Fakultät für Informatik
Technische Universität München

Functional profiles of health: elucidating genetics and microbiome of human disease

Carl Maximilian Miller

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:
Prof. Dr. Stephan Günnemann

Prüfende der Dissertation:
1. Prof. Dr. Burkhard Rost
2. Prof. Dr. Yana Bromberg,
   Rutgers, The State University of New Jersey

Die Dissertation wurde am 10.01.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 23.02.2018 angenommen.

Abstract

Discovering the molecular malfunction patterns leading to disease is one of the major tasks of this century. Multiple studies have explored associations between genetic variation and phenotypic outcomes of human disease. However, progress has been limited and fields like computational variant effect prediction have not seen a significant improvement in method accuracy over the past decade. In this work, we first took a new approach to elucidate the association between genetic variation, functionality and disease. We introduced a new feature, characterizing protein sequence positions by the distribution of variant effects on protein function. We developed a *Machine Learning* approach to predict this feature with high accuracy, providing the foundation for a new and improved variant effect predictor. In parallel we addressed the same question from an alternative perspective. Current research provides substantial evidence that microbiome function is also strongly associated with disease state. We developed two tools that firstly facilitated the analyses of microbiome function profiles between different disease states and secondly, permitted the investigation of microbial functional similarity with respect to different environments. To complement these tools and account for *Big Data* bottlenecks, we additionally developed an automated cluster load balancing software that allowed us to utilize both local and remote compute resources simultaneously and speed up analyses drastically. In summary, we developed new approaches which significantly improve identification of aberrant patterns and their association with disease states.

Zusammenfassung

Das Entschlüsseln jener fehlerhaften molekularen Muster, die unausweichlich zu Krankheiten führen, ist eine der wichtigsten Aufgaben dieses Jahrhunderts. Zahlreiche Studien untersuchten die Zusammenhänge zwischen genetischer Variation und phänotypischen Veränderungen in menschlichen Krankheiten. Bisherige Fortschritte sind jedoch begrenzt, und in Bereichen wie der Effekt-Vorhersage von Mutationen war im letzten Jahrzehnt keine signifikante Zunahme von Präzision festzustellen. In dieser Arbeit verfolgten wir einen neuen Ansatz, um den Zusammenhang zwischen genetischer Variation, Funktion und Krankheit aufzuzeigen. Wir führten ein neues Merkmal ein, welches Positionen in einer Protein Sequenz anhand der Verteilung von Effekten durch Mutationen auf die Protein Funktion charakterisiert. Im Zusammenhang damit entwickelten wir eine auf *Machine Learning* basierende Methode, um dieses Merkmal mit hoher Präzision vorherzusagen. Dies stellt die Grundlage für eine neue und verbesserte Methode zur Effekt-Vorhersage dar. Parallel dazu betrachteten wir die identische Fragestellung aus einer alternativen Perspektive. Der aktuelle Stand der Forschung deutet klar darauf hin, dass das im Mikrobiom enthaltene Funktions-Profil ebenfalls stark mit dem Stadium einer Krankheit assoziiert ist. Wir konzipierten zwei Methoden, die uns zum einen Analysen vollständiger Mikrobiom Funktions-Profile in Bezug auf unterschiedliche Krankheitsstadien ermöglichten, andererseits aber auch die Evaluierung von funktionellen Gemeinsamkeiten zwischen Mikroorganismen in Bezug auf unterschiedliche Milieus erlaubten. Um diese Methoden zu komplementieren und um den Herausforderungen von *Big Data* Analysen zu entsprechen, entwarfen wir zusätzlich eine Software zur automatisierten Lastverteilung zwischen Rechenzentren. Diese erlaubte es uns, lokale und entfernte Ressourcen gleichzeitig zu verwenden und Analysen drastisch zu beschleunigen. Zusammenfassend entwickelten wir neue Ansätze, die eine Identifikation anomaler Muster und deren Assoziation mit Krankheitsstadien signifikant verbessern.

Contents

List of Figures

List of Tables

Acronyms

nsSNP	non-synonymous Single Nucleotide Polymorphism. 1, 2, 13, 30, 65
*fusion*DB	Database relating bacterial fusion functional repertoires to the corresponding environmental niches. 11, 33, 43, 66
mi-faser	MIcrobiome - Functional Annotation of SEquencing Reads. 12, 33, 34, 53, 66
clubber	CLUster-load Balancer for Bioinformatics E-Resources. 11, 30, 33, 34, 43, 66
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool. 33, 43, 52
HPC	high performance computing. 9, 33
fun-TRP	FUNction Toggle-Rheostat Predictor. 28, 31, 33
HGT	horizontal gene transfer. 43
fusion	functional repertoire similarity-based organism network. 43
NMDS	non-metric Multidimensional Scaling. 12, 34, 52, 66
DOI	Digital Object Identifier. 34, 53
KEGG	Kyoto Encyclopedia of Genes and Genomes. 12, 52
CD	Crohn's disease. 6, 52
PWS	Prader-Willi syndrome. 52
GWAS	Genome-Wide Association Studies. 1, 9
MSA	Multiple Sequence Alignment. 3, 4, 5, 30

CV	Cross Validation. 11, 30
faser	functional annotation of sequencing reads. 12
GS	Gold Standard. 12
DMS	deep mutational scanning. 28
IBD	Inflammatory Bowel Disease. 6
MS	Multiple Sclerosis. 6
Pfam	Protein Families. 7
rheostat	rheostat. 13, 28, 30, 65
toggle	toggle. 13, 28, 30, 65
LOO-CV	leave-one-out Cross Validation. 30
RF	Random Forest. 30
AWS	Amazon Web Services. 9

1 Introduction

1.1 Molecular patterns of human disease

The diseases that plague humanity are complex in their etiology, influenced by a variety of genetic and environmental factors. In the last few decades, we have seen significant advances in high-throughput experimental methods, as well as in the sophistication of the resulting data analyses. The Human Genome Project [1] produced the first reference human genome to be used as a baseline for identifying individual-specific sequence variants. Since then we have sequenced many more genomes, *e.g.* for the 1000 Genomes project [2], to increase the resolution of variation baselines across ethnicities and individuals. The many following Genome-Wide Association Studies (GWAS) aimed to uncover specific variants likely to be associated with disease. These associations have been shown for diseases like Schizophrenia, Psoriasis, and Non-alcoholic fatty liver disease among others [3, 4, 5]. However, identifying variants that have a minor effect, but could lead to disease in concert with others, remains difficult if not impossible. Experimental determination of these effects for every single variant is clearly not feasible. Thus, over a hundred computational predictors have been developed to bridge this knowledge gap. Unfortunately, none of these methods solves the question on the level of molecular functionality, much less so to be useful in diagnostic practice or other clinical applications.

Nevertheless, the simultaneous progress in -omics methods have facilitated the study of not only sequence variations but the effect of variants within these sequences at a functional level. Studying these traits in the light of proteomics, epigenetics, diversity and the microbiome (through metagenomics) has led to a much better appreciation of the complexity of human disease. The human microbiome has especially been linked to various host disease phenotypes [6, 7, 8]. GWAS studies have used the variations in the microbiome as a trait, uncovering genetic variations in the host and associating them with microbial variants [9]. Similar to GWAS studies that simply associate sequence variations to disease, we argue that using just microbial variants and associating them to disease is inadequate. Functional genomic approaches provide a more mechanistic as well as realistic perspective of this association.

1.2 Limits of variant effect prediction

Our ability to analyze sequence data currently cannot keep up with the rapidly growing number of genomes and exomes sequenced for research and medical purposes [10, 11, 12]. Experimental assessment of effects caused by all non-synonymous Single Nucleotide Polymorphisms (nsSNPs) is far beyond practical and we require precise specialized computa-

tional methods to substitute for wet-lab evaluations. Distinguishing the roughly 10,000 amino-acid differences observed in protein coding regions of individuals [13] from those associated with disease states is almost like looking for the needle in a haystack. Moreover, substitution effects are not black and white but rather have a gradient range of outcomes [14]. While some variants might only marginally alter a protein structure, *e.g.* leading to a slight change in ligand affinity [15], others alter molecular function and induce drastic changes [15]. Subtle modifications can be very difficult to detect and even though they may have only a minor impact when isolated, these can result in phenotypic changes when co-occurring with other mutations [16, 17]. Those complex epistatic interactions can be found in diseases traits [18, 19] and require alternative approaches than traditional nsSNP analysis.

Figure 1.1: Computational variant effect prediction methods. Around 180 different tools for variant effect predictions are currently available. The different approaches range from conservation and homology to stability analysis and from naive bayes classification to neural networks. Many of them were trained on specifically comprised data sets and thus are meant to predict mutations in very specific use cases (*i.e.* cancer mutations).

In this thesis we focus on the effects of single amino acid substitutions caused through nsSNPs. Single substitutions can frequently be successfully associated to disease traits [20, 21]. Given the broad range of possible applications, it is not surprising that dozens of computational algorithms for the prediction of mutation outcomes have been developed (figure 1.1). However, current approaches still have significant room for improvement [22].

Briefly, computational variant effect prediction methods attempt to assess the impact of an amino acid exchange at a specific position in a protein sequence. This impact can - from a very generic point of view - be scaled into positive, negative or no effect

with regards to protein function. Or more simply, into effect or no effect. Using various computational/mathematical techniques and numerous data inputs, all methods rely, to various extents, on two broad concepts: (i) basic biological principles and (ii) pattern recognition techniques. Despite increasing number and complexity of methods, there has not been a significant improvement in prediction accuracy over the last two decades. This is why we examined the assumptions underlying those concepts with higher scrutiny.

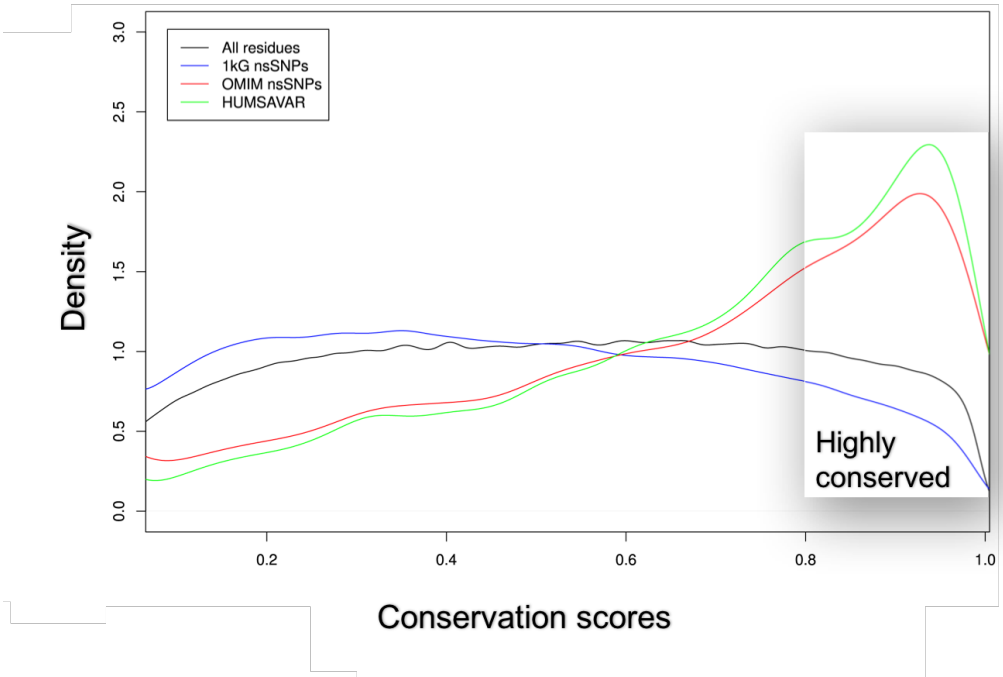

Figure 1.2: Disease causing variants may occur at non-conserved positions. The density of disease causing substitutions is very high within highly conserved regions (0.8 to 1) as expected, however those with low conservation scores (upto 0.6) still demonstrate disease occurrence. Adapted from (de Beer *et al.* 2013 [23])

Basic biological principles range from biochemical amino acid properties to evolutionary information extracted from Multiple Sequence Alignments (MSAs) of homologous sequences. Some approaches use globally-applicable scoring matrices, *e.g.* BLOSUM62 [24], which estimate likelihoods of amino acid substitutions in curated MSAs. Non-disruptive substitutions (*i.e.* those represented in matrices with high probabilities) often match literature based biochemical classifications of side chains. Those in turn are often the starting point for designing experimental mutation studies. Substitutions of amino acids with comparable properties are expected to allow normal or near-normal protein activity. Other substitutions are very likely to alter or abolish function. Variant-effect predictors use different subset of these properties as a base for assessing the potential impact when substituting the wild-type residue with a specific amino acid. One property utilized by most effect prediction models are evolutionary constraints (*i.e.* conserved vs. non-conserved). Generally, if a specific variant is present in functionally-active homologs,

it allows for a 'normal' or 'near-normal' function. In the contrary, if a variant is absent in the MSA, chances are that it has been excluded through evolutionary selection and as such poses an unfavorable substitution. However, many protein positions do change during evolution and therefore are not conserved in MSAs. Those non-conserved positions are often neglected and deemed as less important. This is highly questionable as functional diversification in homologs often emerges through amino acid changes at non-conserved positions [25]. Even though disease variants are often highly conserved, it has been shown that disease-causing substitutions can very well occur at non-conserved positions [26, 27] (figure 1.2). Another implied assumption for prediction models incorporating MSAs is that if a particular variant-effect is known for one homolog, similar outcomes are expected for the same variant in other family members. In strong disagreement with that hypothesis, studies show that pathogenic variations can be compensated for in homologs and thus are not lethal any more on occurrence ([28], figure 1.3).

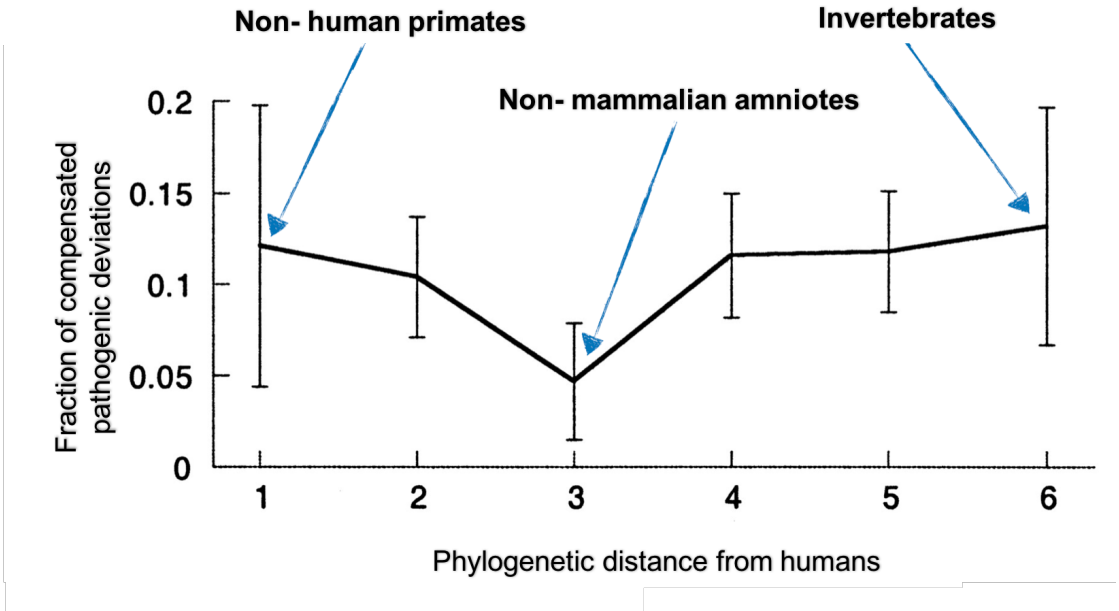

Figure 1.3: Non-conservation does not necessarily indicate neutrality. In a position that is not conserved and/or compensated for within other species, a variant in the human ortholog may still be pathogenic. This indicates that 'non conserved' positions found in MSAs cannot be generically assumed neutral. Adapted from (Kondrashov *et al.* 2002 [28]; Copyright (2002) National Academy of Sciences, U.S.A.)

Prediction methods that apply pattern recognition techniques often require comprehensive, experimentally validated data sets (functional effects of amino acid substitutions) for training and optimization of the underlying models. Even though recently developed deep mutagenesis approaches improve this situation, data sets with sufficient sample size meeting those criteria are still hard to come by. Furthermore, many models work under the assumption that variants used for model training roughly cover the entire

spectrum of variation. The reality is that variants studied in lab conditions are subject to experimental limitations and, even more importantly, to the interests of researchers. This is why currently available sets of experimentally validated variants feature a considerable bias towards conserved amino acid positions [29]. Variants with no (neutral) or weak effects which naturally occur predominantly at non-conserved positions are for the most part not in the main focus of research driven by experimentalists. Without any doubt, the holy grail for experimentalists lies in identifying mutations, which are declarative for a phenotype of interest, rather than all those which do not show any effect. Additionally, embryonically lethal variants are often missing entirely or significantly under-represented in our data sets. We see lethal variants only rarely simply because they are not compatible with life.

As such, one of the reasons for stagnation of prediction accuracy despite advances in methodology is that we are missing essential data in our training sets. Meanwhile, this missing data in training sets is approximated using databases which track variants of known, strong effects, *i.e.* those known to be linked to Mendelian diseases. To approximate neutral variants, MSAs of ortholog sequences represent a starting point, identifying non-conserved residues across species which are assumed to have little or no effect. However, associating the lack of conservation with the likelihood of variants being neutral is problematic as discussed earlier.

It has been shown that some functionally-important, non-conserved positions do not follow any of the evolutionary or biochemical assumptions made for conserved positions [30]. That could explain why prediction algorithms that rely on the conventional substitution rules or on laboratory-derived training sets might not correctly predict the variant outcomes at non-conserved positions. This suggests that key non-conserved positions follow different substitution rules than conserved positions. In this thesis we developed this hypothesis and defined a new concept of sequence position types to compensate for the described mis-assumptions. These concepts may redefine the current prediction models that ultimately aim to clarify our understanding of disease from a host-intrinsic point of view. Additionally we looked extrinsically into other aspects of functional variation that may complement our studies. Since strong associations exist between the microbiome and disease, we describe functional prediction from the perspective of an extrinsic factor-the microbiome.

1.3 Associations with the microbiome

The human microbiome comprises of thousands of species, consisting of more bacterial cells than human cells within the host [31]. Thus, it has even been called an 'organ' by itself. The composition and diversity has been shown to be affected by multiple factors including age, sex, diet, host genetics and most importantly, by disease states [32, 33, 34, 35, 36] (table 1.1).

Disease state
Acne
Auto-immune disease
Allergy
Autism
Cancer
Depression
Diabetes
Inflammatory bowel diseases
Liver disease
Neurological disorders
Obesity

Table 1.1: Some conditions associated with alterations in the human microbiome.

Crohn's disease (CD) for example is a multifactorial illness resulting from a genetic predisposition, environmental influences and changes in the intestinal microbiome. Analysis of the microbiome from CD patients have revealed microbial profiles that are distinct to individuals affected by the disease [37, 38]. These features include an enrichment of certain communities simultaneous to the depletion in others [39, 40]. In order to understand whether disease precedes the change in the microbiome or vice versa, it is important to determine what the functional impact of this change is. For example, whether there is enrichment in the communities associated with 'high inflammation' compared to a depletion in those that are 'protective' in this setting. This is further complicated by the fact that genetic predisposition may play an important role here - firstly in development of the disease directly and therefore an influence on the microbiome subsequently. Alternatively, if host genetics influences microbiome composition and thus consequently leads to CD. It may also be a combination of the two situations. In spite of the advances made in this field, the question of 'causality' still remains. Interestingly, earlier studies demonstrated that the association of the microbiome with CD is more consistent with changes in the functional profile of the microbial communities rather than their diversity [41]. Inflammatory Bowel Disease (IBD), diabetes (type 2) and colorectal cancer are some of the most extensively studied diseases in this context [42, 43, 44, 45]. This may be due to the disease phenotype occurring in close proximity to the gut. However, multiple studies have demonstrated associations between microbial variation during diseases that are not localized to the intestine, for example, during Multiple Sclerosis (MS)

[46, 47]. Interestingly, both host genetic factors and the microbiome status can affect disease state either alone or in combination [48, 49] (figure 1.4).

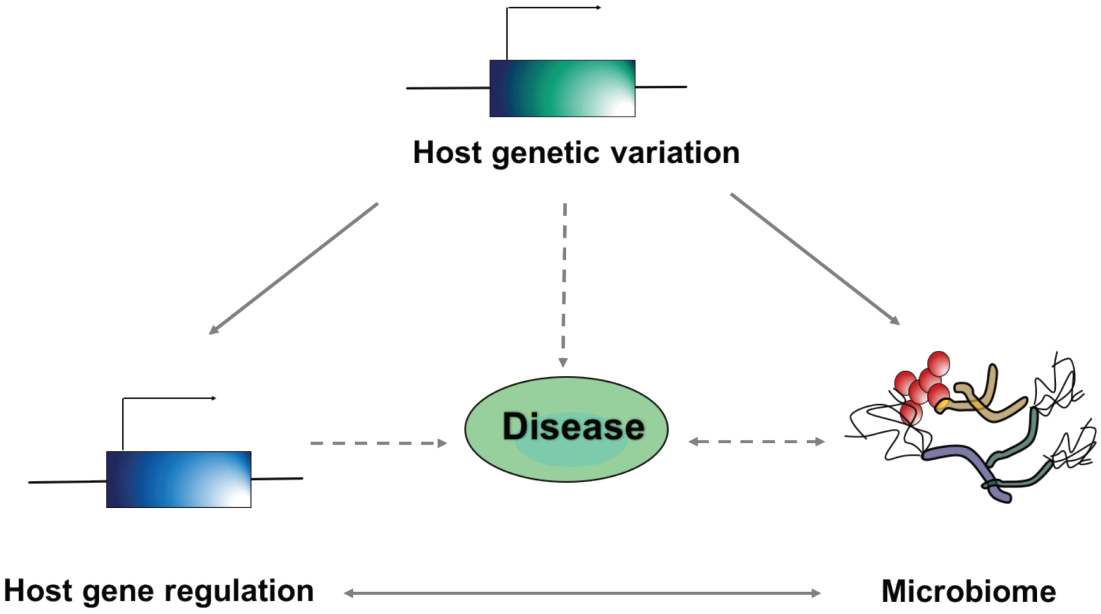

Figure 1.4: Causal interactions between host genetic regulation, variation, the microbiome and disease. Both, the regulation of genes in the host and genetic variation in these genes can affect the development of disease. The microbiome on the other hand can affect disease phenotype but also be affected by the above factors, gene regulation by the host, genetic variation and the disease state itself. Adapted from (Luca *et al.* 2017) [50]

Therefore, recently researchers have begun to consider the microbiome as a complex human trait [9]. However, Luca *et al.* have suggested that the microbiome cannot be considered a traditional quantitative trait but more like an array of complex traits [50]. The human microbiome is a multi-dimensional profile comprising of various features-relative abundance, taxa, diversity molecular/metabolic pathways and other functional aspects. These individual features may be associated with a specific host locus, they may also be influenced by a different environmental factor. Interestingly, human SNPs associated with the microbiome were discovered , which were enriched for genes associated with immunity as well as complex diseases [51]. Nevertheless, environmental factors may indeed have a stronger effect than that of the host genetic effect [52]. Taking this one step further, it has been shown that the environment may have a profound effect on microbial functionality, however a direct link has not been demonstrated. Since the microbiome holds a vast amount of functional information, accessing this in a timely and efficient manner could prove to be extremely useful. Microbial sequences are generally compared based on their evolutionary relatedness, which assists their classification but does not provide a guarantee of their functional relatedness. Based on relatedness, phylogenic trees can be assembled, which are important for re-constructing the evolutionary

history of these organisms. However, how these microorganisms behave functionally in their distinct environments cannot be predicted from this information [53]. Fortunately, the number of sequenced genomes has increased drastically and databases like Protein Families (Pfam) [54] allow us to take a closer look at the functionality through protein sequences. Therefore in 2015, we developed a method that compared microbial functional similarity based on translated proteins from their genomes [53]. Through this tool we observed inconsistencies in the functional diversity at the level of taxa. Additionally, we could use meta-data to underscore definitive environmental factors driving microbial diversification. Our results demonstrated that function annotation methods are more descriptive of organism similarity compared to just gene-sequence identity. Intuitively, microbial communities occupying the same niche can be expected to have similar functional properties than those thriving in a different environment. In line with this, we could identify natural functional clusters of bacteria within a wide range of metabolically, environmentally and phenotypically diverse microorganisms (figure 1.5). Apart from understanding microbial diversity through similar functionality, another important aspect is how this could be translated clinically. Analyzing microbial/microbiome (metagenomic) data by associating it with altered function holds the true key for uncovering mechanistic interactions leading to and from disease.

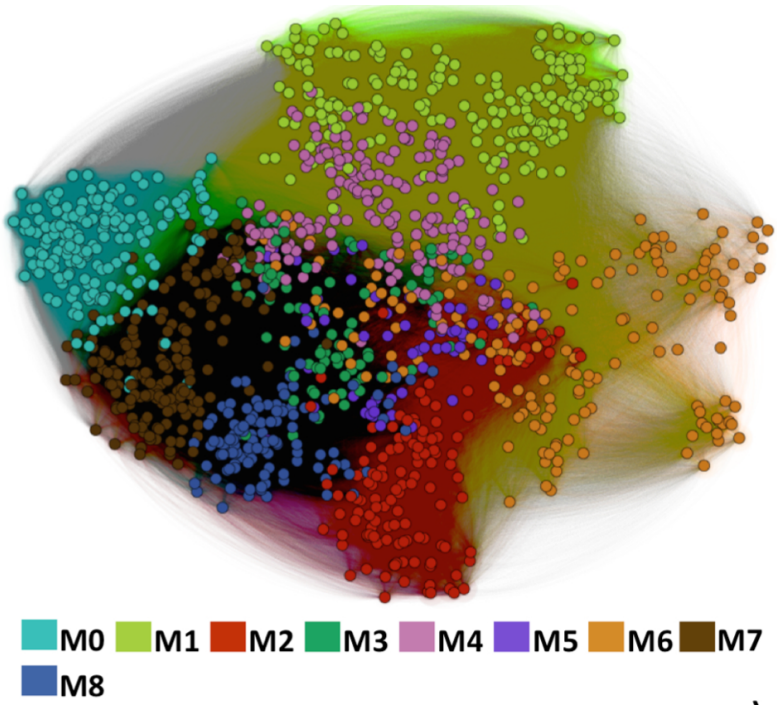

Figure 1.5: Natural organism grouping using the original Fusion module. Source (Zhu *et al.* 2015 [53])

1.4 Big Data in disease

With impressive developments in high-throughput -omics, biologists have recently validated their admission ticket to the growing *Big Data* club [55]. Massive data sets produced in sequencing projects have introduced new challenges for storage and analysis. In the first place, this offered exciting new possibilities and enabled large scale analysis approaches. Fields like infectious diseases have experienced a nearly exponential increase in number of publications (Figure 1.6).

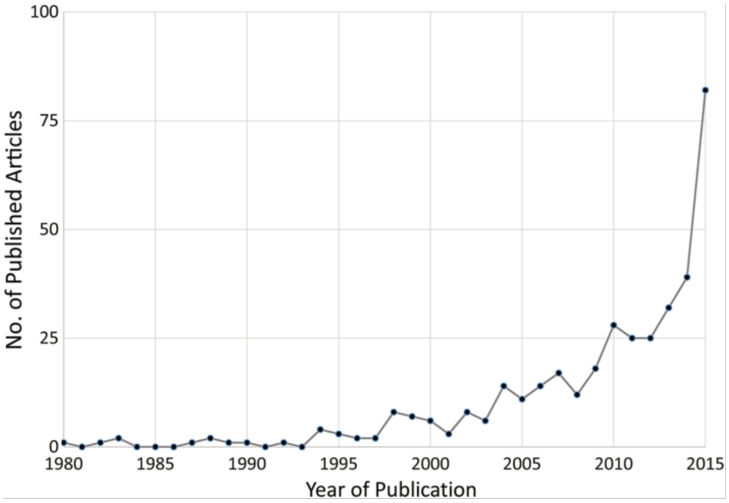

Figure 1.6: Near-exponential increase of published research on infectious diseases with the advent of *Big Data*. Source (Bansal *et al.* 2016 [56])

It is safe to assume that we did not even scratch the surface of what is achievable based on the rapidly growing information produced in various fields. However, if not approached in the right way, this massive increase in data can be more of an obstacle than an opportunity. Developing efficient algorithms has become an essential task, as has detailed knowledge about using cluster or cloud computing. On the first glance, simple issues like slower data-transfer rates evolve into serious obstacles when dealing with vast amounts of data. At this point, a variety of applications and high performance computing (HPC) platforms for biological *Big Data* analysis exist [57]. Commercial solutions like Amazon Web Services (AWS) offer nearly unlimited scaleable on-demand compute resources. In academic settings, initiatives like Jetstream [58] aim to provide researches with adequate tools to meet today's analytical challenges. Identifying disease patterns is often directly associated with the power of a study. Approaches like GWAS require large data sets to identify statistically significant traits. In this context, *Big data* clearly opened new doors in uncovering disease patterns [59]. The methods we describe in this work rely on the processing of large databases to compile reference databases or process new queries. To our advantage, those repositories are consistently growing, which enables us to improve our tools steadily - given the availability of sufficient resources.

Thus, unlocking the possibilities held within biological *Big data* is an important part of the success story of the methods we employ.

1.5 Overview of this work

In section 2.1, I describe a new concept for characterizing protein sequence positions and its impact on computational effect prediction of amino acid substitutions. This concept was based on experimentally evaluated effects of amino acid mutations. We segregated protein sequence positions into two classes based on the range of validated mutational outcomes per position: (i) binary (on/off) effects compared to (ii) progressive changes. We demonstrated that current computational effect predictors fail to correctly assess amino acid substitutions in the context of this classification. We concluded that building more accurate prediction models requires distinguishing between the two position type classes.

In section 2.2, I summarize our progress on building a new machine learning approach for predicting the new position type classes introduced in section 2.1. We extracted experimentally validated substitution effects for five protein sequences from corresponding deep mutagenesis data sets. Based on its distribution of effects, every sequence position could be assigned to a class of binary outcomes, range of effects or no effect (neutral). We developed a 2-step approach to predict those class labels using only sequence based features. We validated our model via Cross Validation (CV) and demonstrated that our approach is resistant to changes in training data as well as variations in the selected features. Overall, we reached an averaged accuracy of 82% and used this model to analyze the distribution of position types in an entire set of human enzymes.

In section 3.1 I outline CLUster-load Balancer for Bioinformatics E-Resources (*clubber*), an automated load balancing software we developed to reduce required computation time and facilitate *Big Data* analyses for all methods described in this thesis. *clubber* was originally intended to bundle compute resources available for our group and allow us to automatically distribute computations to all of them simultaneously. Since then it evolved to be a fully automated cluster load balancer, supporting two job schedulers prominently used in academic settings as well as cloud compute resources. We integrated various pre- and post-processing routines for common approaches enabling analysis of biological data sets. To simplify usage and the integration of *clubber* in existing workflows, we created a stand-alone Docker [60] image which is fully manageable via the provided web interface. *clubber* is connected to all our web services and enables extremely fast processing of user submission.

In section 4.1 I discuss our Database relating bacterial fusion functional repertoires to the corresponding environmental niches (*fusion*DB), a novel database that links functional similarities of microorganisms and environmental preferences. We compiled a reference database consisting of 1374 taxonomically distinct bacteria. Individual microorganisms are represented by the entirety of functions retrieved from their proteome and connected via those functions that they share with other microbes. *fusion*DB can additionally facilitate the functional analysis of unknown microorganisms *e.g.* found in samples from affected patients. Mapping microbial genomes to the functional repertoire of bacteria integrated in *fusion*DB identifies shared functionality. This association may provide crucial information about the characteristics of the query organism. *fusion*DB is available as a web service [61] and allows users to explore functionality graphs inter-

actively. With the increasing number of available microbial genomes and more complete metadata annotations, *fusion*DB has the potential to develop into the 'go-to' workflow for microbial functional analysis.

In section 4.2 I describe MIcrobiome - Functional Annotation of SEquencing Reads (*mi-faser*), an extremely fast and accurate method for annotation of molecular functionality encoded in microbiome sequencing read data without the need for assembly or gene finding. *mi-faser* is comprised of two key components: (i) a Gold Standard (GS) set of reference proteins and (ii) the functional annotation of sequencing reads (*faser*) algorithm. The GS set, used to compile the *mi-faser* reference database, contains only protein sequences with experimentally annotated molecular functions. Thus we avoided erroneous functional annotations due to mis-annotations frequently present in other databases. The *faser* algorithm aligns translated sequencing reads to the set of full-length proteins contained in the *mi-faser* reference database. The molecular functionality is thus determined based on these alignments. Further, *mi-faser* facilitates the comparison between functional profiles of entire metagenomes via non-metric Multidimensional Scaling (NMDS) plot representations. This enabled us to directly compare function abundance profiles between microbiome samples of healthy and disease-related individuals. *mi-faser* is available as a web service [62] additionally offering downstream analysis like mapping metabolic pathways through Kyoto Encyclopedia of Genes and Genomes (KEGG).

2 Improving variant effect prediction

2.1 Computational predictors fail to identify amino acid substitution effects at rheostat positions

2.1.1 Preface

Predicting the effects of amino acid mutations (variants) on protein function is a problem which has been seeing a lot interest and developments in the last two decades. One of the reasons for this attention is based on the observation that nsSNPs are often associated with diseases. Any given individual can have over 10,000 amino-acid differences in their protein coding regions, as compared to the reference genome [13]. Given the large number of variants, it is not feasible to experimentally determine the outcomes/effects for all changes, *i.e.* how they alter molecular function, protein structure, evolutionary fitness, or lead to pathogenesis. Currently, roughly 180 computational methods exist (figure 1.1) approaching this task from various perspectives. While some of them are rather general, others are trained on very selective data sets and predict effects of mutations only within a very narrow range. However, despite the variety of different approaches, current algorithms have significant room for improvement [22]. Especially since we can actually observe that the increase in number and complexity of available methods in the last two decades, is not remotely proportional to an increase in prediction accuracy.

We addressed this by introducing a new feature, characterizing protein sequence positions based on the distribution of experimentally evaluated effects of amino acid substitutions on protein function. We demonstrated that this protein sequence position class - rheostat (*rheostat*) or toggle (*toggle*) - affected computational effect predictions. *toggle* positions were characterized by binary variant outcomes, *i.e.* amino acid substitutions caused either a severe impact or no/weak effects. On the contrary, *rheostat* positions showed a progressive (gradient) change upon mutation. We compared experimentally-evaluated substitutions in the *E. coli* LacI repressor protein with predictions from 16 widely-used computational methods. We focused on how computational predictors scored those substitutions, which were experimentally determined as neutral (no/weak effect) and non-neutral (effect). Explicitly, how those scoring profiles differed in the context of the two positions classes - *toggle* and *rheostat*. All methods failed two key expectations: predicted scores for neutral mutations at *toggle* positions were incorrectly predicted as more non-neutral than scores predicted for non-neutral mutations at *rheostat* positions. Secondly,neutrals at both *toggle* and *rheostat* positions were incorrectly predicted to be different. Further, none of the methods significantly distinguished *toggle* neutrals from *toggle* non-neutrals or *rheostat* neutrals from *rheostat* non-neutrals. However, we observed that toggle non-neutrals were correctly distinguished from rheo-

stat neutrals. This hinted at two conclusions. First, with many toggle positions being conserved, in contrast to most *rheostat*s, all methods appear to annotate position conservation better than mutational effect. This in turn can be explained by the well-known fact that predictors assign disproportionate weight to conservation. Moreover, this clearly poses a limiting factor for improving predictor performance. Second, this behaviour obviously reflects a bias in training data. Due to the experimentalists choice of topic and limitations of biological experimental design, current data sets consist of (for the largest part) either obviously severe - *toggle* - variants (*i.e.* Mendelian disease causing variants) or likely neutral - *rheostat* - variants. This is ultimately what models are trained to distinguish between. With training data unlikely to improve drastically in the near future, the knowledge about *rheostat* and *toggle* positions is a key feature for building more reliable and accurate variant effect predictors.

The entire computational analysis was carried out by me. The experimental work was done by our collaborators, the Liskin Swint-Kruse lab. The manuscript was drafted by all authors.

2.1.2 Journal article. Miller et al., Scientific Reports 2017

Supplementary material can be found online at [63]. The published article is attached below.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Computational predictors fail to identify amino acid substitution effects at rheostat positions

Received: 11 October 2016
Accepted: 15 December 2016
Published: 30 January 2017

M. Miller[1,2,3,*], Y. Bromberg[1,4,5,*] & L. Swint-Kruse[6,*]

Many computational approaches exist for predicting the effects of amino acid substitutions. Here, we considered whether the protein sequence position class – rheostat or toggle – affects these predictions. The classes are defined as follows: experimentally evaluated effects of amino acid substitutions at toggle positions are binary, while rheostat positions show progressive changes. For substitutions in the LacI protein, all evaluated methods failed two key expectations: toggle neutrals were incorrectly predicted as more non-neutral than rheostat non-neutrals, while toggle and rheostat neutrals were incorrectly predicted to be different. However, toggle non-neutrals were distinct from rheostat neutrals. Since many toggle positions are conserved, and most rheostats are not, predictors appear to annotate position conservation better than mutational effect. This finding can explain the well-known observation that predictors assign disproportionate weight to conservation, as well as the field's inability to improve predictor performance. Thus, building reliable predictors requires distinguishing between rheostat and toggle positions.

Recent years have seen an explosion in the number of genomes and exomes sequenced for research and medical purposes, *e.g.* for diagnosing and prognosing predisposition to or progression of disease[1–3]. Unfortunately, our ability to interpret sequence data lags far behind. Take exomes, for example: Any given individual can have over 10,000 amino-acid differences in their protein coding regions, as compared to the reference genome[4]. These differences are often caused by nsSNPs (non-synonymous single nucleotide polymorphisms); in this manuscript, we refer to amino acid substitutions as *variants*. Given the large number of variants, it is not feasible to experimentally determine the *outcomes/effects* for all changes, *i.e.* how they alter molecular function, protein structure, evolutionary fitness, or lead to pathogenesis. Thus, dozens of computational algorithms have been developed to predict outcomes. However, current algorithms have significant room for improvement[5].

Interestingly, underneath the assortment of computational/mathematical techniques and numerous data inputs, all variant-effect predictors rely, to various extents, on two broad concepts: (i) basic biological principles and (ii) pattern recognition techniques, optimized using results from wet-lab experiments. We examined the assumptions underlying these concepts, hoping to identify avenues to improve predictions.

The basic biological principles include biochemical amino acid similarities and evolutionary information about the protein of interest. The latter is usually in the form of a multiple sequence alignment (MSA) of homologs and is used to determine which specific sequence positions show evolutionary constraints (*i.e.* conserved *vs.* non-conserved). This information serves as a proxy to predict which amino acid substitutions are "allowed" (inferred by their presence in functionally-active homologs) and which are "bad" (selected against during evolution and, thus, absent from the multiple sequence alignment). Some variant-effect predictors use globally-applicable scoring matrices, *e.g.* BLOSUM[6], that represent the likelihoods of amino acid substitutions in curated MSAs. The substitutions allowed by these matrices often match the biochemical classifications of side

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08901, USA. [2]Department for Bioinformatics and Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany. [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748 Garching/Munich, Germany. [4]Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA. [5]Institute for Advanced Study at Technische Universität München (TUM-IAS), Lichtenbergstraße 2a, 85748, Garching/Munich, Germany. [6]The Department of Biochemistry and Molecular Biology, The University of Kansas Medical Center, 3901 Rainbow Blvd, Kansas City, KS 66160, USA. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to Y.B. (email: yana@bromberglab.org)

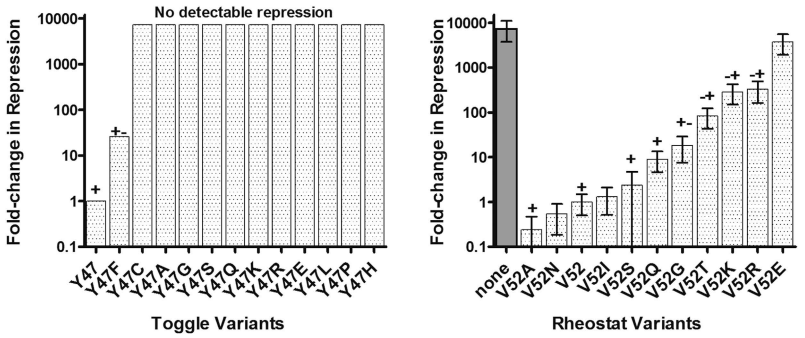

Figure 1. Experimental differentiation of toggle and rheostat positions. The left panel shows an example of a toggle position (tyrosine in position 47): Relative to wild-type (value normalized to 1), most substitutions at LacI position 47 abolish transcription repression of the reporter-gene. The right panel shows an example of a rheostat position (valine in position 52): Variants at this position in LacI exhibit a wide range of repression levels relative to wild-type (value normalized to 1). Data for position 52 (right panel) are adapted from11; the dark gray bar shows the ratio of no-repression (full expression of the reporter gene) to repression by wild-type LacI. Data for position 47 (left panel) were adapted from[14]. Briefly, the earlier study categorized these semi-quantitative data relative to the activity of un-repressed reporter gene (i.e., in the absence of repressor protein). For this figure, we translated the semi-quantitative ranges to the quantitative scale using the "none" value on the right panel.

chains found in textbooks, which, in turn, are often used to design experimental mutation studies. By convention, only substitutions of "similar" amino acids are expected to allow normal or near-normal protein activity, while other substitutions are expected to alter or abolish function.

Variant-effect predictors that incorporate pattern recognition approaches are often training-driven. That is, they use large sets of experimentally verified functional effects of amino acid substitutions to build predictive models. These methods often (wrongly) assume that the variants used for training broadly represent the entire world of variation. Another extrapolation is implicit in any computational method that incorporates MSAs: if a particular variant-effect is known for one homolog, similar outcomes are expected for the same variant in other family members.

Both the conventional substitution rules and training datasets are biased by a gap between unbounded evolutionary reality and limited laboratory work; *i.e.* laboratory variants are subject to experimental limitations and to the interests of the scientists. Indeed, available experimentally-annotated sets of variants are heavily biased to the study of conserved amino acid positions[7].

However, many protein positions are *not* conserved in MSAs; *i.e.* they *do* change during evolution. Non-conserved positions are often ignored as not important, but this need not be true: homologs often evolve functional variance *via* amino acid changes at non-conserved positions[8] and disease-causing substitutions can occur at non-conserved positions[9,10].

We hypothesized that key non-conserved positions follow different substitution rules than conserved positions. If so, prediction algorithms that rely on the conventional substitution rules or on laboratory-derived training sets might not correctly predict the variant outcomes at non-conserved positions. Indeed, we recently showed that some functionally-important, non-conserved positions do *not* follow *any* of the evolutionary or biochemical assumptions made for conserved positions[11]: In that work, we identified 12 positions in the LacI/GalR family of proteins that varied widely among family members[8]. Next, using the natural *E. coli* lactose repressor protein (LacI) and modified (synthetic) versions of seven LacI/GalR homologs (including GalR, PurR, and RbsR), we substituted the native amino acid in each of these positions with 5–13 other amino acids and measured functional outcomes[11]. If these positions were functionally important, the conventional rules would predict that only a few similar substitutions would allow normal function and that most others would abolish function (*e.g.* Fig. 1, left panel). However, at most of the chosen non-conserved positions, variants exhibited a wide range of functional effects (Fig. 1, right panel). Furthermore, these effects did not correlate with evolutionary frequency, side chain similarities, or functional effects of the same substitutions in homologous proteins[11].

We named these positions *rheostats* after their most prominent characteristic: When multiple amino acids were substituted at one rheostat position, functions of the mutant proteins could be rank-ordered to show a progressive effect (Fig. 1, right panel). This contrasts with the *toggle* (on-off) behavior that is frequently observed at conserved positions and is predicted by the conventional rules (Fig. 1, left panel). We have also noted rheostatic behavior in published variant datasets for other proteins (*e.g.* refs 12, 13), which indicates that rheostat positions are likely widespread in the protein universe. Importantly, the progressive functional impact of variants at rheostat positions is all but disregarded by the current variant-effect predictors, which were either developed using variants at toggle positions and/or the thresholded (binary) version of functional outcomes. We also hypothesized

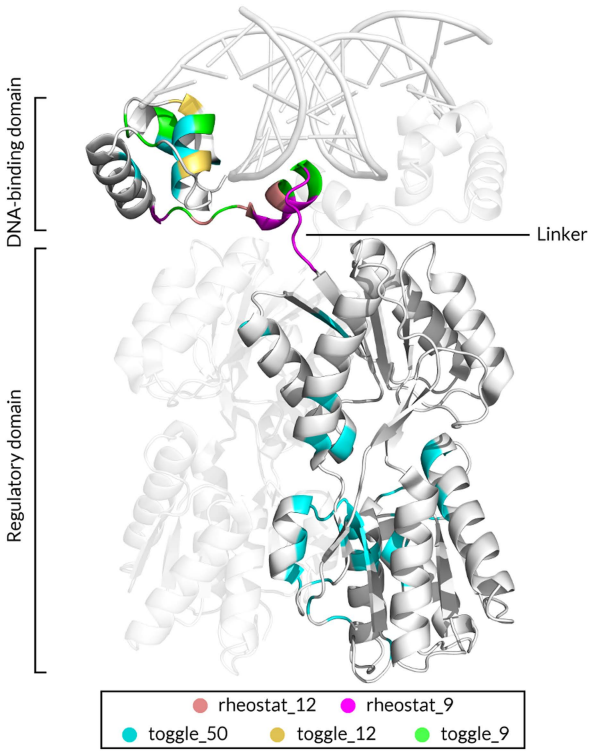

Figure 2. Locations of toggle and rheostat position sets on the structure of the LacI homodimer bound to DNA (PDB 1EFA[34]; visualized with PyMOL)[71]. On one monomer, positions are colored by the sets described in the text. Note that smaller sets are included in the larger sets. For example, *toggle_12* positions are also part of *toggle_50*. Chain B (identical to Chain A) is shown in the background at 50% transparency. DNA is shown as a double helix at the top of the figure.

that the poor correlation observed between functional outcomes of substitutions and biophysical/evolutionary amino acid properties could contribute to erroneous predictions.

Here we compare predictions from 16 widely-used computational methods for experimentally-evaluated substitutions at rheostat and toggle positions in LacI. Each variant at each position was assigned as either non-neutral or neutral based on experimental outcomes (Methods). When experimental outcomes were compared to predictions at rheostat positions, many non-neutral variants were incorrectly predicted to be neutral. At toggle positions, neutral variants were also poorly predicted, as previously observed by Gray *et al.*[7]. We gained insight into this problem by comparing the overall prediction ranges for rheostat and toggle positions, instead of using the default binary neutral/non-neutral thresholds: Our results suggest that current computational predictions could be enhanced by first determining whether the affected position functions as a rheostat or as a toggle.

Finally, our results show that evaluations of predictor performance are misled by bias in the available experimental data: As noted above, there is a dearth of experimental results for variants at rheostat positions; in addition, at toggle positions, the number of experimentally validated non-neutral variants is high and the number of validated neutral variants is very low. The latter is further complicated by an arguably obfuscated definition of "neutrality" that often differs among methods. Together, these biases have had the practical effect that all toggle variants appear to be non-neutral and all rheostat variants are assumed to be neutral. This leads to an artificially low number of variants incorrectly predicted as non-neutral, whereas the number of incorrect, neutral predictions cannot be properly estimated due to the low number of experimentally-validated neutral toggles. We propose that eliminating data bias, *e.g.* by using experimental results that annotate rheostat position variants, and optimizing prediction algorithms to account for position type will lead to more accurate evaluations of variant impact.

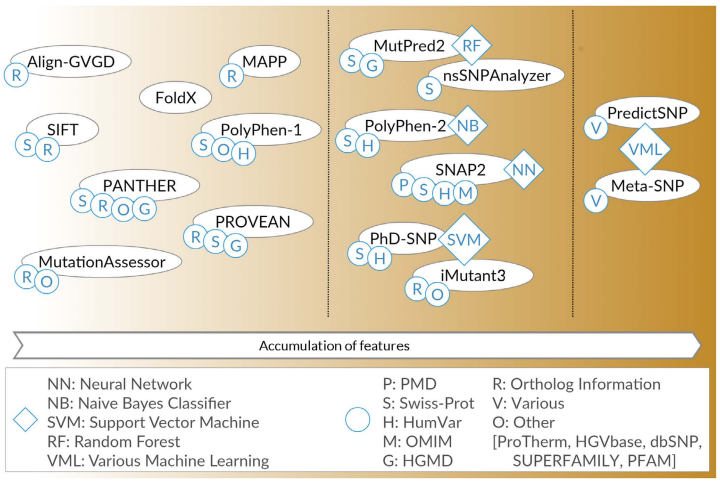

Figure 3. Variant-effect predictors vary in features and development data used. The 16 publicly available variant-effect prediction algorithms can be broadly grouped by use of (i) basic biological principles and evolutionary information, (ii) pattern recognition techniques and machine learning, and (iii) meta/ensemble predictors.

Results

To test the performance of variant-effect predictors at rheostat and toggle positions, we extracted positions that exhibit these behaviors from two experimental datasets for LacI (Methods; Fig. 2)[11,14]. Rheostatic positions were identified by variants whose outcomes showed a progressive distribution; notably the outcomes of individual variants did not correlate with evolutionary frequency, side chain similarities, or outcomes in homologous proteins[11]. Rheostats were identified in both experimental datasets; for this work, predictions were compared to the quantitative measurements of Meinhardt *et al.*[11]. Toggle behavior was identified by *binary* variant outcomes, *i.e.* amino acid substitutions caused either a severe impact or no/weak effects; predictions were compared to data from the semi-quantitative study by the Miller lab[14].

We designated three sets of rheostat and toggle variants (Methods, Supplemental Tables 1 and 2): (i) The *stringent* set comprised the nine rheostat positions that were identified by both experimental studies, and a comparable number of nearby toggle positions. (ii) The *complete* set comprised all 12 rheostat positions from the quantitative experiments and a comparable number of toggle positions. (iii) The *extended* set comprised all 12 rheostat positions and 50 toggles from the semi-quantitative study. For both rheostat and toggle positions, individual variants were classified as either neutral or non-neutral according to their experimentally-determined fold-change relative to wild-type repression.

For all variants in these sets, we predicted outcomes using 16 selected prediction algorithms (Fig. 3; Supplementary Tables 3 and 4). Most prediction trends were similar for all three sets (Fig. 4; Supplementary Figs 1, 2 and 3); differences are highlighted in the text below and likely stem from the small numbers of neutral variants at toggle positions (the *stringent* set contained only four neutral variants and the *extended* set contained 28). Experimentally-determined outcomes were compared to computational predictions in two ways: First, we compared the overall distributions for prediction scores from each of the four classes (rheostat non-neutral, rheostat neutral, toggle non-neutral, and toggle neutral). Second, for algorithms that generate continuous prediction scores, we directly compared predictions and experiments for individual variants.

Differences between rheostat and toggle positions confound variant-effect predictions.

Variant predictors report their results in one of two ways: as a binary decision (neutral/non-neutral) or as a scored value representing the likelihood of non-neutrality, often thresholded to make a binary decision. Ideally, each algorithm would clearly separate the distributions of neutral and non-neutral variants, regardless of their toggle or rheostat position location (Fig. 4a). That is, methods should differentiate toggle neutrals from both toggle and rheostat non-neutrals; moreover, rheostat neutrals should be classified similarly to toggle neutrals. Finally, the progressive effect of rheostat non-neutrals should be different from the binary effect of toggle non-neutrals (Fig. 4a).

However, for the *stringent* comparison set, no algorithm significantly distinguished toggle neutrals from toggle non-neutrals (Fig. 4b and Supplementary Fig. 1, (*B*) *vs.* (*D*): dark red vs. dark blue), as determined by the Kolmogorov-Smirnov test (KS-test) for continuous predictors and the Fisher exact T-test for binary predictors. The *extended* set was better differentiated at toggle positions by 8 of 16 methods (Supplementary Fig. 3). At

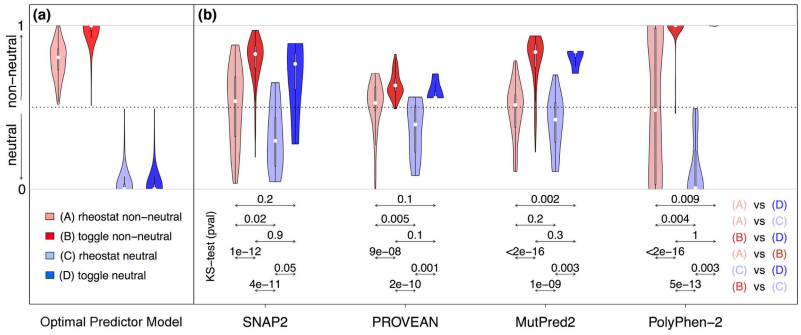

Figure 4. Distributions of variant scores from continuous prediction methods differ between rheostat and toggle positions (*stringent* set). Panel (a) shows the distributions expected from an ideal variant-effect predictor, while panel (b) shows the distributions determined for neutral and non-neutral variants at both rheostat and toggle positions in the stringent set. These four predictors were selected on the basis of top performance in differentiating rheostat non-neutrals from rheostat neutrals. Results for all other predictors are in Supplementary Fig. 1. The violin plot is an augmented box plot where the width at any given Y-axis value indicates the probability density of the data (median, white circles; interquartile range, box outline). The p-values in the legend are from a Kolmogorov-Smirnov (KS) test, indicating whether a method can significantly distinguish between the two distributions pointed to by the respective arrows. Results from the complete and extended sets are in Supplementary Figs 2 and 3.

rheostat positions in the *stringent* set, neutrals and non-neutrals were significantly differentiated by only three of the methods (SNAP2[15], PROVEAN[16], and PolyPhen-2[17]; Fig. 4b and Supplementary Fig. 1, (*A*) *vs.* (*C*), light red vs. light blue).

Note, however, that non-neutral variant PolyPhen-2 scores are nearly uniformly distributed over the entire score range. The *complete* and *extended* sets were differentiated by four methods (the three above and MutPred2, the unpublished successor of MutPred[18]; Supplementary Figs 2 and 3).

In contrast, toggle non-neutrals were well-differentiated from rheostat neutrals (Fig. 4b and Supplementary Fig. 1, (*B*) *vs.* (*C*), dark red vs. light blue) in all three comparison sets by all methods except PANTHER[19]. Note that PANTHER only returned predictions for 30% of all variants; no predictions were made for neutrals in the *stringent* or *complete* sets, and only 8 of 26 neutrals had predictions in the *extended* set. This makes PANTHER an outlier to most trends observed for all other methods. Some of the algorithms also correctly and significantly differentiated rheostat non-neutrals from toggle neutrals (Fig. 4b and Supplementary Fig. 1, (*A*) *vs.* (*D*), light red vs. dark blue). However, all of the continuous prediction methods erroneously assigned higher scores to toggle neutrals than to rheostat non-neutrals – the opposite of the correct prediction. Furthermore, for the larger *extended* set, the distinction between rheostat non-neutrals and toggle neutrals was eliminated (Supplementary Fig. 3); this finding, again, likely indicates the influence of the low number of toggle neutrals in the *stringent* set.

Finally, every method scored toggle neutrals (Fig. 4b and Supplementary Fig. 1, dark blue), on average, above the neutrality threshold and higher than rheostat neutrals (light blue), which on average scored below the neutrality threshold. This difference was significant for 12 of 16 methods. This trend was maintained for the larger comparison sets, suggesting that this issue is inherent to the prediction methods.

Weak correlation between predictions and experimental outcomes at rheostat positions. The main goal of variant-effect prediction is to classify functional outcomes into binary neutral/non-neutral categories. Nevertheless, several predictors calculate continuous prediction scores, allowing us to correlate predictions with experimental effects for individual variants.

For rheostat positions, this would be a more valuable prediction, since the thresholds for biological effects can change with environment (*e.g.*, in response to changes in other proteins or in cellular conditions)[20,21]. Thus, we examined whether the progressive nature of variant outcomes at rheostat positions was captured by the continuous score prediction methods for the *rheostat_9* and *rheostat_12* sets.

For the *rheostat_9* set, only 50% of the continuous prediction methods showed any correlation (Fig. 5; Supplementary Fig. 4). Moreover, only four of the sixteen methods (SNAP2, PROVEAN, MutPred2, and PolyPhen-2) showed statistically-significant differentiation of rheostat neutrals from non-neutrals in the *rheostat_9* or *rheostat_12* sets. Of these, SNAP2 exhibited the highest correlation (Pearson's $r = 0.58$, *rheostat_9*; Fig. 5). Using the *rheostat_12* set did not alter the observed trends (Supplementary Table 5).

Predictions for variants at *neutral positions* have unclear outcomes. The LacI/GalR experimental study carried out experiments at the same positions in multiple homologs[11]. When outcomes were compared, each rheostat position showed a rheostatic profile in ~80% of the homologs studied. Thus, rheostat behavior

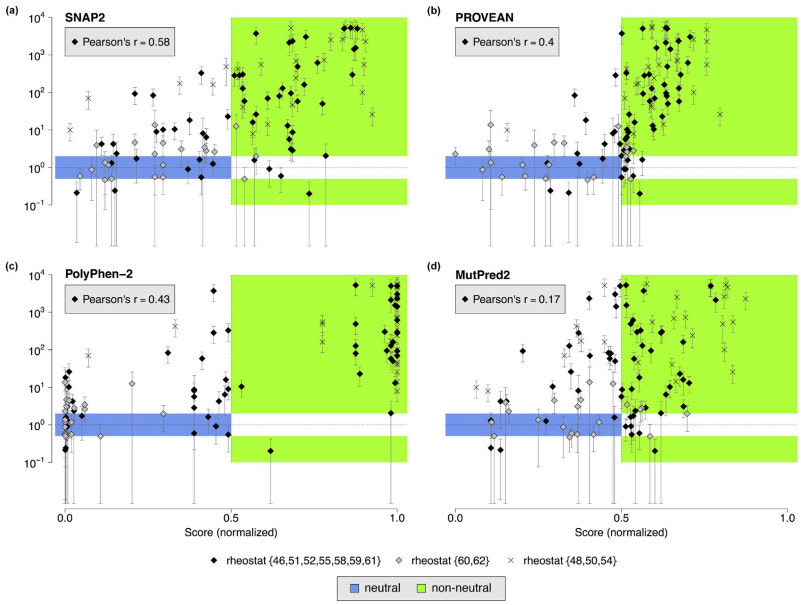

Figure 5. Correlation between experimentally measured fold-changes and predicted variant-effect scores. Panels (a) SNAP2; (b) PROVEAN; (c) MutPred2; (d) PolyPhen-2 show the relationship of the computationally and experimentally derived scores. For each variant at all rheostat positions, fold-change in repression relative to wild-type LacI is shown on log scale (Y axis), whereas predicted scores are normalized to the linear range [0, 1] (X axis). The blue area depicts the scores expected for neutral variants (fold-change between 0.5 and 2.0); the green area depicts scores expected for non-neutral variants. The Pearson product-moment correlation coefficient (Pearson's r) is given for the *rheostat_9* set. Results from other predictors are in Supplementary Fig. 4.

appears to be the baseline for these positions within the protein family, even though individual homologs may evolve the position to different roles. Positions 60 and 62 were among those that appear able to evolve new roles: These positions acted as rheostats in most of the homologs, but in LacI were neither toggle nor rheostat in either dimer or the tetramer datasets[11,14]. Instead, these were *neutral positions* – no variant had a significant effect on repression (in the absence of inducer). Note the distinction between a neutral position and a neutral substitution. Rheostat positions comprise both neutral and non-neutral substitutions, as long as the number of non-neutral variants greatly exceeds the number of neutral variants. Neutral positions comprise only neutral substitutions.

Here, most variants at neutral positions 60 and 62 were indeed predicted to be neutral (Fig. 5, set *rheostat* {*60, 62*}). However, the apparent success may be spurious for two reasons: First, our other current results show that all variants at rheostat positions generally score lower than those at toggle positions, and the results for positions 60/62 could just be a manifestation of that. Second, one of the correctly predicted neutral variants at position 62 was not inducible in experiments. In this work, we generally did not consider the variant effect on LacI induction, because for most variants, significant changes were not experimentally observed (Supplementary Table 2). However, altered response to inducer is classified as dramatic biological impact (the "+s" phenotype in the tetramer dataset)[14]. If this information were taken into consideration, the predictors would be considered to have failed for this non-inducible variant.

To resolve the performance of variant-effect predictors at neutral positions, we ideally need two additional experimental datasets: One comprising variants at positions 60/62 in other natural LacI/GalR homologs (the first study used synthetic homologs, as described in Methods below); and one comprising variants at other neutral positions in a variety of proteins.

Discussion

We previously showed experimentally that amino acid substitutions at rheostat positions have different functional outcomes than those expected for toggle positions[11]. Here, we show that these differences impair the performance of variant-effect prediction algorithms. Overall, the predictors differentiated toggle positions from rheostat positions better than neutral variants from non-neutral variants. This may be due to the fact that, in LacI, rheostat positions are not conserved, whereas most toggle positions are. Many studies have noted that conservation is a key factor in identifying neutral and non-neutral variants[22]. Moreover, the rheostats host the majority of neutral variants. If these characteristics are common to other proteins, then our results demonstrate one possible reason

for the disproportionate importance of conservation for existing tools and for the consistent lack of significant improvement of method performance.

Furthermore, we previously found that predictors disagreed in their annotation for roughly a fifth of the variants[15,23]. Here, we show that this number cannot be accounted for by mis-predictions at toggle positions: Most methods correctly predict toggle non-neutrals, and toggle neutrals are few in number, comprising a small percentage of the total predictions. Thus, mis-predicted rheostat variants are, arguably, a key factor that differentiates prediction methods.

Next, we considered the fact that threshold choice directly affects binary variant-effect predictions. Current variant-effect predictors use various input parameters to calculate a likelihood that a given, single amino acid substitution has an effect on protein function, disease relevance, structure, etc. Many methods use likelihood thresholds to further categorize substitutions as either neutral (tolerant, wild-type, benign) or non-neutral (damaging, pathogenic, deleterious). However, the methods for choosing thresholds vary widely. Some categorization thresholds are established by assessing score distributions across the training sets, and some are determined heuristically.

We further note that thresholding based on experimental training sets can be biased, since different "wet-lab" approaches use different thresholds to classify variant effects and each assay has a technical limit. For example, here we called an experimentally measured two-fold change in repression *neutral*, and everything above that threshold *non-neutral* because biological differences were detected at this level for variants in the quantitative dataset[24]. However, in the semi-quantitative tetramer dataset, everything within ~26-fold of wild-type repression was classified as neutral ("+")[14], and variants that repress *better* than wild-type (also non-neutral in our definition) were also considered to be experimentally neutral. Note that the trends of experiment to prediction comparisons were not changed with altered thresholds (Supplementary Fig. 5).

Adding to the overall confusion in the field is the fact that the "effect" definitions are often unclear; *i.e.* manuscripts detailing predictor implementations and/or performance comparisons often mix different effect terms and types of experimental data used for training/development. Thus, at best, most predictors ultimately differentiate between variants of severe effect (*e.g.* those that abolish function or obviously lead to disease) from those that are unlabeled (*e.g.* variants found in healthy populations, 1000 Genomes[25] or EXAC)[26] or poorly labeled[27,28] (*e.g.* UniProt[29] polymorphisms or variants between orthologous sequences).

Here, we showed that thresholding hides the fact that predictors behave differently for rheostat and toggle positions, *i.e.* they generate different score distributions for the two classes. The nearly identical overall performance of all methods – regardless of the effect to be predicted (disease, functional significance, structure, *etc.*) – reveals that current prediction methods are, in essence, trained to differentiate toggle non-neutrals from rheostat neutrals, rather than for the general differentiation of neutral variants from non-neutral ones. This could explain why all predictors appear to have reached an upper threshold in performance.

However, within the rheostat and toggle sets, several methods were able to differentiate neutrals from non-neutrals (Fig. 4b, Supplementary Figs 1b, 2b and 3b). Thus, our results suggest that, if we could reliably label sequence positions as toggle or rheostat, predictions could be improved by using different neutral/non-neutral thresholds for each position class. In particular, this would circumvent the problems that all methods failed to recognize rheostat non-neutrals as having more effect than toggle neutrals and that toggle neutrals scored higher than rheostat neutrals – a difference that is not biologically feasible, as all neutrals are by definition equivalent.

Finally, we note that binary classifications are insufficient to capture variant effects at rheostat positions. Our results show that thresholding prediction scores into binary classes obscures the progressive effects seen for variants at rheostat positions. Nevertheless, the progressive changes are often biologically significant. For the LacI/GalR homologs, progressive functional outcomes translated into progressive and significant changes on bacterial growth rates[24]. Nor can small changes, classified as neutral, be always disregarded: (i) Less than two-fold differences in the function of the tetracycline resistance protein were biologically-adaptive to bacteria in clinically-relevant conditions[30]. (ii) Neutral substitutions in DNA methyl-transferase were deleterious to a host organism under some conditions[31]. (iii) The wide range of normal human phenotypes appears to arise through combinations of weakly non-neutral protein variants[27]. These observations make a compelling argument for building variant-effect predictors that determine a range of outcomes.

For some current algorithms, the magnitude of the prediction scores correlated with the size of the effects. We illustrated this behavior earlier for our method, SNAP[15,23,27] and, additionally here, for SNAP2. Further, PROVEAN and PolyPhen-2 showed significant, though weak, correlations for non-neutral variants at rheostat positions. Note that none of the methods was explicitly trained to recognize the severity of effects. Instead, they were trained to differentiate binary effects: neutral from non-neutral. Thus, their prediction scores are indications of a statistical likelihood that a variant of a particular effect will occupy a particular scoring space. Indeed, high impact substitutions at toggle positions, which make up a disproportionately large fraction of the available training sets[7], were predicted with higher statistical likelihood; *i.e.* toggle non-neutrals score distributions were more dense and significantly higher than rheostat non-neutral scores. In contrast, a statistical likelihood of a variant occupying a neutral scoring space has no equivalent meaning in biology – neutral (no effect) variants cannot be less or more neutral.

We conclude with the acknowledgement that change in protein functionality is not consistently predictive of disease. Regardless of the accuracy of any particular prediction, annotating outcomes is just the first step in a series of inquiries that must be made when trying to map pathogenicity. Each protein must ultimately be considered in the context of its biological role. For example, an interacting protein can change to offset a pathogenic variant to restore normal function[32]. Moreover, functionally deficient proteins may cause disease in some contexts yet protect in others. For example, the variant in hemoglobin that leads to sickle cell anemia in homozygotic humans is protective from malaria for heterozygotes[33]. Thus, variant pathogenicity predictor scores are but one step in modeling the specific mechanisms of disease. Nevertheless, to provide a reliable foundation

for quantitative models that predict changes in larger biological systems, we must build consistently-reliable variant-effect predictors.

Methods

Experimental characteristics of rheostat positions. In our earlier work[11], we used a dimeric version of the *E. coli* LacI repressor and synthetic versions of seven other LacI/GalR homologs as hosts for amino acid substitutions at twelve non-conserved positions. All of these positions were located within the linker region that joins the N-terminal DNA binding domain and the allosteric regulatory domain (Fig. 2; PDB 1EFA[34]; positions 46–62). These positions were experimentally shown to be rheostats: At each position, the progressive functional effects of multiple amino acids substitutions were quantified by determining ability to repress transcription of a reporter gene *in vivo*. These studies were extensively validated: For all variants, we confirmed that the protein was expressed at comparable levels, folded, and capable of binding DNA[11,35]. We also benchmarked the *in vivo* repression data against *in vitro* biophysical measurements of protein-DNA interactions; in most cases, the repression changes resulted from altered K_d for DNA binding[36–38]. Finally, we determined the impact of altered repression on bacterial growth rates to show that the changes were biologically significant[24].

For the current work, we only used experimental results for *E. coli* LacI (UniProt accession number: P03023). While the synthetic LacI/GalR homologs were critical for disproving the assumptions discussed in the introduction about amino acid interchangeability[11], these proteins were not naturally evolved and we excluded them to avoid the possibility that they are not properly evaluated by available computational techniques. We also considered whether amino acid substitutions in dimeric LacI had equivalent outcomes in wild-type, tetrameric LacI. The latter is a dimer of dimers, with each dimer serving as a functional unit capable of binding the DNA operator and inducer molecules[39]. Dimeric LacI was created by truncating the C-terminal tetramerization domain[40]. Aside from its lessened ability to simultaneously bind and "loop" two DNA operators[41], dimeric LacI is extremely similar to tetrameric LacI[40,42]. Dimeric LacI was chosen in the 2013 study[11] so that substitution outcomes could be directly compared with the synthetic homologs, all of which lack a tetramerization domain. For nine of the twelve LacI rheostat positions, the quantitative substitution outcomes experimentally measured for dimeric variants[11] were in strong agreement with the semi-quantitative *in vivo* measurements previously made for the tetrameric LacI[14] (Fig. 1, right panel) and with *in vitro* measurements of LacI/DNA variant binding affinities[38].

Disagreements between the dimeric and tetrameric datasets were only observed for positions 48, 50, and 54. These three positions showed toggle behavior in the tetrameric study, *i.e.* most substitutions abolished function[14], and rheostat outcomes in the dimeric study[11]. This difference is *opposite* any artifacts expected from truncating the dimerization domain: The tetramerization domain enhances LacI stability relative to the dimer[43] and tetramer looping enhances repression[44]; either outcome would enhance DNA binding and repression, concealing diminished function of the variants. Thus, we propose the differences between the datasets are due to very low (or zero) LacI protein expression in the tetramer study, which relied upon suppression of amber codons in mutated bacterial strains to create the protein variants[45]. The latter can be an inefficient process that obscures the true outcome of the protein variation. Since the tetramer data are widely used to benchmark computational predictions, developers should be aware that this is a potential experimental bias of this dataset.

For this work, analyses were carried out in parallel using two sets of rheostat positions: all twelve identified in our 2013 study[11], and the nine that showed agreement between the dimer and tetramer forms of LacI. To determine the categories of neutral/non-neutral, we used the two-fold technical limit of the quantitative repression assay. Thus, variants exhibiting fold-changes in the range of [0.5, 2] relative to wild-type were assigned to the category of *rheostat neutrals*. All other variants were designated as *rheostat non-neutrals*. The fold change used for determining rheostatic behavior was calculated relative to the wild-type repression of 0.124 Miller units. The *rheostat_12* set comprised a total of 103 variants across 12 positions, of which 18 (17%) were neutral and 85 were non-neutral. The *rheostat_9* set comprised 78 variants across 9 positions, of which 18 (23%) were neutral and 60 were non-neutral.

For this study, we considered only the functional impact on repression in the absence of allosteric inducer, since most variants did not show significant changes in allosteric response. For each variant, we re-cast repression in the absence of inducer as fold-change with respect to repression by dimeric wild-type LacI, using equation (1) (Supplementary Table 1):

$$Fold.change\,(AxB) = \frac{(AxB)_{measured}}{A_{measured}} \tag{1}$$

where (AxB) stands for a substitution of amino acid A by amino acid B at position x and having the amino acid A corresponds to the wild-type protein. Experimental errors associated with the wild-type and variant functional data (standard deviations from the average of ~8 technical and biological replicates) were propagated using equation (2):

$$Error\,(AxB) = \sqrt{\left(\frac{\sigma^2_{(AxB)_{measured}}}{(AxB)^2_{measured}} + \frac{\sigma^2_{(A)_{measured}}}{(A)^2_{measured}}\right) * Fold.change\,(AxB)^2} \tag{2}$$

where $\boldsymbol{\sigma}$ is the standard deviation. This equation is derived for correlated variables. For uncorrelated variables, covariance terms present in the original formula equal zero. Solving that formula for the non-squared ratio of two variables results in equation (2).

Selection and characteristics of toggle positions. Our 2013 study focused on non-conserved positions, and none of the tested LacI positions showed toggle behavior. Thus, to obtain a set of toggle positions for this study, we used the semi-quantitative variant data for tetrameric LacI generated by the Miller lab[14]. In that work, positions 2–329 were substituted with 12–13 amino acids each (Supplementary Table 2) and functional outcomes were broadly grouped into several phenotypes, including those with severe effects (non-neutral) and those with no effects (neutral; Supplementary Table 6). In this work, we defined toggle positions as those with at least 8 severe variants and at most 2 neutral variants. This definition identified 53 toggle positions distributed over the LacI protein. As noted above, rheostat positions 48, 50 and 54 fell in this set of toggles positions. As noted above, we hypothesized that these results are due to extremely low protein concentrations, perhaps related to inefficient suppression of the amber codon, and we thus excluded these positions to yield the final *toggle_50* set (618 variants, 26 neutral and 592 non-neutral).

To create a toggle set of comparable size to the rheostat set, we selected a subset of toggle positions based on the availability of neutral variants and structural proximity to rheostat positions: First, we selected all positions within the DNA-binding domain with at least one neutral variant (10, 13, 30, 49). Second, toggle positions 47, 49, 53, 56, and 57 were chosen because they interdigitate with rheostat positions in the linker region (46–62). Of the remaining 13 toggles within the DNA-binding domain, we selected positions 16, 18, 21, and 22. The resulting *toggle_12* set comprised 145 variants across 12 positions, of which 4 (2.8%) were neutral. We also designated the *toggle_9* set (*toggle_12* minus variants at positions 16, 21 and 22), which comprised 109 variants (4 *toggle neutrals* and 105 *non-neutrals*) across 9 positions.

In Results, we labeled the *rheostat_9 vs. toggle_9* comparison as the *stringent* set (see Fig. 4, Supplementary Fig. 1). For a more comprehensive analysis, we also computed results for *complete* (Supplementary Fig. 2) and *extended* (Supplementary Fig. 3) sets; which included the *rheostat_12* set and, respectively, the *toggle_12* or the full *toggle_50* set (Supplementary Table 2). By using the three comparison sets, we hoped to minimize the impact of mis-assigned toggle positions due to poor protein expression.

Variant-effect prediction algorithms. To predict variant effects at rheostat and toggle positions, we used 16 publicly available computational methods (Fig. 3, Supplementary Table 7). These were selected to cover a wide variety of computational techniques and training sets. Note that not all publications explicitly mention what is meant by the word "effect." Some predict disease variants, others focus on evolutionary conservation or evolutionary fitness, and still others evaluate functional or structural impacts. To be able to use all tools, here we broadly use the term *outcome* without further identifying differences between methods. For all methods that require a 3D structure, we used 1EFA (PDB ID: 1EFA)[34], consisting of dimeric LacI bound to DNA operator and an "anti-inducer" allosteric ligand. All reported graphs/statistics were generated using R[46].

1. SIFT[22] uses PSI-BLAST[47] (position-specific iterated Basic Local Alignment Search Tool; BLAST) to query a sequence database (*e.g.* NCBInr)[48] and generates a position-specific scoring matrix (PSSM) based on the retrieved sequences. Note that we used SIFT with a manually curated MSA (Supplementary File 1)[49] rendering the initial PSI-BLAST query obsolete. Combined with known generic likelihoods of amino acid substitutions (BLOSUM62 substitution matrix), this approach allows estimating probabilities of effect for any position-specific amino acid substitution.
2. PROVEAN[16] uses BLAST[50] to collect homologous and distantly related sequences from NCBInr, which are then clustered by sequence identity. To measure the effect of a variation, the algorithm calculates the average divergence score between the cluster sequences and the query sequence using the BLOSUM62 substitution matrix.
3. PANTHER[19] uses BLAST to identify the best match to the input query in the PANTHER database[51]. This matched protein is linked to a pre-computed phylogenetic tree of the specific protein family. To evaluate the effect of the substitution, the mutated residue is traced back through increasingly older ancestral proteins in this tree.
4. MutationAssessor[52] uses BLAST to query the UniProt and identify related protein sub-families, which are used to extract characteristic conservation patterns. The latter are used to calculate the effect of mutating a specific residue in a protein family and, separately, in each of its sub-families.
5. FoldX (PositionScan)[53] uses protein 3D structures and an empirical force field to evaluate the effects (free energy changes) due to variation.
6. Align-GVGD[54] is an extension of the Grantham Difference[55], combining a conservation score based on a given MSA (Grantham Variation) with a measure of the biochemical difference between the mutant and the wild-type amino acids (Grantham Deviation).
7. MAPP[56] constructs a phylogenetic tree based on substitution frequency per site within an MSA of orthologs or closely related paralogs. Topology and branch lengths of the tree are used to calculate weights for each sequence respectively. These weights are used to generate an alignment summary that is interpreted using a matrix of physiochemical properties resulting in an estimate of the physiochemical constraints on each position of the MSA. Deviations from these constraints are calculated for each position of the query sequence and transformed into an effect prediction score.
8. PolyPhen-1[57] classifies variants via empirically derived rules using various sequence-based characteristics of the substitution site (*e.g.* UniProt annotations), along with structure and homology descriptors.
9. PolyPhen-2[17] trains a Naïve Bayes classifier on HumDiv (set of single amino acid substitutions from UniProt known to cause human Mendelian diseases and non-damaging variants found in closely related mammalian homologs) and HumVar[58] (set of disease- and neutral polymorphism-annotated single amino acid substitutions of human proteins from Swiss-Prot)[29] variants. PolyPhen-2 uses structure-based

(*e.g.* accessible surface area and conformational mobility of the wild-type amino acid residue) and sequence-based features (*e.g.* MSA-based conservation and depth, CpG context and residue volume).

10. SNAP2[15] uses an artificial neural network, trained on experimentally-obtained variant functional effect data, with a variety of precomputed biochemical and evolutionary amino acid substitution rules, as well as conservation and predicted sequence-derived protein features, *e.g.* secondary structure and solvent accessibility.
11. PhD-SNP[58] applies a decision tree to predict effects either using a profile based support vector machine (SVM; sequence profile calculated by BLAST against UniRef90) or by a single sequence based SVM. Training data are extracted from Swiss-Prot (disease *vs.* polymorphism variants) and enriched with OMIM[59] annotations.
12 iMutant3[58,60] offers different SVMs to predict (i) stability changes, using SVMs trained on the ProTherm database[61], from sequence only or from structural information, and (ii) disease associated variants from sequence only using the PhD-SNP prediction pipeline.
13. nsSNPAnalyzer[62] uses a random forest classifier, trained on a curated dataset of variants (ModSNP)[63] using (i) variant structural environment, (ii) position conservation within the MSA, and (iii) similarity between variant and original amino acid. If no structural information is provided, the ASTRAL database[64] is queried for a homolog structure.
14. PredictSNP[65] is a meta-predictor incorporating eight methods (MAPP, nsSNPAnalyzer, PANTHER, PhD-SNP, PolyPhen-1/-2, SIFT, SNAP2) into a consensus classifier based on a majority vote weighted by the method-specific confidence scores. PredictSNP is trained on a benchmark dataset compiled from five different sources (training datasets of four variant-effect prediction tools not selected for the PredictSNP pipeline: SNPs&GO, MutPred[18], PON-P and HumVar; the fifth source is a subset of UniProt variants). Testing datasets are derived from the Protein Mutant Database[66] (PMD) and from experimental studies.
15. Meta-SNP[67] is a random forest-based binary classifier meta-predictor, combining the predictions of four methods (SNAP2, SIFT, PANTHER, PhD-SNP) and four features extracted from the PhD-SNP protein sequence profile; the training dataset is derived from Swiss-Prot (disease *vs.* polymorphism variants).
16. MutPred2 (unpublished; the successor of MutPred)[18] consists of an ensemble of bagged neural networks, trained on amino acid substitutions from HGMD[68], Swiss-Prot, dbSNP[69], and ortholog alignments. In addition to sequence, conservation, and physicochemical features in and around the variant position, MutPred2 uses predictions of change due to amino acid variation in over 50 local structural and functional properties (*e.g.* post-translational modification sites, macromolecular binding, among others).

For all variants in the LacI rheostat and toggle sets, predictions were generated using the 16 algorithms listed above. When no publicly available web-service was present, prediction methods were installed and run locally. Input parameters were set to default values. To obtain comparable predictions between the different algorithms, all predictor scores were transformed and normalized: Some tools provide a probability or some other score for the likelihood of variant non-neutrality. For these, we converted pre-defined, method-specific binary thresholds to a value of 0.5 and normalized the neutral and non-neutral score ranges separately to [0, 0.5] and (0.5, 1], respectively. For methods that predict the classes of functional outcomes, scores were assigned manually. Details of scoring and thresholds used for normalization are as follows:

1. SNAP2 scores are [−100, 100], threshold at 0, neutrals below threshold.
2. SIFT scores are [0, 1], threshold at 0.05, neutrals above threshold. Scores were reversed (equation (3) prior to normalization.

$$Score_{reversed} = 1 - Score_{raw} \tag{3}$$

3. MutationAssessor score range was not defined by the authors, but available data[52] suggests a [−4, 5] range, which we use in normalization; default threshold is 1.9, neutrals below threshold. Note, however, in this work we used a threshold of 0.8, as described in ref. 16, to more accurately differentiate neutrals.
4. PROVEAN score range was not defined by the authors, but predicted scores for our variants occurred in the [−14.875, 1.908] range, which we use in normalization; threshold at −2.5, neutrals above threshold.
5. MAPP scores are [0, 1], threshold at 0.5, neutrals above threshold. Scores were reversed using equation (3).
6. iMutant3 score range was not defined by the authors, but predicted scores for our variants occurred in the [−3.5, 0.63] range, which we use in normalization; threshold at −0.5, neutrals above threshold. Note that we used this threshold to transform predictions into a binary form.
7. FoldX (PositionScan) score range is not defined by the authors, but predictions for free energy changes below 0.05 kcal/mol (neutrals) are not reported. The maximum predicted score for our variants was 3.76102. Here we assigned a score of 0 to missing predictions, resulting in [0, 3.76102] range, and set the threshold at 0.5 (as in iMutant, above), neutrals below the threshold. Note that we used this threshold (as in iMutant, above), to transform predictions into a binary form.
8. PolyPhen-1 classifications of [benign, possibly damaging, probably damaging], were converted to [0, 0.5, 1]
9. Align-GVGD classifications are [C0, C15, C25, C35, C45, C55, C65], assigned to corresponding risk estimates, ranging [1.16, 3.12][70]. The authors did not define the threshold, but C0 was suggested to be the only neutral class. Thus, threshold was set at the corresponding risk estimate of 1.16, neutrals below threshold. Note that we used this threshold to transform predictions into a binary form.
10. PolyPhen-2 scores are [0, 1], threshold at 0.92, neutrals below threshold.

11-14. MutPred2, PANTHER, PhD-SNP, and Meta-SNP obtain probability scores, which range [0, 1] with a threshold at 0.5, neutrals below threshold. Note that scores for PANTHER and PhD-SNP reported

here were each obtained from meta-predictors (PredictSNP and Meta-SNP, respectively).
15,16. nsSNPAnalyzer and PredictSNP are binary classifiers [neutral, non-neutral], which were converted to [0, 1].

To analyze the performance of prediction tools in differentiating rheostat and toggle position non-neutrals and neutrals, we applied the Kolmogorov-Smirnov (KS) test (two-sided) for continuous predictors and Fisher exact T-test (two-sided) for binary predictors. To compare the correlation between experimentally-measured fold-changes for rheostat variants and predicted variant-effect scores, we computed the Pearson product-moment correlation coefficient (Pearson's r). For this analysis, we made two assumptions: (i) In addition to the variants with diminished repression, five variants showed enhanced repression relative to wild-type (fold-change values less than 0.5) and thus were treated as non-neutral. We used the reciprocal of the fold-change to correlate them to prediction scores. (ii) Neutral variants are, by definition, all equivalent to wild-type. However, the continuous predictors usually assign a range of scores that fall below their neutrality threshold (here normalized to 0.5). Thus, we assigned all neutrally predicted variants a score of 0.5.

References

1. Bruse, S. *et al.* Whole exome sequencing identifies novel candidate genes that modify chronic obstructive pulmonary disease susceptibility. *Hum Genomics* **10,** 1, doi: 10.1186/s40246-015-0058-7 (2016).
2. Ellinghaus, D. *et al.* Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* **145,** 339–347, doi: 10.1053/j.gastro.2013.04.040 (2013).
3. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am J Hum Genet* **98,** 58–74, doi: 10.1016/j.ajhg.2015.11.023 (2016).
4. Bromberg, Y. Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol* **425,** 3993–4005, doi: 10.1016/j.jmb.2013.07.038 (2013).
5. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24,** 2125–2137, doi: 10.1093/hmg/ddu733 (2015).
6. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89,** 10915–10919 (1992).
7. Gray, V. E., Kukurba, K. R. & Kumar, S. Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics* **28,** 2093–2096, doi: 10.1093/bioinformatics/bts336 (2012).
8. Swint-Kruse, L., Larson, C., Pettitt, B. M. & Matthews, K. S. Fine-tuning function: correlation of hinge domain interactions with functional distinctions between LacI and PurR. *Protein Sci* **11,** 778–794, doi: 10.1110/ps.4050102 (2002).
9. Pendergrass, D. C., Williams, R., Blair, J. B. & Fenton, A. W. Mining for allosteric information: natural mutations and positional sequence conservation in pyruvate kinase. *IUBMB Life* **58,** 31–38, doi: 10.1080/15216540500531705 (2006).
10. de Beer, T. A. *et al.* Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput Biol* **9,** e1003382, doi: 10.1371/journal.pcbi.1003382 (2013).
11. Meinhardt, S., Manley, M. W. Jr., Parente, D. J. & Swint-Kruse, L. Rheostats and toggle switches for modulating protein function. *PLoS One* **8,** e83502, doi: 10.1371/journal.pone.0083502 (2013).
12. Ishwar, A., Tang, Q. & Fenton, A. W. Distinguishing the interactions in the fructose 1,6-bisphosphate binding site of human liver pyruvate kinase that contribute to allostery. *Biochemistry* **54,** 1516–1524, doi: 10.1021/bi501426w (2015).
13. Weaver, Y. M. & Hagenbuch, B. Several conserved positively charged amino acids in OATP1B1 are involved in binding or translocation of different substrates. *J Membr Biol* **236,** 279–290, doi: 10.1007/s00232-010-9300-3 (2010).
14. Suckow, J. *et al.* Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *J Mol Biol* **261,** 509–523, doi: 10.1006/jmbi.1996.0479 (1996).
15. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16** Suppl 8, S1, doi: 10.1186/1471-2164-16-S8-S1 (2015).
16. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7,** e46688, doi: 10.1371/journal.pone.0046688 (2012).
17. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7,** 248–249, doi: 10.1038/nmeth0410-248 (2010).
18. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25,** 2744–2750, doi: 10.1093/bioinformatics/btp528 (2009).
19. Tang, H. & Thomas, P. D. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, doi: 10.1093/bioinformatics/btw222 (2016).
20. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol* **31,** 1581–1592, doi: 10.1093/molbev/msu081 (2014).
21. Swint-Kruse, L. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys J* **111,** 10–18, doi: 10.1016/j.bpj.2016.05.030 (2016).
22. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res* **11,** 863–874, doi: 10.1101/gr.176601 (2001).
23. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35,** 3823–3835, doi: 10.1093/nar/gkm238 (2007).
24. Meinhardt, S. *et al.* Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression. *Nucleic Acids Res* **40,** 11139–11154, doi: 10.1093/nar/gks806 (2012).
25. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74, doi: 10.1038/nature15393 (2015).
26. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291, doi: 10.1038/nature19057 (2016).
27. Bromberg, Y., Kahn, P. C. & Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci USA* **110,** 14255–14260, doi: 10.1073/pnas.1216613110 (2013).
28. Rost, B., Radivojac, P. & Bromberg, Y. Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett* **590,** 2327–2341, doi: 10.1002/1873-3468.12307 (2016).
29. UniProt, C. UniProt: a hub for protein information. *Nucleic Acids Res* **43,** D204–212, doi: 10.1093/nar/gku989 (2015).
30. Walkiewicz, K. *et al.* Small changes in enzyme function can lead to surprisingly large fitness effects during adaptive evolution of antibiotic resistance. *Proc Natl Acad Sci USA* **109,** 21408–21413, doi: 10.1073/pnas.1209335110 (2012).
31. Rockah-Shmuel, L., Toth-Petroczy, A. & Tawfik, D. S. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput Biol* **11,** e1004421, doi: 10.1371/journal.pcbi.1004421 (2015).
32. Urano, D., Dong, T., Bennetzen, J. L. & Jones, A. M. Adaptive evolution of signaling partners. *Mol Biol Evol* **32,** 998–1007, doi: 10.1093/molbev/msu404 (2015).
33. Williams, T. N. Human red blood cell polymorphisms and malaria. *Curr Opin Microbiol* **9,** 388–394, doi: 10.1016/j.mib.2006.06.009 (2006).

34. Bell, C. E. & Lewis, M. A closer view of the conformation of the Lac repressor bound to operator. *Nat Struct Biol* **7,** 209–214, doi: 10.1038/73317 (2000).
35. Meinhardt, S. & Swint-Kruse, L. Experimental identification of specificity determinants in the domain linker of a LacI/GalR protein: bioinformatics-based predictions generate true positives and false negatives. *Proteins* **73,** 941–957, doi: 10.1002/prot.22121 (2008).
36. Tungtur, S., Skinner, H., Zhan, H., Swint-Kruse, L. & Beckett, D. *In vivo* tests of thermodynamic models of transcription repressor function. *Biophys Chem* **159,** 142–151, doi: 10.1016/j.bpc.2011.06.005 (2011).
37. Zhan, H., Taraban, M., Trewhella, J. & Swint-Kruse, L. Subdividing repressor function: DNA binding affinity, selectivity, and allostery can be altered by amino acid substitution of nonconserved residues in a LacI/GalR homologue. *Biochemistry* **47,** 8058–8069, doi: 10.1021/bi800443k (2008).
38. Zhan, H., Swint-Kruse, L. & Matthews, K. S. Extrinsic interactions dominate helical propensity in coupled binding and folding of the lactose repressor protein hinge helix. *Biochemistry* **45,** 5896–5906, doi: 10.1021/bi052619p (2006).
39. Lewis, M. *et al.* Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science* **271,** 1247–1254 (1996).
40. Chen, J. & Matthews, K. S. Subunit dissociation affects DNA binding in a dimeric lac repressor produced by C-terminal deletion. *Biochemistry* **33,** 8728–8735 (1994).
41. Muller, J., Barker, A., Oehler, S. & Muller-Hill, B. Dimeric lac repressors exhibit phase-dependent co-operativity. *J Mol Biol* **284,** 851–857, doi: 10.1006/jmbi.1998.2253 (1998).
42. Chen, J. & Matthews, K. S. Deletion of lactose repressor carboxyl-terminal domain affects tetramer formation. *J Biol Chem* **267,** 13843–13850 (1992).
43. Barry, J. K. & Matthews, K. S. Thermodynamic analysis of unfolding and dissociation in lactose repressor protein. *Biochemistry* **38,** 6520–6528, doi: 10.1021/bi9900727 (1999).
44. Oehler, S., Eismann, E. R., Kramer, H. & Muller-Hill, B. The three operators of the lac operon cooperate in repression. *EMBO J* **9,** 973–979 (1990).
45. Markiewicz, P., Kleina, L. G., Cruz, C., Ehret, S. & Miller, J. H. 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* **240,** 421–433, doi: 10.1006/jmbi.1994.1458 (1994).
46. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2015).
47. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25,** 3389–3402 (1997).
48. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42,** D756–763, doi: 10.1093/nar/gkt1114 (2014).
49. Tungtur, S., Parente, D. J. & Swint-Kruse, L. Functionally important positions can comprise the majority of a protein's architecture. *Proteins* **79,** 1589–1608, doi: 10.1002/prot.22985 (2011).
50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410, doi: 10.1016/S0022-2836(05)80360-2 (1990).
51. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res* **41,** D377–386, doi: 10.1093/nar/gks1118 (2013).
52. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39,** e118, doi: 10.1093/nar/gkr407 (2011).
53. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* **33,** W382–388, doi: 10.1093/nar/gki387 (2005).
54. Mathe, E. *et al.* Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* **34,** 1317–1325, doi: 10.1093/nar/gkj518 (2006).
55. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).
56. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* **15,** 978–986, doi: 10.1101/gr.3804205 (2005).
57. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30,** 3894–3900 (2002).
58. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22,** 2729–2734, doi: 10.1093/bioinformatics/btl423 (2006).
59. McKusick, V. A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80,** 588–604, doi: 10.1086/514346 (2007).
60. Capriotti, E., Fariselli, P., Calabrese, R. & Casadio, R. Predicting protein stability changes from sequences using support vector machines. *Bioinformatics* **21** Suppl 2, ii54–58, doi: 10.1093/bioinformatics/bti1109 (2005).
61. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res* **32,** D120–121, doi: 10.1093/nar/gkh082 (2004).
62. Bao, L., Zhou, M. & Cui, Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res* **33,** W480–482, doi: 10.1093/nar/gki372 (2005).
63. Yip, Y. L. *et al.* The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum Mutat* **23,** 464–470, doi: 10.1002/humu.20021 (2004).
64. Chandonia, J. M. *et al.* The ASTRAL Compendium in 2004. *Nucleic Acids Res* **32,** D189–192, doi: 10.1093/nar/gkh034 (2004).
65. Bendl, J. *et al.* PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput Biol* **10,** e1003440, doi: 10.1371/journal.pcbi.1003440 (2014).
66. Kawabata, T., Ota, M. & Nishikawa, K. The Protein Mutant Database. *Nucleic Acids Res* **27,** 355–357 (1999).
67. Capriotti, E., Altman, R. B. & Bromberg, Y. Collective judgment predicts disease-associated single nucleotide variants. *BMC Genomics* **14** Suppl 3, S2, doi: 10.1186/1471-2164-14-S3-S2 (2013).
68. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* **21,** 577–581, doi: 10.1002/humu.10212 (2003).
69. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29,** 308–311 (2001).
70. Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E. & Thomas, A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Hum Mutat* **29,** 1342–1354, doi: 10.1002/humu.20896 (2008).
71. Schrödinger, L. L. C. *The PyMOL Molecular Graphics System, Version 1.8*. URL https://www.pymol.org (2015).

Acknowledgements

We thank Vikas R. Pejaver, Predrag Radivojac (Indiana University, Bloomington) and Sean Mooney (University of Washington, Seattle) for sharing their not-yet-published MutPred2 code and for helpful advice on how to implement it; Yannick Mahlich, Chengsheng Zhu, Yanran Wang (all Rutgers University) Konrad Schreiber (Ludwig-Maximilians-Universität of Munich, LMU) and Sonakshi Bhattacharjee (Technical University of Munich, TUM) for discussions and technical support, and Inga Weise (TUM) for other support. Particular thanks are due to Burkhard Rost (TUM) for his hospitality and valuable discussions. Last but not least, we thank all those

who deposit their experimental data in public databases, those who maintain these databases, and all researchers who maintain public accessibility to their variant-effect predictors. L.S.K. was supported by an internal Lied basic science grant from the KUMC Research Institute, with support from a NIH Clinical and Translational Science Award grant (UL1TR000001, formerly UL1RR033179) and by private funds. Y.B. and M.M. were supported by the NIH/NIGMS grant U01 GM115486. Y.B. was additionally supported by an Informatics Research Starter grant from the PhRMA foundation, NIH 01 GM 115486, and USDA-NIFA 1015:0228906 grants.

Author Contributions

L.S.K. and Y.B. conceived of the study; all three authors contributed to parameter definitions; M.M. carried out the analyses; and all three authors contributed to data interpretation and writing the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Miller, M. *et al.* Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci. Rep.* **7**, 41329; doi: 10.1038/srep41329 (2017).

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

2.2 fun-TRP: accurate annotation of protein position classes (in preparation)

2.2.1 Introduction

One of the main objectives of my research was to study the association between disease patterns and altered protein function caused by amino acid substitution. As mentioned before (section 1.2), currently available variant effect prediction methods generally lack the required accuracy for this endeavour. Consequently, the development of an improved variant effect prediction model was a vital aspect to achieve this objective. In section 2.1 we established a new concept of protein sequence position classes - *toggle* and *rheostat*. We demonstrated that existing computational predictors fall short on accurately differentiating between neutral and non-neutral mutations between those two classes. We concluded that for improving prediction accuracy, new models require the implementation of the position class as an additional feature. *toggle*s and *rheostat*s are characterized based on the distribution of experimentally validated variant effect scores per protein sequence position. Consequently, as experimental data is still very limited, we required a tool that assigns those class labels without requiring wet-lab work. We developed a new machine learning approach, FUNction Toggle-Rheostat Predictor (fun-TRP), to predict position classes using a manually curated set of sequence based features. This was the first step towards our main goal of establishing an improved variant effect predictor.

2.2.2 Methods

We extracted experimentally evaluated amino acid substitution effect scores from five deep mutational scanning (DMS) data sets [64, 65, 66, 67, 68]. These sets were explicitly selected to cover a wide range of species (table 2.1).

Gene	**sub-region**	**organism**	**method**	**variants**	**score**
BRCA1	RING domain	*H. sapiens*	DMS	3169	E3 ligase activity
PAB1	RRM domain	*S. cerevisiae*	DMS	1244	Ampicillin resistance
UBE4B	U-box domain	*H. sapiens*	DMS	993	E3 ligase activity
TEM-1	-	*E. coli*	DMS	5469	Ampicillin resistance
SPG1	GB1	*Streptococcus sp*	exp. assay	417	Binding affinity to IgG

Table 2.1: List of experimentally validated data sets used for training and Cross Validation of prediction models.

From all variants contained in the described data sets we removed those which were not 'SNP-possible', *i.e.* those amino acid substitutions which required more than one nucleotide to be altered with respect to the wildtype residue. An overview of the fun-TRP pipeline is depicted in figure 2.1.

We normalized the data sets to the wildtype score. Thus, neutral mutations (*i.e.* mutations which exhibited the same scores as the wildtype) were automatically assigned a score of 0. Subsequently we applied K-Means Clustering (n=3) to each of these data

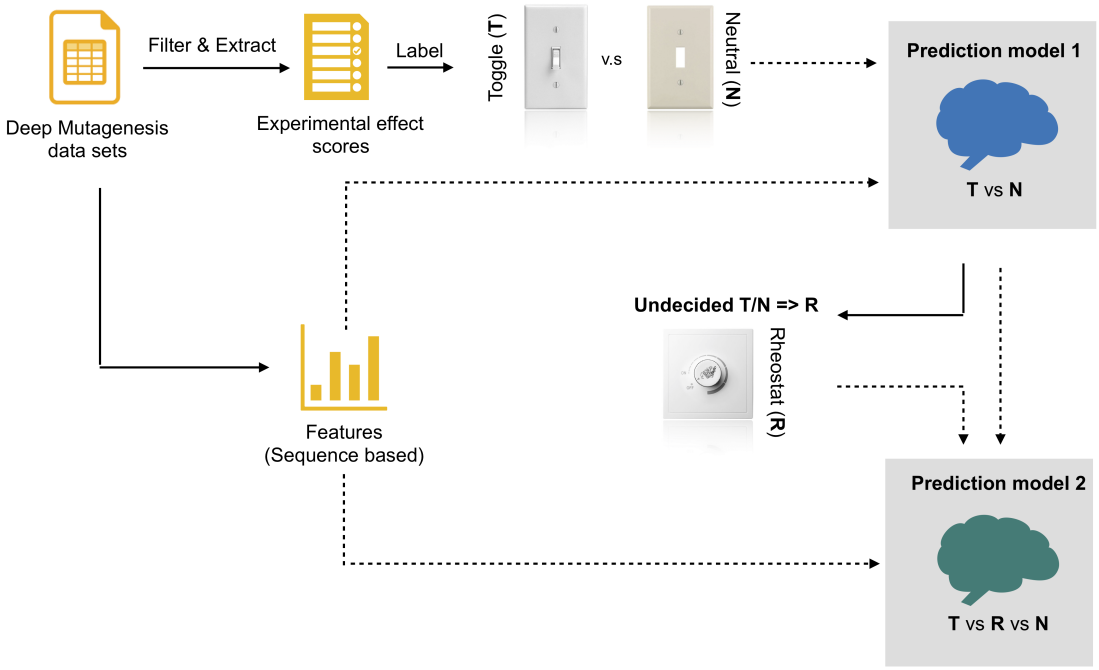

Figure 2.1: fun-TRP Workflow.

sets separately, including 0 as wildtype (*wt*) score and a dataset specific knock-out (*ko*) score (score threshold for complete loss of function). Thus each of the experimentally validated scores was assigned to one of three clusters, namely *wildtype* (containing the *wt* score), *severe* (containing the *ko* score) or *functional*. A residue was labeled as either *toggle* or *neutral* based on the clustering of associated effect scores. If, for a specific residue, the majority of experimental scores was assigned to the *wildtype* cluster and not more than one attributed to any other cluster, we labeled this residue position as *neutral*. On the other hand, if more than one experimental score was assigned to the *severe* cluster but not more than two attributed to any other cluster, we labeled this residue position as *toggle*. If none of the two scenarios held true, residue positions were labeled as *unknown*. Distributions of experimentally validated effect scores per residue for PAB1 (colored by assigned class labels) are shown in figure 2.2. *toggle* and *neutral* positions are clearly separated from each other. Positions labeled as *unknown* (not shown here), which exhibit a progressive range of effects, are clearly separated as well. Combining all five data sets, our initial set contained 820 instances, *i.e.* residues. We excluded 228 residues, which did not possess a sufficient number of experimentally measured values ($>= 6$). 66 positions were labeled as *toggle* and 153 positions as neutral. Upon manual

re-assessment we excluded six *toggle* positions and twelve neutral positions (shown in grey in figure 2.2) to generate a more conservative data set. Thus our toggle-neutral training set (*TNT*) finally comprised 60 *toggle* and 141 neutral positions. A total of 373 positions were labeled as *unknown*.

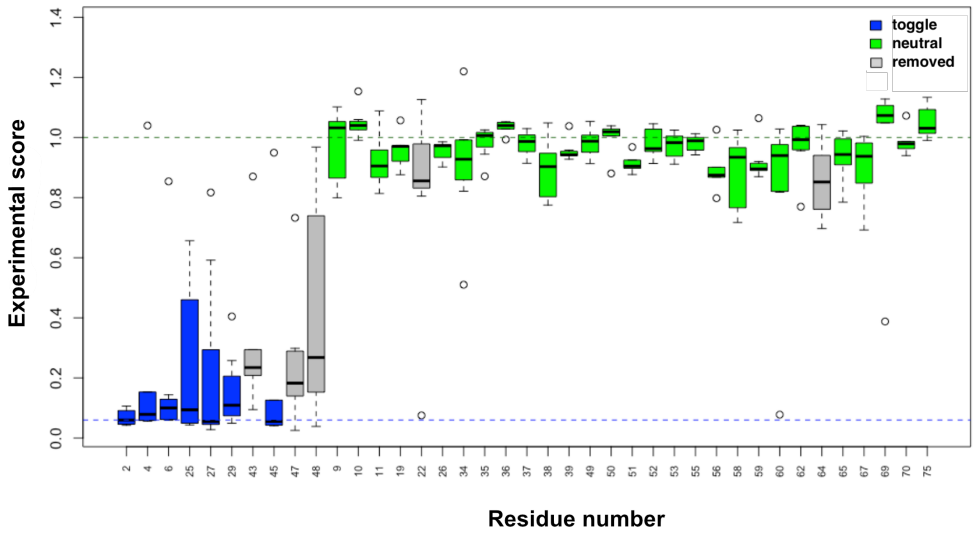

Figure 2.2: Distributions of experimentally validated effect scores per residue colored by assigned class labels for PAB1. Residues with box plots showed in grey were excluded upon manual re-assessment based on position of distribution median. Dotted lines represent data set specific thresholds for severe (blue) and neutral (green) variant effects.

Next we identified *rheostat* positions in the set of residues labeled as *unknown*. We trained a machine learning model (*model 1*) using a Random Forest (RF) classifier on the above-described *TNT* set and a manually curated set of sequence-based features (table 2.2). To account for bias in class labels we re-sampled the set and trained on a balanced set, comprised of 1000 instances. We then used this model to predict *toggle* vs. *neutral* labels for the *unknown* set of positions. Based on the resulting scores of *toggle* vs. *neutral* predictions, we defined those positions as *rheostat*s where the prediction model could not decide between *toggle* and *neutral* classes. Specifically, if the predicted score fell in the range of 0.37-0.57 (with predictions ranging from 0 (*neutral*) to 1 (*toggle*)), we considered the prediction unreliable and, as such, the position class - a *rheostat*. This resulted in 84 positions changing their label from *unknown* to *rheostat*. Together with the previous *TNT* set, our final training set consisted of 285 labeled positions.

We used our final set to train a second machine learning model (*model 2*; again using a RF classifier with re-sampling and the same set of sequence based features) to successfully predict *toggle*s, *neutral*s and *rheostat*s. We used an implementation of RF Classification available in the WEKA library [71] and R [72] for K-Means Clustering. The general workflow was implemented in Python.

id	feature	tool	description
1	PACC	PROF (*)	predicted solvent accessibility
2	pH	PROF (*)	'probability' for assigning helix
3	pE	PROF (*)	'probability' for assigning strand
4	pL	PROF (*)	'probability' for assigning neither helix, nor strand
5	consurf_score	consurf (*)	predicted conservation
6	PROFbval	PROFbval (*)	predicted residue flexibility
7	MD_raw	MD (*)	predicted protein disorder
8	msa_frequency	-	calculated MSA entropy
9	amino acid	-	amino acids encoded as a vector of length 20
10	small	-	basic amino acid property
11	polar uncharged	-	basic amino acid property
12	negatively charged / acidic	-	basic amino acid property
13	positively charged / basic	-	basic amino acid property
14	possible_snps	-	number of possible nsSNPs

Table 2.2: Set of sequence based features used by prediction model. (*) tools are applied via the PredictProtein pipeline [69]. We created a dockerized version of PredictProtein [70] which allows us to run predictions paralellized in cluster environments using *clubber* (see section 3.1).

2.2.3 Preliminary Analysis

We evaluated our models extensively via CV. We used leave-one-out Cross Validation (LOO-CV) to assess the performance for both our prediction models. Note, that we re-sampled our training set for each validation run (after removing the test instance) to account for bias in class labels. Evaluating *model 1* on the *TNT* set (see Methods; 201 instances, 60 *toggle*s & 141 *neutrals*) resulted in an overall accuracy of 90% (correctly classified *neutrals*: 133/141 and *toggle*s 48/60) and F-measure of 0.93. An identical evaluation of *model 2* resulted in an averaged accuracy of 82.1% (correctly classified *neutrals*: 125/141, *toggle*s 44/60 and *rheostat*s 65/84). We further evaluated our models by leave-one-dataset-out as well as and 10-fold CV resulting in overall comparable performances (data not shown).

We retrieved the entire set of human enzymes from the UniProt KnowledgeBase [73] and applied our fun-TRP pipeline to predict position classes for all 12,362 protein sequences. Our analysis shows the following distribution of class labels for human enzymes: *toggles* (18.4%), *neutrals* (45.5%) and *rheostats* (36.1%). Surprisingly, the distribution of residue classes shows that charged residues are most interchangeable of the entire amino acid alphabet (*neutrals*). As expected, smaller aliphatic residues can often be *rheostats* and cysteines act as *toggles* (figure 2.3). It is interesting to note that proline, a residue that is usually considered to be immutable, is sometimes still a *rheostat*, suggesting that further insight is necessary to understand its role in specific proteins. We also observe differences in distributions across broad enzyme classes; particularly, oxidoreductases show distinct patterns of toggles and rheostats, which are different from all other enzyme classes. We suspect, that due to their ancient origins, importance to organism life and function, and corresponding ubiquitous presence, oxidoreductases are likely to allow for a larger spectrum of functional tuning (more *rheostats* than *neutrals* and *toggles*) (figure 2.4).

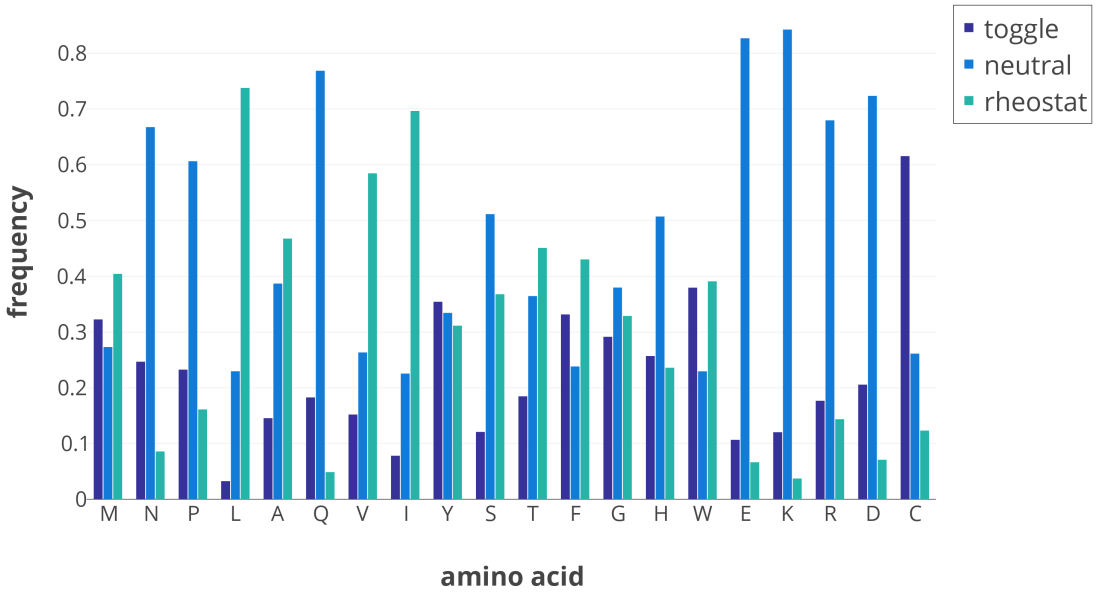

Figure 2.3: Distribution of *toggles*, *neutrals* and *rheostats* of human enzymes grouped by amino acid.

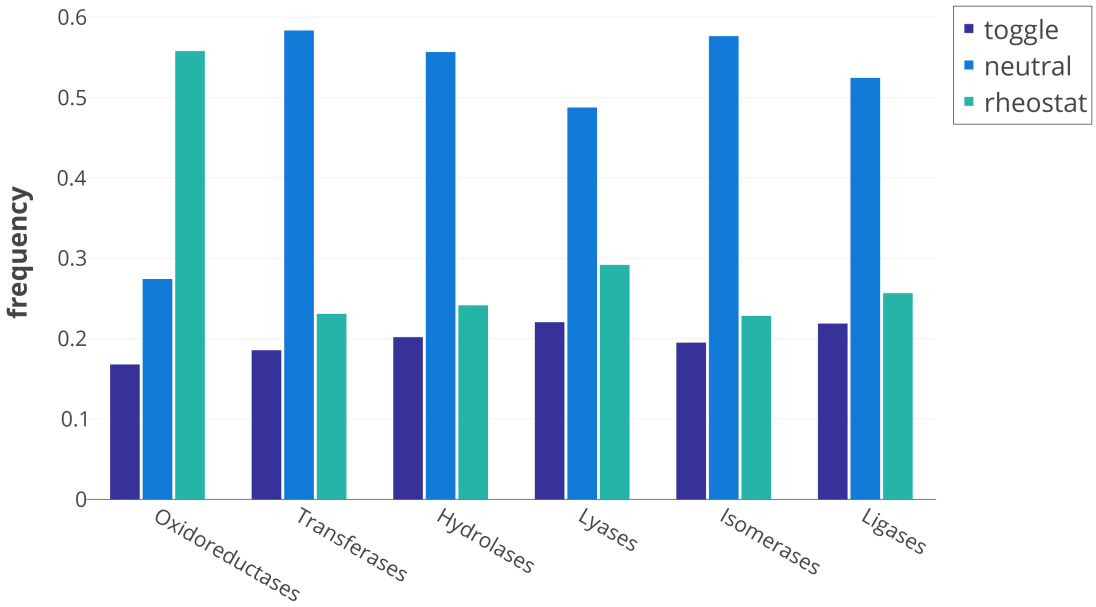

Figure 2.4: Distribution of *toggles*, *neutrals* and *rheostats* of human enzymes grouped by enzyme class.

3 Efficient Big Data analysis through load balancing

3.1 clubber: removing the bioinformatics bottleneck in big data analyses

3.1.1 Preface

In modern day bioinformatics, *Big Data* related challenges have become a recurring obstacle in the path of efficacious analysis. When designing new algorithms, efficiency has to play a major role during method development. Yet, even the most efficient tools require large resources when applied to exponentially growing biological data sets. Thus the bottleneck in biological discovery has gradually shifted from the cost of doing experiments to that of analyzing results. Access to sufficient computational resources to deal with *Big Data* is therefore essential. The methods we developed and which are described in this thesis enable the processing of a vast amount of input data. *fusion*DB maps new microbial genomes to the functional spectrum of reference bacteria. This requires thousands of Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [74] runs to compare the entire proteome of the new organism to the reference database. *mi-faser* annotates read-'parent protein' molecular functionality for an entire metagenome. Tens of millions of sequencing reads have to be translated and aligned to a reference database of full-length proteins. The fun-TRP predictor requires a set of fourteen features to be calculated, many of which involve resource intensive prediction methods. Large scale analyses like assessing the distribution of class labels in the entire set (12.362) of enzymes present in the human genome are therefore very compute intensive.

Complementing our analytical methods I developed *clubber*, an automated cluster load balancing software. Using *clubber* enabled us to extensively speed up those methods by parallelizing and distributing computations among all (heterogeneous) compute resources accessible for our group. By doing so, we had access to up to 7000 cores shared among six HPC resources at a time. Further, we avoided long queuing times on specific resources (*i.e.* due to a high workload or scheduled maintenance) as *clubber* distributes computations prioritized to resources with the least expected total processing time. The advantage of using *clubber* is particularly obvious in a scenario where only one cluster is available for computation *vs.* having two clusters – a local and one additional remote cluster. Computations are sped up by up to 100 %. Simply logging in to another cluster and submitting job subsets manually is a tedious task, which would not, even in the best case scenario, achieve a comparable speed up. Simplicity was a key feature for de-

veloping *clubber*. Using the software is as simple as downloading and installing Docker [60], a software container platform available for every environment, and using a single command to run the Docker-cloud clubber container [75]. From there, any interaction for basic configuration, job submission, monitoring and displaying results is achieved via the *clubber* web interface or the provided RESTful [76] API. For integration into larger projects we provide *clubber* as a *Python* package. Finally, *clubber* facilitates upstream and downstream processing of input data as well as results. It was designed specifically to simplify and accelerate common computational biology experimental workflows.

Using *clubber* as back-end for all our applications enabled us to offer extremely fast (web) services for the research community. Our *mi-faser* web service functionally annotates microbiomes in less then 20min per 10GB of reads. In this work we used this service to rapidly analyze the Deepwater Horizon oil-spill study data [77] (BioProject PRJNA260285; 16 samples, 73GB sequence reads). We used NMDS for visualization of the similarity between individual result vectors by mapping their similarity as a function of two-dimensional Euclidean space. We could quantitatively show that the beach sands have not yet entirely recovered. Further, our analysis of the CAMI challenge [78] (five Hiseq samples, 15 Gbp each) data revealed that microbiome taxonomic shifts do not necessarily correlate with functional shifts. These examples (21 metagenomes processed in 172 min) clearly illustrates the impact of clubber in the everyday computational biology environment.

Concept, implementation and design of the software was done by me. *clubber* performance analysis was done by me. The analysis of the Horizon oil-spill study and CAMI challenge data was carried out by Chengsheng Zhu. The manuscript was drafted by Yana Bromberg and me. *clubber* is available as Docker container [79], Git repository [80] and is archived via Digital Object Identifier (DOI) [81].

3.1.2 Journal article. Miller et al., Journal of Integrative Bioinformatics 2017

The published article is attached below.

DE GRUYTER Journal of Integrative Bioinformatics. 2017; 20170020

Maximilian Miller[1,2,3] / Chengsheng Zhu[1] / Yana Bromberg[1,4,5]

clubber: removing the bioinformatics bottleneck in big data analyses

[1] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08901, USA, E-mail: mmiller@bromberglab.org, yana@bromberglab.org
[2] Department for Bioinformatics and Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany, E-mail: mmiller@bromberglab.org
[3] TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748 Garching/Munich, Germany, Tel.: +1 848 932 5638, Fax: +1 732 932 8965, E-mail: mmiller@bromberglab.org
[4] Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, Piscataway, NJ 08854, USA, E-mail: yana@bromberglab.org
[5] Institute for Advanced Study at Technische Universität München (TUM-IAS), Garching/Munich, Germany, E-mail: yana@bromberglab.org

Abstract:
With the advent of modern day high-throughput technologies, the bottleneck in biological discovery has shifted from the cost of doing experiments to that of analyzing results. *clubber* is our automated cluster-load balancing system developed for optimizing these "big data" analyses. Its plug-and-play framework encourages re-use of existing solutions for bioinformatics problems. *clubber*'s goals are to reduce computation times and to facilitate use of cluster computing. The first goal is achieved by automating the balance of parallel submissions across available high performance computing (HPC) resources. Notably, the latter can be added on demand, including cloud-based resources, and/or featuring heterogeneous environments. The second goal of making HPCs user-friendly is facilitated by an interactive web interface and a RESTful API, allowing for job monitoring and result retrieval. We used *clubber* to speed up our pipeline for annotating molecular functionality of metagenomes. Here, we analyzed the Deepwater Horizon oil-spill study data to quantitatively show that the beach sands have not yet entirely recovered. Further, our analysis of the CAMI-challenge data revealed that microbiome taxonomic shifts do not necessarily correlate with functional shifts. These examples (21 metagenomes processed in 172 min) clearly illustrate the importance of *clubber* in the everyday computational biology environment.

Keywords: cluster job scheduler, high performance computing, job management, load balancing
DOI: 10.1515/jib-2017-0020
Received: March 23, 2017; **Revised:** April 19, 2017; **Accepted:** April 27, 2017

1 Introduction

Fast-paced growth of high performance computing (HPC), coupled with the recent appearance of new cloud computing solutions, created a new scope of possibilities for applications in today's science. At the same time, more advanced and less expensive high throughput experimental assays have led to exponential growth of new biological datasets. Having access to sufficient computational resources to deal with the growing "big data" is therefore essential not only for computational, but also for experimental biology research labs, particularly those working in genomics. Less than two decades ago the first human genome took 13 years and $2.7 billion to sequence [1]. Today sequencing a genome takes a day and $1000, with costs projected to go even lower in the near future. Recent projects like the 1000 Genomes Project [2] and others currently under way [3], [4] will provide the field with an unprecedented amount of data, opening up new possibilities to significantly improve current models and tools.

These developments come at a cost, as traditional HPC is quite expensive both in purchase and maintenance. Research labs espouse different models for dealing with this computing need – some have their own computational power, others share machines across an institute or outsource their computing to collaborators. Although usability varies significantly across setups, compute nodes rarely reach the often-targeted utilization rates of 75–85 % consistent workload. Usage usually peaks with a specific high priority project running on the

Maximilian Miller, Yana Bromberg are the corresponding authors.

cluster for a limited time or with short-term jobs submitted through a web interface, where timing and responsiveness are essential. Cloud computing offers new alternatives, but is not always an adequate replacement for traditional HPC. The nature of cloud solutions often creates new challenges, such as the transfer of enormous amounts of data to and from the remote cloud storage.

Both from time and performance points of view, there is a clear advantage in making use of all available computational resources when necessary. However, this is a considerable challenge as the, often distributed and setup-disparate, clusters have distinct runtime pre-requisites. Ideally all resources would "speak" the same language, *i.e.* have a shared common base (OS, executables, job scheduler, etc.) Existing tools [5], [6] for bringing together disjoint computational resources and for distributing jobs among them require significant IT-related knowledge to get up and running. Moreover, none of these were designed explicitly for evaluations and approaches common in computational biology. Their capability is mostly limited to retrieval of job results from compute clusters and does not extend to downstream processing. Thus, post-processing and publishing of results is not automated and has to be dealt with individually.

Here we describe our novel *clubber* (CLUster-load Balancer for Bioinformatics E-Resources) framework, available at http://services.bromberglab.org/clubber. *clubber* is designed specifically to facilitate and accelerate common computational biology experimental workflows and used in conjunction with existing methods or scripts to efficiently process large-scale datasets. Using *clubber* is as simple as downloading and installing Docker [7], a software container platform available for every environment, and using a single command to run the Docker-cloud *clubber* container [8]. From there, any interaction for basic configuration, job submission, monitoring and displaying results is achieved via the *clubber* web interface. Note that *clubber* can also be run from command-line using an interactive console, or from within a Python project by importing the *clubber* package. Due to our method's modular design, all of its main components (Manager, Database, Web Interface) can run separately on different environments/machines. Further, *clubber* can be easily configured to use any of the databases or webservers and thus to directly integrate into existing external services. Results can be accessed directly from the *clubber* web interface, either as downloadable files or as searchable data tables (given an appropriate output format). A RESTful [9] API provides programmatic access to the jobs managed by *clubber*, enabling other frameworks to monitor individual job progress and retrieve and display the final results. Very importantly, the *clubber* API facilitates integration into existing and new web services; i.e. tasks submitted through a web interface can be simply "handed over" to *clubber* and results queried once available. *clubber* can be set up on a dedicated server to be accessible by all members of a research group or by a selected few authenticated via a built-in user authentication module.

Existing workflow frameworks like Galaxy [10] and Nextflow [11] allow users to create computational pipelines to process and analyze biological data. Although both environments are highly usable, they have some limitations. Galaxy, for example, requires some time for setup of all components and limits the selection of available tools to those for which corresponding plugins have been written. Nextflow, on the other hand, has limited data filtering and visualization capabilities. Further, both tools can be configured to run jobs on a remote cluster, and Galaxy additionally provides means to make results accessible via a web interface. However, in both cases, jobs are submitted sequentially to only one previously configured cluster. Distributing jobs to multiple resources requires manual interaction and, potentially, adaptation of the necessary submission scripts. This leads to extensive computation times, directly correlated with the amount of processed data.

We designed *clubber* to deal with the challenges of growing datasets, which are particularly obvious in genome research. The current *clubber* package includes built-in methods to simplify parallelized job submission, e.g. splitting a single multi-sequence input file to submit parallel jobs, each containing a user-defined number of sequences. All of these features make *clubber* an essential tool for processing and analyses of vast amounts of biological data in a parallel, efficient, and (very) fast fashion.

2 Methods

***clubber* works in all environments and integrates seamlessly with existing workload managers.** We made *clubber* available as a ready-to-launch Docker image. Adding a computing resource (an HPC cluster) requires only a valid username and password combination for a user who is eligible to submit jobs on this specific resource. Note that there is no need for any additional software to be installed on these resources. The standalone *clubber python* installation has only two requirements: (i) access to a MySQL database (version 5.x) and (ii) availability of *python* (version 3.x). The optional web interface additionally requires access to a webserver with installed PHP module (version 5.6.x). Detailed installation instructions and sources can be found online [8]. Figure 1 illustrates the *clubber* workflow. *clubber*'s three components, Manager, Database, and Web Interface, are independent from each other. The Manager accesses registered clusters via Secure Shell (SSH) and commu-

nicates with the Database using MySQL queries. The Web Interface interacts with the Database to register jobs, monitor their progress and retrieve results.

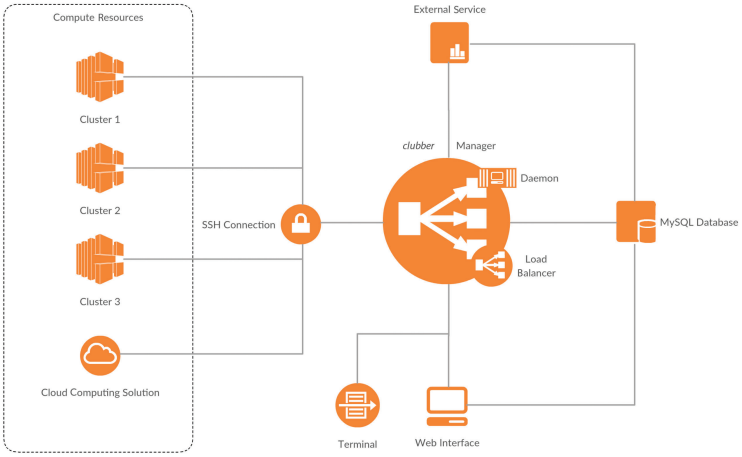

Figure 1: The *clubber* pipeline. Jobs can be submitted either through a web interface or via command-line to the *clubber* manager. These are registered and managed using a relational database. The manager uses an automated balancing approach to distribute jobs among available clusters; the manager daemon runs locally and communicates with available clusters, transferring completed job results and storing them locally or, optionally, in the database.

***clubber* bundles computational resources, providing an interface for a simple centralized submission.** *clubber* can be used in two different ways: through an interactive web interface or via command-line. First, a *clubber* project is created, defining basic parameters like project name, selection of clusters to use and the environment variables necessary for job submission. Projects can contain binaries or database files required by the associated jobs. Note that single jobs can be submitted without creating a project; these will automatically be assigned to a default project with no environment variables set. After a project has been created and automatically initiated on the specified clusters, jobs can be submitted using the web interface or from command-line. Additional environment and job specific variables are defined in a simple syntax described in the *clubber* documentation. The manager uses an auto balancing approach to automatically distribute new jobs between registered clusters. Three factors determine how many jobs are submitted to each cluster during the auto balancing process. These are, in decreasing priority: (i) the cluster workload, (ii) the expected queuing time and (iii) the average job runtime. Cluster workload is calculated as a percentage of total possible workload, with 100 % representing a fully occupied cluster. The expected queuing time and the average job runtime are normalized to a [0,1] range, with one representing the maximum amount of time spent in either queue or run state, respectively, over all jobs of the same project among all active clusters. Both factors are set to one by default and are updated automatically during the progression of a project. In order to obtain the cluster specific load balancing factor (LBF) they are combined with the respective cluster workload (Eq. 1).

$$(1 - \text{workload}) \times 0.5 + (1 - \text{queuing}_{\text{factor}}) \times 0.3 + (1 - \text{runtime}_{\text{factor}}) \times 0.2 \quad (1)$$

clubber communicates with clusters exclusively via encrypted SSH. The rsync [12] utility and Secure Copy (SCP) are used to transfer files to and from the clusters. Since some of the inquiries sent to the many clusters take minutes to process, all communication is threaded to avoid blocking faster transactions. This architecture enables *clubber* to efficiently distribute and retrieve jobs in a highly parallelized fashion.

To track and update current job states *clubber* relies on a relational database. This approach results in very robust job exception handling, both regarding errors on remote clusters and exceptions like lost connections on the machines running the *clubber* manager. The database also allows independent services, which use *clubber* as a job manager, to monitor current job states and retrieve results. Job success is continuously and extensively validated, ensuring that a project with millions of jobs is completed correctly even after allowing for power failures and compute node breakdowns. Once a job is identified as finished, the validation pipeline ensures that the expected results are present and correctly retrieved from the clusters. In case of errors, jobs are reset

and re-sent out for computation. A detailed logging and notification module tracks these processes and notifies the user if specific jobs produce recurring errors.

***clubber* is designed to be used with existing software tools.** Our plug-and-play framework makes it possible to use any existing tools or scripts within the *clubber* environment. User-defined specific pre- and post-processing actions can also be re-used with *clubber* projects. This allows for manipulation of input data prior to batch processing (e.g. converting fastq to fasta format) and for automatic processing of job results once they have been retrieved from the clusters. In its initial release, *clubber* includes two built-in methods for specific pre- and post-processing to simplify parallelized job submission. They allow to automatically split a single multi-sequence input file to submit parallel jobs and merge results once all jobs have been computed. The number of sequences used for each parallel job is user-defined. We expect that with increasing use of *clubber* (available as Git repository hosted on bitbucket) [8], the community will produce a larger repertoire of common pre- and post-processing tools, e.g. file conversion, filtering, etc., commonly applied in every-day computational biology.

3 Results and Discussion

***clubber* significantly reduces the "real-world" compute time by parallelizing and optimizing the workload distribution across available resources.** We evaluated *clubber* performance by measuring the time required to complete one thousand individual jobs, requiring 1-min CPU time each. Note, that these jobs did not require any data to be transferred to remote clusters. The evaluation was performed in various scenarios. We compared the required time at different cluster workloads when using *clubber* with one to five separate clusters available vs. a standard job submission (Figure 2). A standard job submission is defined as a manual submission of a single shell script running all thousand jobs on a single local HPC cluster. Note that workloads for remote clusters registered with *clubber* are conservatively estimated to be consistently at 50 %; the actual gain in computation efficiency could be substantially higher. Also note that 0 % workload is here defined as the ability to run at most 100 jobs in parallel. For the (ideal, but also rare) case of no (0 %) workload on the local cluster, only two additional registered clusters, both exhibiting a workload average, reduce the overall computation time by approximately 50 %. The total gain in computation time is directly correlated to the current workloads on the remote clusters. *clubber*'s auto-balancing job submission ensures that clusters with a low workload are preferentially selected, optimizing and reducing to a minimum the total required computation time. As expected, the more clusters are registered with *clubber* the less effect single clusters with a high load have on the final computation time. The advantage of using *clubber* is particularly obvious in a scenario where only one cluster is available for computation vs. having two clusters – a local and one additional remote cluster. Using *clubber* speeds up computation by up to 100 %. Note that simply logging into another cluster and submitting job subsets is tedious task, which would not, even in the best case scenario, achieve similar speed up – as one cluster finishes, the other is still only somewhat through its assigned computation.

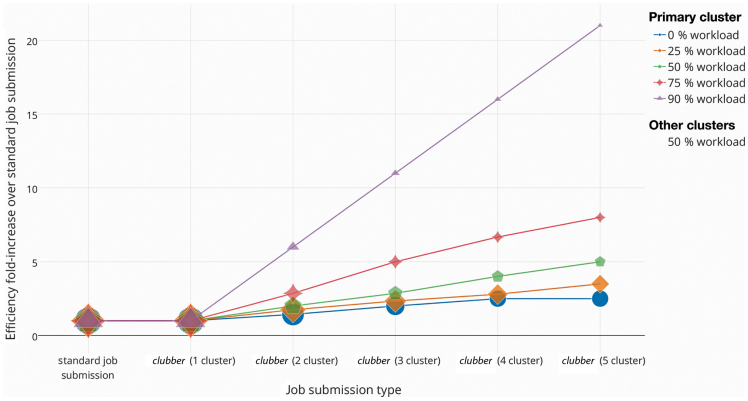

Figure 2: Efficiency fold-change of *clubber* vs. standard job submission: Efficiency fold increase in submitting jobs using *clubber* as compared to a standard job submission. Primary cluster workload is varied between 0 % and 90 %, where 0 % workload is defined here as the ability to run at most 100 jobs in parallel (100 CPU cluster). Compute time is measured for a submission of 1000 jobs, each requiring 1-min CPU time and no data transfer. Active workloads for remote clusters registered with *clubber* are conservatively estimated to be consistently at 50 % of possible total. None of these clusters dropped below that threshold in our use experience. They have, however, gone significantly higher. Thus, the actual gain in computation efficiency could be even higher than that displayed.

***clubber* facilitates fast evaluation of millions of sequences.** Our recent work required a total of 19.4 million bacterial sequences to be analyzed for all-to-all pairwise similarity using BLAST [13]. We estimated that our single local cluster of 640 compute cores in its entirety would have taken roughly 4 months to perform the computation. This estimate is based on a 24 day-long 3,797,793 job BLAST run against the 19.4 M sequence database. Using *clubber* to run on three additional clusters (800, 1536, and 3120 cores, respectively; of varied load, but no more than 50 % of any one cluster available at any given time), speeds up this time to a bit over 2 months (70 days, a factor of 1.8).

Deepwater oil spill metagenome analysis using mi-faser. Our lab's recently created web service [14], mi-faser [15], uses *clubber* to rapidly annotate gigabytes of genomic sequence read data for the molecular functionality encoded by the "read-parent genes" without the need for assembly. For every input metagenome, mi-faser computes a function profile – a list of Enzyme Commission (EC) numbers and the associated read abundances. To illustrate *clubber* functionality, we ran mi-faser on 16 beach sand metagenomes from four phases of the Deepwater Horizon oil spill [16] (BioProject PRJNA260285) study – Pre-oil, two samples of Oil, and Recovery phases (available at http://services.bromberglab.org/mifaser/example). Analysis of this data (73GB sequence reads) using the mi-faser web interface with a *clubber* back-end was done in only 1 h, with *clubber* distributing a total of 4.5 k jobs among three compute clusters. Note that running these jobs using only our local cluster (640 cores) with an average workload (unavailability of nodes) of 30 %, took 170 min – 3-fold slower than *clubber*.

For further analysis, we removed sample-specific functions and normalized the individual entries of the function profile vectors by the total number of annotated reads. We found that microbiome functional profiles of samples from different phases significantly differ from each other (Figure 3, non-metric multidimensional scaling (NMDS) analysis [17]; $P < 0.001$, permanova test [18]). Interestingly, the samples from the Oil phases show higher variation than the samples from the Pre-oil phase and the Recovery phase, suggesting that "normal" ecosystem microbiomes are functionally more consistent than those in the disturbed ecosystems. The samples from Oil phases are functionally closer to the samples from the Recovery phase than to the Pre-oil phase, indicating that the beach sands have likely not entirely recovered.

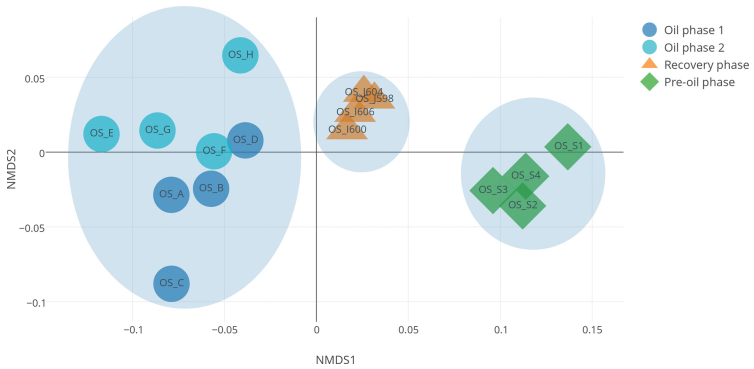

Figure 3: Microbiome functional capabilities of beach sand metagenomes from a study of the Deepwater Horizon oil spill (16) (BioProject PRJNA260285) differ across phases. The samples were collected from four phases, including Pre-oil phase (OS-S1, OS-S2, OS-S3 and OS-S4), Oil phase 1 (OS-A, OS-B, OS-C and OS-D), Oil phase 2 (OS-E, OS-F, OS-G and OS-H) and Recovery phase (OS-I600, OS-I606, OS-J598 and OS-J604). The distances between samples in this non-metric multidimensional scaling (NMDS) graph represent the variation between sample function profiles. Samples from Pre-oil phase, Oil phases and Recovery phase localize separately. Oil phase samples are closer to Recovery phase samples than to Pre-oil phase samples.

Regardless of the significant differences between phases, the fraction of housekeeping functions (compiled from [19]) was highly consistent across samples (22.1±0.5 %); *e.g.* DNA-directed RNA polymerase (2.7.7.6) is the most abundant function in all samples (about 4∼5 %). As the number of reads encoding a particular functionality is highly correlated to the number of individual cells performing said functionality, these results are not

very surprising – all bacterial phyla, no matter how different, carry housekeeping genes. This finding serves as a confirmation of mi-faser's accuracy, while highlighting its ability to estimate functional diversity in a non-taxon dependent level.

Critical Assessment of Metagenomic Interpretation (CAMI) challenge analysis using mi-faser. We further used mi-faser to evaluate a high complexity data set from the CAMI [20] challenge. The data set contains a time series of five Hiseq samples (15 Gbp each) with small insert sizes sampled from a complex microbial community. With *clubber* optimizing job submissions, the total computation time for 500 M sequence reads was only 1 h 59 min. Note that the CAMI challenge did not evaluate runtimes for the submitted tools/predictions, but they note that this evaluation is a necessary feature of future method development [20]. Metagenome comparative analysis revealed that the microbiome functional profiles remain highly consistent (Table 1), regardless of a clear community composition shift (Table 2). Interestingly, these results indicate that, over time, microbial species were exchanged, while maintaining the same functional capacity. Thus, the time effect on the microbial community is not as striking as what the taxonomical changes would suggest. This example highlights the fact that inferring microbiome function from its taxonomy composition is misleading. Thus, metagenomic analysis tools such as mi-faser are essential for a deeper understanding of microbiome functional potentials. Note that *clubber* is uniquely responsible for allowing our lab to make the mi-faser web interface available to the general public for the purposes of extremely fast (and accurate) functional annotation of millions of raw sequence reads.

Table 1: Spearman correlation between taxonomic profiles[a] of CAMI metagenomes.

	RH_S001	RH_S002	RH_S003	RH_S004	RH_S005
RH_S001	1	–	–	–	–
RH_S002	0.78	1	–	–	–
RH_S003	0.64	0.75	1	–	–
RH_S004	0.51	0.59	0.73	1	–
RH_S005	0.45	0.51	0.54	0.71	1

[a]The taxonomic profiles were obtained from http://cami-challenge.org.

Table 2: Spearman correlation between functional profiles[a] of CAMI metagenomes.

	RH_S001	RH_S002	RH_S003	RH_S004	RH_S005
RH_S001	1	–	–	–	–
RH_S002	0.99	1	–	–	–
RH_S003	0.99	0.99	1	–	–
RH_S004	0.99	0.99	0.99	1	–
RH_S005	0.99	0.99	0.99	0.99	1

[a]The functional profiles were annotated by mi-faser (15).

Dealing with tool heterogeneity in *clubber*-accessible resources. Even though *clubber* is highly successful in facilitating HPC use, there may be still scenarios, which require manual interaction with the individual compute clusters. When creating a *clubber* project that includes binaries, the user has to validate these binaries on each of the cluster resources. When using pre-installed tools local to each resource, all installs have to be of the same version and produce identical results given identical input. To prevent erroneous results in these scenarios, *clubber* offers the option to automatically compare cluster environments and submit test jobs before starting a project run on different computing resources. Note that virtualization solutions, *e.g.* Docker, offer a simple solution to these problems by guaranteeing identical environments on every resource. In this scenario (planned for the next release of our software) *clubber* distributes a user provided Docker image to the clusters and relays job parameters when starting a Docker container.

Impact of dataset size on *clubber* performance. *clubber* was developed to process extremely large datasets using remotely accessed resources. The remoteness of these resources, thus, poses a bottleneck in transferring data between compute clusters. For the larger compute centers, it is safe to assume that an appropriately fast connection is available. For smaller set-ups, data transfer speeds may vary. In testing to evaluate the contribution of transfer times for our collection of clusters, some smaller and some larger ones, we found that times did not vary across remote and local machines and did not affect the relative performance. For all five of our clusters the transfer times varied by as little as 6 %, despite being located in different places of the world (New Brunswick, NJ, USA and Garching, Germany); the speed of transfer of 1Gb of data was 146 $\pm$ 8 s. Note that

jobs requiring large data transfers would necessarily be slowed down, but roughly in equal measure for local or remote machines. The slow-down is especially visible in cases where the computation time for a single job is fairly short. Increasing the number of jobs processed reduces this initial impact as performance improves by use of additional resources.

Better resource management and faster processing speeds with *clubber*. Our *clubber* framework provides a simple way to bundle available, possibly heterogeneous, computational resources and to distribute computations minimizing the required processing time. This approach avoids long computation times associated with an overloaded local cluster when there are in fact additional resources available elsewhere. Simple job submission/monitoring and automated exception handling make *clubber* easy-to-use and ideal for handling projects with millions of jobs. Its ability to use cloud-computing services like Amazon Web Services (AWS) with *clubber* on-demand, additionally allows for temporary, large-scale increases in computational resources. With all of these features, web services, the bread-and-butter of the computational biology community, are made extremely responsive with *clubber*.

With the exponential growth of available data in computational biology waiting to be analyzed, bioinformatics, not experimental analysis, has unexpectedly become the progress bottleneck. By combining the available resources and using them in the most optimal fashion, *clubber* offers a new approach to tackling this challenge.

Acknowledgements

We thank Max Haggblom, Yannick Mahlich, Alexandra Pushkar, and Yanran Wang (all Rutgers University, New Brunswick, NJ, USA) for many discussions and manuscript review. We thank Bill Abbott and Kevin Abbey (both Rutgers) for technical support. We are grateful to Sonakshi Bhattacharjee (Technical University of Munich, TUM) for advice during the initial development phase. Particular thanks are due to Burkhard Rost (TUM) for his hospitality, valuable discussions, and letting us use the rostlab cluster! We also thank Timothy Karl (TUM), for his help with the rostlab cluster setup within the *clubber* environment and the rostlab members for quickly adapting *clubber* after evaluating its functionality. Last but not least, we thank all those who deposit their experimental data in public databases and those who maintain these databases. YB, MM, and CZ were partially supported by the NIH/NIGMS grant U01 GM115486 (to YB). YB and CZ were also supported by the NSF CAREER grant 1553289 (to YB). YB was additionally supported by USDA-NIFA 1015:0228906 grant and the TU München – Institute for advanced study Hans Fischer fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme, grant agreement 291763.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] Gyles C. The DNA revolution. Canadian Vet J. 2008;49:745–6.
[2] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56–65.
[3] 100K Food Pathogen Project: Bart Weimer. 2016. Available from: https://www.ncbi.nlm.nih.gov/bioproject/186441.
[4] McGrath JA. Rare inherited skin diseases and the Genomics England 100 000 Genome Project. Br J Dermatol. 2016;174:257–8.
[5] Thain D, Tannenbaum T, Livny M. Distributed computing in practice: the Condor experience. Concurr Comp Pract Ex. 2005;17:323–56.
[6] Weitzel D, Sfiligoi I, Bockelman B, Frey J, Wuerthwein F, Fraser D, et al. Accessing opportunistic resources with Bosco. J Phys Conf Ser. 2014;513:032105.
[7] Docker, the world's leading software container platform: the Docker open source project. 2017. Available from: https://www.docker.com/. Accessed on: April 12th, 2017.
[8] clubber: Yana Bromberg Lab, Rutgers University. 2017. Available from: https://bitbucket.org/bromberglab/bromberglab-clubber/. Accessed on: April 12th, 2017.
[9] Web Services Architecture: World Wide Web Consortium. 2004. Available from: https://www.w3.org/TR/2004/NOTE-ws-arch-20040211/–relwwwrest. Accessed on: April 12th, 2017.
[10] Galaxy: The Galaxy Project. 2017. Available from: https://galaxyproject.org/. Accessed on: April 12th, 2017.

[11] Nextflow: Comparative Bioinformatics group, Barcelona Center for Genomic Regulation (CRG). 2016. Available from: https://www.nextflow.io/. Accessed on: April 12th, 2017.
[12] rsync. 2015. Available from: https://rsync.samba.org/. Accessed on: April 12th, 2017.
[13] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
[14] Zhu C, Miller M, Marpaka S, Vaysberg P, Rühlemann M, Heinsen F-A. Functional sequencing read annotation for high precision microbiome analysis. Submitted, 2017.
[15] mi-faser: Yana Bromberg Lab, Rutgers University. 2017. Available from: http://services.bromberglab.org/mifaser. Accessed on: April 12th, 2017.
[16] Rodriguez-R LM, Overholt WA, Hagan C, Huettel M, Kostka JE, Konstantinidis KT. Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. ISME J. 2015;9:1928–40.
[17] Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964;29:115–29.
[18] Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral Ecology. 2001;26:32–46.
[19] Gil R, Silva FJ, Pereto J, Moya A. Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev. 2004;68:518–37 Table of Contents.
[20] Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droege J, et al. Critical Assessment of Metagenome Interpretation — a benchmark of computational metagenomics software, 20171 9. DOI:10.1101/099127.

4 Comprehensive microbiome function analyses

4.1 fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks

4.1.1 Preface

In recent years, there has been an increased interest in microbial organisms and their associated functional repertoire. In fact, the molecular functionality found in microorganisms is relevant to a range of human interests, including health, industrial production, and bio-remediation. A main driving force for microbial functional diversification are environmental factors. Microorganisms inhabiting the same environmental niche tend to be more functionally similar than those from different environments. In some cases, even closely phylogenetically related microbes differ more across environments than across taxa. Experimental study of these microbes to optimize their uses is expensive and time-consuming. Experimental assessment of bacterial functional capacity is very challenging [82]. Our functional repertoire similarity-based organism network (fusion) algorithm allows the comparison of microbial functional similarities based on their proteome. However, while those similarities are often reported in terms of taxonomic relationships, no existing databases directly links microbial functions to the environment.

This gap is closed by *fusion*DB. *fusion*DB assesses microbial functional similarity on the basis of their corresponding proteome, connecting individual microbes via common functions. *fusion*DB uses the fusion protocol [83], an organism functional similarity network. It contains 1374 taxonomically distinct bacteria annotated with available metadata: habitat/niche, preferred temperature, and oxygen use. An interactive (web) service allows mapping of new microbial genomes to the functional spectrum of reference bacteria, rendering interactive similarity networks that highlight shared functionality. This often includes matching proteins of yet unannotated function across organisms. *fusion*DB provides a fast means of comparing microbes, identifying potential horizontal gene transfer events, and highlighting key environment-specific functionality. *fusion*DB also provides quantitative support to the fact that environmental factors drive microbial functional diversification. We mapped a recently sequenced genome of a freshwater *Synechococcus* bacterium to *fusion*DB and demonstrated that this microorganism is more functionally related to other fresh water *Cyanobacteria* than to the marine *Synechococcus* [84]. In a case study on *Bacillus* microbes, we used *fusion*DB to track organism-unique functions and illustrated the detection of core-function repertoires that capture traces of environmentally driven horizontal gene transfer (HGT). Building the reference database

and mapping new microbial genomes to *fusion*DB is compute intensive. For the current reference database we had to align 4.284.540 proteins from 1374 bacterial genomes in an all-vs-all comparison. Genomes were retrieved from the NCBI GenBank [85]. In the current version of this database the number of protein sequences grew more than 3-fold. To manage the growing complexity in generating the reference database and achieve reasonable processing times for user submitted genomes, we integrated *clubber* [86] (section 3.1) as *fusion*DB back-end. Further, we are evaluating a new alignment approach (MMseqs2 [87]) to replace PSI-BLAST in the long run. Those improvements will enable us to provide considerably faster mappings of new genomes to the database. More importantly, we will be able to incorporate significantly more bacterial genomes for building the reference database. That will in turn allow us to better and more comprehensively annotate functions within unknown microorganisms, *e.g.* isolated from patient samples. Identifying shared functionalities within the reference database can provide important information like pathogenic characteristics or antibiotic resistance status. Finally, we expect that *fusion*DB will additionally facilitate the study of environment-specific microbial molecular functionalities.

The web service front- and back-end was developed by Yannick Mahlich and me. Network visualization and interactivity was done by me. Evaluation was done by Yannick Mahlich. The project was designed by Chengsheng Zhu, Yannick Mahlich and Yana Bromberg. The manuscript was drafted by all four authors.

4.1.2 Journal article. Zhu, Mahlich & Miller et al., Journal of Nucleic Acids Research 2017

Supplementary material can be found online at [88]. The published article is attached below.

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1060

*fusion*DB: assessing microbial diversity and environmental preferences via functional similarity networks

Chengsheng Zhu[1,†], Yannick Mahlich[1,2,3,4,*,†], Maximilian Miller[1,2,3,†] and Yana Bromberg[1,4,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA, [2]Computational Biology & Bioinformatics - i12 Informatics, Technical University of Munich (TUM), Boltzmannstrasse 3, 85748 Garching/Munich, Germany, [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technical University of Munich (TUM), 85748 Garching/Munich, Germany and [4]Institute for Advanced Study, Technical University of Munich (TUM), Lichtenbergstrasse 2 a, 85748 Garching/Munich, Germany

Received August 11, 2017; Revised September 24, 2017; Editorial Decision October 12, 2017; Accepted October 22, 2017

ABSTRACT

Microbial functional diversification is driven by environmental factors, i.e. microorganisms inhabiting the same environmental niche tend to be more functionally similar than those from different environments. In some cases, even closely phylogenetically related microbes differ more across environments than across taxa. While microbial similarities are often reported in terms of taxonomic relationships, no existing databases directly link microbial functions to the environment. We previously developed a method for comparing microbial functional similarities on the basis of proteins translated from their sequenced genomes. Here, we describe *fusion*DB, a novel database that uses our functional data to represent 1374 taxonomically distinct bacteria annotated with available metadata: habitat/niche, preferred temperature, and oxygen use. Each microbe is encoded as a set of functions represented by its proteome and individual microbes are connected via common functions. Users can search *fusion*DB via combinations of organism names and metadata. Moreover, the web interface allows mapping new microbial genomes to the functional spectrum of reference bacteria, rendering interactive similarity networks that highlight shared functionality. *fusion*DB provides a fast means of comparing microbes, identifying potential horizontal gene transfer events, and highlighting key environment-specific functionality.

INTRODUCTION

Microorganisms are capable of carrying out much of molecular functionality relevant to a range of human interests, including health, industrial production, and bioremediation. Experimental study of these microbes to optimize their uses is expensive and time-consuming; e.g. as many as three hundred biochemical/physiological tests only reflect 5–20% of the bacterial functional potential (1). The recent drastic increase in the number of sequenced microbial genomes has facilitated access to microbial molecular functionality from the gene/protein sequence side, via databases like Pfam (2), COG (3), TIGRfam (4), RAST (5) and others. Note that the relatively low number of available experimental functional annotations limits the power of these databases in recognizing microbial proteins that provide novel functionality. Additional information about microbial environmental preferences can be found, e.g. in GOLD (6). While it is well known that environmental factors play an important role in microbial functionality (7), none of the existing resources directly link environmental data to microbial function.

We mapped bacterial proteins to molecular functions and studied the functional relationships between bacteria in the light of their chosen habitats. We previously developed *fusion* (8), an organism functional similarity network, which can be used to broadly summarize the environmental factors driving microbial functional diversification. Here, we describe *fusion*DB – a database relating bacterial *fusion* functional repertoires to the corresponding environmental niches. *fusion*DB is explorable via a web-interface by querying for combinations of organism names and environments. Users can also map new organism proteomes to the functional repertoires of the reference organisms in *fusion*DB; including, notably, matching proteins of yet unannotated function across organisms. The submitted organisms are vi-

*To whom correspondence should be addressed. Tel: +1 848 932 5638; Fax +1 848 932 8965; Email: ymahlich@bromberglab.org
Correspondence may also be addressed to Bromberg Yana. Tel: +1 646 220 3290; Email: yanab@rci.rutgers.edu
†These authors contributed equally to this work as first authors.

sualized, and can be further explored, interactively as *fusion* networks in the context of selected reference genomes. Additionally, the web interface generates *fusion*+ networks, *i.e.* views that explicitly indicate shared microbial functions.

Our overall analyses of the *fusion*DB data for the first time give quantitative support to the fact that environmental factors drive microbial functional diversification. To demonstrate *fusionDB* functionality for individual organisms, we mapped a recently sequenced genome of a freshwater *Synechococcus* bacterium to *fusion*DB. In line with our previous findings (8), we demonstrate that this microorganism is more functionally related to other fresh water Cyanobacteria than to the marine *Synechococcus*. In a case study on *Bacillus* microbes, we use *fusionDB* to track organism-unique functions and illustrate the detection of core-function repertoires that capture traces of environmentally driven horizontal gene transfer (HGT). *fusionDB* is a unique tool that provides an easy way of analysing the, often unannotated, molecular function spectrum of a given microbe. It further places this microbe into a context of other reference organisms and relates the identified microbial function to the preferred environmental conditions. Our approach allows for detection of microbial functional similarities, often mediated via horizontal gene transfer, that are difficult to recover via phylogenetic analysis. We note that, in the near future, *fusionDB* may also be useful for the analysis of functional potentials encoded in microbiome metagenomes. We expect that *fusionDB* will facilitate the study of environment-specific microbial molecular functionalities, leading to improved understanding of microbial lifestyles and to an increased number of applied bacterial uses.

METHODS

Database setup

*fusion*DB is based on alignments of 4 284 540 proteins from 1374 bacterial genomes (December 2011 NCBI GenBank (9). For each bacterium, we store its (a) NCBI taxonomic information (10) and, where available, (b) environmental metadata (temperature, oxygen requirements, and habitat; GOLD (6). The environments are generalized, e.g. *thermophiles* include hyper-thermophiles. '*No data*' is used to indicate missing annotations (Supplementary Online Material, SOM Table S1, SOM Figure S1). The general *fusion* (functional repertoire similarity-based organism network) protocol is described in our previous work (8). Briefly, all proteins in our database are aligned against each other using three iterations of PSI-BLAST (11) and the alignment length and sequence identity are used to compute Homology-derived Secondary Structure of Proteins (HSSP) distances (12). A network of protein similarities is then clustered using the Markov Clustering Algorithm (MCL) (13). For *fusion*DB the original *fusion* algorithm was modified to use less stringent protein functional similarity criteria (with HSSP distance cutoff = 10), which resulted in 457 576 functions (protein clusters; Table 1). Each bacterium was thus mapped to a set of functions, its functional repertoire (∼2400 functions on average, ranging from 118 to 6134 functions). Note that our functional repertoires include all the bacterial functions, regardless of annotation. We are thus able to make function predictions for proteins in new bacteria, even if these functions have not been annotated before.

Mapping new organisms to fusion

User submitted microbial proteomes and the associated functions are stored in a separate database (SOM Figure S2). For each query protein of the new organism, the mapping pipeline (SOM Figure S3, SOM Methods) (a) runs PSI-BLAST (reporting e-value 1e–10, inclusion e-value 1e–3, three iterations) against reference proteins in *fusion*DB and (b) maps the query to a *fusion* functional cluster, which contains the reference with the highest hit HSSP score. Note that novel proteins that cannot be assigned to existing functional groups (do not match any reference at HSSP distance $\geq$10) are reported as functional singletons even if they are similar among themselves. Additionally, protein alignments that exceed 12 CPU hours of run-time are currently eliminated from further consideration. In testing, we found that no $>$0.1% of the proteins fall into this category. Although long run-times usually indicate that query proteins likely align to many others in our database, they contribute only a small fraction to the overall bacterial similarity and are eliminated for the sake of a faster result turnaround. Note that we also evaluated a number of other algorithms for mapping organism functional repertoires, of which the above-described algorithm performed best (SOM Methods).

All functional cluster assignments of proteins in the query proteome are then combined into a functional repertoire where each functional cluster is unique; i.e. if two query proteins are assigned to the same functional cluster, this cluster is listed only once in the final repertoire.

Evaluating *fusion*DB performance

We evaluated the accuracy of functional mapping of new proteomes by iteratively mapping each of the *fusion*DB organisms back to the remaining 1373. We aligned each protein of the query organism to all proteins in other organisms and selected the alignment with highest HSSP score. We then assigned the query protein to the functional cluster of its match as described above for mapping new organisms.

The performance of this approach was evaluated on a per-function basis, i.e. for each function of each 'newly added' organism we retrieved counts of true positives (TP, proteins correctly assigned to this *fusion*DB function), false positives (FP, proteins falsely assigned to this *fusion*DB function), and false negatives (FN, proteins that are part of this *fusion*DB function in the reference database, but not correctly assigned). Note that reference singleton proteins that were not assigned to any *fusion*DB function were considered true positives. Averaged across all functions, the mean per-function precision and recall of correctly assigning proteins were 97.2% and 96.6%, respectively (3.1×10^{-8} mean per function false positive rate, FPR), while the overall precision of assigning any protein to a function was 98.2% (Eq. 1).

Individual organisms were assigned to their functional repertoires with 99.5% precision and 98.9% recall (Eq. 1,

Table 1. Annotation status of (HSSP-based) function groups

	Function groups (>1 sequence)	Function groups (1 sequence)	Total
Known (Kn)	54 522	15 738	70 260
Hypothetical (Hy)	85 252	89 282	174 534
Unknown (Un)	22 802	189 980	212 782
Total	162 576	295 000	457 576

SOM Figures S4 and S5). For this estimate we evaluated to overlap between reference and assigned repertoire; i.e. functional clusters that appear in both the reference and mapped functional repertoire are true positives. False positives are functional clusters in the mapped functional repertoire but not the reference repertoire, false negatives vice versa. The reported precision and recall are the mean precision and recall values averaged over all organism submissions.

$$precision = \frac{TP}{TP+FP},\ recall = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN} \quad (1)$$

Web interface

*fusion*DB web interface has two functions: *explore* and *map new organisms*. The *explore* section contains access to all the 1374 bacteria and their metadata. Users can search these with (combinations of) organism names and environmental preferences by using text box input or built in filters. A user-selected organism set can be used to create a *fusion* network, in which organism nodes are connected by functional similarity edges. The *fusion* network can be viewed in an interactive display, as well as downloaded as network data files or static images. The user-defined color labels of the organism nodes reflect microbial taxonomy or environment. In the interactive display clicking an organism node reveals its taxonomic information and environmental preferences, while clicking an edge between two organisms yields a list of their shared functions. A *fusion*+ network can further be generated from the same list of organisms. There are two types of vertices (nodes) in *fusion*+: organism nodes and function nodes. Organism nodes are connected to each other only through the function nodes they share. The number of edges (degree) of an organism node represents the total number of functions of the organism; the relative position of each organism node is determined by the pull *toward* other organisms via common functions and *away* from others via unique functions (8). Like *fusion*, *fusion*+ can be interactively displayed, downloaded, and colored by the users' choices. For both network types, users can further retrieve the functions shared by the selected organisms—the core-functional repertoire of the set. Note that the primary function annotation of each functional cluster is the myRAST (5) description most commonly assigned to the cluster members. For each cluster we also include the corresponding Pfam (2) families. This feature is an efficient tool for investigating functions underlying organism diversification, particularly within different environment conditions.

In the *map* section, users can submit their own new organism proteomes (in fasta format) to our server (SOM Figure S3). The server sends out emails to users when mapping is finished. The *map* result page contains two tables containing (a) functional annotations, including the associated *fusion*DB reference sequences and proteins of the query organism that mapped to each functional cluster, as well as (b) similarity (Eq. 2) to the reference organisms in *fusion*DB, including functional repertoire size, functional overlap with the query, and metadata. Tables can be easily sorted, searched and exported as comma-separated files. The submitted proteome is further mapped to user-selected reference organisms with *fusion* and/or *fusion*+ as described above (Figure 1).

$$\text{similarity} = \frac{\text{shared functions}}{\text{the larger functional repertoire size}} \quad (2)$$

Analysis of environment-driven organism similarity

For each environmental condition in *fusionDB*, we sampled organism pairs where organisms were from (a) the same condition (SC, e.g. both mesophiles) and (b) different conditions (DC, e.g. thermophile versus mesophile). To alleviate the effects of data bias, the organisms in one pair were always selected from different taxonomic groups (different families). The smallest available set of pairs, SC-psychrophile contained 33 organisms from 17 families (SOM Table S1; 136 pairs—48 same phylum, 88 different phyla; due to high functional diversity of *Proteobacteria,* its classes were considered independent phyla). For all other environmental factors we sampled (bootstrap with 100 resamples) 136 organism pairs for both SC and DC sets, covering this same minimum taxonomic diversity. We calculated the pairwise functional similarity (Eq. 2) distributions and discarded organism pairs with <5% similarity.

RESULTS AND DISCUSSION

Mapping a new Synechococcus genome to fusionDB

We downloaded the full genome of *Synechococcus* sp. PCC 7502 (GCA_000317085.1) as translated protein sequence fasta (.faa file) from the NCBI Genbank (9) and submitted it to our web interface. This 3,318 protein fresh water Cyanobacteria is isolated from a Sphagnum (peat moss) bog (6). 86% (2,853) of the bacterial proteins mapped to 2208 *fusion*DB functions, while 462 (14%) were functional singletons; three proteins exceeded runtime and were excluded (Methods). The whole process from submission to results notification e-mail took under three and a half hours. The mapping indicates that *Synechococcus* sp. PCC 7502 is most functionally similar (56%) to *Synechocystis* PCC 6803, a fresh water organism evolutionarily closely related to *Synechococcus*. It also shares a high functional similarity with a mud *Synechococcus* (*S.sp.* PCC 7002; 53%) and with other fresh water *Synechococcus* (*S. elongatus* PCC 7942 and *S. elongatus* PCC 6301; 52%). Notably, but not surprisingly, *Synechococcus* sp. PCC 7502 shares much less functional

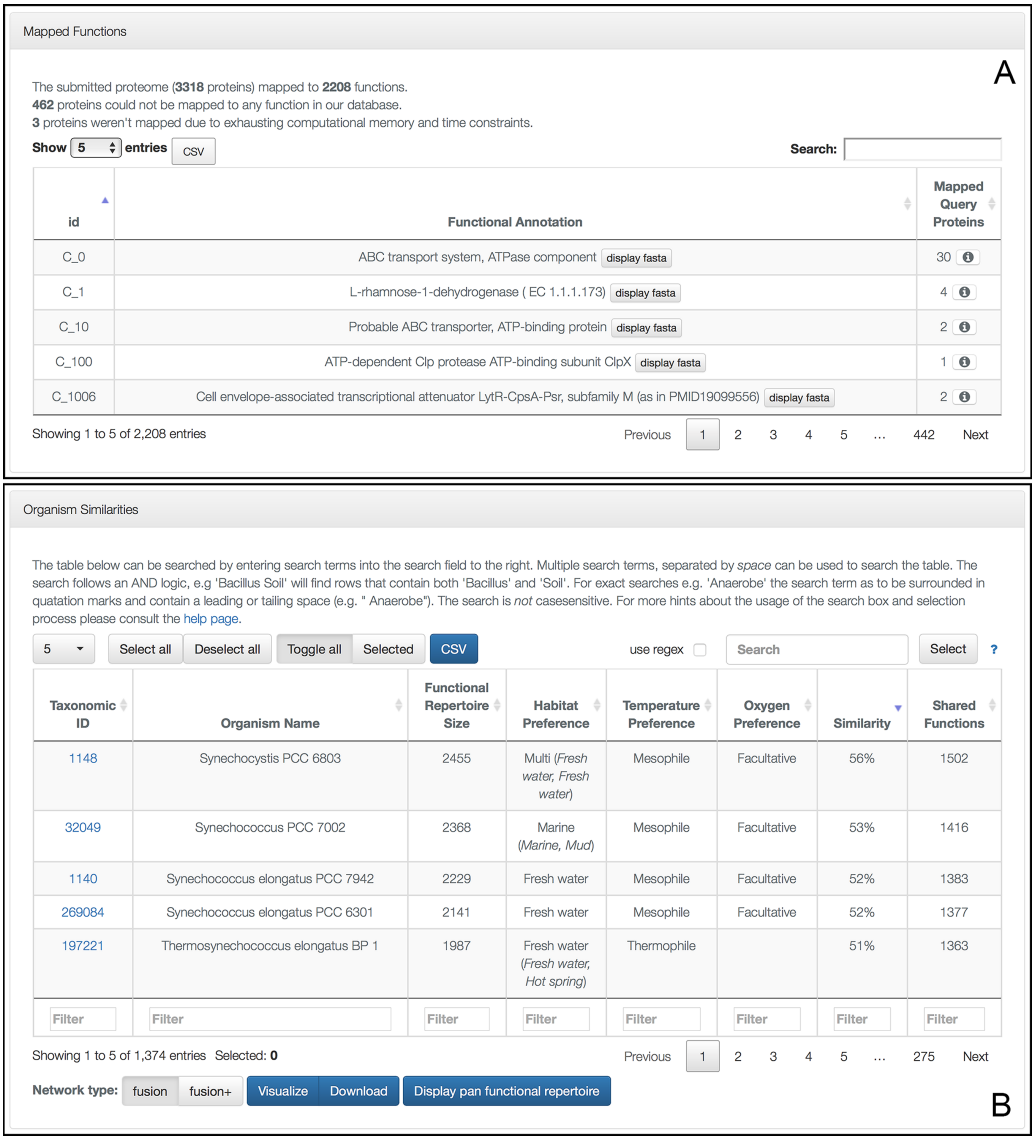

Figure 1. Screenshot of the organism mapping result page. (**A**) The 'Mapped Functions' table lists the functions that the submitted organism is mapped to. For each function, associated proteins from *fusion*DB and mapped query proteins can be displayed. (**B**) The 'Organism Similarities' table displays, all 1374 *fusion*DB organisms and their functional similarities to the query organism, including additional information such as environmental metadata; the view can be toggled between all and user-selected organisms. *fusion*(+) networks of the query and user-selected organisms can be created for on-site visualization (see Figure 2) or download.

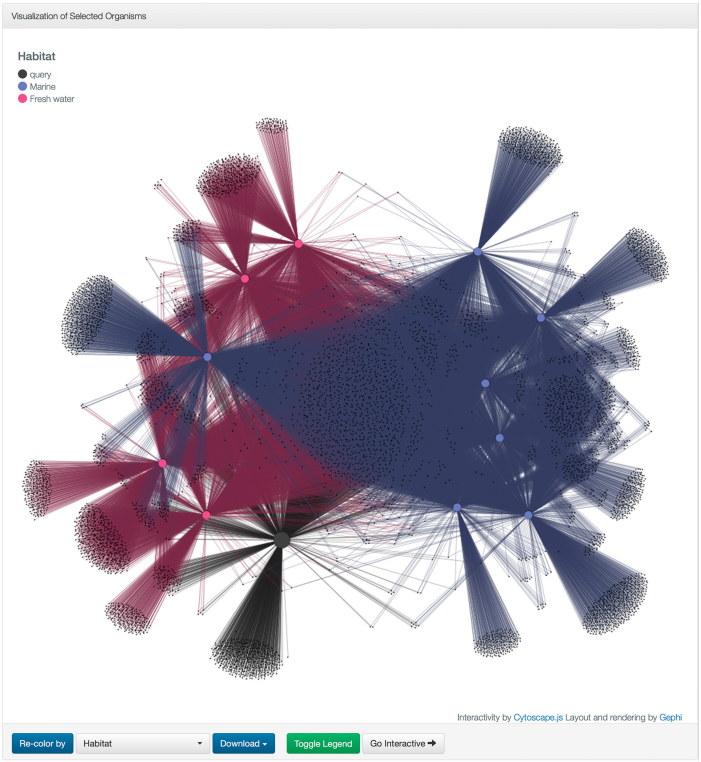

Figure 2. Screenshot of the fusion+ visualization of all Synechoccocus genomes. The submitted *Synechococcus* sp. PCC 7502 (query, black) clusters with the fresh water *Synechococcus* organisms (magenta). Note that *Synechococcus* sp. PCC 7002 – clustered among fresh water organisms; colored dark blue (marine) – is isolated from marine mud. It is salt tolerant but does not require salt for growth).

similarity (40–42%) with the marine *Synechococcus* bacteria. This relationship is clearly demonstrated by the *fusion+* networks (Figure 2). There are 874 functions shared by all the twelve *Synechococcus* (SOM Data 1), the core-function repertoire for this genus, and 1128 functions shared among only the fresh water *Synechococcus* (SOM Data 2). These differential 254 functions (SOM Data 3) are likely important for living in fresh water, as opposed to marine, environment, e.g. low salinity and low osmotic pressure.

Environment significantly affects microbial function

In our evaluation of the effects of environmental pressures on microbial functionality we found that, in general, same environmental condition (SC) organisms across all environmental factors are more functionally similar than DC organisms (from different environments; Figure 3; with some exceptions mentioned below, Kolmogorov-Smirnov test (14) P-value $<$ 2.5e–6). This finding is intuitive and many studies have demonstrated the presence of horizontal gene transfer (HGT) within environment-specific microbiomes (15–17). Our results, however, for the first time, quantify on a broad scale the environmental impact on microorganism function diversification.

SC-thermophile and SC-psychrophile pairs demonstrate significantly higher similarities when compared to DC pairs (Figure 3A). Notably, the higher functional similarity between thermophiles than between psychrophiles suggests that protein functional adaptation to low temperature may be less taxing than to high temperature – an interesting finding in itself. When contrasted with the extremophiles, mesophiles seem to have much larger functional diversity; in fact, SC-mesophile similarities are comparable to those of DC pairs (Figure 3A).

Different molecular pathways of aerobic-respiration and anaerobic-respiration/fermentation may explain the high level of dissimilarity between the aerobes and anaerobes (DC-anaerobe-aerobe; Figure 3B). Interestingly, the SC-anaerobe similarities are higher than the SC-aerobe similarities, likely because the more ancient anaerobic-respiration/fermentation machinery tends to be simpler (fewer reactions) (18) and more conserved.

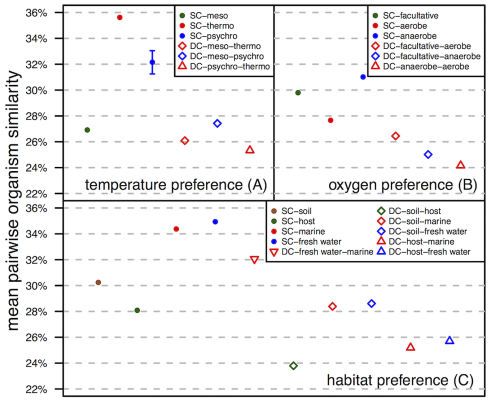

Figure 3. Organism pairwise similarity is higher among organisms living in the same environmental conditions. The mean pairwise similarity for same (SC) and different (DC) condition organisms according to (**A**) temperature, (**B**) oxygen and (**C**) habitat preferences. For all points without error bars, the standard errors are vanishingly small.

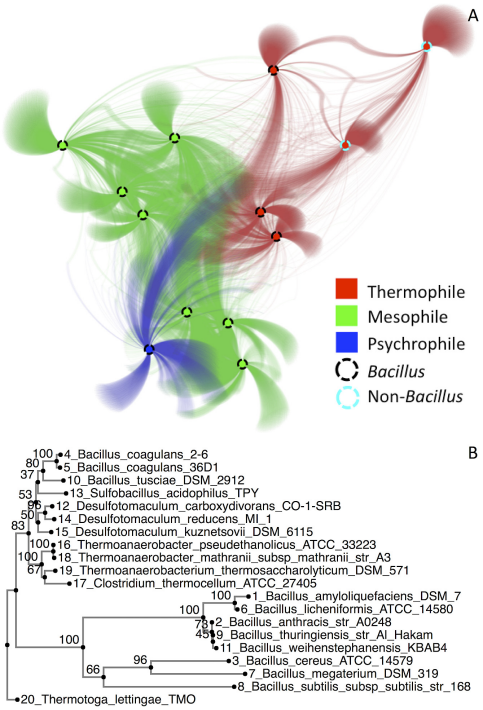

Figure 4. *fusion*DB reveals an HGT event between thermophilic *Bacilli* and thermophilic *Clostridia*. (**A**) *fusion*+ visualization of Bacillus and thermophilic Clostridia. Large organism nodes are connected via small function nodes. The two thermophilic *Clostridia* are connected to the thermophilic *Bacilli* via functions that are possibly horizontally transferred; (**B**) phylogenetic analysis of pyruvate, phosphate dikinase (PPDK) gene suggests HGT between thermophilic *Bacilli* and thermophilic *Clostridia*. The PPDK genes in thermophilic *Bacilli* are evolutionarily more related to those in thermophilic *Clostridia* than those in other *Bacilli*.

Different habitat (DC) samples show lower pairwise organism similarity than SC samples as well (Figure 3C). Interestingly fresh water and marine organism similarity (DC-fresh water-marine) is fairly high, likely due to overlaps in requirements of the aquatic conditions. Note however, that the dissimilarity across fresh water and marine conditions is still high enough to differentiate organisms of the same taxa (e.g. strains of *Synechococcus* in Figure 2). SC-host has the lowest mean organism similarity of the habitat SC samples; we speculate this to be a result of differential adaptations necessary to deal with diverse host defense mechanisms (19). The soil organisms also share low functional similarity, which is likely due to soil heterogeneity at physical, chemical, and biological levels, from nano- to landscape scale (20).

Case study of a temperature driven HGT event

Using the *fusion*DB *explore* functionality, we extracted thermophilic, mesophilic, and psychrophilic species representatives (one per species) of the *Bacillus* genus. We also added two other thermophilic *Clostridia*, *Desulfotomaculum carboxydivorans* CO-1-SRB and *Sulfobacillus acidophilus* TPY, to generate a *fusion*+ network (SOM Table S2; Figure S4A). As expected, note here that overall thermophilic bacteria are further removed from psychrophiles than from mesophiles. Moreover, the thermophilic *Bacilli* were more closely related to the non-*Bacillus* thermophiles than to other *Bacilli*. The three *Bacilli* thermophiles share 29 functions (SOM Data 4) that are not found in other *Bacilli* in this organism set, three of which also exist in the two thermophilc *Clostridia*. One is a likely pyruvate phosphate dikinase (PPDK) that, in extremophiles, works as a primary glycolysis enzyme (21). The thermophilic *Bacilli*'s PPDK proteins are more similar to those in thermophilic *Clostridia* (sequence identity $= 0.65 \pm 0.03$), than to those in mesophilic/psychrophilic *Bacilli* (sequence identity $= 0.17 \pm 0.05$). Phylogenetic analysis of the genes with additional thermophilic organisms (SOM Methods) suggests a likely HGT event between the thermophilic organisms (Figure 4B). The other two shared functions are carried out by proteins translated from mobile genetic elements (MGEs) that mediate the movement of DNA within genomes or between bacteria (22). Shared closely-related MGEs in distant organisms imply HGT (23). We thus suggest that *fusion*DB offers a fast and easy way to trace likely functionally necessary HGT events within niche-specific microbial communities.

In this work, we have highlighted the importance of environmental factors for microbial function, and demonstrated the capability of *fusion*DB to not only annotate functions, but also directly link function to environment. Although it was developed for mapping new microbial genomes, *fusion*DB also has the potential for microbiome

annotations. By mapping metagenome assemblies to *fusion*DB, both the functional and taxonomical annotations can be obtained. Moreover, our recent work (Zhu *et al.* 2017, Functional sequencing read annotation for high precision microbiome analysis, *submitted*) suggests that accurate functional annotations can also be obtained without assembly. We thus also expect to make *fusion*DB useful in this type of analyses in the near future.

CONCLUSIONS

*fusion*DB links microbial functional similarities and environmental preferences. Our analysis reveals environmental factors driving microbial functional diversification. By mapping new organisms to the reference functional space, our database offers a novel, fast, and simple way to detect core-function repertoires, unique functions, as well as traces of HGT. With more microbial genome sequencing and further manual curation of environmental metadata, we expect that *fusion*DB will become an integral part of microbial functional analysis protocols in the near future.

AVAILABILITY

*fusion*DB is publicly available at http://services.bromberglab.org/fusiondb/

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Drs Burkhard Rost (TU Munich), Max Haggblom, Tamar Barkay (both Rutgers), and Tom O. Delmont (U Chicago) for all help with interpreting our data and understanding the community needs. Big thanks to Yanran Wang and Dr Anton Molyboha (both Rutgers) for all discussions. We want to thank the anonymous reviewers for their thorough review and suggestions to improve this manuscript. We are also grateful to all those who deposit their data in public databases – fusionDB wouldn't be possible without them.

FUNDING

NSF CAREER Award [1553289 to Y.B. and C.Z.]; USDA-NIFA [1015:0228906]; TU München – Institute for Advanced Study Hans Fischer Fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme [291763 to Y.B. and Y.M.]; German Research Foundation (DFG) and the Technical University of Munich (TUM) in the framework of the Open Access Publishing Program.
Conflict of interest statement. None declared.

REFERENCES

1. Garrity,G.M., Boone,D.R. and Castenholz,R.W. (eds). (2001) *Bergey's Manual of Systematic Bacteriology*. 2nd edn. Springer, NY, Vol. **1**.
2. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
3. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
4. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
5. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
6. Pagani,I., Liolios,K., Jansson,J., Chen,I.M., Smirnova,T., Nosrat,B., Markowitz,V.M. and Kyrpides,N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, **40**, D571–D579.
7. Sun,W., Yu,G., Louie,T., Liu,T., Zhu,C., Xue,G. and Gao,P. (2015) From mesophilic to thermophilic digestion: the transitions of anaerobic bacterial, archaeal, and fungal community structures in sludge and manure samples. *Appl. Microbiol. Biotechnol.*, **99**, 10271–10282.
8. Zhu,C., Delmont,T.O., Vogel,T.M. and Bromberg,Y. (2015) Functional basis of microorganism classification. *PLoS Comput. Biol.*, **11**, e1004472.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, **37**, D26–D31.
10. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
11. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
12. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
13. Dongen,S.V. (2000) Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht.
14. Massey,F.J. (1951) The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Statist. Assoc.*, **46**, 68–78.
15. Kim,S.E., Moon,J.S., Choi,W.S., Lee,S.H. and Kim,S.U. (2012) Monitoring of horizontal gene transfer from agricultural microorganisms to soil bacteria and analysis of microbial community in soils. *J. Microbiol. Biotechnol.*, **22**, 563–566.
16. Liu,L., Chen,X., Skogerbo,G., Zhang,P., Chen,R., He,S. and Huang,D.W. (2012) The human microbiome: a hot spot of microbial horizontal gene transfer. *Genomics*, **100**, 265–270.
17. Saye,D.J., Ogunseitan,O., Sayler,G.S. and Miller,R.V. (1987) Potential for transduction of plasmids in a natural freshwater environment: effect of plasmid donor concentration and a natural microbial community on transduction in Pseudomonas aeruginosa. *Appl. Environ. Microbiol.*, **53**, 987–995.
18. Raymond,J. and Segre,D. (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science (New York, N.Y.)*, **311**, 1764–1767.
19. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
20. Lehmann,J., Solomon,D., Kinyangi,J., Dathe,L., Wirick,S. and Jacobsen,C. (2008) Spatial complexity of soil organic matter forms at nanometre scales. *Nat. Geosci.*, **1**, 238–242.
21. Chastain,C.J., Failing,C.J., Manandhar,L., Zimmerman,M.A., Lakner,M.M. and Nguyen,T.H.T. (2011) Functional evolution of C4 pyruvate,orthophosphate dikinase. *J. Exp. Bot.*, **62**, 3083–3091.
22. Frost,L.S., Leplae,R., Summers,A.O. and Toussaint,A. (2005) Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Micro*, **3**, 722–732.
23. Krupovic,M., Gonnet,M., Hania,W.B., Forterre,P. and Erauso,G. (2013) Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. *PLoS ONE*, **8**, e49044.

4.2 Functional sequencing read annotation for high precision microbiome analysis

4.2.1 Preface

Current research provides substantial evidence that the repertoire of functions embodied in microbiomes is strongly associated with disease state [89]. Therefore, the ability to uncover functional profiles of microbiome samples with high accuracy and efficiency is extremely beneficial. This is especially gaining attention for diagnostic purposes. In the majority of cases it is practically impossible to separate individual microorganisms within a microbiome community. Consequently, the most common approach to analyze microbial functionality is to sequence the entire metagenome instead. Functional annotation of the metagenome can be achieved with or without gene assembly. Commonly, the high number of unassembled reads, chimeric assemblies and issues encountered by gene finding tools are obstacles when choosing the latter route. Existing methods like MG-RAST [90] can access molecular functionality of the entire community. However, as annotations are usually based on homology, high frequency of short read lengths pose a problem. Those reads are likely to be functionally misannotated as they may be mapped to an incorrect sequence. Another, often neglected complication that causes faulty annotations is the inaccurate computational annotation of most genes in the compiled reference databases [91]. In summary, currently available pipelines for microbiome functional annotation are either lacking in precision or speed to be applicable for large scale analysis.

To address these issues we developed *mi-faser*, an extremely fast and accurate ($> 90\%$ precision) method for annotation of molecular functionality encoded in microbiome sequencing read data. To avoid erroneous annotation mappings, we compiled a new reference database from which annotations are transferred. We explicitly used only protein sequences with experimentally annotated molecular functions. *mi-faser* does not require time consuming assembly or error prone gene finding pre-processing steps. To further speed up the alignment of translated reads against the reference database without missing potential hits, *mi-faser* uses DIAMOND [92] instead of PSI-BLAST. The publicly available web service allows the user to process 10GB of reads in less then 20 minutes using our *clubber* load balancer (described in section 3.1) as back-end. In a soon to be released update, this time frame will be further reduced to about half. We built a comprehensive web interface, which allows for further in depth analyses. Once an input metagenome is functionally annotated, any subsets of retrieved enzymatic functions can be mapped to the KEGG database [93] to uncover common metabolic pathways. *mi-faser* includes means to compare functional profiles of microbiomes, either against each other or versus a set of reference metagenomes (*i.e.* sequenced microbiome of healthy individuals). The distances between function profiles can be easily visualized by the provided NMDS plots. This allows for direct comparison of functional abundance profiles between microbiome samples *i.e.* from healthy and disease-associated individuals. We demonstrated this ability analyzing microbial functions associated with CD using samples available in the Biobank popgen [94]. We used the *mi-faser* pipeline to process

11 microbiomes from individuals of the same extended family - two CD affected patients and nine first-degree relatives. The nine healthy individuals shared highly similar microbiome functional profiles, whereas the two affected patients exhibited distinct profiles from those of their healthy relatives. However, the two affected profiles also differed from each other, hinting at either different microbial pathogenesis mechanisms of CD or a diverse impact of the disease on microbiome functionality. Further we were able to identify functional signatures of individual-specific gut microbiome responses to a dietary intervention in children affected by Prader-Willi syndrome (PWS) [95]. Finally we uncovered previously unseen oil degradation-specific functions in metagenomic data collected from beach sands in different stages of oil contamination [96]. Overall, we developed an extremely fast method for microbiome function annotation which outperforms other approaches not only by processing speed but also by coverage at the same precision. The comprehensive web interface allows for large scale batch analysis and offers various means for further in-depth analysis.

The original *mi-faser* algorithm was developed by Chengsheng Zhu. Rewriting for optimization, paralellization and the standalone version was done by me. The *mi-faser* web service front- and back-end was developed by me. Analysis and evaluation was done by Chengsheng Zhu. The project was designed by Chengsheng Zhu and Yana Bromberg. The manuscript was drafted by Chengsheng Zhu and Yana Bromberg. *mi-faser* is available as Git repository [97] and is archived via DOI [98].

4.2.2 Journal article. Zhu, Miller & Marpaka et al., Journal of Nucleic Acids Research 2017

Supplementary material can be found online at [99]. The published article is attached below.

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1209

Functional sequencing read annotation for high precision microbiome analysis

Chengsheng Zhu[1,*], Maximilian Miller[1,2,3], Srinayani Marpaka[1], Pavel Vaysberg[1], Malte C. Rühlemann[4], Guojun Wu[5], Femke-Anouska Heinsen[4], Marie Tempel[6], Liping Zhao[1,5,7], Wolfgang Lieb[6], Andre Franke[4] and Yana Bromberg[1,8,9,*]

[1]Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Dr, New Brunswick, NJ 08873, USA, [2]Department for Bioinformatics and Computational Biology, Technische Universität München, Boltzmannstr. 3, 85748 Garching/Munich, Germany, [3]TUM Graduate School, Center of Doctoral Studies in Informatics and its Applications (CeDoSIA), Technische Universität München, 85748 Garching/Munich, Germany, [4]Institute of Clinical Molecular Biology, Kiel University, Kiel, Germany, [5]State Key Laboratory of Microbial Metabolism and Ministry of Education Key Laboratory of Systems Biomedicine, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China, [6]Institue of Epidemiology, Kiel University, Kiel, Germany, [7]Canadian Institute for Advanced Research, Toronto, Canada, [8]Department of Genetics, Rutgers University, Human Genetics Institute, Life Sciences Building, 145 Bevier Road, Piscataway, NJ 08854, USA and [9]Institute for Advanced Study, Technische Universität München (TUM-IAS), Lichtenbergstrasse 2 a, D-85748 Garching, Germany

Received May 30, 2017; Revised November 15, 2017; Editorial Decision November 16, 2017; Accepted November 27, 2017

ABSTRACT

The vast majority of microorganisms on Earth reside in often-inseparable environment-specific communities—microbiomes. Meta-genomic/-transcriptomic sequencing could reveal the otherwise inaccessible functionality of microbiomes. However, existing analytical approaches focus on attributing sequencing reads to known genes/genomes, often failing to make maximal use of available data. We created *faser (functional annotation of sequencing reads)*, an algorithm that is optimized to map reads to molecular functions encoded by the read-correspondent genes. The *mi-faser* microbiome analysis pipeline, combining *faser* with our manually curated reference database of protein functions, accurately annotates microbiome molecular functionality. *mi-faser*'s minutes-per-microbiome processing speed is significantly faster than that of other methods, allowing for large scale comparisons. Microbiome function vectors can be compared between different conditions to highlight environment-specific and/or time-dependent changes in functionality. Here, we identified previously unseen oil degradation-specific functions in BP oil-spill data, as well as functional signatures of individual-specific gut microbiome responses to a dietary intervention in children with Prader–Willi syndrome. Our method also revealed variability in Crohn's Disease patient microbiomes and clearly distinguished them from those of related healthy individuals. Our analysis highlighted the microbiome role in CD pathogenicity, demonstrating enrichment of patient microbiomes in functions that promote inflammation and that help bacteria survive it.

INTRODUCTION

Microorganisms inhabit every available niche of our planet, and our bodies are no exception. Microbes that survive and thrive in the environments at the extremes of temperature, pH, and chemical or radiation contamination possess unique molecular functions of high industrial, clinical, and bioremediation value. The human body microbiome critically impacts our health. For example, Crohn's disease (CD) is a multifactorial disorder resulting from the interplay of individual genetic susceptibility, the gastrointestinal (GI) microbiome and other environmental factors. Taxonomic surveys of the GI microbiome have revealed microbial community features that are unique to CD patients, e.g. overall loss of microbial diversity (1,2), as well as depletion and enrichment of certain bacterial taxa (3–6). Establishing whether these observed microbial community shifts contribute to pathogenesis or, instead, correlate with or result from the disease onset, requires understanding not only what are the microbes involved, but also what they do. Ear-

*To whom correspondence should be addressed. Email: czhu@bromberglab.org
Correspondence may also be addressed to Yana Bromberg. Tel: +1 848 932 5638; Fax: +1 848 932 8965; Email: yanab@rci.rutgers.edu

lier studies indicate that in association with CD, the microbiome molecular function potential is more consistently disturbed than taxonomic makeup (7). More thorough functional analyses, e.g. based on deep metagenomic sequencing, are necessary to elucidate these findings.

Metagenome functional annotation can be performed with or without genome assembly. If the reads can be assembled into large contigs, existing annotation pipelines, such as RAST (8) and IMG (9), can be applied. However, assembly is difficult and often plagued by a large fraction of unassembled reads or short length contigs, which belong to the minor microbiome members, and by chimeric assemblies, which are especially common for complex and highly diverse samples (see Sczyrba *et al.,* 2017, doi: https://doi.org/10.1101/099127). Downstream gene finding algorithms are further faced with incomplete and erroneously assembled sequences, complicating statistical model constructions. Read-based annotation, *e.g.* using a platform such as MG-RAST (10), can access molecular functionality of the entire community. However, reads are usually annotated via function transfer by homology that, due to the short read length, is lacking in precision. This inaccuracy is additionally compounded by the erroneous computational annotations of most genes in the reference databases (11).

Here, we compiled a gold standard set of reference proteins (GS), with experimentally annotated molecular functions. We further developed *faser* (functional annotation of sequencing reads), an algorithm that uses alignments of translated sequencing reads to full-length proteins to annotate read-'parent protein' molecular functionality. *faser* annotates reads with higher precision at higher resolution, *i.e.* more specific functionality, than BLAST or PSI-BLAST. In a benchmark test, the functional annotations produced by the combination of the *faser* algorithm with the GS database were 12% more accurate than MG-RAST. Note that this performance may be an overestimate because the benchmark metagenome included the GS database. However, when GS was replaced with md5nr, MG-RAST's reference database, *faser* annotated 20% more reads than MG-RAST at a comparable precision level. These results illustrate that the GS and *faser* combination improves on MG-RAST capabilities.

Our *mi-faser* pipeline implementation (Figure 1), combining *faser* and GS, is highly parallelized, making use of all available compute cores and processing a (∼10GB/70M read) meta-genomic/-transcriptomic file in under half an hour (using 400 compute cores, on average). Note that if multiple microbiomes are submitted for annotation in parallel, the time scales favourably; in testing, 17 metagenomes were processed within 66 minutes. *mi-faser* results for all microbiomes analysed in this manuscript are available at http://services.bromberglab.org/mifaser/results/example. The standalone version of the pipeline, along with the *mi-faser* source code, is available at https://bitbucket.org/bromberglab/mifaser, as well as on the bromberglab website.

We applied our *mi-faser* to metagenomic data collected from beach sands in different stages of oil contamination (12). Here, *mi-faser* was able to identify oil degradation functionality that was missed by MG-RAST. We further performed large-scale analysis of 68 metagenomic datasets

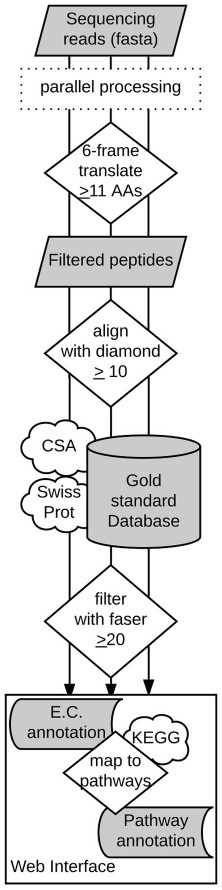

Figure 1. *mi-faser* pipeline. *mi-faser* is parallelized and runs a load balancer to submit jobs to available [1–2000] compute cores. Under normal functioning conditions (∼400 available cores, on average), it takes ∼30 min to process a single (10G/70M read) meta-genome/-transcriptome.

from a study of dietary intervention in Prader-Willi syndrome (PWS) affected obese children. Each dataset was processed in approximately 16 minutes, highlighting *mi-faser*'s processing speed. We identified previously unseen individual-specific patterns in microbiome changes induced by the treatment. Finally, we also analyzed the GI tract microbiome data from Crohn's Disease (CD) patients and their relatives. We found the microbiome functional profiles were similar between healthy individuals but different across patients and between patients and their healthy relatives. Particularly, our analysis revealed that CD patients' microbiomes were enriched in functions that help bacteria survive inflammation, i.e. glutathione metabolism and RNA modification, and in functions that cause inflam-

mation, i.e. lipopolysaccharide and acetaldehyde production. These results suggest the microbiome's role in CD-associated pathogenicity.

MATERIALS AND METHODS

Datasets

To compile the *PE1-set*, we extracted from SwissProt (Oct. 2015) (13) proteins that are (i) bacterial, (ii) with evidence of existence, i.e. SwissProt protein evidence is 1, and (iii) explicitly assigned an E.C. (Enzyme Commission) number (14); note that we excluded proteins with incomplete annotations, e.g. 1.1.1.-, as well as those with multiple annotations. From the *PE1-set*, we further extracted proteins whose functions are experimentally verified (Evidence = 'any experimental assertion'; *EXP-set*).

From the Catalytic Site Atlas database (*CSA-set*) (15) we extracted all proteins that had literature-based annotations. We identified the overlap between the *PE1-set* and these proteins, and defined our gold-standard dataset (*GS-set*; Supplementary Data 1) as the combination of *CSA-set* and *EXP-set*, with 100% identical sequences removed.

For each protein of the *PE1-set* and *GS-set*, we extracted the corresponding gene from ENA (European Nucleotide Archive) (16) (including 5′ UTR and 3′ UTR) and randomly generated 10 DNA reads (50–250 nucleotides) that overlap by at least one nucleotide of the coding region. We further performed 6-frame translations of the reads and excluded peptides shorter than 11 amino acids. We defined the corresponding peptide collection as *rPE1-set* and *rGS-set*.

We downloaded from MG-RAST the *md5nr* database and defined its proteins as the *md5nr-set*.

We obtained six beach sand metagenomes from a previous study of the Deepwater Horizon oil spill (12). Here, metagenomic DNA was sequenced using Illumina MiSeq with paired-end strategy to produce 151 bp reads. The samples reside in NCBI (BioProject PRJNA260285), including (i) pre-oil phase samples, OS-S1 (SRX692936) and OS-S2 (SRX695904), (ii) oil phase samples, OS-A (SRX696142) and OS-B (SRX696240) and (iii) post-oil recovered phase samples, OS-I600 (SRX696250) and OS-I606 (SRX696254).

We also obtained 68 gut metagenomic sequencing datasets (SRA (17) accession number SRP045211) from a study of dietary intervention in Chinese children affected by PWS (18). Fecal DNA samples before and after the treatment ($n = 17$, at Day 0, 30, 60 and 90) were sequenced using Illumina HiSeq 2000 with paired-end strategy to produce 100 bp reads. The quality control was performed as described in the previous study (18).

We additionally obtained 11 human gut (fecal) microbiome samples from a family affected by CD from the PopGen biobank (Schleswig-Holstein, Germany; accessible via a Material Data Access Form. Information and application procedures for data access can be found at http://www.uksh.de/p2n/Information±for±Researchers.html). Of these, nine members were self-reported as healthy and two were affected. Metagenomic data were generated using the Illumina Nextera DNA Library Prep Kit and sequenced 2 × 125 bp on an Illumina HiSeq2500. In total, 424.8 million paired-end reads were generated with a median number of 38.9 million read pairs per sample. Adapter trimming was performed using Trimmomatic (19) in paired-end mode, discarding reads shorter than 60 bp. Quality filtering was done using Sickle (20) run in paired-end mode, with a quality threshold of 20 and a minimum length of 60 bp. To remove contaminating host sequences from the dataset, DeconSeq (v0.4.3) (21) was run with the human reference genome (GRCh38) as database. Only read-pairs where both sequences survived quality control were retained. On average 11.76% of raw reads were discarded, leaving 374.8 million read pairs for downstream analysis.

faser curve optimization

We PSI-BLASTed the *rGS-set* against the *GS-set* (parameters: evalue $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3; max_target_seqs 1 000 000), excluding self-hits, i.e. peptide hits of their 'parent' proteins. For any peptide, functional annotation (E.C. number) was inherited from the 'parent' protein; one nucleotide overlap required to transfer annotation. A peptide-protein alignment is considered positive if the functional annotations of the peptide and the aligned protein match exactly at the selected number of E.C. digits, and negative otherwise. Any given alignment can be plotted in an L (alignment length) *vs.* Id (alignment sequence identity) two-dimensional space. Further, an exponential decay curve (as for HSSP calculations, (22)) can be used to identify the alignments in this space as true positives (alignments of peptides to proteins of identical function that fall above or on the curve), false positives (different functions above or on the curve), true negatives (different functions below the curve) and false negatives (identical functions below the curve). From these values, we calculated precision (positive accuracy; Equation 1) and recall (positive coverage; Equation 2) for different curve parameters (a and b in Equation 4), optimizing the latter to fit a curve best separating positive from negative alignments in terms of the highest F-measure (Equation 3).

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

$$F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

$$b \times L^{-a \times \left(1 + e^{-\frac{L}{1000}}\right)} \quad (4)$$

To avoid overestimating performance of *faser*, we clustered the *GS-set* with CD-hit at 40% sequence identity and split the clusters into ten subsets. We further optimized *faser* curve parameters in 10-fold cross-validation, i.e. we iteratively optimized the curve on nine subsets and tested it on the remaining one, repeating this process 10 times for a different subset as the test set. We evaluated the performance reported here by summing the numbers of true and false positives and negatives in each test set. As all ten curves were

very similar in parameters, we took the average of these to establish the final *faser* curve.

To summarise, the *faser* curve is meant to predict from a peptide-protein alignment, whether the 'parent' protein of the peptide and the aligned protein share the same function (E.C. annotation). Additionally, the distance of the alignment point to the curve along the sequence identity (Id) axis indicates the reliability of the prediction.

Evaluating *faser* using DIAMOND results

We extracted the proteins from the *GS-set* and *md5nr-set* that had identical UniProt IDs. We performed searches against the *md5nr* database using PSI-BLAST (parameters: evalue $1e^{-3}$; inclusion ethresh $1e^{-10}$; num iterations 3; max_target_seqs 1 000 000), BLASTP (parameters: evalue $1e^{-3}$; max_target_seqs 1 000 000), and DIAMOND (parameters: min-score 10; k 1 000 000). We further excluded from the results the alignments to subject proteins that were not in the overlap set. We compared the *faser* values calculated from the results of different alignment algorithms by performing a 100-fold bootstrap, sampling ∼20% of the results at each iteration. Note that we used the bootstrap approach to assess the consistency of the observed performance differences.

Comparison to other methods

We submitted the artificial metagenome as well as the six sand metagenomes for processing to MG-RAST via its website and downloaded the resulting function annotations via the MG-RAST API (23). We used the KEGG (24) annotations from the *md5nr* database to establish the annotated E.C.s. Note that although proteins can carry out multiple functions, in this study we, conservatively, only included proteins with unique and complete E.C. annotations; i.e. we excluded proteins with incomplete or multiple E.C. annotations.

We compared different database/algorithm combinations for the annotation of the same sample (Supplementary Figure S2). The Venn diagrams of the numbers of E.C.s annotated by different such combinations were generated by Venny (25). When comparing across sand metagenome samples from different phases, sample-specific E.C.s were removed as uninformative (<1% of total E.C.s in both cases). The correlation between samples was calculated with Spearman's rho, ρ, offered in the R package, Hmisc (26).

Two other tools, Fun4Me and ShotMAP, were installed locally and run on the artificial metagenome with default parameters; for both, we compared the precision of the methods (Equation 1) as well as the number of correctly annotated reads.

Functional analysis of PWS dietary intervention metagenomes

We performed NMDS (Non-metric multidimensional scaling) (27) analysis and the subsequent permanova test using the Vegan R package (28) and calculated the Euclidean distance between samples in the NMDS graph. Within the untreated Day0 group of samples, we identified outliers (individuals with inter-sample distance two standard deviations away from the average distance; 3% of all distances). All time-point samples of these individuals were removed from subsequent analysis. For remaining individuals, we compared the distances within each time-point group, as well as the distances of all the time points from Day0 for every individual separately.

Functional analysis of CD metagenomes

As described above, NMDS analysis (Shepard plot in Supplementary Figure S10), along with the subsequent permanova test was carried out using the Vegan R package (28). From the distributions of E.C.s in the microbiomes of healthy individuals, we calculated the 'confidence range' for each E.C. as Q1 – 3*IQR (three interquartile ranges below the first quartile) to Q3 + 3*IQR (three interquartile ranges above the third quartile). Patient E.C.s that fell outside this range were identified as significantly depleted or enriched, respectively. Pathway analysis was performed with the KEGG Mapper tool (24). Jaccard Index was calculated as the size of intersection divided by the size of union of the two sample sets.

RESULTS AND DISCUSSION

Few proteins have experimentally verified function annotation

Among the 332 193 bacterial proteins in SwissProt (Oct. 2015) (13,29), only 18 240 (∼5%) are annotated as *existing* with evidence at protein level. Of these, we extracted 5 965 that have unique (one per protein) and explicit (all four digits) Enzyme Commission (E.C.) annotations (*PE1-set*; Materials and Methods). From our *PE1-set*, we further selected proteins whose *functions* were experimentally verified, as noted in the Catalytic Site Atlas (*CSA-set*) (15) or SwissProt (*EXP-set*) (13,29). After filtering, our set contained 2 848 (2 810 non-redundant at 100% sequence identity; *GS-set*) bacterial proteins of experimentally verified function. Note that analysis of available mass-spectrometry databases (30,31) is likely to retrieve a much larger set of verified existing proteins; however, these are not yet experimentally annotated for molecular functionality. Thus, our collection is the cleanest available dataset of functional annotations; i.e. functional annotations in public databases are usually based on (many rounds of) function transfer by homology and are, as such, often questionable.

faser is more accurate for function transfer by homology than PSI-BLAST

We created artificial reads from the gene nucleotide sequences corresponding to the proteins in *GS-set* and *PE1-set* (6-frame translated to peptides, *rGS-set* and *rPE1-set*, Materials and Methods). We further PSI-BLASTed (32) the *rGS-set* against *GS-set,* excluding self-hits, to determine the equation of the curve (Equation 5) separating the correct alignments (same function) from the incorrect ones (different functions) in the *L* (alignment length) versus *Id* (sequence identity) space. Our approach was modeled after the HSSP metric for function transfer between full-length proteins (22,33). We optimized the curve parameters to

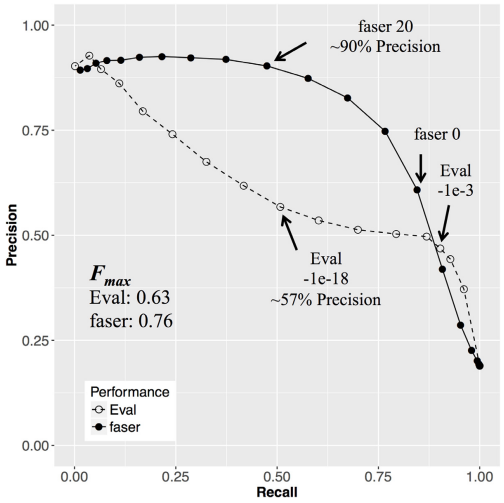

Figure 2. *faser* outperforms PSI-BLAST in annotating read functions. At most cutoffs, *faser* (filled circles) is more precise than PSI-BLAST (empty circles). For example, for nearly half the reads, it provides as much as 90% annotation accuracy as compared to 57% attained by PSI-BLAST (arrows at *faser* score = 20 and e-value = e^{-18}). At the default cutoff of 0, *faser* attains similar accuracy as PSI-BLAST at e-value = e^{-18}, but for ∼35% more reads.

maximize the *F* measure (Materials and Methods), representative of best separation of peptide–protein alignments of the same function (E.C. annotation) from those of different functions (Methods). Thus, if a given alignment is above the curve, the 'parent protein' of the peptide and the aligned reference protein are predicted to share function. The *faser* score (the distance from the curve along the *Id* axis) indicates the reliability of such predictions. This measure clearly outperforms PSI-BLAST e-value in annotating function (F_{max} of 0.76 versus 0.63, respectively; Equation (3), the highest *F* measure as in (34); recall in Figure 2 was calculated with the background of all PSI-BLAST results at e-value = 10^{-3}). For example, at recall levels of ∼50%, the *faser* score (=20) is nearly 90% accurate, which is >30% more than e-value (=10^{-18}; Figure 2). E-value reaches ∼90% precision at cut-offs <10^{-36}, which corresponds to recall of <7% (Figure 2).

The number of matching E.C. digits reflects the level of resolution of function annotation; i.e. proteins that share only the first three E.C. digits have similar functions with slight differences. For example, both 1.1.1.1 and 1.1.1.2 are alcohol dehydrogenases, but with different electron acceptors: NAD+ and NADP+, respectively. PSI-BLAST exhibits comparable performance to *faser* when matching the first three E.C. digits (Supplementary Figure S1A), but fails to differentiate functions at the fourth digit resolution level, producing a large number of false positives (Figure 2). *faser* resolves the fourth E.C. digit at >90% precision with >40% recall. At all cut-offs, when compared to PSI-BLAST, *faser* consistently offers as much as ∼50% higher recall at same precision level and up to ∼25% higher precision at same recall level (Figure 2).

Table 1. Artificial metagenome (rPE1-set) annotation by F_G, F_M and M_M

	F_G	F_M	M_M
Annotated reads	34 851	48 481	30 800
Multi-E.C. reads[a]	1004	11 373	200
Erroneously annotated reads	416	5705	4237
Correctly annotated reads	33 431	31 103	26 363
Precision	99%	85%	86%

[a]Reads with multiple E.C. annotations were excluded from the analysis.

$$\text{faser score} = \begin{cases} -100, & L < 11 \\ Id - 352.3L^{-0.302\times\left(1+e^{-\frac{L}{1000}}\right)}, & L \geq 11 \end{cases} \quad (5)$$

Note that a previous study has shown that PSI-BLAST is not necessarily the best alignment method for function transfer, e.g. it was inferior to BLAST (34). Although faser was developed using PSI-BLAST, it can also be calculated via other alignment mechanisms. To alleviate the long alignment runtimes, we exhaustively tested our options (including comparing BLAST performance to PSI-BLAST) and ended up switching to DIAMOND (35) (Supplementary Text S1).

faser outperforms MG-RAST

We compared *faser* performance to that of MG-RAST (10), one of the most popular public metagenome annotation platforms. We considered both algorithm and database levels using the: (i) *faser* algorithm with the *GS-set* database (F_G, the *mi-faser* pipeline); (ii) *faser* algorithm with the *md5nr* database (36) (F_M; *faser*-md5nr); (iii) MG-RAST algorithm with *md5nr* database (M_M, the MG-RAST pipeline) (Supplementary Figure S2; Methods). Note that we could not run the MG-RAST algorithm with the *GS-set* database because the MG-RAST developers advised against it, citing complicated installation.

When the *rPE1-set* is used as the artificial metagenome, the F_G and M_M annotations are significantly different (Table 1), although both pipelines annotate a similar number of reads (Figure 3A). This variation in performance is not biased toward any specific E.C. class (Supplementary Figure S3). Note that the *rPE1-set* is a superset of *GS-set*, which likely contributes to the improved performance of the F_G pipeline. The differences between F_G and M_M annotations (Figure 3B, first column) stem from the differences between the databases (*GS-set* vs. *md5nr*) and/or algorithms (*faser* versus MG-RAST). The divergence between F_G and F_M annotations (Figure 3B, second column) indicates that the database differences contribute significantly to the F_G/M_M variation. Note that this difference is not surprising as the *GS-set* and *md5nr* share only 779 E.C.s (62% and 29%, respectively).

The comparison between F_M and M_M results is more interesting (Figure 3B, third column), as it highlights the differences between the *faser* and MG-RAST algorithms. Using the same *md5nr* database, *faser* (F_M) annotated ∼20% more reads than MG-RAST (M_M, Figure 3A) with comparable precision (Table 1). Note that the precision reported in

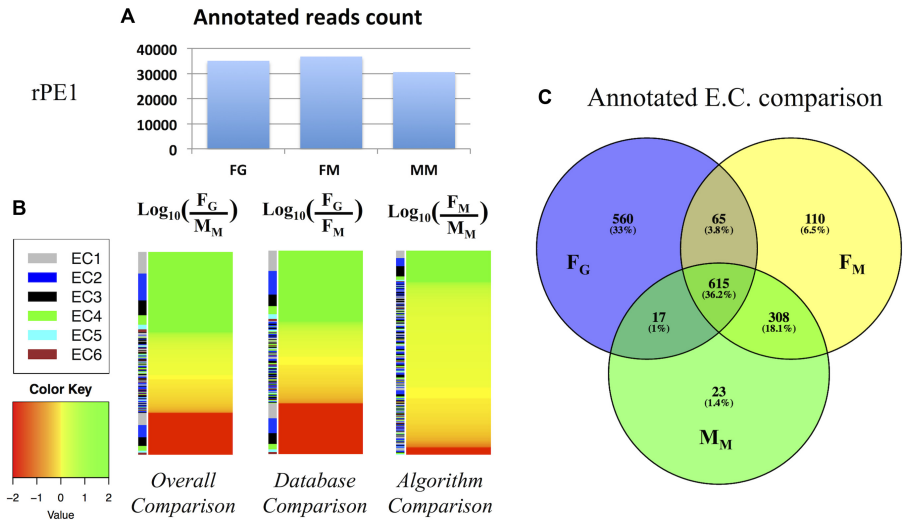

Figure 3. The *faser* algorithm in combination with the GS database annotates the artificial metagenome functions in a manner complementary to MG-RAST. (**A**) The number of reads annotated by each combination of algorithms and databases; (**B**) the read abundance by E.C. annotated via each combination of algorithm/database; (**C**) the total E.C. count annotated via each combination of algorithm/database.

these comparisons is affected by the misannotation (∼14%), i.e. UniProt proteins in both the *GS-set* and *md5nr* annotated with different E.C. numbers – a finding, which is in line with a previous study (11). F_M and M_M identified 923 E.C.s in common, while 175 and 40 E.C.s were uniquely identified by *faser* and MG-RAST, respectively (Figure 3C). In other words, for the same artificial metagenome, *faser* annotates ∼14% more functions (E.C.s) than MG-RAST algorithms. After exclusion of the database-specific E.C.s, the database impact was reduced (F_G/F_M, Supplementary Figure S4), yet we still observed substantial F_G/M_M differences largely due to the *faser vs.* MG-RAST algorithms. Notably, F_M still annotates ∼8% more functions than M_M (Supplementary Figure S4).

To summarize, the *faser* method comprises an exponential decay curve separating the two-dimensional space of alignment length versus sequence identity into 'same function' and 'different functions' peptide–protein alignments. The distance from a given alignment to the curve along the sequence identity axis is the final *faser* score. Implicitly, *faser* tries to capture homology of the peptide's 'parent' protein to the subject protein of the alignment. In *faser* development we used the database of experimentally described proteins (*GS-set*) to optimize and evaluate performance. We continue to use the GS database in the *mi-faser* implementation. However, *faser* alignment scoring can be applied to any other database as well. Note that we set the default cut-off of *faser* score at 20 for high precision (90%).

We further extended the comparison of the annotation methods to six metagenomic samples from the Deepwater Horizon oil spill beach sand study (12) (Methods). Note that in this real-life case, there was no 'correct' annotation to use for comparing annotation results. However, it appears that F_M and M_M results are orthogonal. For example, for OS-A (*oil phase*) F_M annotated >50% more reads than M_M (Supplementary Figure S5A); moreover, there were 220 E.C.s unique to F_M and 42 E.C.s unique to M_M (Supplementary Figure S5C). Annotation of other samples followed a similar pattern. Database differences resulted in a significant disparity between the number of reads annotated in each sample by F_G and M_M (e.g. Supplementary Figure S5B). However, both pipelines agreed that: (i) samples taken in the same phase were highly functionally correlated (Supplementary Tables S1 and S2), (ii) samples in *oil phase* were functionally more correlated with samples in *recovered phase* than *pre-oil phase* (Supplementary Tables S1 and S2, which may indicate that the environment has not fully recovered from the contamination) and (iii) ∼20% of reads in all samples mapped to housekeeping functions (housekeeping E.C.s complied from (37)). This agreement across methods suggests that F_G reflects true variation in functionality between samples from a perspective complementary to M_M.

We further searched for functions enriched in *oil phase* metagenomes as compared to either *pre-oil* or *recovered phases*. F_G returned 909 E.C.s (65%, 588 E.C.s, are *GS-set* specific), while M_M returned 1 627 E.C.s (65%, 1 062 E.C.s, are *md5nr* specific). Note that even for the E.C.s present in both databases, F_G and M_M revealed considerable discrepancies in abundance fold-changes across phases; $\rho = 0.46$ (Spearman's rho) for *oil-to-recovered* phase and only $\rho = 0.09$ for *oil-to-pre-oil* phase (Supplementary Figure S6). We explored E.C.s annotated by F_G as highly enriched (≥5 times) in the *oil phase* as compared to other phases, yet un-

changed or even decreased by M_M. There are nine of these E.C.s in *oil-to-pre-oil* comparison and ten in *oil-to-recovered* comparison, with three E.C.s overlapping across comparisons; i.e. enriched in the *oil phase* as compared to either *pre-oil* or *recovered phases* (Supplementary Tables S3 and S4). Of the three overlapping E.C.s, two are particularly notable: 1.3.11.1 (catechol 1,2-dioxygenase) directly associates with BTEX (benzene, toluene, ethylbenzene and xylenes) degradation, while 1.8.99.1 (assimilatory sulfite reductase) is essential for sulfur reducing bacteria, known to degrade BTEX. Note that there were also three E.C.s annotated only by M_M that were enriched in the *oil phase*; however, we were not able to identify them as being directly related to oil degradation (Supplementary Tables S5 and S6).

We also compared our pipeline to two recently published metagenome annotation tools, Fun4Me (38) and ShotMAP (39), using the above-described artificial metagenome. Note that Fun4Me includes its own reference database, which cannot be changed on demand. ShotMAP allowed using our *GS-set* as reference. F_G correctly annotated 4 900 (17%) more reads than Fun4Me. Additionally, when the multi-EC-annotated reads were excluded, F_G attained 7% higher precision than Fun4Me (99% versus 92%, respectively; Supplementary Table S7). While results were not as striking, *faser* still outperformed ShotMAP (using our GS database) with 1 160 (4%) more correctly annotated reads and 2% higher precision (99% to 97%, respectively; Supplementary Table S7). Notably, the entire run took *mi-faser* (standalone version) 42 seconds, while Fun4Me required more than 25 minutes. The speed evaluation for ShotMAP was not possible via command-line due to installation issues, but the virtual machine implementation was able to finish in 3 minutes.

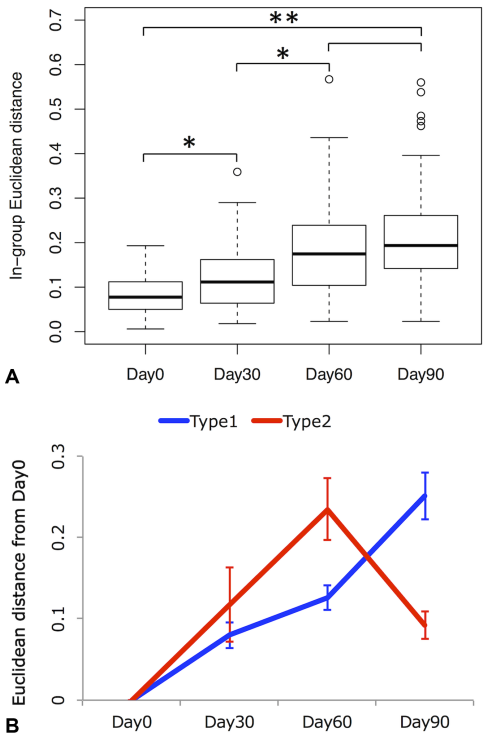

Figure 4. Functional capabilities of microbiomes of PWS patients shift in the course of dietary intervention. (**A**) The boxplot of Euclidean distance between samples of the same group (in-group distances), i.e. Day 0, 30, 60 or 90, on the NMDS diagram (Supplementary Figure S7). The in-group diversity increases significantly with time; * indicates *P*-value <1e−4; ** indicates *P*-value <1e−14; there is no significance between Day 60 and Day 90; *t*-test. (**B**) Two types of long term diet intervention effect on PWS patients: type1 individuals (GD02, GD03, GD15, GD40, GD42, GD43, GD47, GD51 and GD58) with gut microbiome functional capacity furthest removed from Day 0 at Day 90; type 2 individuals (GD04, GD18, GD52 and GD59) with gut microbiome functional capacity reversed at Day 90 toward their Day 0.

mi-faser facilitates novel functional discovery, while accelerating large-scale metagenomic analysis

The online service of *mi-faser* uses *clubber* (Cluster Load Balancer for Bioinformatics e-Resources (40)) for faster processing. To demonstrate our method's performance we obtained and analysed with *mi-faser,* 68 gut metagenomic datasets from a study of Chinese children affected by the PWS and treated via dietary intervention (Methods) (18). The analysis was automatically distributed to three clusters (640, 800, and 3400 cores with load-dependent access) by *clubber* via the *mi-faser* interface, for an average of 16 minutes of user-wait time (11.8 CPU hours) per metagenome.

Note that after NMDS Euclidean distance analysis of microbiomes of untreated individuals (Day0), four individuals (GD12, GD39, GD41 and GD50) were identified as outliers and removed (Methods). While we do not expect that all PWS-affected children share the same microbiome features, we felt that treatment effect and progression could be better evaluated from a narrow starting point.

For the remaining individuals ($n = 13$), it was clear that the dietary intervention significantly altered gut microbiome functionality (Supplementary Figure S7; Day 0 versus Day>0, *P*-value = 0.001, permanova test). More precisely, the intervention gradually increased the functional beta-diversity among the patients' gut microbiomes (Figure 4A; Supplementary Figure S7), which was in line with the results of the original study (18).

We further investigated the treatment progress of each patient using the Euclidean distance of the Day 30, 60 and 90 samples from the Day 0 sample of the same individual. Overall, the distances increased with the treatment progress (Day 30, 0.09 ± 0.02; Day 60, 0.16 ± 0.02; Day 90, 0.2 ± 0.03; Supplementary Figure S7), indicating the progressive changes of gut microbial functional potentials correlated with the diet time-line. Although Day 90 samples showed the highest dissimilarity from Day 0 samples in most cases, four patients (GD04, GD18, GD52 and GD59) reached the highest dissimilarity at Day 60, showing reversal of diet effects at Day 90 (Figure 4B). Follow-up studies on these differential trajectories could contribute to a more thorough

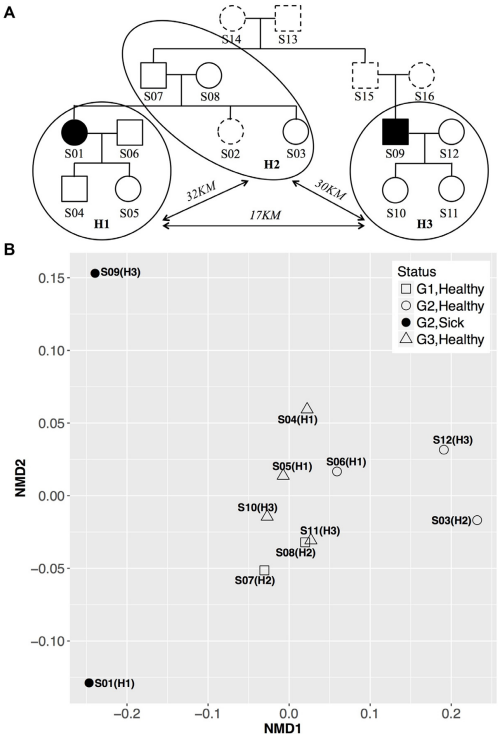

Figure 5. Functional capabilities of microbiomes of CD-affected individuals differ from healthy individuals and from each other. (**A**) The pedigree of the family in our study. Filled markers indicate CD affected individuals and empty markers are healthy individuals; dashed outline markers indicate individuals not included in this study. Individuals grouped by circles live in the same household. (**B**) The non-metric multidimensional scaling (NMDS) graph represents the distribution of individual microbiome functional profiles. Samples are labeled with identifiers (S1-S11) and household numbers (H1, H2, or H3, in parenthesis). Legend marker numbers (G1—grandparents, G2—parents, G3—children) represent generations, while marker shapes relate generations and CD status. Sick individuals (filled markers) localize separately from each other and from the cluster of healthy individuals (empty markers).

understanding of the effectiveness of the dietary intervention in PWS children.

mi-faser reveals microbial functions associated with Crohn's disease (CD)

We used our *mi-faser* pipeline (Figure 1) to analyse 11 microbiomes from individuals of the same extended family—two CD affected patients and nine first-degree relatives (Figure 5A). The members of this family live in three households that are no more than 32km apart from each other, with the CD affected individuals living in households 17 km away. No statistically significant distinction between functional profiles of individuals in the study was observed on the basis of generational or household differences (Figure 5B; *P*-value = 0.55 and 0.60 respectively, permanova test (41)). The nine healthy individuals shared highly similar microbiome functional profiles (rho, $\rho = 0.93 \pm 0.03$; Figure 5B; Supplementary Table S8). This finding is in line with previous studies that show that microbiome functional profiles across healthy individuals are more consistently maintained than bacterial species profiles (7). On the other hand, the microbiome functional profiles of the two CD patients are not only distinct from those of their healthy relatives (Figure 5B; $\rho = 0.75 \pm 0.11$; *P*-value = 0.02, permanova test), but also between themselves ($\rho = 0.72$; Figure 5B; Supplementary Table S8). Note that the former holds true even within the same household. In concert, these findings indicate that either there are different microbiome pathogenesis mechanisms of CD or that CD has a diverse impact on microbiome functionality.

We identified those E.C.s in our microbiomes whose abundance significantly changed in each patient compared to healthy individuals (Methods). S01 and S09 both have a large fraction of such E.C.s (45% and 31% respectively, sum of enriched and depleted, Supplementary Table S9). For example, nine E.C.s enriched in both S01 and S09 are annotated as rRNA methyltransferases (Supplementary Table S10), which are known to be essential for microbial response to environmental stresses (42). Another three E.C.s enriched in both patients are annotated as RNA pseudouridine synthase. RNAs with modified nucleotides, such as pseudouridine, have been shown to suppress host innate immune system (43). Thus, RNA modification may be an important bacterial strategy of surviving the CD-associated inflammation. We further explored these E.C.s to identify pathways uniquely altered in each patient; e.g. more than half of Biotin metabolism pathway E.C.s are altered in S01, while Xylene degradation is enriched only in S09 (Figure 6). There are also pathways that are similarly changed in both patients, i.e. they are enriched in the same E.C.s; for example, glutathione metabolism and lipopolysaccharide biosynthesis (Figure 6, Supplementary Figure S8). Given the distant microbiome functional profiles between S01 and S09 (Figure 5B), these similarities are unlikely to occur by chance. Glutathione is known to help bacteria survive oxidative stress, thus the enriched glutathione pathway could be a response to inflammation (44); a previous study has reported enrichment in abundance of genes associated with glutathione transportation in CD patients (7). However, the latter study (7) also suggested a decrease in propanoate and butanoate metabolism, both of which showed overall enrichment in S01 and S09 (Figure 6). Finally, to the best of our knowledge, the role of the lipopolysaccharide (LPS) biosynthesis pathway in CD patient microbiomes has not yet been reported. However, bacterial LPS is previously reported to increase intestinal tight junction permeability in mouse modules (45). Tight junctions normally form a selective seal between adjacent intestinal epithelial cells. Its increased permeability induces luminal pro-inflammatory molecules, resulting in sustained inflammation and tissue damage (46). Additionally, we also observed differences within individual pathway changes between patients. For example in the glycolysis/gluconeogenesis pathway, S01 is depleted in proteins necessary to convert glucose to pyru-

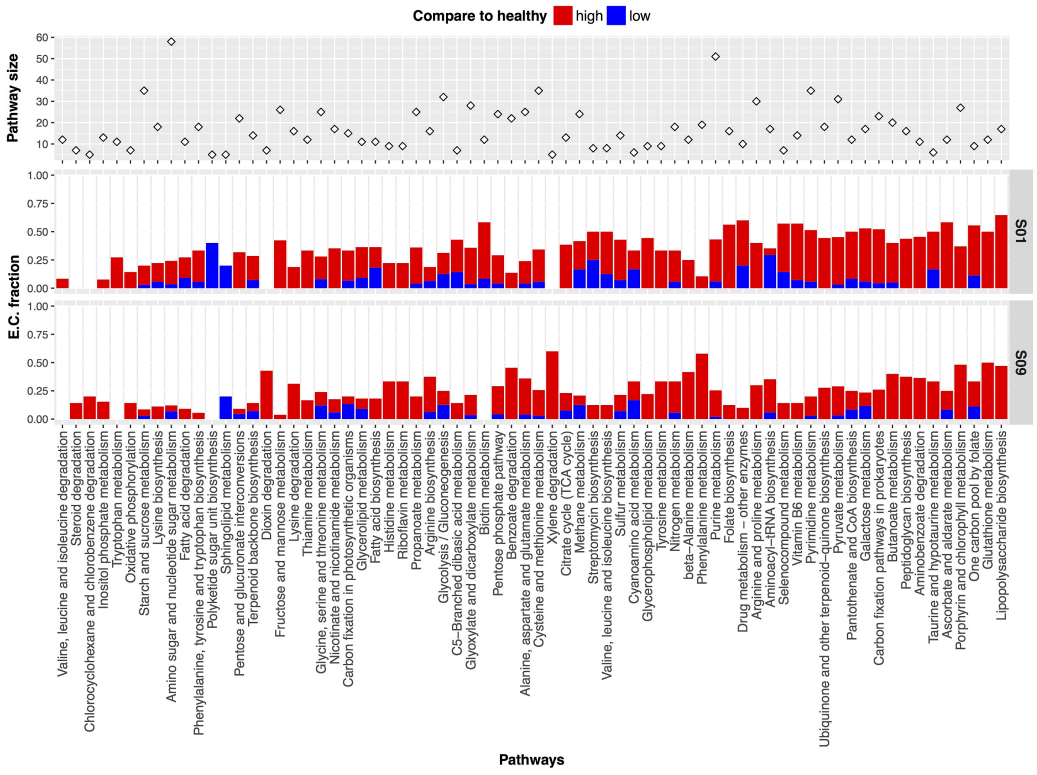

Figure 6. Enriched or depleted molecular pathways in microbiomes of CD-affected individuals. Changes in molecular pathways were obtained by counting the numbers of enriched or depleted E.C.s as compared to microbiome functional profiles of the healthy family members.

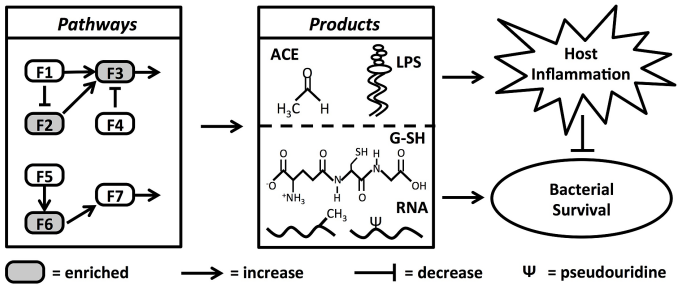

Figure 7. Microbial function shift in CD patients is involved in inflammation. Functions that are associated with inflammation inducers (acetaldehyde and lipopolysaccharide) are enriched in CD patient microbiomes, as are the functions that help bacteria survive inflammation conditions (glutathione metabolism, rRNA methytransferase and RNA pseudouridine synthase). Note that pathways above are toy examples for illustration purposes only; light gray nodes indicate enriched functions and white nodes indicate unchanged or undetected functions. Products are: ACE = acetaldehyde, LPS = lipopolysaccharide, G-SH = glutathione, RNA = RNAs with methylation or pseudouridine.

vate, while the pyruvate metabolism pathways are enriched (Supplementary Figure S9A). S09 shows a similar pattern, while enriching an alternative route from glyceraldehyde-3P to glycerate-3P (Supplementary Figure S9B). Interestingly, in both patients, most enriched E.C.s in pyruvate metabolism lead to acetaldehyde production (Supplementary Figure S9), a metabolite also known to induce tight junction disruption in intestinal epithelial cells (47). Thus, our result indicates the microbiome function shift in CD patients contributes to pathogenicity, while helps the bacteria survive host inflammation (Figure 7).

CONCLUSION

In this study, we compiled a 'clean' protein dataset with experimentally confirmed E.C. annotations (gold standard, GS-set), and trained the *faser* algorithm to optimise transfer of function annotation from reference proteins to short peptides translated from sequencing reads. The *faser* algorithm significantly outperforms PSI-BLAST in differentiating functions at high-resolution levels. It also offers ∼20% more annotations at comparable precision levels than the function annotation algorithm of MG-RAST. The (highly-parallelized and fast) *mi-faser* pipeline (*faser* in combination with GS*)* was able to identify, in BP oil spill data, unique candidate functions associated with oil-degradation, which were missed by the MG-RAST pipeline. Analysis of 68 metagenomic datasets from a dietary intervention study in PWS patients highlighted previously unseen individual-specific trajectories of functional changes in the gut microbiomes. Our pipeline also revealed that gastrointestinal microbiomes of related CD patients are functionally very different. We observed two types of functions enriched in CD patients: those that cause inflammation and those that help bacteria survive inflammatory stress; these may highlight the possible role of the microbiome in CD pathogenicity. Note that all *mi-faser* annotations, although highly informative, are based on the proteins making up the, currently limited, *GS-set*. On the other hand, *faser* itself is a robust read annotation algorithm that can be used with any reference database supplied. We also expect the growth in the number of proteins with experimentally verified functions to make our approach even more powerful in the near future.

AVAILABILITY

mi-faser is available online at http://services.bromberglab.org/mifaser/.

The standalone version of the pipeline, along with the *mi-faser* source code, is available at https://bitbucket.org/bromberglab/mifaser. The DOI for the source code used in this manuscript is https://doi.org/10.5281/zenodo.1045582, and the DOI for the current GS database is https://doi.org/10.5281/zenodo.1048268.

The fasta file for *GS-set* is available at http://bromberglab.org/sites/default/files/SOM_Data1_gold_standard.fasta.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Tamar Barkay, Max Haggblom, Yannick Mahlich, Alexandra Pushkar, Huan Qiu and Yanran Wang (all Rutgers University, New Brunswick, NJ) for many discussions and manuscript review. We thank Bill Abbott and Prentice Bisbal (both Rutgers) for technical support. We are grateful to Sonakshi Bhattacharjee (Technical University of Munich, TUM) for advice on the user interface. Particular thanks are due to Burkhard Rost (TUM) for his hospitality and valuable discussions. Last but not least, we thank all those who deposit their experimental data in public databases and those who maintain these databases.

Author contributions: C.Z. and Y.B. conceived and designed the experiments. C.Z., S.M. and P.V. performed the experiments. C.Z. and Y.B. analyzed the data. M.M. implemented the web service and made *mi-faser* publically available. Y.B. contributed materials/analysis tools. M.C.R., F.-A.H., M.T., W.L. and A.F. provided the Crohn's Disease data. G.W. and L.Z. provided the Prader–Willi syndrome data. C.Z. and Y.B. wrote the paper. All authors have seen and approved the final manuscript.

FUNDING

NSF CAREER [1553289 to Y.B. and C.Z.]; TU Munchen – Institute for advanced study Hans Fischer fellowship, funded by the German Excellence Initiative and the EU Seventh Framework Programme [291763 to Y.B.]; NIH/NIGMS [U01 GM115486 to Y.B., M.M., C.Z.]; The CD data collection was supported by the biobank popgen (the popgen 2.0 network is financed by the German Ministry for Education and Research [01EY1103]; German Research Foundation (DFG) Excellence Cluster 306/2, 'Inflammation at Interfaces' and by German Federal Ministry of Education and Research (BMBF) project 'SysINFLAME' [01ZX1306A]. Funding for open access charge: German Research Foundation (DFG) and Technical University of Munich within the funding programme Open Access Publishing.

Conflict of interest statement. None declared.

REFERENCES

1. Manichanh,C., Rigottier-Gois,L., Bonnaud,E., Gloux,K., Pelletier,E., Frangeul,L., Nalin,R., Jarrin,C., Chardon,P., Marteau,P. *et al.* (2006) Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, **55**, 205–211.
2. Dicksved,J., Halfvarson,J., Rosenquist,M., Jarnerot,G., Tysk,C., Apajalahti,J., Engstrand,L. and Jansson,J.K. (2008) Molecular analysis of the gut microbiota of identical twins with Crohn's disease. *ISME J.*, **2**, 716–727.
3. Frank,D.N., St. Amand,A.L., Feldman,R.A., Boedeker,E.C., Harpaz,N. and Pace,N.R. (2007) Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 13780–13785.
4. Frank,D.N., Robertson,C.E., Hamm,C.M., Kpadeh,Z., Zhang,T., Chen,H., Zhu,W., Sartor,R.B., Boedeker,E.C., Harpaz,N. *et al.* (2011) Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflamm. Bowel Dis.*, **17**, 179–184.
5. Sokol,H., Seksik,P., Furet,J.P., Firmesse,O., Nion-Larmurier,I., Beaugerie,L., Cosnes,J., Corthier,G., Marteau,P. and Doré,J. (2009) Low counts of Faecalibacterium prausnitziiin colitis microbiota. *Inflamm. Bowel Dis.*, **15**, 1183–1189.

6. Martinez-Medina,M., Aldeguer,X., Lopez-Siles,M., González-Huix,F., López-Oliu,C., Dahbi,G., Blanco,J.E., Blanco,J., Garcia-Gil,J.L. and Darfeuille-Michaud,A. (2009) Molecular diversity of Escherichia coli in the human gut: New ecological evidence supporting the role of adherent-invasive E. coli (AIEC) in Crohn's disease. *Inflamm. Bowel Dis.*, **15**, 872–882.
7. Morgan,X.C., Tickle,T.L., Sokol,H., Gevers,D., Devaney,K.L., Ward,D.V., Reyes,J.A., Shah,S.A., LeLeiko,N., Snapper,S.B. *et al.* (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, **13**, R79.
8. Aziz,R.K., Bartels,D., Best,A.A., DeJongh,M., Disz,T., Edwards,R.A., Formsma,K., Gerdes,S., Glass,E.M., Kubal,M. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.
9. Markowitz,V.M., Chen,I.-M.A., Palaniappan,K., Chu,K., Szeto,E., Pillay,M., Ratner,A., Huang,J., Woyke,T., Huntemann,M. *et al.* (2014) IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.*, **42**, D560–D567.
10. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
11. Schnoes,A.M., Brown,S.D., Dodevski,I. and Babbitt,P.C. (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput. Biol.*, **5**, e1000605.
12. Rodriguez-R,L.M., Overholt,W.A., Hagan,C., Huettel,M., Kostka,J.E. and Konstantinidis,K.T. (2015) Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. *ISME J.*, **9**, 1928–1940.
13. Bairoch,A., Boeckmann,B., Ferro,S. and Gasteiger,E. (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinformatics*, **5**, 39–55.
14. EC,W. (1992) *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Academic Press, San Diego.
15. Furnham,N., Holliday,G.L., de Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.
16. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tárraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.*, **39**, D28–D31.
17. Leinonen,R., Sugawara,H., Shumway,M. and on behalf of the International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
18. Zhang,C., Yin,A., Li,H., Wang,R., Wu,G., Shen,J., Zhang,M., Wang,L., Hou,Y., Ouyang,H. *et al.* (2015) Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*, **2**, 968–984.
19. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
20. Joshi,N.A. and Fass,J.N. (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. http://www.citeulike.org/user/mvermaat/article/13260426.
21. Schmieder,R. and Edwards,R. (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE*, **6**, e17288.
22. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **318**, 595–608.
23. Wilke,A., Bischof,J., Harrison,T., Brettin,T., D'Souza,M., Gerlach,W., Matthews,H., Paczian,T., Wilkening,J., Glass,E.M. *et al.* (2015) A RESTful API for accessing microbial community data for MG-RAST. *PLOS Comput. Biol.*, **11**, e1004008.
24. Kanehisa,M., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.
25. Oliveros,J.C. (2007) VENNY. An interactive tool for comparing lists with Venn Diagrams. http://www.citeulike.org/user/hroest/article/6994833.
26. Harrell,F.E. Jr (2016) *Hmisc: Harrell Miscellaneous*. R package version 3.17-4.
27. Kruskal,J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, **29**, 115–129.
28. Oksanen,J., Friendly,M., Kindt,R., Legendre,P., McGlinn,D., Minchin,P.R., O'Hara,R.B., Simpson,G.L., Solymos,P., Henry,M. *et al.* (2016) *vegan: Community Ecology Package*. R package version 2.4-0.
29. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol. (Clifton, N.J.)*, **1374**, 23–54.
30. Lam,H., Deutsch,E.W., Eddes,J.S., Eng,J.K., King,N., Stein,S.E. and Aebersold,R. (2007) Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, **7**, 655–667.
31. Stein,S. (2012) Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.*, **84**, 7274–7282.
32. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
33. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **25**, 226–230.
34. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
35. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
36. Wilke,A., Harrison,T., Wilkening,J., Field,D., Glass,E.M., Kyrpides,N., Mavrommatis,K. and Meyer,F. (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. *BMC Bioinformatics*, **13**, 1–5.
37. Gil,R., Silva,F.J., Pereto,J. and Moya,A. (2004) Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.: MMBR*, **68**, 518–537.
38. Sharifi,F. and Ye,Y. (2017) From gene annotation to function prediction for metagenomics. *Methods Mol. Biol. (Clifton, N.J.)*, **1611**, 27–34.
39. Nayfach,S., Bradley,P.H., Wyman,S.K., Laurent,T.J., Williams,A., Eisen,J.A., Pollard,K.S. and Sharpton,T.J. (2015) Automated and accurate estimation of gene family abundance from shotgun metagenomes. *PLOS Comput. Biol.*, **11**, e1004573.
40. Miller,M., Zhu,C. and Bromberg,Y. (2017) clubber: removing the bioinformatics bottleneck in big data analyses. *J. Integrative Bioinformatics*, **14**, doi:10.1515/jib-2017-0020.
41. Anderson,M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, **26**, 32–46.
42. Baldridge,K.C. and Contreras,L.M. (2014) Functional implications of ribosomal RNA methylation in response to environmental stress. *Crit. Rev. Biochem. Mol. Biol.*, **49**, 69–89.
43. Durbin,A.F., Wang,C., Marcotrigiano,J. and Gehrke,L. (2016) RNAs containing modified nucleotides fail to trigger RIG-I conformational changes for innate immune signaling. *mBio*, **7**, e00833-16.
44. Masip,L., Veeravalli,K. and Georgiou,G. (2006) The many faces of glutathione in bacteria. *Antioxid. Redox Signal.*, **8**, 753–762.
45. Guo,S., Al-Sadi,R., Said,H.M. and Ma,T.Y. (2013) Lipopolysaccharide causes an increase in intestinal tight junction permeability in vitro and in vivo by inducing enterocyte membrane expression and localization of TLR-4 and CD14. *Am. J. Pathol.*, **182**, 375–387.
46. Lee,S.H. (2015) Intestinal permeability regulation by tight junction: implication on inflammatory bowel diseases. *Intestinal Res.*, **13**, 11–18.
47. Atkinson,K.J. and Rao,R.K. (2001) Role of protein tyrosine phosphorylation in acetaldehyde-induced disruption of epithelial tight junctions. *Am. J. Physiol. - Gastrointest. Liver Physiol.*, **280**, G1280.

5 Conclusion

In this work, we explored new avenues for facilitating identification of molecular malfunction patterns linked to disease. We considered two important perspectives of functional alteration, coding genetic variation and microbiome changes. Gene sequence variants, leading to alteration of protein sequence, structure, and, ultimately, function, are often linked to disease. Similarly, changes in the diversity and/or levels of functionality of the human-associated microbiomes are often markers, and sometimes causes, of many pathologies.

Predicting genetic variant effects on protein function and, thus, potentially their relationship to disease is an endeavor taken up by many. In the first part of this thesis, we highlighted the limited progress and issues in the field of computational variant effect prediction over the last years. To address this problem, we introduced a new concept that characterizes protein sequence positions into two classes, tuneable *rheostat*s and on-off *toggle*s based on the distribution of substitution effects caused by nsSNPs (coding variants leading to single amino acid substitutions). We first showed that current computational predictors fail to accurately differentiate between non-neutral (functionally disruptive) and neutral mutations within each of the two classes. Note that it can expected that mutations in *toggle* positions are significantly more rare and more likely to be associated with Mendelian disorders, due to their overly deleterious nature. On the flip side, the bulk of common diseases is very likely to be due to some combinations of variants in rheostatic positions. To study these concepts in more detail, we developed a new model to predict the class of each protein residue using a state of the art *Machine Learning* approach. This model classifies residues into three classes *toggle*, *rheostat*, and *neutral* (where no mutation has an effect) using 14 sequence based features and reaching a combined accuracy of over 82%. Preliminary analyses of the distribution of residue classes for the entire set of human enzymes suggests that charged residues are, surprisingly, most interchangeable of the entire amino acid alphabet (*neutrals*). We observe other distinct patterns, *e.g.* that smaller aliphatic residues can, as expected, often be *rheostats* and cysteines act as *toggles*. It is interesting to note that proline, a residue that is usually considered to be immutable, is sometimes still a *rheostat*, suggesting that further insight is necessary to understand its role in specific proteins. We also observe differences in distributions across broad enzyme classes; particularly, oxidoreductases show distinct patterns of toggles and rheostats, which are different from all other enzyme classes. We suspect, that due to their ancient origins, importance to organism life and function, and corresponding ubiquitous presence, oxidoreductases are likely to allow for a larger spectrum of functional tuning (more *rheostats* than *neutrals* and *toggles*).

We intend to continue this work to build residue class-aware functional effect predictors, which can further be useful for larger disease gene evaluations and, eventually, for developing diagnostic/prognostic tools.

In the second part of this dissertation, I put forward a new tool- *clubber*, automated cluster load balancing software to facilitate *Big Data* analyses. We developed this tool to complement and enhance the efficiency of all the analytical methods described in this thesis. Integrating new tools into an already established and complicated workflow is a tedious task, which we simplified by equipping *clubber* with multiple interfaces that cover a wide range of applications. Aside from its user-friendliness, the main purpose of *clubber* is to bundle local and remote compute resources and distribute jobs with the goal of minimal queuing and processing times. Thus, it integrated seamlessly into our existing pipelines, including web services, and currently allows us to process a large amounts of user requests in an efficient and rapid manner. We expect *clubber* will become more and more useful as availability of large-scale shared compute resources available to researchers grows. For example, we are currently in the process of implementing our web-available pipelines on Jetstream resources, speeding up data processing and assuring redundancy/availability of our tools.

In the last part of this thesis I describe our approach to seeking out disease patterns in functions of human-associated microbes and microbiomes. We developed two methods to analyze of the microbial functional repertoires. The first, *fusion*DB is a novel database that comprises functional descriptors of 1,374 taxonomically distinct bacteria, annotated with available experimentally determined metadata. Each microbe is encoded as a set of functions carried out by the complete set of proteins it its genome encodes (its proteome) and individual microbes are connected via common functions. By mapping newly acquired disease-sample microbial genomes to the reference functional repertoires in *fusion*DB, we can highlight shared functionality and draw conclusions of the possible mechanisms of infection and pathogenesis. The second method, *mi-faser*, is a fast and accurate method for annotation of molecular functionality encoded in microbiome sequencing read data without the need for assembly or gene finding. *mi-faser* allows for the comparison of functional profiles between microbiomes, e.g. healthy and disease-associated, visualizing the differences via NMDS plot representations. This approach enables identifying functional patterns specifically related to disease.

The results from this dissertation contribute significantly to the field of variant effect prediction and functional microbial/microbiome analysis; both in the context of disease vs. healthy states. All computational models also boast drastically improved processing speeds. From web-service use observations since publication of individual tools, we suggest that usability and high efficiency strongly encourage utilization of our methods. Our objective is to improve these models even further, with an aim to finally translate these into a clinical setting.

Bibliography

[1] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001. URL: `https://www.ncbi.nlm.nih.gov/pubmed/11237011`, `doi:10.1038/35057062`.

[2] C. Genomes Project, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. URL: `http://www.ncbi.nlm.nih.gov/pubmed/26432245`, `doi:10.1038/nature15393`.

[3] R. Birnbaum and D. R. Weinberger. Genetic insights into the neurodevelopmental origins of schizophrenia. *Nat Rev Neurosci*, 18(12):727–740, 2017. URL: `https://www.ncbi.nlm.nih.gov/pubmed/29070826`, `doi:10.1038/nrn.2017.125`.

[4] F. Capon. The genetic basis of psoriasis. *Int J Mol Sci*, 18(12), 2017. URL: `https://www.ncbi.nlm.nih.gov/pubmed/29186830`, `doi:10.3390/ijms18122526`.

[5] M. Eslam, L. Valenti, and S. Romeo. Genetics and epigenetics of nafld and nash: Clinical impact. *J Hepatol*, 2017. URL: `https://www.ncbi.nlm.nih.gov/pubmed/29122391`, `doi:10.1016/j.jhep.2017.09.003`.

[6] E. Ierardi, C. Sorrentino, M. Principi, F. Giorgio, G. Losurdo, and A. Di Leo. Intestinal microbial metabolism of phosphatidylcholine: a novel insight in the cardiovascular risk scenario. *Hepatobiliary Surg Nutr*, 4(4):289–92, 2015. URL: `https://www.ncbi.nlm.nih.gov/pubmed/26312245`, `doi:10.3978/j.issn.2304-3881.2015.02.01`.

[7] M. Simren, G. Barbara, H. J. Flint, B. M. Spiegel, R. C. Spiller, S. Vanner, E. F. Verdu, P. J. Whorwell, E. G. Zoetendal, and C. Rome Foundation. Intestinal microbiota in functional bowel disorders: a rome foundation report. *Gut*, 62(1):159–76, 2013. URL: `https://www.ncbi.nlm.nih.gov/pubmed/22730468`, `doi:10.1136/gutjnl-2012-302167`.

[8] Y. Haberman, T. L. Tickle, P. J. Dexheimer, M. O. Kim, D. Tang, R. Karns, R. N. Baldassano, J. D. Noe, J. Rosh, J. Markowitz, M. B. Heyman, A. M. Griffiths, W. V. Crandall, D. R. Mack, S. S. Baker, C. Huttenhower, D. J. Keljo, J. S. Hyams, S. Kugathasan, T. D. Walters, B. Aronow, R. J. Xavier, D. Gevers, and L. A. Denson. Pediatric crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest*, 124(8):3617–33, 2014. URL: `https://www.ncbi.nlm.nih.gov/pubmed/25003194`, `doi:10.1172/JCI75436`.

[9] A. K. Benson. The gut microbiome-an emerging complex trait. *Nat Genet*, 48(11):1301–1302, 2016. URL: `https://www.ncbi.nlm.nih.gov/pubmed/27787511`, `doi:10.1038/ng.3707`.

[10] S. Bruse, M. Moreau, Y. Bromberg, J. H. Jang, N. Wang, H. Ha, M. Picchi, Y. Lin, R. J. Langley, C. Qualls, J. Klensney-Tait, J. Zabner, S. Leng, J. Mao, S. A. Belinsky, J. Xing, and T. Nyunoya. Whole exome sequencing identifies novel candidate genes that modify chronic obstructive pulmonary disease susceptibility. *Hum Genomics*, 10:1, 2016. URL: `http://www.ncbi.nlm.nih.gov/pubmed/26744305`, `doi:10.1186/s40246-015-0058-7`.

[11] D. Ellinghaus, H. Zhang, S. Zeissig, S. Lipinski, A. Till, T. Jiang, B. Stade, Y. Bromberg, E. Ellinghaus, A. Keller, M. A. Rivas, J. Skieceviciene, N. T. Doncheva, X. Liu, Q. Liu, F. Jiang, M. Forster, G. Mayr, M. Albrecht, R. Hasler, B. O. Boehm, J. Goodall, C. R. Berzuini, J. Lee, V. Andersen, U. Vogel, L. Kupcinskas, M. Kayser, M. Krawczak, S. Nikolaus, R. K. Weersma, C. Y. Ponsioen, M. Sans, C. Wijmenga, D. P. Strachan, W. L. McArdle, S. Vermeire, P. Rutgeerts, J. D. Sanderson, C. G. Mathew, M. H. Vatn, J. Wang, M. M. Nothen, R. H. Duerr, C. Buning, S. Brand, J. Glas, J. Winkelmann, T. Illig, A. Latiano, V. Annese, J. Halfvarson, M. D'Amato, M. J. Daly, M. Nothnagel, T. H. Karlsen, S. Subramani, P. Rosenstiel, S. Schreiber, M. Parkes, and A. Franke. Association between variants of prdm1 and ndp52 and crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*, 145(2):339–47, 2013. URL: `http://www.ncbi.nlm.nih.gov/pubmed/23624108`, `doi:10.1053/j.gastro.2013.04.040`.

[12] T. N. Turner, F. Hormozdiari, M. H. Duyzend, S. A. McClymont, P. W. Hook, I. Iossifov, A. Raja, C. Baker, K. Hoekzema, H. A. Stessman, M. C. Zody, B. J. Nelson, J. Huddleston, R. Sandstrom, J. D. Smith, D. Hanna, J. M. Swanson, E. M. Faustman, M. J. Bamshad, J. Stamatoyannopoulos, D. A. Nickerson, A. S. McCallion, R. Darnell, and E. E. Eichler. Genome sequencing of autism-affected families reveals disruption of putative noncoding regulatory dna. *Am J Hum Genet*, 98(1):58–74, 2016. URL: `http://www.ncbi.nlm.nih.gov/pubmed/26749308`, `doi:10.1016/j.ajhg.2015.11.023`.

[13] Y. Bromberg. Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol*, 425(21):3993–4005, 2013. URL: `http://www.ncbi.nlm.nih.gov/pubmed/23928561`, `doi:10.1016/j.jmb.2013.07.038`.

[14] L. Swint-Kruse. Using evolution to guide protein engineering: The devil is in the details. *Biophys J*, 111(1):10–8, 2016. URL: `https://www.ncbi.nlm.nih.gov/pubmed/27410729`, `doi:10.1016/j.bpj.2016.05.030`.

[15] I. H. Walker, P. C. Hsieh, and P. D. Riggs. Mutations in maltose-binding protein that alter affinity and solubility properties. *Appl Microbiol Biotechnol*, 88(1):187–97, 2010. URL: `https://www.ncbi.nlm.nih.gov/pubmed/20535468`, `doi:10.1007/s00253-010-2696-y`.

[16] R. Zabalza, A. Nurminen, L. S. Kaguni, R. Garesse, M. E. Gallardo, and B. Bornstein. Co-occurrence of four nucleotide changes associated with an adult mitochondrial ataxia phenotype. *BMC Res Notes*, 7:883, 2014. URL: `https://www.ncbi.nlm.nih.gov/pubmed/25488682`, `doi:10.1186/1756-0500-7-883`.

[17] A. Kowarsch, A. Fuchs, D. Frishman, and P. Pagel. Correlated mutations: a hallmark of phenotypic amino acid substitutions. *PLoS Comput Biol*, 6(9), 2010. URL: `https://www.ncbi.nlm.nih.gov/pubmed/20862353`, `doi:10.1371/journal.pcbi.1000923`.

[18] F. Hartwig. Snp-snp interactions: Focusing on variable coding for complex models of epistasis. *Journal of Genetic Syndromes & Gene Therapy*, 4(9):189, 2013. `doi:10.4172/2157-7412.1000189`.

[19] A. Upton, O. Trelles, and J. Perkins. Epistatic analysis of clarkson disease. *Procedia Computer Science*, 51(Supplement C):725–734, 2015. URL: `http://www.sciencedirect.com/science/article/pii/S1877050915009990`, `doi:https://doi.org/10.1016/j.procs.2015.05.191`.

[20] J. de Ligt, J. A. Veltman, and L. E. Vissers. Point mutations as a source of de novo genetic disease. *Curr Opin Genet Dev*, 23(3):257–63, 2013. URL: `https://www.ncbi.nlm.nih.gov/pubmed/23453690`, `doi:10.1016/j.gde.2013.01.007`.

[21] R. Kumar, C. Arioz, Y. Li, N. Bosaeus, S. Rocha, and P. Wittung-Stafshede. Disease-causing point-mutations in metal-binding domains of wilson disease protein decrease stability and increase structural dynamics. *Biometals*, 30(1):27–35, 2017. URL: `https://www.ncbi.nlm.nih.gov/pubmed/27744583`, `doi:10.1007/s10534-016-9976-7`.

[22] C. Dong, P. Wei, X. Jian, R. Gibbs, E. Boerwinkle, K. Wang, and X. Liu. Comparison and integration of deleteriousness prediction methods for nonsynonymous snvs in whole exome sequencing studies. *Hum Mol Genet*, 24(8):2125–37, 2015. URL: `http://www.ncbi.nlm.nih.gov/pubmed/25552646`, `doi:10.1093/hmg/ddu733`.

[23] T. A. de Beer, R. A. Laskowski, S. L. Parks, B. Sipos, N. Goldman, and J. M. Thornton. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput Biol*, 9(12):e1003382, 2013. URL: `https://www.ncbi.nlm.nih.gov/pubmed/24348229`, `doi:10.1371/journal.pcbi.1003382`.

[24] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–9, 1992. URL: http://www.ncbi.nlm.nih.gov/pubmed/1438297.

[25] L. Swint-Kruse, C. Larson, B. M. Pettitt, and K. S. Matthews. Fine-tuning function: correlation of hinge domain interactions with functional distinctions between laci and purr. *Protein Sci*, 11(4):778–94, 2002. URL: http://www.ncbi.nlm.nih.gov/pubmed/11910022, doi:10.1110/ps.4050102.

[26] D. C. Pendergrass, R. Williams, J. B. Blair, and A. W. Fenton. Mining for allosteric information: natural mutations and positional sequence conservation in pyruvate kinase. *IUBMB Life*, 58(1):31–8, 2006. URL: http://www.ncbi.nlm.nih.gov/pubmed/16540430, doi:10.1080/15216540500531705.

[27] T. A. de Beer, R. A. Laskowski, S. L. Parks, B. Sipos, N. Goldman, and J. M. Thornton. Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 genomes project dataset. *PLoS Comput Biol*, 9(12):e1003382, 2013. URL: http://www.ncbi.nlm.nih.gov/pubmed/24348229, doi:10.1371/journal.pcbi.1003382.

[28] A. S. Kondrashov, S. Sunyaev, and F. A. Kondrashov. Dobzhansky-muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A*, 99(23):14878–83, 2002. URL: https://www.ncbi.nlm.nih.gov/pubmed/12403824, doi:10.1073/pnas.232565499.

[29] V. E. Gray, K. R. Kukurba, and S. Kumar. Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*, 28(16):2093–6, 2012. URL: http://www.ncbi.nlm.nih.gov/pubmed/22685075https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3413386/pdf/bts336.pdf, doi:10.1093/bioinformatics/bts336.

[30] S. Meinhardt, J. Manley, M. W., D. J. Parente, and L. Swint-Kruse. Rheostats and toggle switches for modulating protein function. *PLoS One*, 8(12):e83502, 2013. URL: http://www.ncbi.nlm.nih.gov/pubmed/24386217, doi:10.1371/journal.pone.0083502.

[31] R. Sender, S. Fuchs, and R. Milo. Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol*, 14(8):e1002533, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27541692, doi:10.1371/journal.pbio.1002533.

[32] A. Zhernakova, A. Kurilshikov, M. J. Bonder, E. F. Tigchelaar, M. Schirmer, T. Vatanen, Z. Mujagic, A. V. Vila, G. Falony, S. Vieira-Silva, J. Wang, F. Imhann, E. Brandsma, S. A. Jankipersadsing, M. Joossens, M. C. Cenit, P. Deelen, M. A. Swertz, s. LifeLines cohort, R. K. Weersma, E. J. Feskens, M. G. Netea, D. Gevers, D. Jonkers, L. Franke, Y. S. Aulchenko, C. Huttenhower, J. Raes, M. H. Hofker, R. J. Xavier, C. Wijmenga, and J. Fu. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–9, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27126040, doi:10.1126/science.aad3369.

[33] G. Falony, M. Joossens, S. Vieira-Silva, J. Wang, Y. Darzi, K. Faust, A. Kurilshikov, M. J. Bonder, M. Valles-Colomer, D. Vandeputte, R. Y. Tito, S. Chaffron, L. Rymenans, C. Verspecht, L. De Sutter, G. Lima-Mendez, K. D'Hoe, K. Jonckheere, D. Homola, R. Garcia, E. F. Tigchelaar, L. Eeckhaudt, J. Fu, L. Henckaerts, A. Zhernakova, C. Wijmenga, and J. Raes. Population-level analysis of gut microbiome variation. *Science*, 352(6285):560–4, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27126039, doi:10.1126/science.aad3503.

[34] L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, S. B. Biddinger, R. J. Dutton, and P. J. Turnbaugh. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–63, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24336217, doi:10.1038/nature12820.

[35] J. K. Goodrich, E. R. Davenport, J. L. Waters, A. G. Clark, and R. E. Ley. Cross-species comparisons of host genetic associations with the microbiome. *Science*, 352(6285):532–5, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27126034, doi:10.1126/science.aad9379.

[36] C. A. Thaiss, S. Itav, D. Rothschild, M. Meijer, M. Levy, C. Moresi, L. Dohnalova, S. Braverman, S. Rozin, S. Malitsky, M. Dori-Bachash, Y. Kuperman, I. Biton, A. Gertler, A. Harmelin, H. Shapiro, Z. Halpern, A. Aharoni, E. Segal, and E. Elinav. Persistent microbiome alterations modulate the rate of post-dieting weight regain. *Nature*, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27906159, doi:10.1038/nature20796.

[37] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca, and J. Dore. Reduced diversity of faecal microbiota in crohn's disease revealed by a metagenomic approach. *Gut*, 55(2):205–211, 2006. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1856500/, doi:10.1136/gut.2005.073817.

[38] J. Dicksved, J. Halfvarson, M. Rosenquist, G. Jarnerot, C. Tysk, J. Apajalahti, L. Engstrand, and J. K. Jansson. Molecular analysis of the gut microbiota of identical twins with crohn's disease. *ISME J*, 2(7):716–727, 2008. URL: http://dx.doi.org/10.1038/ismej.2008.37, doi:http://www.nature.com/ismej/journal/v2/n7/suppinfo/ismej200837s1.html.

[39] D. N. Frank, C. E. Robertson, C. M. Hamm, Z. Kpadeh, T. Zhang, H. Chen, W. Zhu, R. B. Sartor, E. C. Boedeker, N. Harpaz, N. R. Pace, and E. Li. Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases. *Inflammatory bowel diseases*, 17(1):10.1002/ibd.21339, 2011. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3834564/, doi:10.1002/ibd.21339.

[40] D. N. Frank, A. L. St. Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13780–13785, 2007. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1959459/, doi:10.1073/pnas.0706625104.

[41] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biology*, 13(9):R79, 2012. URL: http://dx.doi.org/10.1186/gb-2012-13-9-r79, doi:10.1186/gb-2012-13-9-r79.

[42] A. D. Kostic, R. J. Xavier, and D. Gevers. The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–99, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24560869, doi:10.1053/j.gastro.2014.02.009.

[43] N. T. Baxter, J. P. Zackular, G. Y. Chen, and P. D. Schloss. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome*, 2:20, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24967088, doi:10.1186/2049-2618-2-20.

[44] M. B. Burns, J. Lynch, T. K. Starr, D. Knights, and R. Blekhman. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome Med*, 7(1):55, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26170900, doi:10.1186/s13073-015-0177-8.

[45] D. Gevers, S. Kugathasan, L. A. Denson, Y. Vazquez-Baeza, W. Van Treuren, B. Ren, E. Schwager, D. Knights, S. J. Song, M. Yassour, X. C. Morgan, A. D. Kostic, C. Luo, A. Gonzalez, D. McDonald, Y. Haberman, T. Walters, S. Baker, J. Rosh, M. Stephens, M. Heyman, J. Markowitz, R. Baldassano, A. Griffiths, F. Sylvester, D. Mack, S. Kim, W. Crandall, J. Hyams, C. Huttenhower, R. Knight, and R. J. Xavier. The treatment-naive microbiome in new-onset crohn's disease. *Cell Host Microbe*, 15(3):382–392, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24629344, doi:10.1016/j.chom.2014.02.005.

[46] A. K. Probstel and S. E. Baranzini. The role of the gut microbiome in multiple sclerosis risk and progression: Towards characterization of the "ms microbiome". *Neurotherapeutics*, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/29147991, doi:10.1007/s13311-017-0587-y.

[47] E. M. Mowry and J. D. Glenn. The dynamics of the gut microbiome in multiple sclerosis in relation to disease. *Neurol Clin*, 36(1):185–196, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/29157399, doi:10.1016/j.ncl.2017.08.008.

[48] M. J. Bonder, A. Kurilshikov, E. F. Tigchelaar, Z. Mujagic, F. Imhann, A. V. Vila, P. Deelen, T. Vatanen, M. Schirmer, S. P. Smeekens, D. V. Zhernakova, S. A. Jankipersadsing, M. Jaeger, M. Oosting, M. C. Cenit, A. A. Masclee, M. A. Swertz, Y. Li, V. Kumar, L. Joosten, H. Harmsen, R. K. Weersma, L. Franke, M. H. Hofker, R. J. Xavier, D. Jonkers, M. G. Netea, C. Wijmenga, J. Fu, and A. Zhernakova. The effect of host genetics on the gut microbiome. *Nat Genet*, 48(11):1407–1412, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27694959, doi:10.1038/ng.3663.

[49] J. Wang, L. B. Thingholm, J. Skieceviciene, P. Rausch, M. Kummen, J. R. Hov, F. Degenhardt, F. A. Heinsen, M. C. Ruhlemann, S. Szymczak, K. Holm, T. Esko, J. Sun, M. Pricop-Jeckstadt, S. Al-Dury, P. Bohov, J. Bethune, F. Sommer, D. Ellinghaus, R. K. Berge, M. Hubenthal, M. Koch, K. Schwarz, G. Rimbach,

P. Hubbe, W. H. Pan, R. Sheibani-Tezerji, R. Hasler, P. Rosenstiel, M. D'Amato, K. Cloppenborg-Schmidt, S. Kunzel, M. Laudes, H. U. Marschall, W. Lieb, U. Nothlings, T. H. Karlsen, J. F. Baines, and A. Franke. Genome-wide association analysis identifies variation in vitamin d receptor and other host factors influencing the gut microbiota. *Nat Genet*, 48(11):1396–1406, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27723756, doi:10.1038/ng.3695.

[50] F. Luca, S. S. Kupfer, D. Knights, A. Khoruts, and R. Blekhman. Functional genomics of host-microbiome interactions in humans. *Trends Genet*, 34(1):30–40, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/29107345, doi:10.1016/j.tig.2017.10.001.

[51] R. Blekhman, J. K. Goodrich, K. Huang, Q. Sun, R. Bukowski, J. T. Bell, T. D. Spector, A. Keinan, R. E. Ley, D. Gevers, and A. G. Clark. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*, 16:191, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26374288, doi:10.1186/s13059-015-0759-1.

[52] R. N. Carmody, G. K. Gerber, J. Luevano, J. M., D. M. Gatti, L. Somes, K. L. Svenson, and P. J. Turnbaugh. Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe*, 17(1):72–84, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/25532804, doi:10.1016/j.chom.2014.11.010.

[53] C. Zhu, T. O. Delmont, T. M. Vogel, and Y. Bromberg. Functional basis of microorganism classification. *PLoS Comput Biol*, 11(8):e1004472, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26317871, doi:10.1371/journal.pcbi.1004472.

[54] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, 44(D1):D279–85, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/26673716, doi:10.1093/nar/gkv1344.

[55] V. Marx. Biology: The big challenges of big data. *Nature*, 498(7453):255–60, 2013. URL: https://www.ncbi.nlm.nih.gov/pubmed/23765498, doi:10.1038/498255a.

[56] S. Bansal, G. Chowell, L. Simonsen, A. Vespignani, and C. Viboud. Big data for infectious disease surveillance and modeling. *J Infect Dis*, 214(suppl_4):S375–S379, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/28830113, doi:10.1093/infdis/jiw400.

[57] Z. Yin, H. Lan, G. Tan, M. Lu, A. V. Vasilakos, and W. Liu. Computing platforms for big biological data analytics: Perspectives and challenges. *Comput Struct Biotechnol J*, 15:403–411, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28883909, doi:10.1016/j.csbj.2017.07.004.

[58] URL: https://jetstream-cloud.org/.

[59] I. D. Dinov, B. Heavner, M. Tang, G. Glusman, K. Chard, M. Darcy, R. Madduri, J. Pa, C. Spino, C. Kesselman, I. Foster, E. W. Deutsch, N. D. Price, J. D. Van Horn, J. Ames, K. Clark, L. Hood, B. M. Hampstead, W. Dauer, and A. W. Toga. Predictive big data analytics: A study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLoS One*, 11(8):e0157077, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27494614, doi:10.1371/journal.pone.0157077.

[60] 2017. URL: https://www.docker.com.

[61] URL: http://services.bromberglab.org/fusiondb.

[62] 2017. URL: http://services.bromberglab.org/mifaser.

[63] URL: http://www.nature.com/articles/srep41329#supplementary-information.

[64] L. M. Starita, D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud, R. J. Hause, D. M. Fowler, J. D. Parvin, J. Shendure, and S. Fields. Massively parallel functional analysis of brca1 ring domain variants. *Genetics*, 200(2):413–22, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/25823446, doi:10.1534/genetics.115.175802.

[65] D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields. Deep mutational scanning of an rrm domain of the saccharomyces cerevisiae poly(a)-binding protein. *RNA*, 19(11):1537–51, 2013. URL: https://www.ncbi.nlm.nih.gov/pubmed/24064791, doi:10.1261/rna.040709.113.

[66] L. M. Starita, J. N. Pruneda, R. S. Lo, D. M. Fowler, H. J. Kim, J. B. Hiatt, J. Shendure, P. S. Brzovic, S. Fields, and R. E. Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc Natl Acad Sci U S A*, 110(14):E1263–72, 2013. URL: `https://www.ncbi.nlm.nih.gov/pubmed/23509263`, `doi:10.1073/pnas.1303309110`.

[67] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol Biol Evol*, 31(6):1581–92, 2014. URL: `https://www.ncbi.nlm.nih.gov/pubmed/24567513`, `doi:10.1093/molbev/msu081`.

[68] N. C. Wu, C. A. Olson, and R. Sun. High-throughput identification of protein mutant stability computed from a double mutant fitness landscape. *Protein Sci*, 25(2):530–9, 2016. URL: `https://www.ncbi.nlm.nih.gov/pubmed/26540565`, `doi:10.1002/pro.2840`.

[69] G. Yachdav, E. Kloppmann, L. Kajan, M. Hecht, T. Goldberg, T. Hamp, P. Honigschmid, A. Schafferhans, M. Roos, M. Bernhofer, L. Richter, H. Ashkenazy, M. Punta, A. Schlessinger, Y. Bromberg, R. Schneider, G. Vriend, C. Sander, N. Ben-Tal, and B. Rost. Predictprotein–an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res*, 42(Web Server issue):W337–43, 2014. URL: `https://www.ncbi.nlm.nih.gov/pubmed/24799431`, `doi:10.1093/nar/gku366`.

[70] URL: `https://store.docker.com/community/images/bromberglab/predictprotein`.

[71] T. C. Smith and E. Frank. Introducing machine learning concepts with weka. *Methods Mol Biol*, 1418:353–78, 2016. URL: `https://www.ncbi.nlm.nih.gov/pubmed/27008023`, `doi:10.1007/978-1-4939-3578-9_17`.

[72] 2015. URL: `https://www.R-project.org/`.

[73] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, P. Bansal, A. J. Bridge, S. Poux, L. Bougueleret, and I. Xenarios. Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view. *Methods Mol Biol*, 1374:23–54, 2016. `doi:10.1007/978-1-4939-3167-5_2`.

[74] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997. URL: `http://www.ncbi.nlm.nih.gov/pubmed/9254694`.

[75] 2017. URL: `https://bitbucket.org/bromberglab/clubber/`.

[76] 2004. URL: `https://www.w3.org/TR/2004/NOTE-ws-arch-20040211/#relwwwrest`.

[77] L. M. Rodriguez-R, W. A. Overholt, C. Hagan, M. Huettel, J. E. Kostka, and K. T. Konstantinidis. Microbial community successional patterns in beach sands impacted by the deepwater horizon oil spill. *ISME J*, 9(9):1928–1940, 2015. URL: `http://dx.doi.org/10.1038/ismej.2015.5https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4542042/pdf/ismej20155a.pdf`, `doi:10.1038/ismej.2015.5`.

[78] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Droege, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. Sparholt Jorgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, L. Hestbjerg Hansen, S. J. Sorensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. D. Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. Gueiros Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Goeker, N. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy. Critical assessment of metagenome interpretation - a benchmark of computational metagenomics software. *bioRxiv*, 2017. URL: `http://biorxiv.org/content/early/2017/01/09/099127.abstract`.

[79] URL: `https://store.docker.com/community/images/bromberglab/clubber`.

[80] URL: `https://bitbucket.org/bromberglab/clubber`.

[81] URL: `https://doi.org/10.5281/zenodo.1127743`.

[82] C. R. e. Garrity GM, Boone DR. *Bergey's Manual of Systematic Bacteriology, Volume 1*. Springer, New York (NY), 2nd edition, 2001.

[83] C. Zhu, T. O. Delmont, T. M. Vogel, and Y. Bromberg. Functional basis of microorganism classification. *PLoS Comput Biol*, 11(8):e1004472, 2015. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1004472, doi:10.1371/journal.pcbi.1004472.

[84] C. Zhu, Y. Mahlich, M. Miller, and Y. Bromberg. fusiondb: assessing microbial diversity and environmental preferences via functional similarity networks. *Nucleic Acids Res*, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/29112720, doi:10.1093/nar/gkx1060.

[85] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic Acids Res*, 37(Database issue):D26–31, 2009. doi:10.1093/nar/gkn723.

[86] M. Miller, C. Zhu, and Y. Bromberg. clubber: removing the bioinformatics bottleneck in big data analyses. *J Integr Bioinform*, 14(2), 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28609295, doi:10.1515/jib-2017-0020.

[87] M. Steinegger and J. Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 2017.

[88] URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkx1060/4588109#102570775.

[89] A. B. Shreiner, J. Y. Kao, and V. B. Young. The gut microbiome in health and in disease. *Curr Opin Gastroenterol*, 31(1):69–75, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/25394236, doi:10.1097/MOG.0000000000000139.

[90] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. Edwards. The metagenomics rast server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386, 2008. URL: http://dx.doi.org/10.1186/1471-2105-9-386, doi:10.1186/1471-2105-9-386.

[91] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt. Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol*, 5(12):e1000605, 2009. URL: http://dx.doi.org/10.1371%2Fjournal.pcbi.1000605, doi:10.1371/journal.pcbi.1000605.

[92] B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using diamond. *Nat Meth*, 12(1):59–60, 2015. URL: http://dx.doi.org/10.1038/nmeth.3176, doi:10.1038/nmeth.3176http://www.nature.com/nmeth/journal/v12/n1/abs/nmeth.3176.html#supplementary-information.

[93] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(Database issue):D457–D462, 2016. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702792/, doi:10.1093/nar/gkv1070.

[94] URL: https://www.epidemiologie.uni-kiel.de/biobanking/biobank-popgen.

[95] C. Zhang, A. Yin, H. Li, R. Wang, G. Wu, J. Shen, M. Zhang, L. Wang, Y. Hou, H. Ouyang, Y. Zhang, Y. Zheng, J. Wang, X. Lv, Y. Wang, F. Zhang, B. Zeng, W. Li, F. Yan, Y. Zhao, X. Pang, X. Zhang, H. Fu, F. Chen, N. Zhao, B. R. Hamaker, L. C. Bridgewater, D. Weinkove, K. Clement, J. Dore, E. Holmes, H. Xiao, G. Zhao, S. Yang, P. Bork, J. K. Nicholson, H. Wei, H. Tang, X. Zhang, and L. Zhao. Dietary modulation of gut microbiota contributes to alleviation of both genetic and simple obesity in children. *EBioMedicine*, 2(8):968–84, 2015. doi:10.1016/j.ebiom.2015.07.007.

[96] L. M. Rodriguez-R, W. A. Overholt, C. Hagan, M. Huettel, J. E. Kostka, and K. T. Konstantinidis. Microbial community successional patterns in beach sands impacted by the deepwater horizon oil spill. *ISME J*, 9(9):1928–1940, 2015. URL: http://dx.doi.org/10.1038/ismej.2015.5, doi:10.1038/ismej.2015.5.

[97] URL: https://bitbucket.org/bromberglab/mifaser.

[98] URL: https://doi.org/10.5281/zenodo.1045583.

[99] URL: https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkx1209/4670955#102571262.

A List of Pulications

1. **Miller, M.**, Bromberg, Y., & Swint-Kruse, L. (2017). Computational predictors fail to identify amino acid substitution effects at rheostat positions. Sci Rep, 7, 41329. doi:10.1038/srep41329

2. **Miller, M.**, Zhu, C., & Bromberg, Y. (2017). clubber: removing the bioinformatics bottleneck in big data analyses. J Integr Bioinform, 14(2). doi:10.1515/jib-2017-0020

3. Zhu, C.*, Mahlich, Y.*, **Miller, M.***, & Bromberg, Y. (2017). fusionDB: assessing microbial diversity and environmental preferences via functional similarity networks. Nucleic Acids Res. doi:10.1093/nar/gkx1060. **equal contribution*

4. Zhu, C., **Miller, M.**, Marpaka, S., Vaysberg, P., Rühlemann, M. C., Wu, G. H. F.-A., . . . Bromberg, Y. (2017). Functional sequencing read annotation for high precision microbiome analysis. Nucleic Acids Res. doi:10.1093/nar/gkx1209

B Manuscripts in Preparation

1. **Miller, M.**, Vitale, D., Swint-Kruse, L., Rost, B. & Bromberg, Y. fun-TRP: accurate annotation of protein position classes *(in prep)*

C Declaration

I, Carl Maximilian Miller, hereby declare that I independently prepared the present thesis, using only the references and resources stated. This work has not been submitted to any examination board, yet. Parts of this work will or have been published in scientific journals.

Munich, December 2017

D Acknowledgements

Firstly, I would like to thank Prof. Dr. Yana Bromberg and Prof. Dr.Burkhard Rost, they have always believed in me, pushed me to achieve more than I thought was possible and made me the scientist I am today. I am especially grateful to Yana for her guidance and support during the last two years. I could not have come this far without her. Next, I would like to thank the entire BrombergLab, Yanran Wang, Chengsheng Zhu, Yannick Mahlich, Zishuo Zeng and Anton Molyboha for their continued support and friendship. I would also like to extend my gratitude towards our collaborator, Prof. Dr Liskin Swint-Kruse, for her data, critical feedback, candid comments, as well as laying the foundation that led us to successfully publish the rheostat story. I further want to thank Daniel Vitale for his contribution on the development of our new rheostat predictor. Of course I also want to thank the entire Rostlab, especially Tim Karl, Tatyana Goldberg and Inga Weise for their support and help. Further, I want express my gratitude to Kevin Abbey, who keeps our infrastructure at Rutgers running and his assistance with technical issues any time of the day/night. The one person who I can't thank enough is Sonakshi Bhattacharjee. Her scientific expertise, patience, friendship and love supported me through the last two years. I can't think of a more important person in my life. Finally, I want to thank my family, who have stood by me and supported every decision I made. Mum and Dad, without your support, I would not be where I am today. Thank you for everything.