 48: 4548–4556.

42. Urfer R, Kirschner K (1992) The importance of surface loops for stabilizing an eightfold beta alpha barrel protein. Protein Sci 1: 31–45.

43. Lokeshwar VB, Schroeder GL, Carey RL, Solorway MS, Iida N (2002) Regulation of hyaluronidase activity by alternative mRNA splicing. J Biol Chem 277: 33654–33663.

Acknowledgments

The authors thank Filiberto Sánchez, Karina Verdél-Aranda, and Francisco Barona-Gómez for technical insights on the in vivo and in vitro PRA isomerase activity assays.

Author Contributions

Conceived and designed the experiments: AOL XS. Performed the experiments: AOL GMM GSR. Analyzed the data: AOL GMM GSR. Contributed reagents/materials/analysis tools: LGB. Wrote the paper: AOL XS.
44. Pitkanen LT, Heiskala M, Andersson LC (2001) Expression of a novel human ornithine decarboxylase-like protein in the central nervous system and testes. Biochem Biophys Res Commun 287: 1051–1057.

45. Privitera S, Procya CA, Callahan JW, Hinek A (1998) The 67-kDa enzymatically inactive alternatively spliced variant of beta-galactosidase is identical to the elastin/laminin-binding protein. J Biol Chem 273: 6319–6326.

46. Saito A, Ozaki K, Fujiwara T, Nakamura Y, Tanigami A (1999) Isolation and mapping of a human lung-specific gene, TSA1902, encoding a novel chitinase family member. Gene 239: 325–331.

47. Puisac B, Ramos M, Arnedo M, Menao S, Gil-Rodriguez MC, et al. (2012) Characterization of splice variants of the genes encoding human mitochondrial HMG-CoA lyase and HMG-CoA synthase, the main enzymes of the ketogenesis pathway. Mol Biol Rep 39: 4777–4785.

48. Böttge RG, Rentkema GH, Verhoek M, Strijland A, Blick J, et al. (1998) The human chitotriosidase gene. Nature of inherited enzyme deficiency. J Biol Chem 273: 25680–25685.

49. Pucci L, Perozzi S, Cinamadore F, Orsomando G, Raffaelli N (2007) Tissue expression and biochemical characterization of human 2-amino 3-carboxymuconate 6-semialdehyde decarboxylase, a key enzyme in tryptophan catabolism. FEBS J 274: 827–840.

50. Jezyk MR, Snyder JT, Gershberg S, Worthylake DK, Harden TK, et al. (2006) Crystal structure of Rac1 bound to its effector phospholipase C-beta2. Nat Struct Mol Biol 13: 1135–1140.

51. Wu M, Gu S, Xu J, Zou X, Zheng H, et al. (2005) A novel splice variant of human gene NPL, mainly expressed in human liver, kidney and peripheral blood leukocyte. DNA Seq 16: 137–142.

52. Salaberi JK, Li XP, Petarash JM, Bhatnagar A, Barski OA (2011) Functional expression of novel human and marine AKR1B genes. Chem Biol Interact 191: 184–188.

53. Lopez AJ (1998) Alternative splicing of pre-mRNA: developmental consequences and mechanisms of regulation. Annu Rev Genet 32: 279–305.

54. Smith CW, Valcarcel J (2000) Alternative pre-mRNA splicing: the logic of combinatorial control. Trends Biochem Sci 25: 301–308.

55. Blencowe BJ (2006) Alternative splicing: new insights from global analyses. Cell 126: 37–47.

56. Toth-Petroczy A, Tavšík DS (2013) Protein insertions and deletions enabled by neutral roaming in sequence space. Mol Biol Evol 30: 761–771.

57. Bertolact BL, Knowles JR (1995) Complementation of fragments of triosephosphate isomerase defined by exon boundaries. Biochemistry 34: 5736–5743.

58. Hocker B, Claren J, Sterner R (2004) Mimicking enzyme evolution by generating new (betaalpha)4-half-barrels from (betaalpha)4-half-barrels. Proc Natl Acad Sci U S A 101: 16448–16453.

59. Hatakeyama K, Ohshima K, Fukuda Y, Oguma S, Terashima M, et al. (2011) Identification of a novel protein isoform derived from cancer-related splicing variants using combined analysis of transcriptome and proteome. Proteomics 11: 2275–2282.

60. Ge K, DuHadaway J, Du W, Herlyn M, Rodeck U, et al. (1999) Mechanism for elimination of a tumor suppressor: aberrant splicing of a brain-specific exon causes loss of function of Bin1 in melanoma. Proc Natl Acad Sci U S A 96: 9609–9614.

61. Barbour AP, Reeder JA, Walsh MD, Fawcett J, Antalis TM, et al. (2003) Expression of the CD44v2-10 isoform confers a metastatic phenotype: importance of the heparan sulfate attachment site CD44v5. Cancer Res 63: 897–892.

62. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, et al. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112: 535–542.

63. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, et al. (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 39: D465–474.

64. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) UCSF Chimera – a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.

65. Van Durme J, Delgado J, Stricher F, Serrano L, Schymkowitz J, et al. (2011) A graphical interface for the FoldX forcefield. Bioinformatics 27: 1711–1712.

66. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. (2005) The FoldX web server: an online force field. Nucleic Acids Res 33: W382–388.

67. Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18: 2714–2723.