A Profile-Based Method for Measuring the Impact of Genetic Variation

Short title:

Measuring the Impact of Genetic Variation

Nicole E. Wheeler[1¶*], Lars Barquist[2¶], Fatemeh Ashari Ghomi[1], Robert A. Kingsley[3,4], Paul P. Gardner[1,5]

Author affiliations:

[1]School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.

[2]Institute for Molecular Infection Biology, University of Wuerzburg, Wuerzburg, Germany.

[3]Institute of Food Research, Norwich Research Park, Norwich, UK.

[4]Wellcome Trust Sanger Institute, Hinxton, UK.

[5]Biomolecular Interaction Centre, University of Canterbury, Christchurch, New Zealand.

* Corresponding author

Email: nicole.wheeler@pg.canterbury.ac.nz

[¶] These authors contributed equally to this work

Abstract

Advances in our ability to generate genome sequence data have increased the need for fast, effective approaches to assessing the functional significance of genetic variation. Traditionally, this has been done by identifying single nucleotide polymorphisms within populations, and calculating derived statistics to prioritize candidates, such as dN/dS. However, these methods commonly ignore the differential selective pressure acting at different positions within a given protein sequence and the effect of insertions and deletions (indels). We present a profile-based method for predicting whether a protein sequence variant is likely to have functionally diverged from close relatives, which takes into account differences in residue conservation and indel rates within a sequence. We assess the performance of the method, and apply it to the identification of functionally significant genetic variation between bacterial genomes. We demonstrate that this method is a highly sensitive measure of functional potential, which can improve our understanding of the evolution of proteins and organisms. An implementation can be found at https://github.com/UCanCompBio/deltaBS.

Author Summary

Next generation sequencing projects are producing vast amounts of sequence data, however our ability to produce this data is outpacing our ability to derive valuable information from it. We present a sequence analysis method for predicting whether genome variation is likely to result in phenotypic differences, and an application of the method across a variety of analysis scales.

Introduction

Genome sequencing technologies allow us to explore the wealth of genetic variation between and within species, and as these technologies advance this data is becoming progressively cheaper and faster to produce [1–3]. As a result, comparative sequence analyses have become a popular approach for exploring biological concepts. A typical investigation involves determining the phylogenetic relationship between organisms of interest and identifying key genetic differences, in the form of rearrangements, gene gains and losses, and variation within orthologous genes. Exploring genetic variation between closely-related organisms has provided key insights into the evolution of many species [4,5], but how can the significance of this genetic variation be quantified and prioritized for investigation? We know that variation in the genome can cause changes in phenotype, however predicting the nature and degree of these changes is challenging. In some instances the functional impact of a variation in sequence can be negligible, while in other cases a single nucleotide change can have dramatic fitness consequences. The development of fast and accurate ways of assessing the functional significance of sequence variation is an important step in extracting meaning from comparative analyses.

A number of methods have attempted to predict the impact of sequence variation on protein function. One such example is PROVEAN, which uses a BLAST-based approach to score sequence variants against closely related sequences [6]. Another is PolyPhen-2, which uses a combination of 11 predictive measures to calculate the likelihood that a change is deleterious [7]. These predictive measures include sequence conservation, substitution probabilities, CpG context, structural features and the domain architecture of the protein. The SIFT algorithm uses position-specific scoring matrices based on sequence homology and known patterns of common amino acid substitutions to predict the functional consequences of non-synonymous single nucleotide polymorphisms (nsSNPs) [8]. As a final example, MutationAssessor computes both the conservation and specificity of residues within protein subfamilies to assess the impact of a mutation [9].

An assumption underlying these methods is that functionally important residues in a protein are conserved during evolution, while those that are less important to function are able to tolerate more mutations. Our method relies on this same assumption, but applies more sophisticated scoring methods to the task of assessing the degree to which a mutation deviates from evolutionary patterns that the protein has shown in the past.

Unlike BLAST-based methods, which use constant penalties for substitutions and indels, our approach uses profile hidden Markov models (HMMs) to score sequence variation. A profile hidden Markov model (HMM) is a statistical representation of an alignment of protein sequences that incorporates the position-specific probabilities of substitutions, insertions and deletions for each column of the alignment. In addition to the information provided by the alignment, the models incorporate sequence weighting schemes to ensure an even representation of different protein lineages, and Dirichlet mixture priors [10] to factor in prior knowledge of the influence of biochemical and structural constraints placed on amino acid substitution patterns. These robust models of protein families can predict the sequence space that could potentially be explored by functional variants of a given protein family [11-13]. Due to this more sophisticated scoring approach, HMM-based methods should, in principle, perform better than both BLAST-based methods and position-specific approaches that ignore indels.

Our approach, first introduced in [14] takes the difference in bitscore for profile HMM matches between variant sequences and uses the magnitude of this difference as an estimate of the functional impact of the mutation. The HMMs can be downloaded from databases such as Pfam or Treefam [15,16] or can be built for specific target proteins [17,18]. The approach performs better than the more complex PROVEAN and PolyPhen methods, and scales easily to screen entire genomes for variation of note, while this option is not yet offered by other methods. We have named our scoring metric “delta-bitscore” (DBS), as it is the difference in bitscores between the two comparator proteins (e.g. wild-type vs variant, or species A vs species B).

In the following pages we present the results of testing our method across multiple analysis scales. Firstly, we consider the effects of systematic mutational analyses on several well characterised proteins. Secondly, we consider population-level variation and the ability of DBS to discriminate between disease associated and polymorphic human nsSNPs. Finally, we use the method to identify functionally significant sequence variation between bacterial genomes. We demonstrate that DBS is a sensitive measure of the functional impacts of protein sequence variation, and is applicable to a broad range of biological questions.

Results

Comparison of a given protein sequence to a profile HMM produces a bitscore value, which gives an indication of the likelihood that a given protein sequence belongs to the protein family the profile HMM has been built for. In subtracting the bitscore of one protein or domain from that of another, we produce a measure of the relative quality of the match between each of the two proteins and the model of the protein family they evolved from.

If sequence 1 (S1) and sequence 2 (S2) are orthologous, then the difference in bitscores gives an indication of how divergent S1 and S2 are that we call delta-bitscore or DBS. High positive values indicate that S1 is a strong match to the profile-model and that S2 has diverged, high negative values imply the reverse, and values near zero generally imply that that there is little difference between the sequences, but could also indicate that neither sequence is a strong match to the profile-model.

Highly conserved positions in a model alignment will receive higher scores than poorly conserved positions. Functionally neutral variation is likely to result in individual bitscore differences that are small in magnitude and cancel out over the length of the protein, while functionally significant change in one protein will likely produce one or more DBS values of high magnitude that have a greater impact on overall DBS (see S1 Fig.). For a more detailed discussion of the method, see Additional Note.

Identifying mutations that impair the functioning of a protein

In order to test the ability of profile HMMs to identify loss-of-function mutations we used three independent datasets from protein mutagenesis studies on phage lysozyme, *E. coli* LacI and HIV protease (see Materials and Methods). These experiments systematically mutated residues in the protein and measured the impact of these mutations on protein function. We tested two different reference HMM sets: curated HMMs from the Pfam database, and automatically constructed HMMs built from sequences with a range of residue identities. We also compared these results to predictions made by the PROVEAN and PolyPhen-2 methods. In order to evaluate the accuracy of each method we computed the ‘Area Under the Curve’ (AUC) for a series of Receiver Operating Characteristic (ROC) plots. These values range between one and zero: an AUC equal to one indicates a perfect prediction tool, while an AUC of 0.5 indicates the method performs no better than chance.

The performance of PROVEAN and Polyphen-2 is practically indistinguishable in Fig. 1A and B. The performance of the Pfam HMMs was respectable, however in 2/3 cases they were outperformed by PROVEAN and PolyPhen-2. As indicated in Fig. 1B and S2 Fig., at conservative scoring thresholds Pfam HMMs are able to make highly confident predictions, but these decline in accuracy as scoring thresholds become more permissive. In contrast, the custom HMMs built using filtering for 40%

sequence identity between the query sequence and potential homologs outperformed PROVEAN and PolyPhen on all three of the protein mutagenesis data sets.

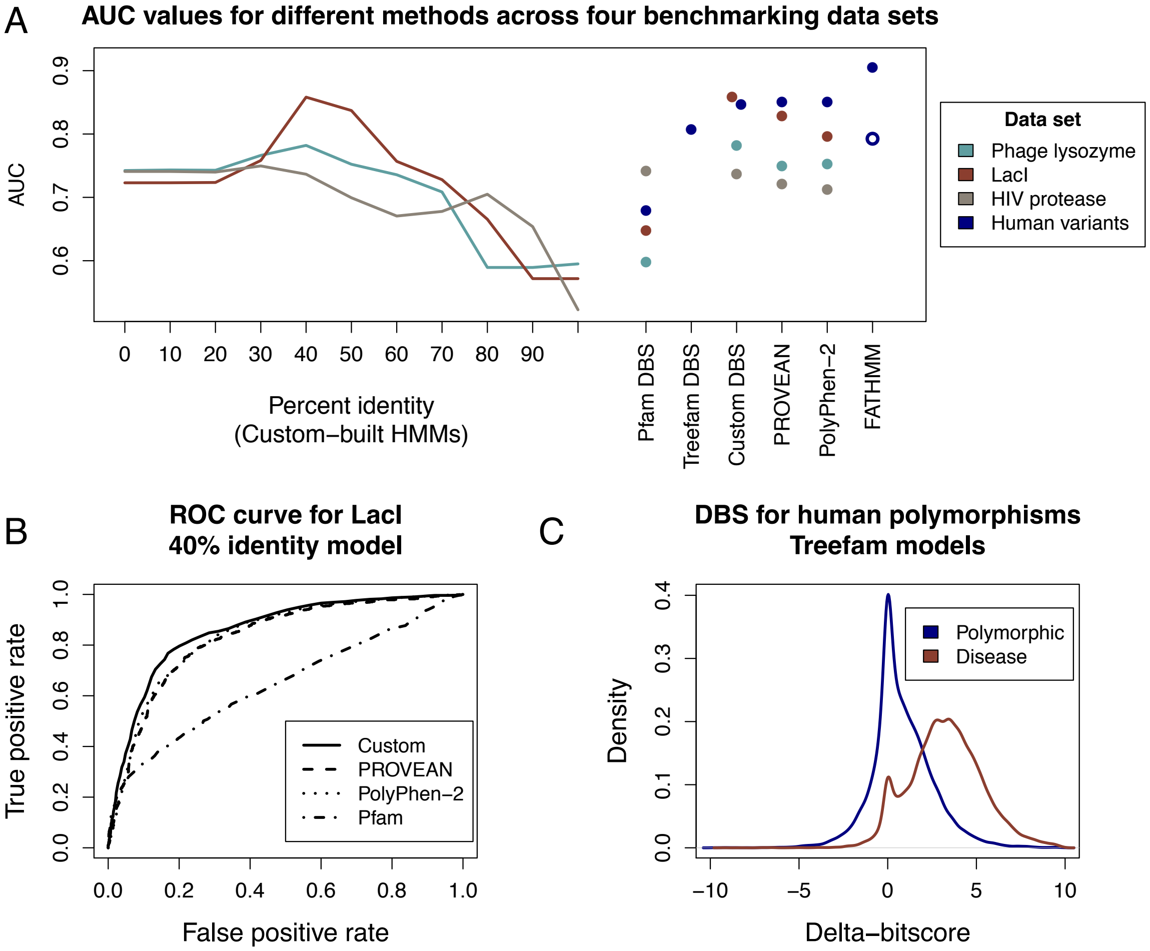

Fig. 1: Comparison of the performance of DBS and similar measures.

A: AUC plot for performance of custom HMMs built from a range of percent sequence identity cutoffs, Pfam, Treefam, PROVEAN, PolyPhen-2 and FATHMM. Filled point for FATHMM is the weighted method, while the empty point is the unweighted method.

B: ROC curve for the custom LacI model (40% sequence identity), PROVEAN and PolyPhen-2 methods and Pfam models.

C: Distribution of DBS scores for disease and polymorphic SNPs in human genes.

DBS as a measure of clinically significant genetic variation

If DBS is an accurate predictor of functionally significant sequence variation, one would expect it to distinguish between human polymorphisms and disease-causing SNPs. This is both a useful benchmark for methods aiming to predict functionally significant changes, and an indication of the value these methods have in a clinical context in identifying mutations that cause congenital disorders (e.g. [19]). In order to investigate this assumption, we used the humsavar database (http://www.uniprot.org/docs/humsavar), which catalogs human polymorphisms and disease mutations as a test dataset.

For this analysis, the Treefam database was used as an additional source of profile HMMs. This database collects profile HMMs of vertebrate genes, resulting in models that span entire genes rather than only functional domains. We also tested a similar HMM-based predictive method, FATHMM [20]. FATHMM has two relevant modes of use - a naïve, unweighted method and a weighted method trained to discriminate between human polymorphisms and disease variants.

As shown in Fig. 1A, Treefam-based DBS scoring outperforms Pfam-based scoring, but does not outperform PROVEAN, PolyPhen-2 or weighted FATHMM. There appears to be a clear shift in the distribution of DBS values for the disease variants compared with the polymorphisms when variants are scored using the Treefam profile HMMs (Fig. 1C). There also appears to be a positive skew in the polymorphism data (P < 0.001, exact binomial test), which could suggest variants that impact phenotypic traits not associated with disease, or that may contribute to disease but have not been implicated due to a number of possible reasons (low penetrance, low frequency in the population, high homozygote lethality [21]). We investigated enrichment of GO terms in the high-scoring polymorphisms in the DBS = 5 to DBS = 10 range. The molecular functions that were enriched in these genes are shown in S1 Table. The highest ranked GO terms relate to molecular binding, catalysis and signal transduction. There was also a long tail of high DBS values for typically problematic sequences for profile hidden Markov model-based analyses - these are proteins that show a bias in sequence composition, such as repetitive sequences (collagen, hornerin, zinc finger proteins), membrane proteins, or charged proteins such as myosin [22].

Building custom models for each of the human proteins using sequences from the Treefam database improved performance. In spite of using a slightly more permissive sequence identity cutoff of 35%, a number of proteins had fewer than 10 hits to the sequence database, so would have yielded models with poor predictive ability (AUC value when these were included was 0.81). Increasing the sequence search space to a more diverse database may further improve results. Models built from fewer than 10 sequences were excluded from the analysis, improving performance to an AUC of 0.84, but lowers the number of variants that can be analysed. See S1 Table for counts of the variants that could be analyzed for all methods tested.

Overall, the HMM-based approach FATHMM performed the best in discriminating human disease variants, with its built-in weighting scheme offering a significant boost in performance over our naïve methods. That being said, our naïve custom models performed better than the un-weighted FATHMM approach, suggesting that even greater performance could be achieved through tuning custom model construction in combination with a purpose-trained weighting scheme. While the development of our method is focussed on nonhuman genome comparisons, there is certainly scope to improve HMM-based detection of clinically significant genetic variation.

Identification of functionally significant variation in pathogen evolution

We have established that Pfam models can detect deleterious mutations with great specificity when conservative thresholds are used (S1 Fig.). Because of the general nature of the models we postulate that they can be used for detecting functionally significant genetic variation between organisms of any species. To explore this, we developed a tool that takes whole proteome files as input (FASTA, GenBank or EMBL format) and identifies functionally significant genetic variation between the two proteomes. Using this tool, we compared the proteomes of two closely related *Salmonella enterica* serovars with markedly different lifestyles.

Host-restriction is a common phenomenon in highly adapted invasive pathogens, often resulting in characteristic genomic features such as the proliferation of transposable elements and the degradation of substantial fractions of coding sequences [23,24]. Within the salmonellae such restriction events have occurred independently multiple times, in various hosts. *Salmonella enterica* serovar Enteritidis is a broad host range pathogen, capable of infecting humans, cattle, rodents and a variety of birds, while serovar Gallinarum is restricted to infecting galliform birds [25]. *S.* Gallinarum and *S.* Enteritidis have recently evolved from a common ancestor, however the *S.* Gallinarum genome has undergone extensive degradation since divergence [26]. In addition to being restricted to a narrow host range, *S.* Gallinarum has lost motility and causes a systemic, typhoid-like infection in birds, unlike Enteritidis which usually causes gastroenteric infections [27]. A recent analysis of pseudogenes within this lineage identified signatures of host-restriction in Gallinarum isolates, characterized by loss of metabolic genes required for survival in the intestine [26]. We expect DBS to add an additional layer of information to such an analysis, identifying genes which have shifted in or lost function due to non-synonymous mutations or small indels occurring since the restriction event, but have not yet succumbed to obvious disruption events such as large truncations, frameshifts, or complete deletions.

Identifying functionally significant variation in orthologous genes

We identified orthologous proteins between the two serovars using a reciprocal HMMER3 search, incorporating a screen for identical Pfam domain architectures (see Materials & Methods). We compared our ortholog list to a manually curated ortholog list [28]. We computed DBS for this comparison, by subtracting the summed domain bitscores of orthologous proteins, using S. Enteritidis as a reference. Consequently, large positive DBS values indicate a potential loss of function (LOF) in *S.* Gallinarum, and large negative values indicate a potential LOF in *S.* Enteritidis. We expect that variation in DBS will depend on evolutionary distance between organisms, and that proteins under negative selection will have DBS values that show a normal distribution centered around zero. Based on this assumption we calculate a Z-score from the DBS, and count proteins with a Benjamini-Hochberg corrected p-value of < 0.05 as candidates for LOF (Materials & Methods).

As expected, the distribution of DBS values centers around zero (Fig. 2A), indicating predominantly neutral sequence variation. The distribution of DBS values shows an enrichment for positive DBS (exact binomial test, $P = 1.553 \times 10^{-13}$), indicating greater divergence from the domain models in protein coding genes in *S.* Gallinarum when compared to *S.* Enteritidis. As shown in Fig. 2B, our loss-of-function predictions included many genes not identified as hypothetically disrupted coding sequences (HDCs) by manual inspection in Nuccio and Bäumler (2014) [28]. This indicates that DBS is able to rapidly identify loss-of-function mutations that have been missed by more time-intensive searches. To look at loss-of-function mutations in a functional context, we grouped genes into functional categories based on their annotation in the KEGG database. Not only does *S.* Gallinarum have fewer genes than *S.* Enteritidis for most of the functional categories we considered, but it also has a greater number of functional losses across these groupings (Fig. 2B). Previous work that found the presence of non-ancestral pseudogenes in *S.* Enteritidis was limited while *S.* Gallinarum had accumulated a large number of pseudogenes since divergence is consistent with our results [26]. Using DBS, 140 genes passed the threshold for potential LOF mutations in *S.* Gallinarum, corresponding to 53 functional roles in the metabolic pathway database generated by Langridge and colleagues [26]. Only 16 reached the cutoff for LOF in *S.* Enteritidis. Of these, 3 could be assigned to functional roles in metabolic pathways.

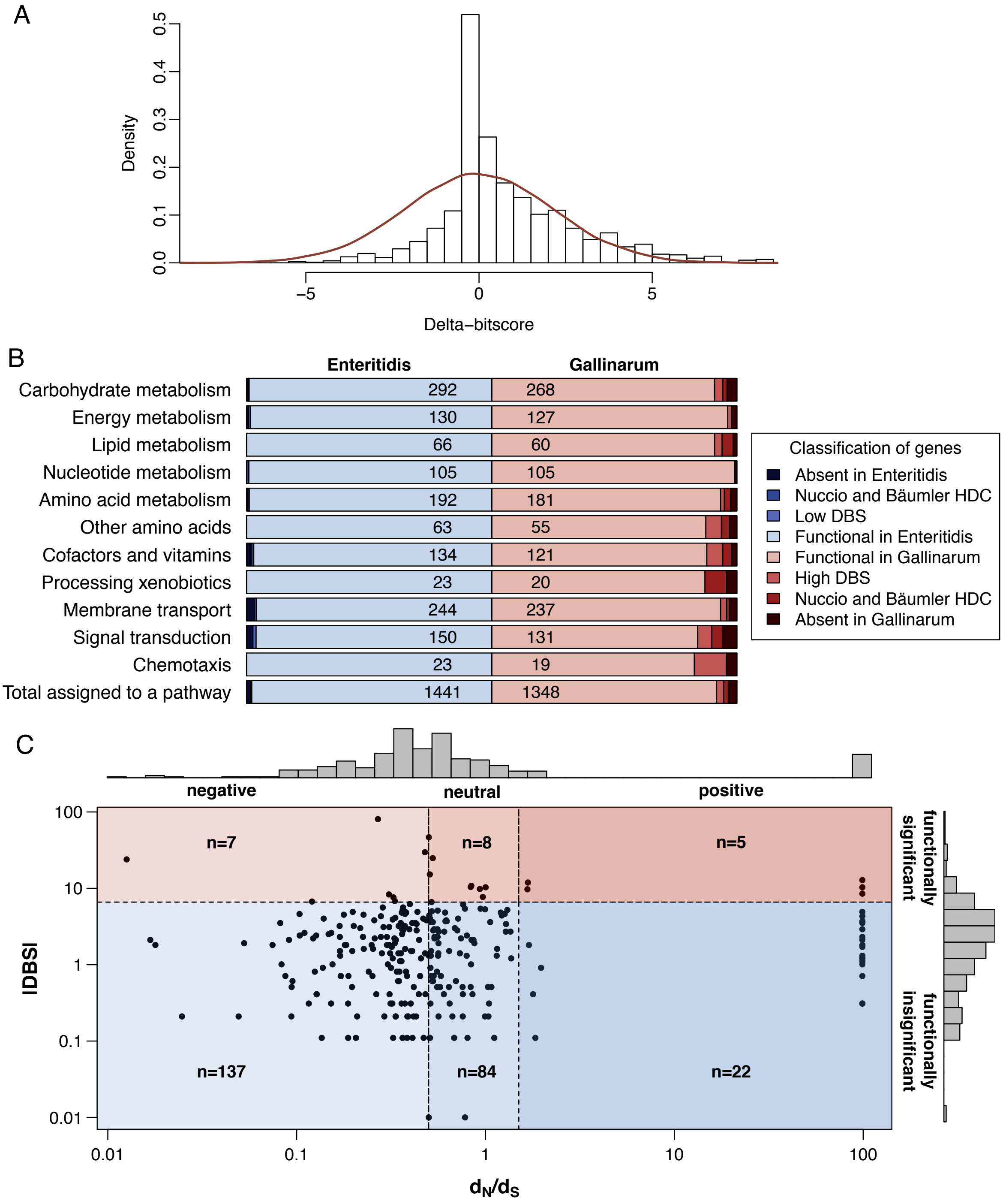

Fig. 2: Results of DBS comparison of the proteomes of *S.* Enteritidis and *S.* Gallinarum.

A: Distribution of delta-bitscores for non-synonymous mutations between *S.* Enteritidis and *S.* Gallinarum. Normal distribution fitted to the data is shown in red.

B: Functional changes in orthologous protein-coding genes of *S.* Enteritidis and *S.* Gallinarum grouped into functional categories according to the KEGG pathways database. Genes included in the pathway that have no ortholog in the other serovar are indicated in the darkest colour, followed by genes previously identified as HDCs, then genes with significant DBS values that had not already been identified as HDCs. Genes with orthologs in both species that have not been identified as non-functional by either manual inspection or DBS are indicated in the lightest colour, with a count shown on the bar.

C: Distribution of |DBS| vs dN/dS for orthologous genes in *S.* Enteritidis and *S.* Gallinarum, filtered for DBS>0, dN>0 and dS>0. Cutoffs set at 0.5 and 1.5 for dN/dS, and Benjamini-Hochberg adjusted p-value of 0.05 for DBS.

To test whether this skewed distribution was a common feature of host adaptation in *Salmonella*, we performed DBS analysis on three broad host range and three host-adapted *Salmonella* serovars. We found that comparisons of a generalist to a host-restricted serovar generally showed a significant positive skew compared to broad-host range serovars (S3 Table). We previously observed a similar, albeit less extreme, phenomenon in host-restricted strains of *S*. Typhimurium [14], suggesting that the analysis of obvious pseudogenes alone may underestimate the degree of functional decay and change during host-adaptation.

A full discussion of the functionally significant differences between the two serovars is outside the scope of this investigation, however we did see some exciting trends in our data. Genes relating to chemotaxis showed the greatest proportion of losses of all KEGG functional categories we investigated. This finding is consistent with our previous finding that restriction of host range in a pathovar of S. Typhimurium was accompanied by a loss-of-function mutation in the chemotaxis gene *tar* [14]. Genes involved in chemotaxis have been found to have also been found to be degraded in the host-restricted serovars Typhi and Paratyphi, and *tar* mutations in *S*. Typhimurium result in a hyper-invasive phenotype [29]. In addition, we found a number of LOF mutations in genes identified as being involved in the utilization of nutrients derived from the inflamed host gut environment colonized by gastrointestinal serovars of *Salmonella* [28] (S4 Table). This is consistent with the recent adaptation of *S.* Gallinarum from an ancestral gastrointestinal pathogen to an extra-intestinal environment.

These findings demonstrate that DBS is a more sensitive measure of loss of gene function than gene deletion when investigating serovars that have recently diverged. As time since divergence increases it is expected that non-functional genes will be deleted and the ratio of non-functional and deleted genes will change. We anticipate that our method will be most useful in comparisons of organisms which have recently diverged, as it offers the best opportunity to identify loss-of-function mutations that occur as an immediate response to a new environment, before deletion of entire genes occurs [30].

Comparing DBS and the commonly used measure of selection, dN/dS

In comparing genome sequences, we have focussed on identifying functionally significant variation, whereas a common approach is to classify genes from an evolutionary point of view, i.e. which genes are under negative selection, positive selection or are evolving neutrally. Our expectation is that DBS will not distinguish between genes that are under positive selection or are evolving neutrally - both classes will have high DBS values, while genes under negative selection will have low DBS values. A commonly used measure of selection is dN/dS (ω), which compares the rate of nonsynonymous changes in protein-coding sequences to the rate of synonymous changes to classify genes as either negative selection ($\omega < 1$), positive selection ($\omega > 1$) or neutrally evolving ($\omega \approx 1$) [31]. While this can be a useful approach across long time scales on shorter timescales it has been shown to be inaccurate. Studies have found that dN/dS only has effective discriminatory power across inter-species comparisons [32]. However, this method is commonly used to identify genes which show an enrichment of nonsynonymous mutations in closely-related isolates from the same species. Our method provides an alternative approach to identifying genes that have varied significantly, which is more appropriate for the study of closely related species/strains.

In comparing the results from dN/dS analysis of our two proteomes and DBS scoring, one of the most striking results is that there is no correlation between the two measures (R ~0.1, see Fig. 2C), and that the correlation between DBS and dN is also low (R ~0.14), suggesting a wide range of effect sizes for non-synonymous variation. A similar pattern can be seen in a comparison of human and chimpanzee genes, with a stronger correlation between |DBS| and dN, but still indicating a high degree of conflict between the two measures (see S3 Fig.). 5/27 genes reported to be under positive selection by dN/dS are predicted by DBS to show functionally insignificant variation, while 7/137 genes reported to be under negative selection by dN/dS show functionally significant variation according to DBS. Of the genes with high DBS and low dN/dS, most of these can be attributed to indels within protein domains. Genes with low DBS and high dN/dS tend to carry a small number of nonsynonymous changes, most of which are between chemically similar residues, and an even smaller number of synonymous mutations. There are also 92 genes classified by dN/dS as being under neutral selection,of which only 8 show functionally significant change according to DBS.

A limitation of dN/dS is that it ignores insertions and deletions, in spite of the fact that these changes may significantly impact the functioning of a protein. This is a major contributor to the discrepancies in classification between the two methods. As a result there is a spread of genes (N = 16) with dN/dS values equal to zero across a range of positive DBS values, due to indels incurring penalties according to DBS but being ignored by dN/dS (data not shown). This is the greatest limitation we see in dN/dS-based analysis, as it ignores a key form of genetic variation which is likely to have significant impacts on protein function.

Discussion

In our analysis, we first demonstrated the ability of DBS to detect deleterious amino acid substitutions in protein coding sequences, then presented the utility of this application in detecting clinically significant sequence variation. We next demonstrated the expansion of this approach to a tool that allows comparison of whole proteomes to identify sequence variation of functional importance.

DBS for detection of loss-of-function mutations

We have demonstrated that DBS performs comparably to or better than other methods designed to perform similar analyses, and that the Pfam-based profiles are well suited to analyses of entire genomes. The key benefits of this approach are the broad-scale applicability of the method to any species comparison and the ability of the method to score most types of protein sequence variation -

substitutions, insertions, deletions and large frameshifts and truncations. The use of domain annotations additionally provides insight into the type of function that will be lost as a result of the mutation.

While the speed and scale of analysis using Pfam models is admirable, we suspect the tuning of these models towards sensitivity (ability to detect novel instances of these domains) limits their ability to accurately predict protein sequence variations that are deleterious in a more specific context. The reason behind the significant difference in performance between the Pfam models and custom models is likely to be two-fold.

Firstly, the Pfam models capture sequence data from within domains, ignoring the spacer regions between domains. A number of publications have pointed to the functional importance of these spacer regions in the correct folding and interaction of individual domains within proteins [33][34]. The custom, full protein models capture additional sequence data from these spacer regions that may be essential to the functioning of the protein as a whole.

Secondly, the purpose of the Pfam HMM database is to detect novel occurrences of a domain across vast evolutionary distances. However, in assessing the functional potential of a domain, greater specificity is required to distinguish between a domain that is functional in the biological context in question, and one which is undergoing loss or adaptation of function.

Building specific protein models using a sequence identity cutoff of ~40% provides optimal performance of the models for the benchmarking data. This balances gathering enough information about amino acid substitution rates, and restricting the sources of this information to only those sequences that share the same function. This figure is in agreement with Tian and Skolnick's finding [35], that enzyme function is generally conserved up to 40% sequence identity, but tends to decrease past that point. There is potential to significantly improve the accuracy and sensitivity of genome-wide scans by building custom models for a database of query proteins. This would involve an additional jackhmmer stage for our delta-bitscore method, which performs a sequence search based on the query proteome against a sequence database. These sequences could then be filtered for sequence identity and used to build models. This approach is analogous to that taken by PROVEAN, except with the benefit of then having a profile-based comparison rather than pairwise sequence comparisons. FATHMM also builds custom models as part of its analysis, however the models are constructed without filtering for percent identity. A script for custom model building for genome-wide scans is likely to be the next stage of development of the software.

Overall, the method performs well as a naïve classifier of deleterious mutations in protein coding sequences and scales well to perform genome-wide comparisons.

A comparison of DBS and other methods

Overall, delta-bitscore has areas of strength and weakness when compared to similar methods. While the weighted, profile-based method FATHMM offers the best performance for human proteins, this approach is limited to only human nsSNPs. For a single protein of interest of non-human origin, building a custom profile HMM holds the most promise in terms of performance. With respect to proteome comparisons, alternative methods present some serious limitations, the greatest of all being limited options for scoring of indels. PROVEAN is able to score indels, but offers batch analysis for only human and mouse proteins.

Genome-wide analysis of *Salmonella* serovars

We set out to demonstrate the utility of our method in the comparative analysis of genome sequence data. We chose to compare *S*. Enteritidis and *S*. Gallinarum, as extensive manual inspection of ortholog calls and gene disruptions by Nuccio and Bäumler (2014) [28] allowed us to compare our high-throughput method to a more labour-intensive approach. In addition, literature suggests that loss-of-function has been a strong feature in the evolution of Gallinarum from its ancient ancestor.

Our ortholog calling method performed very well compared to Nuccio and Bäumler’s more stringent method which included manual inspection of alignments of orthologs [28]. The loss-of-function predictions generated in our analysis are largely in agreement with trends observed by other studies of the genomic signatures that accompany adaptation to a more invasive, host adapted lifestyle in *Salmonella* [14][26]. A major theme is that adaptation to a niche involves greater degradation of the genome of the niche-adapted isolate than that seen in closely related generalist isolates. This can be observed in the Enteritidis-Gallinarum comparison as a clear positive skew in the distribution of DBS values. *Salmonella* Gallinarum also showed a greater number of functional losses over most functional categories as defined by the KEGG pathway database.

Our method only identified 25% of the pseudogenes that were picked up in the re-annotation performed by Nuccio and Bäumler. Hypothetically disrupted coding sequences (HDCs) identified by Nuccio and Bäumler were identified on the basis of frameshifts and truncations. The discrepancy in LOF calls is mostly due to some truncations and frameshifts occurring near the end of the protein, outside of domains. We also identified a number of LOF mutations that were not identified in Nuccio and Bäumler's analysis. This result is not suprising, as our analysis allows for LOF mutations to have been caused by a small number of nonsynonymous changes, while their analysis relied on deletions and truncations for the identification of potentially disrupted genes. Our loss-of-function predictions added an additional layer of information on genes that are likely to have lost their function which can improve our understanding of which metabolic pathways are most susceptible to degradation during host adaptation. Our predictions indicate that degradation of genes involved in chemotaxis, signal transduction and the metabolism of non-proteinogenic amino acids, cofactors and vitamins was greater than previously predicted.

Our investigation compared the predictions made by our analysis with predictions of selection pressures on genes using dN/dS. We showed that there was little correlation between this measure of selection and our measure of the functional significance of nonsynonymous changes. dN/dS has been shown to be inappropriate for comparisons over short evolutionary timescales, such as those involving the evolution of strains within a species. Nozawa, Suzuki and Nei (2009) conducted a computer simulation that demonstrated poor predictive ability of dN/dS when the number of substitutions per gene is less than ~80 [36]. Their study showed that the method failed to pick up a number of experimentally determined functional changes, and predicted positive selection in sites where amino acid substitutions were unlikely to have a functional effect. In spite of this, dN/dS is still commonly used in comparisons of closely related bacteria [37-39].

For analyses such as this, where closely related strains are being compared, dN/dS values are considered to be unreliable due to the higher frequency of chance nonsynonymous mutations compared to chance synonymous mutations, and the lag in the removal of slightly deleterious mutations. This leads to high dN/dS ratios being commonplace in comparisons of closely related strains, suggesting positive or relaxed selection where there is none [40].

Due to the unreliability of dN/dS measures at short evolutionary timescales and the inability of dN/dS based methods to score indels, we propose that DBS is a much more suitable analysis tool for the study of recently diverged organisms.

Concluding statement

Delta-bitscore analysis of protein sequence variants performed with tuned, custom built models has been demonstrated to be the best naïve classifier of deleterious mutations tested in our study. Additionally, using delta-bitscore with Pfam models facilitates a fast, easy workflow for analysing proteome sequence variants, requiring only the provision of proteome files. The results are estimates of functional significance for each variant protein and a direct mapping to functional annotations from Pfam. In spite of this simplicity, the approach separates the wealth of genetic variation we have to investigate into that which is likely to have functional consequences and that which is not, making the task of pointing to genetic determinants of phenotypic differences simpler and more precise.

Materials and methods

Delta-bitscore

We define delta-bitscore using the following equation:

$$DBS = x_{var} - x_{can} \tag{1}$$

Where DBS is delta-bitscore and, x_{can} and x_{var} are bitscores for canonical and variant sequences derived from alignments to the same profile HMM. See the supplementary text for further discussion of this metric and comparisons with related previously published metrics [20,41,42].

Benchmarking

Data were obtained from three independent protein mutagenesis experiments, documenting functional consequences for mutations in HIV-1 protease (336 sequences) [43], *E. coli* Lac I (4041 sequences) [44] and phage lysozyme (2015 sequences) [45]. The data were downloaded from the SIFT website (http://sift.bii.a-star.edu.sg/) [46]. For binary classification of mutants, proteins with a classification of "+" were termed functional mutants and "+-", "-+" and "-" were termed loss of function mutants. Data were also obtained form the Humsavar database (http://www.uniprot.org/docs/humsavar) for human sequence variants [47]. The ROCR package (www.cran.r-project.org/package=ROCR) was used to calculate AUC scores for each method [48]. GO term enrichment of high-scoring human polymorphisms was calculated using AmiGO 2 (http://amigo.geneontology.org/rte). PROVEAN results were downloaded pre-computed from their website. Otherwise, all methods were tested using their default settings.

Building custom models

Query sequences were searched against the Uniref90 database [49] using a single iteration of jackhmmer. Sequence ID was calculated as number of matches divided by total length of the alignment after the removal of gap-gap columns. HMMs were built using hmmbuild, then DBS was calculated using hmmscan. jackhmmer, hmmbuild and hmmscan are all part of the HMMER3 package, which can be downloaded at http://hmmer.janelia.org/.

Sequences from the Treefam database (http://www.treefam.org/) were used to build custom models for the human proteins. jackhmmer was used to search each query human protein against the Treefam sequence database. Hits were then filtered for 35% sequence identity, and the remaining sequences were used to build profile HMMs. Models were only used if they had been built from more than 10 sequences. This reduced the number of models to 7090 from 8777.

Comparison of *Salmonella* genomes

Genomes for *Salmonella* Enteritidis str P125109 (AM933172.1, GI:206707319) and *Salmonella* Gallinarum str. 287/91 (AM933173.1, GI:205271127) were retrieved from the EBI Bacterial Genomes page (http://www.ebi.ac.uk/genomes/bacteria.html). Orthologs were identified using a reciprocal hmmer search with a screen for identical Pfam domain architectures. Scores for proteomes were calculated using hmmscan, compared for each domain individually, then summed across the multiple domains in the protein. The predicted distribution of scores was calculated by assuming a normal distribution with mean = 0. Standard deviation for the predicted distribution was calculated by trimming scores lying 5 standard deviations outside the overall distribution of scores and re-calculating standard deviation until it stabilised. This distribution was then used to calculate p-values for DBS, and p-value for enrichment of positive DBS values in the data (exact binomial test performed using R function binom.test, after checking for no shift in mode). p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure (R package p.adjust). Functional classification was performed by using KEGG functional annotation of genes. We created pathway groupings according to the groupings found at http://www.genome.jp/kegg/pathway.html. We grouped genes into those present in a pathway but with no ortholog in the other serovar, genes identified as HDCs by Nuccio and Bäumler, remaining genes identified by our DBS method as loss-of-function mutations, and genes in the pathway not identified as non-functional by either method. dN/dS values were calculated using PAML [50], and for the comparison of DBS and dN/dS, genes were filtered for those with DBS>0, dN>0 and dS>0. Correlations between measures were computed using a Spearman's rho statistic (R package cor).

Software availability

The software used for the Salmonella investigation can be found at https://github.com/UCanCompBio/deltaBS. Analysis from the other sections was performed using HMMER3. Scripts and data for reproducing the main figures in this manuscript can be found at https://github.com/nwheeler443/DBSmanuscript.

Acknowledgements

BIOL430 class of 2013 for preliminary testing of DBS on medically relevant genetic variants. Gemma Langridge for supplying her *Salmonella* Enteritidis pathway database. Sean R. Eddy for writing the HMMER package and supplying additional technical details regarding the HMMER3 implementation.

Author contributions

Script design: LB, NW; Testing: NW, FA; Analysis design: NW, LB, PG, RK; Writing: NW, PG, LB

References

1. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. Nature Publishing Group; 2009;11: 31–46.

2. Loman NJ, Constantinidou C, Chan JZM, Halachev M, Sergeant M, Penn CW, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. nature.com; 2012;10: 599–606.

3. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. Curr Opin Microbiol. 2015;23C: 110–120.

4. Croucher NJ, Didelot X. The application of genomics to tracing bacterial pathogen transmission. Curr Opin Microbiol. 2014;23C: 62–67.

5. Bryant J, Chewapreecha C, Bentley SD. Developing insights into the mechanisms of evolution of bacterial pathogens from whole-genome sequences. Future Microbiol. Future Medicine; 2012;7: 1283–1296.

6. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7: e46688.

7. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7: 248–249.

8. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4: 1073–1081.

9. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011;39: e118.

10. Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, et al. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci. 1996;12: 327–345.

11. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14: 755–763.

12. Krogh A, Brown M, Mian IS, Sjölander K, Haussler D. Hidden Markov Models in Computational Biology: Applications to Protein Modeling. J Mol Biol. 1994;235: 1501–1531.

13. Eddy SR. What is a hidden Markov model? Nat Biotechnol. selab.janelia.org; 2004; Available: http://selab.janelia.org/publications/Eddy-ATG4/Eddy-ATG4-preprint.pdf

14. Kingsley RA, Kay S, Connor T, Barquist L, Sait L, Holt KE, et al. Genome and transcriptome adaptation accompanying emergence of the definitive type 2 host-restricted Salmonella enterica serovar Typhimurium pathovar. MBio. 2013;4: e00565–13.

15. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40: D290–301.

16. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 2014;42: D922–5.

17. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics. 2010;11: 431.

18. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.

19. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. J Med Genet. 2012;49: 353–361.

20. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden

Markov models. Hum Mutat. 2013;34: 57–65.

21. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008;40: 695–701.

22. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41: e121.

23. Moran NA, Plague GR. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 2004;14: 627–633.

24. Goodhead I, Darby AC. Taking the pseudo out of pseudogenes. Curr Opin Microbiol. 2015;23C: 102–109.

25. Rabsch W, Andrews HL, Kingsley RA, Prager R, Tschäpe H, Adams LG, et al. Salmonella enterica serotype Typhimurium and its host-adapted variants. Infect Immun. 2002;70: 2249–2255.

26. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al. Patterns of genome evolution that have accompanied host adaptation in Salmonella. Proc Natl Acad Sci U S A. 2015;112: 863–868.

27. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, et al. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 2008;18: 1624–1637.

28. Nuccio S-P, Bäumler AJ. Comparative analysis of Salmonella genomes identifies a metabolic network for escalating growth in the inflamed gut. MBio. 2014;5: e00929–14.

29. McClelland M, Sanderson KE, Clifton SW, Latreille P, Porwollik S, Sabo A, et al. Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. Nat Genet. 2004;36: 1268–1274.

30. Kuo C-H, Ochman H. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 2010;6. doi:10.1371/journal.pgen.1001050

31. Yang, Bielawski. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15: 496–503.

32. Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. PLoS Genet. dx.plos.org; 2008; Available: http://dx.plos.org/10.1371/journal.pgen.1000304

33. Bhaskara RM, de Brevern AG, Srinivasan N. Understanding the role of domain-domain linkers in the spatial orientation of domains in multi-domain proteins. J Biomol Struct Dyn. 2013;31: 1467–1480.

34. George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. Protein Eng. 2002;15: 871–879.

35. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003;333: 863–882.

36. Nozawa M, Suzuki Y, Nei M. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A. 2009;106: 6700–6705.

37. Roumagnac P, Weill F-X, Dolecek C, Baker S, Brisse S, Chinh NT, et al. Evolutionary history of Salmonella typhi. Science. 2006;314: 1301–1304.

38. Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains. J Bacteriol. 2002;184: 5479–5490.

39. Holden MTG, Feil EJ, Lindsay JA, Peacock SJ, Day NPJ, Enright MC, et al. Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci U S A. 2004;101: 9786–9791.

40. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, et al. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239: 226–235.

41. Clifford RJ, Edmonson MN, Nguyen C, Buetow KH. Large-scale analysis of non-synonymous coding region single nucleotide polymorphisms. Bioinformatics. 2004;20: 1006–1014.

42. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR. Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics. 2013;29: 1504–1510.

43. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA 3rd. Complete mutagenesis of the HIV-1 protease. Nature. 1989;340: 397–400.

44. Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. Genetic studies of the lac repressor. XIV. Analysis of 4000 altered Escherichia coli lac repressors reveals essential and non-essential residues, as well as“ spacers” which do not require a specific sequence. J Mol Biol. Elsevier; 1994;240: 421–433.

45. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol. 1991;222: 67–88.

46. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4: 1073–1081.

47. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res. 2006;34: D187–91.

48. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21: 3940–3941.

49. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31: 926–932.

50. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 1997;13: 555–556.

Supporting Information

S1 Fig. Illustration of DBS calculation on a per-residue basis. This illustration uses two proteins that have accumulated a number of non-synonymous changes between *E. coli* and *S.* Enteritidis, showing a protein that scored a high DBS (*E. coli* gene b0589, *S.* Enteritidis gene SEN0560) and a protein that scored a low DBS (*E. coli* gene b0064, *S.* Enteritidis gene SEN0105). The high scoring protein has a number of positions with high positive DBS which strongly influence the overall DBS for the protein. The low scoring protein has DBS values that are smaller in magnitude and have little individual impact on the overall score. DBS values in the positive and negative directions are of a similar magnitude, so cancel out to some extent.

S2 Fig. Closer view of the ROC curve for the LacI predictions, FPR up to 5%. Pfam performs well to begin with, then performance relative to other methods declines with increasingly permissive scoring thresholds.

S3 Fig. Distribution of |DBS| vs dN/dS for orthologous genes in human and chimpanzee, filtered for |DBS|>0, dN>0 and dS>0. Cutoffs set at 0.5 and 1.5 for dN/dS, and Benjamini-Hochberg adjusted p-value of 0.05 for DBS. Correlation between |DBS| and dN/dS is 0.20. Correlation between |DBS| and dN is 0.39.

S1 Table. Number of human variants from the humsavar database that could be scored by each predictive method.

S2 Table. GO terms showing enrichment/depletion in human polymorphisms with DBS between 5 and 10 when scored against the Treefam HMMs. Data generated using AmiGO 2 (http://amigo.geneontology.org/rte).

S3 Table. *P* values for exact binomial test of whether there is an enrichment for positive DBS values across comparisons of *Salmonella enterica* serovars.
* Significant enrichment for negative DBS values (*P* = 0.01 for Enteritidis-Heidelberg and 0.005 for Gallinarum-Paratyphi)

S4 Table. Putative LOF mutations in genes involved in anaerobic metabolism of *S.* Enteritidis and *S.* Gallinarum. Inclusion is according to the classification in Table S7 of Nuccio and Bäumler, 2014, using our scoring method. Newly identified LOFs are in bold.

S1 Text: Technical details for: A profile-based method for measuring the impact of genetic variation.