Computational analysis of metagenomic data: delineation of compositional features and screens for desirable enzymes

Dissertation zur Erlangung
des naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Konrad Ulrich Förstner
Heidelberg

Würzburg 2008

Eingereicht am:

Mitglieder der Promotionskommission:
- Vorsitzender: Prof. Dr. Martin J. Müller
- 1. Gutachter: Dr. habil. Peer Bork
- 2. Gutachter: Prof. Dr. Thomas Dandekar

Tag des Promotionskolloquiums:

Doktorurkunde ausgehändigt am:

License

Dedicated to the free access to humankind's knowledge.

Contents

Abbreviations and acronyms

DNA	Deoxyribonucleic acid
GC	Guanine/Cytosine
HGT	Horizontal gene transfer
HMM	Hidden Markov Model
NHase	Nitrile hydratase
ORF	Open reading frame
PCR	Polymerase chain reaction
PKS	Polyketide synthase
USE	Upstream sequence element

Chapter 1

Summary/Zusammenfassung

1.1 Summary

The topic of my doctorial research was the computational analysis of metagenomic data. A metagenome comprises the genomic information from all the microorganisms within a certain environment. The currently available metagenomic data sets cover only parts of these usually huge metagenomes due to the high technical and financial effort of such sequencing endeavors. During my thesis I developed bioinformatic tools and applied them to analyse genomic features of different metagenomic data sets and to search for enzymes of importance for biotechnology or pharmaceutical applications in those sequence collections. In these studies nine metagenomic projects (with up to 41 subsamples) were analysed. These samples originated from diverse environments like farm soil, acid mine drainage, microbial mats on whale bones, marine water, fresh water, water treatment sludges and the human gut flora. Additionally, data sets of conventionally retrieved sequence data were taken into account and compared with each other. The results of these studies were published in six publications in diverse scientific journals:

The first publication described the comparative analysis of the GC-value distribution (percentage of Guanine and Cytosine in a DNA sequence) in the unassembled sequence reads of different environments [1] (Appendix A).

It was shown that despite the enormous species diversity in the different environments there were certain GC preferences that differed between the habitats. For example, the sequences from a Minnesota farm soil sample unexpectedly had a much higher average GC value than the sequences of samples taken from Sargasso Sea surface water. The trend was even stronger for the third codon base and had an influence on the amino acid recruitment of the organism in the particular environment.

In a review that covered the burgeoning field of metagenomics and shed light on its challenges and potential we presented the results of a DNA complexity study (measurements of the nonamere distribution) and protein similarity comparisons of available metagenomic samples with conventional protein databases [2] (Appendix B). We could show the influence of an environment's complexity on the complexity of its inhabitants metagenome and that a huge fraction of predicted open reading frames (ORFs) in the metagenomic samples had no counterpart in conventional protein data bases and could therefore be classified as new.

In a second review we discussed the general methodology of the computational analysis of metagenomes. Additionally, we presented an extension of the previously published study of GC values on further samples that had become available in the meantime [3] (Appendix C). Among others, it contained the so far biggest published metagenomic data set – the sequences of the *Global Ocean Sequencing Expedition*. The extended view confirmed the previously discovered trend regarding the GC value distributions. The review also covered the results of a screening of biotechnologically relevant enzymes in metagenomic data: Nitrilases are a group of enzymes that are intensively used in the chemical industry to hydrolase nitriles to their corresponding carboxylic acids and ammonia. With the help of a Hidden Markov Model (HMM), members of nitrilases were searched for in a collection of predicted proteins from metegenomic data sets and conventional protein databases (*UniRef*). Maximum-likelihood trees were then generated to verify the membership of the detected sequence and to investigate the classification of this

group of enzymes. By doing this, we detected new nitrilase members and could unexpectedly define previously unknown subclasses of nitrilases.

The discovery that the habitat influences the GC content of the species living there was used in a subsequent study to detect gene transfer between geographically distant environments [4] (Appendix D). Based on this, we analysed synonymous nucleotide codon composition, the frequency of DNA oligomers and sequence similarity between the predicted genes in sequences gained from two environments. Based on this, we assumed that the detected transfer events took place mainly from soil habitats to marine habitats.

We used the same method that was applied to detect nitrilases (see above) to screen for nitrile hydratases (NHase). They are another group of widely applied biotechnologically enzymes that hydrolyse nitriles to their corresponding amides [5] (Appendix F). In contrast to the nitralses that needed only a single domain to be searched for, we screened for two subunits (α- and β-subunit) of the nitrile hydratase with two generated Hidden Markov Models. Numerous members were identified in metagenomic and conventional data sets and were verified by maximum-likelihood trees. Additionally we discovered that nitrile hydratases which were previously only described in bacterial organisms, to also exist in an eukaryote: The two subunits that are usually located in separate proteins in bacteria were detected in a single protein of the marine choanoflagellate *Monosiga brevicollis*.

In a further study [6] (Appendix E) the combination of HMMs and maximum-likelihood trees was used to detect type I polyketide synthases (PKS) in metagenomic sequences. Polyketides are a very heterogeneous group of secondary metabolites with miscellaneous functions e.g. as antibiotics. During their biosynthesis which is similar to fatty acid synthesis, short acyl units are added stepwise to the polyketide growing molecule. Every step is catalysed by a member of the diverse polyketide synthases. In our study, we focused on the type I polyketide synthases that contain a dedicated domain module for each step. Due to the enormous length of the type I PKS proteins and

the short sequences of metagenomic sequences, the main aim was not to find complete PKS but to estimate the potential of the different environments to serve as a source of type I PKS proteins. For each of the eight domains a Hidden Markov Model was generated and used to detect PKS members in metagenomic and conventional data sets (*UniRef*). After that, maximum-likelihood trees of the extracted sequences were calculated to distinguish the type I PKS members from the closely related type I fatty acid synthases and other enzymes. The information aquired from the eight domain searches was integrated with other data into a database and evaluated. We showed that the farm soil has the highest PKS density, but also that other environments like acid mine drainage unexpectedly contain species with PKS functionality. Additionally, we were able to allocate proteins from the *UniRef* database which were not annotated as type I polyketides to this group to date.

In addition to my work in the field of metagenomics I was involved in a study with a focus on splicing factors (Appendix G). Due to the thematic distance of it to my core research interest it is not further discussed here.

1.2 Zusammenfassung

Das Thema meiner Doktorarbeit war die bioinformatische Analyse von metagenomischen Sequenzdaten. Ein Metagenom umfasst die genomische Information aller Mikroorganismen eines Biotops. Die bisher durchgeführten metagenomische Projekte sequenzierten auf Grund des technischen und finanziellen Aufwands einer solchen Unternehmung nur kleine Teile dieser im allgemeinen sehr großen Metagenome. Im Zuge meiner Doktorarbeit, die auf solchen Sequenzierungprojekten aufbaut, wurden bioinformatische Werkzeuge entwickelt und angewandt um genomische Eigenschaften verschiedener metagenomische Datensätze zu analysieren und um biotechnologisch und pharmakologisch relevante Enzyme exemplarisch in diesen Datensätzen zu suchen. In den Analysen wurden neun publizierte, metagenomische Projektedatensammlungen (teilweise mit bis zu 41 Subproben) untersucht. Die Probem stammen von zahlreichen unterschiedlichen Habitaten wie Farmerde,

sauerer Minendrainage, dem mikrobiellen Belag auf Walknochen, Meerwasser, Süßwasser, Abwasseraufbereitungssschlamm und der menschlichen Darmflora. Zusätzlich wurden in den meisten Analysen konventionell gewonnene Sequenzdaten vergleichend hinzugezogen und analysiert. Die Ergebnisse dieser Forschungstätigkeit wurden in sechs Artikeln in unterschiedlichen Fachzeitschriften publiziert:

Die erste Publikation umfasste eine vergleichende Analyse der Verteilung des GC-Gehaltes – der prozentuale Anteil der Basen Guanin und Cytosin in einer DNS-Sequenz – von unassemblierten Sequenz-Reads von verschiedenen Standorten [1] (Anhang A). Es konnte gezeigt werden dass, trotz der enormen Speziesdiversität innerhalb einer Probe GC-Präferenzen existieren und diese Präferenzen sich zwischen den verschiedenen Biotopen stark unterscheiden. Die Sequenzen einer Farmbodenprobe aus Minnesota haben einen bei weitem höheren durchschnittlichen GC-Wert als Sequenzen von Oberflächenwasserproben, die der Sargassoseegebiet entnommen wurden. Dieser Trend, der in der dritten Codonbase sogar verstärkt zu beobachten ist, hat Einfluss auf die Aminosäurerekrutierung der Organismen in dem jeweiligen Biotop.

In einem Übersichtsartikel, der das junge Feld der Metagenomik beschrieb und dessen Herausforderungen und Möglichkeiten beleuchtete, brachten wir Ergebnisse von DNA-Komplexitätsanalysen (gemessen an der Nonamer-Verteilung) und Protein-Ähnlichkeitsvergleiche von zu der Zeit verfügbaren Metagenomen mit konventionellen Proteindatenbanken ein [2] (Anhang B). Wir konnten zeigen, dass die Komplexität der Biotope sich in der Komplexität der Metagenome widerspiegelt und dass ein Großteil der vorausgesagten offenen Leserahmen (ORF - open reading frame) der Metagenome kein Pendant in konventionellen Protein-Datenbanken besitzt und somit als neu einzustufen ist.

In einem zweiten Übersichtsartikel wurde neben der Diskussion der generellen Herangehensweise für die bioinformatische Analyse von Metagenomen die im ersten Artikel beschriebene GC-Gehalt-Analyse auf weitere, zu dem Zeit-

punkt neu veröffentlichte Proben, ausgeweitet [3] (Anhang C). Unter anderem umfasste diese Analyse den bis heute größten metagenomischen Datensatz – die Sequenzen der *Global Ocean Sampling Expedition*. Der erweiterte Blick bestätigte den zuvor festgestellten Trend bezüglich der GC-Wert-Verteilungen. Der Übersichtsartikel beschrieb zudem die Ergebnisse einer Suche nach biotechnologisch relevanten Proteinen in metagenomischen Daten: Mittels eines Hidden-Markov-Modells wurden Vertreter der Nitrilasen, eine Enzymgruppe die in der chemischen Industrie intensiv für die Hydrolyse von Nitrilen in ihre Carbonsäuren und Ammoniak angewandten wird, in Sammlungen der vorausgesagten Proteinsequenzen von Metagenomdaten und konventionellen Proteindatenbanken (*UniRef*) ausfindig gemacht. Maximum-Likelihood-Bäume (maximale Wahrscheinlichkeitsbäume) wurden generiert um die Zugehörigkeit der gefundenen Sequenzen zu den Nitrilasen zu verifizieren und die Klassifikation dieser Enzymgruppe zu untersuchen. Wir fanden hierbei zahlreiche bisher unbekannte Unterklassen der Nitrilasen.

Die Entdeckung, dass Biotope den GC-Gehalt der bewohnenden Spezien in eine bestimmte Richtung lenken wurde in einer Studie genutzt um Gentransfer zwischen geographisch weit auseinander liegenden Biotopen nachzuweisen [4] (Anhang D). Wir untersuchten dazu die synonymen Codonpositionen, die DNA-Oligomer-Vorkommen und Sequenzähnlichkeiten zwischen den vorausgesagten Genen von zwei Habitaten. Die nachgewiesen Tranferereignisse fanden hauptsächlich von Erdhabitaten zu marinen Habitaten statt und erklären möglicherweise das Vorkommen bestimmter genetischer Funktionen in diesen Biotopen.

Die Methode, die für die Detektion von Nitrilase angewandt wurde (siehe oben), nutzten wir um Nitril-Hydratasen aufzufinden [5] (Anhang F). Diese sind ebenfalls eine industriell intensive genutzte Gruppe von Biokatalysator, die Nitrile zu ihrem entsprechenden Amiden hydrolysieren. Im Gegensatz zu der Nitrilase, bei der nur eine einzelne Domäne zu detektieren war, wurden α- und β-Untereinheit der Nitril-Hydratase jeweils mit einem selbstgenerierten HMM gesucht. Es wurden zahlreiche Vertreter in metagenomischen

und konventionellen Proteindatenbanken gefunden und durch Maximum-Likelihood-Bäume bestätigt. Zudem entdeckten wir das Vorkommen dieser Nitril-Hydratasen, die bisher nur in Bakterien beschrieben waren, in einem Eukaryoten: Die beiden üblicherweise auf separaten Proteinen lokalisierten Untereinheiten, wurden zusammen auf einem Protein des marinen Choanoflagellaten *Monosiga brevicollis* detektiert.

Die Kombination von Hidden-Markov-Modellen und Maximum-Likelihood-Bäumen diente in einer weiteren Untersuchung dem Auffinden von Typ-I-Polyketid-Synthasen (PKS) in metagenomischen Daten [6] (Anhang E). Polyketide sind eine sehr heterogene Gruppe von Sekundärmetaboliten, die zahlreiche Funktionen, u.a. als Antibiotika, wahrnehmen. Für ihrer Biosynthese werden, ähnlich wie bei der Synthese von Fettsäuren, kleine Kohlenstoffketteneinheiten schrittweise an ein wachsendes Molekül gehängt. Jeder Syntheseschritt wird von einem Mitglied der diversen Polyketidsynthasen katalysiert, die dazu zahlreiche Domänen besitzen. In unserer Untersuchung fokussierten wir uns auf eine Unterklasse, die Typ I Polyketid-Synthasen, bei denen für jeden einzelnen Additionsschritt ein separates Domänenmodul in dem Protein existiert. Auf Grund der enormen Länge von PKS-Proteinen und der geringen Länge der metagenomischen Sequenzen war es nicht das Hauptziel, vollständige Type-I-Polyketidsynthasen zu finden. Stattdessen sollte das Potenzial der Biotope als PKS-I-Quelle zu fungieren abgeschätzt werden. Es wurde für jede der acht Domänen jeweils ein Hidden-Markov-Modell generiert und für die Suche von Polyketidsynthase-Vertretern in metagenomischen Daten als auch konventionellen Proteindatenbanken (*UniRef*) genutzt. Die danach berechneten Maximum-Likelihood-Bäume der extrahierten Sequenzen dienten hier der Unterscheidung der Typ-I-PKS von dem verwandten Typ-I-Fettsäuresynthasen. Wir konnten zeigen, dass die Farmerde die höchste PKS-Dichte aufweist, aber zudem, dass Habitate wie die saure Mienendrainge unerwarteter Weise Spezies mit PKS enthält. In der Protein-Datenbank *UniRef* konnten einige Proteine, die vorher nicht als Typ-I-Polyketid-Vertreter annotiert waren, dieser Gruppe an Hand unserer Methode zugeordnet werden.

Zusätzlich zu meiner Arbeit auf dem Gebiet der Metagenomik habe ich zu einer Studie über Splicing-Faktoren (Anhang G) beigetragen. Da diese thematisch nicht mit meinem Kernforschungsgebiet zusammenhängt, wird sie hier nicht weiter diskutiert.

Chapter 2

Introduction

2.1 The advent of metagenomics and its implications

The genome of an organism comprises its entire genetic information coded in its DNA [7]. Having the sequence of an organisms genome at hand enables deep insights into its physiology and genetics. The full genome sequencing boosted the generation of biological knowledge intensively. In 1995 the first full genome was published [8] and according to *Genomes OnLine Database* (GOLD) [9] more than 800 mainly microbial, complete genomes are currently available. Due to improvements in present techniques and the invention of new generation non-Sanger platforms [10, 11, 12] sequencing is becoming quicker and cheaper which leads to a rapidly growing amount of genomes that are available in public databases.

Yet the classical microbial genome sequencing has the frequently underestimated disadvantage that it relies on the culturing of the target species. As is evident from 200 years of microscopy [13] most habitats contain thousands of different species, however it was shown that only a small fraction of the microbial community can be cultivated under standard laboratory conditions. This phenomenon was coined the *great plate count anomaly* [14]. Due to this, culture independent sequence approaches were developed to explore the

diversity of environments. In the beginning microbiologists focused on the study of the highly conserved ribosomal RNA to get a first impression of the plethora of organisms that were present [15, 16]. Then single genes of biotechnological importance or with phylogenetic marker functions were selectively extracted using the polymerase chain reaction (PCR) with distinct primers [17]. Later methods for cloning DNA fragments taken from environmental samples were developed. This was done the first time in 1991 [18] and the name metagenomics [19] was dubbed seven years later [20].

Environmental clone libraries can be screened for phylogenetic markers or enzymatic functions. Alternatively, the inserts of random clones can be sequenced [13]. In 2004 the results from the first large random shotgun sequencing of such libraries were published [21, 22] and commenced a wave of environmental sequencing projects. These endeavours generally generate enormous amounts of sequence data. A publication of a metagenomic analysis of the Sargasso Sea [21] and later the Global Ocean Sampling Expedition publication which covered samples from many marine sites [23] each almost doubled the amount of sequence data available in public databases. The huge dimensions also in addition to the strong fragmentation, high diversity of the sequences, differing sampling protocols and other factors created challenges when analysising this new kind of data.

In general these studies were data driven instead of hypothesis driven. The longterm aim was and still is to understand the species in the context of their abiotic and biotic environment. Additionally, questions like “How many species are out there?”, “Are there unknown protein families and new pathways?” or more complex ones like “How do these organisms adapt to their habitats?” are expected to be easier solved by insights of this approach.

2.2 The metagenomic workflow

The first step in the construction of a metagenomic library is to extract DNA from the environmental sample. The strain isolation which is done in sequencing projects of single microbial species is not taking place. Depending on the nature of the sample preceding treatments like the disruption of the cells have to be performed. The recovered DNA is digested by restriction enzymes and during a ligation step it is integrated into vectors (e.g. plasmids). After that, these vectors are transformed into a host organism. The resulting library of clones can then be functionally screened or sequenced [13].

In the case of a sequencing approach the next step is an attempt to assemble the resulting sequence reads into larger fragments, so called contigs, as done in conventional genome sequencing efforts. Due to hampering factors like the large amount of different species in the samples, virus sequences and low sequence coverage of the metagenome this is a non-trivial task [3] and established methods must be adapted to be able to successfully process this data. The assembly is usually followed by gene prediction methods like in single organism sequencing projects. In metagenomic projects these necessary programs might struggle with the short sequences and low sequence quality. The successfully detected genes can then undergo functional and phylogenetic prediction based on their sequence similarities with proteins in conventional databases. Besides these standard techniques many different analyses can be performed to answer project specific questions. Figure 1 illustrates the workflow of a sequence-focused metagenomic project.

The studies presented here build upon the results generated in these previously performed steps.

2.3 Data sets used

For our comparative analyses we selected metagenomic data sets originating from various sources. We used samples taken from a Minnesota farm

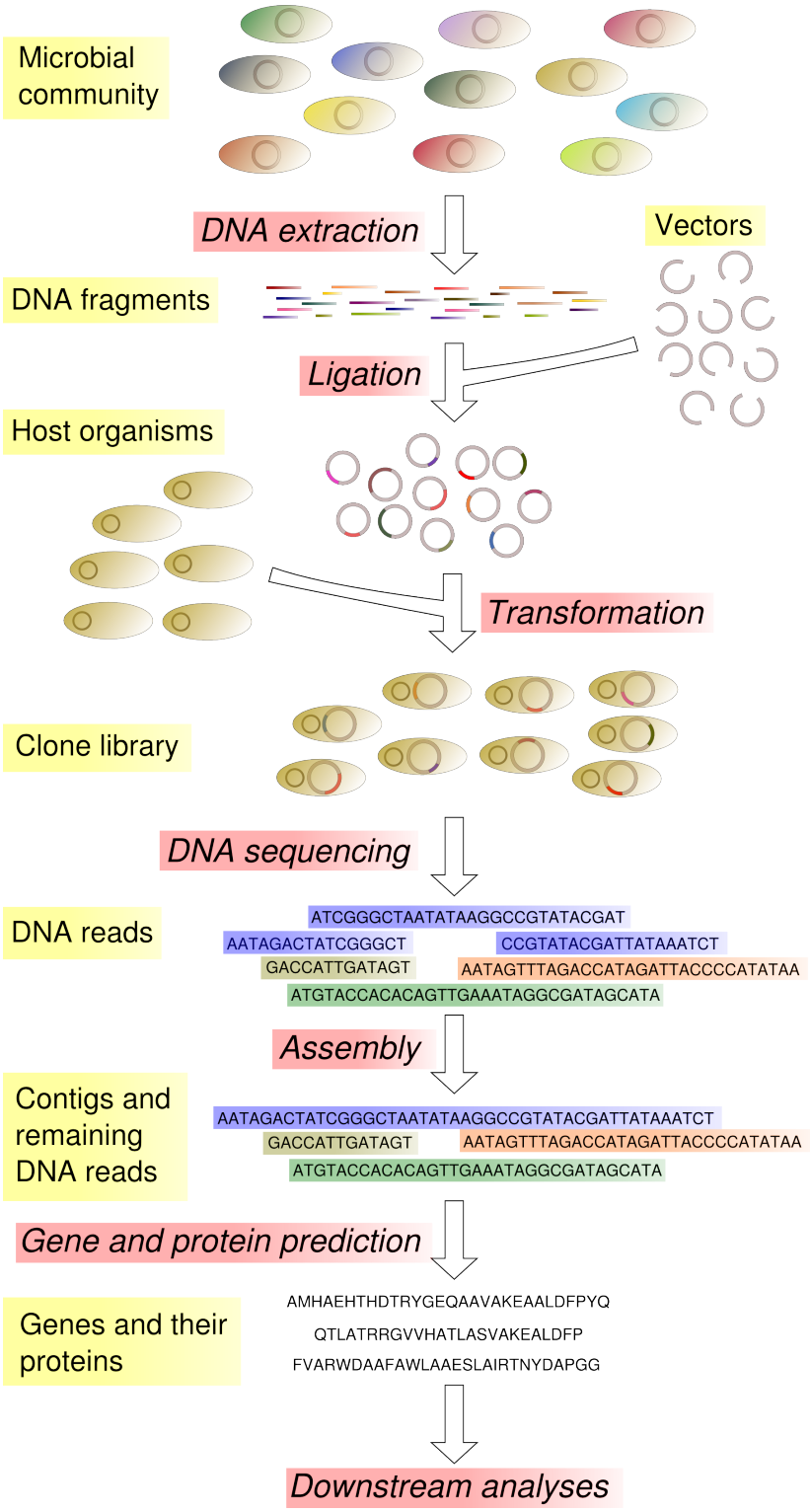


Figure 1: Workflow of a metagenomic sequencing project.

soil with around 220 Mbp of raw reads and sunken whale bones, which were published together [24]. Another collection contained sequences from a German soil sample [25]. In the earlier studies we made use of the Sargasso Sea samples sequences [21] (with around 2 Gbp of raw sequence in 7 subsamples), later when the complete Global Ocean Sampling data collection (6.3 Gpb of raw sequence taken in 41 different places) [23, 26] was available we applied our method to that instead. Further marine sequences were taken from an environmental sequencing project with sample points at different depths in a North Pacific Subtropical Gyre [27]. Furthermore, the sequence data of a microbial community living in an acid mine drainage site [22], sequences of the microbial flora of human guts [28], and data of enhanced biological phosphorus removal sludge communities [29] were included in our analyses. Additionally, we used conventionally retrieved sequences from the *String* database [30], the [31] *UniRef* database and the *Monosiga brevicollis* genome sequencing project [32].

2.4 Analysis of genomic features of metagenomic samples

The GC content – the relative abundance of the nucleotides guanine and cytosine in a DNA sequence – of microbial genomes is a intensively studied attribute that can differ strongly between different species. Many possible environmental factors might drive the GC content of an organism to a certain value [33]. Among others, the optimal growth temperature [34], the dependency on oxygen [35] or the ability to fix nitrogen [36] are parameters that can potentially influence the GC content of an organism in a particular habitat. Also genome size [34, 37] and ultraviolet light exposure [38] could affect the GC content.

In our analyses [1, 3] (Appendix A and C) we compared the GC content of reads of several environmental samples and also simulated the expected GC distributions based on the particular species composition.

2.5 Screening for enzymes in metagenomic samples

Microbial enzymes are heavily used in the chemical industry and other fields such as food production. Environmental DNA libraries are established sources for new members of such enzymes [39, 20] and there are three commonly used approaches for screening: sequence-based, function-based and substrate-induced gene-expression screening (SIGEX) [40]. The increasing amount of available metagenomic sequence data represents a new source of enzymes with potential biotechnology applications. They can be identified by using computation screening methods.

We developed a pipeline for such a computational screening (Figure 2) and illustrated its functionality by applying it to search for three types of enzymes as case studies. The main components of the pipeline are Hidden Markov Model searches (performed with programs from the *HMMER* package [41]) with high sensitivity and the construction of maximum-likelihood trees (using the program *PHYML* [42]) to increase the selectivity of the screening.

Hidden Markov Models are statistical models which use state probability and state transition probabilities to find patterns in a sequence of tokens [43, 41]. In bioinformatics they are often applied to find a certain pattern in DNA or protein sequences after training the HMMs with a set of sequences containing these patterns. For the construction of our HMM we used manually curated sets of sequences.

Among many other applications, maximum likelihood approaches are often used to construct phylogenetic trees. With the help of these methods a statistical models can be developed based on a representative data sample. In the case of phylogenetic analyses the most likely tree is presented at the end

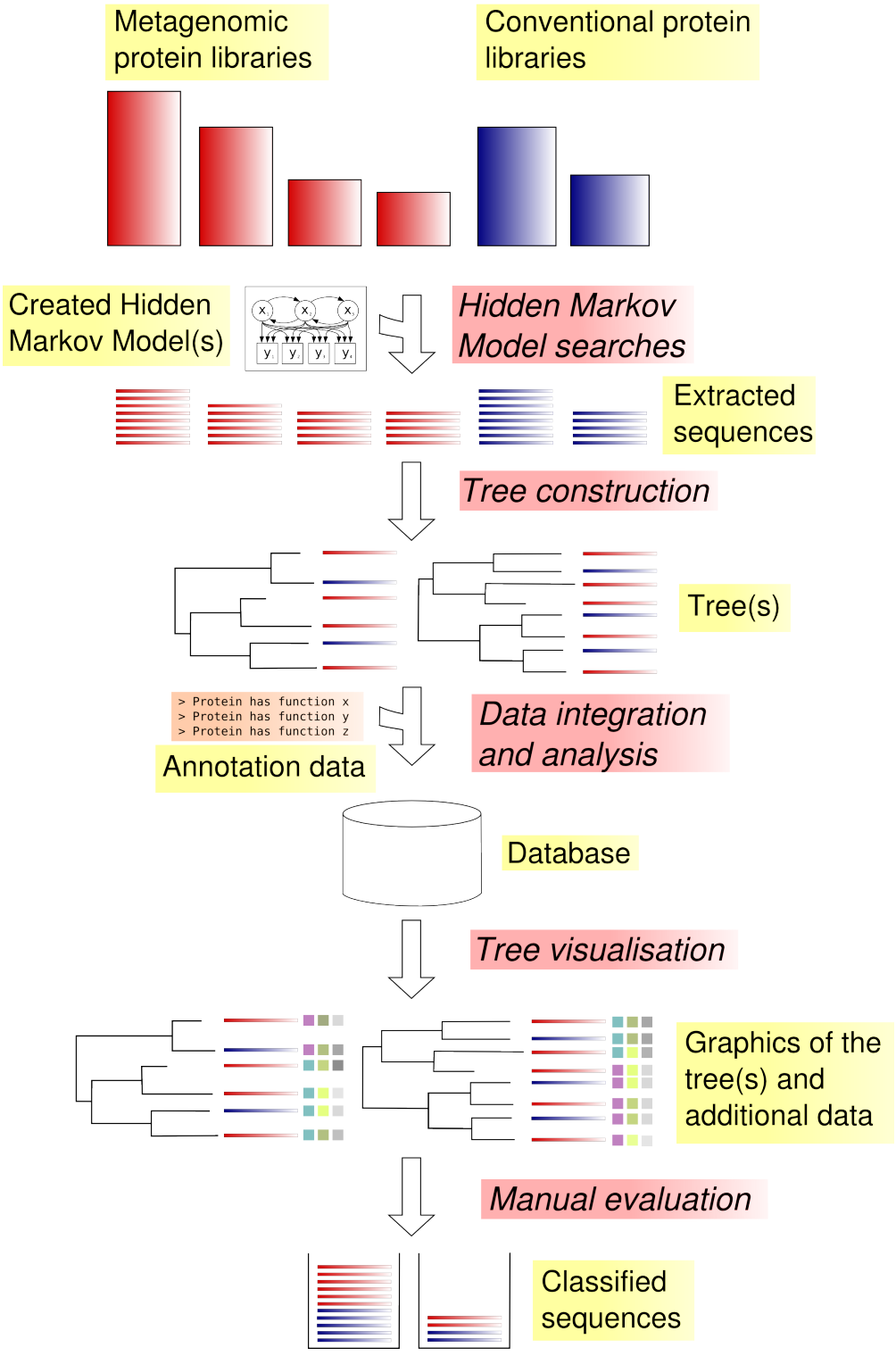


Figure 2: Workflow of the protein screening method.

of the process.

In addition to HMMs and maximum-likelihood trees, the integration and comparative analyses of sequence and annotation data from conventional sequence databases were performed to improve the quality of the results.

With the help of this pipeline we could exemplarily show how to overcome the intrinsic obstacles of metagenomic data screening. It was applied to search for nitrilases [3] (Appendix C), nitrile hydratase [5] (Appendix F) and type I polyketide synthases [6] (Appendix E). The enzymes were chosen due to their different complexity and their importance in biotechnological processes or their pharmaceutical relevance.

Nitrilases (nitrile aminohydrolase, EC 3.5.5.1) catalyse the hydrolysis of nitriles to their corresponding carboxylic acids and ammonia [44] while nitrile hydratases (NHase, EC 4.2.1.84) catalyse the hydrolysis of nitriles to their corresponding amides [45]. Both groups of enzymes are widely applied in the biotechnological production of commodity chemicals [46, 47]. The nitrilases consist of a single peptide, the NHases need two domains that are usually located on two separated subunits.

Polyketide synthases (PKS) are responsible for the synthesis of a large and heterogenous group of secondary metabolites called polyketides that can function as antibiotics, pigments or immunosuppressants [48]. There are many polyketide antibiotics with medical applications (e.g. Erythromycins, Rifamycins) [48]. In our study we focused on type I PKS proteins which usually contain numerous domains of eight different types [49]. These domains are found on one single protein and are arranged in modules. Similar to the synthesis of fatty acids each module fuses in a Claisen condensation step, thus adding a small acetyl unit to the growing polyketide molecule and optionally modifies the newly added part [48].

From the perspective of complexity the nitrilases with their single domain

were the most straight forward ones in our case studies. Because of the need to search for two subdomains NHases were slightly harder to analyse. In contrast, the analysis of the type I PKS protein was quite challenging due to the high number of domains involved and their complicated evolution. Still, the combination of Hidden Markov Models, maximum-likelihood trees and the integration of knowledge from other databases was able to handle these challenges.

Chapter 3

Discussion

3.1 Genomic features

In two studies we were able to show that the analysed environments strongly influence the GC content of their inhabiting species and favour distinct values [1, 3] (Appendix A and C). This is especially true in the case of soil and marine surface water samples. In these samples the trend is even stronger for the third codon base which is under fewer functional restrictions. The GC preferences are reflected in the recruitment of amino acids: the ones which are mainly coded by codons with a GC content that is close to the prefered value (e.g. Alanin has mainly high GC codons, Lysine has mainly low GC codons) are present in an overrepresented manner.

In general there are different scenarios that might explain the underlying selection mechanism independently from the driving factors. The nucleotide composition could be actively steered into a certain direction or the environments could select for newly entering species that already carry genomes with the favoured GC content.

The previously stated correlation between GC content and genome size [34, 37] can be underpinned by these two studies in combination with a later published estimation of effective genome sizes of species from environmen-

tal sequencing projects [50]. It seems that larger genomes as possessed by soil organism have in general a higher GC value than smaller those ones like found in marine environments.

However, this connection is just a correlation and we cannot pinpoint the environmental feature responsible for it. It might be that many environmental and intrinsic factors are involved in the selection of certain GC values [33]. Despite this, these two studies built the foundation for the detection of gene transfer between environments which was done in another of our studies [4] (Appendix D).

3.2 Screening for enzymes

As demonstrated in three of our publications, metagenomic data sets can be treasuries of new enzymes with applications in the chemical and pharmaceutical industry. We were able to detect members of enzymes with intensive use in biotechnological processes or in the production of antibiotics [3, 6, 5] (Appendix C, E and F).

In all three studies we successfully extracted protein sequences with highly sensitive HMMs and used maximum-likelihood trees to decide if the detected proteins belong to the group of targeted enzymes or if this was not the case. New members of the nitrilases were discovered in metagenomic sequence collections and conventional sequences databases [3] (Appendix C). They revealed the need to extend the previoulsy proposed classification of these hydrolytic enzymes evidenced by the presence of new significant subgroups in the trees. Nitrile hydrates were screened in a similar manner and numerous new members were discovered including the first described eukaryotic NHase which carries the usually separated subunit fused in one single protein. For both enzyme classes these results are not only interesting from an evolutionary perspective. Due to their application in the production of commodity chemicals these findings might also have biotechnological impact.

In the screening of type I PKS proteins [6] the maximum-likelihood tree also played an essential role. All of the HMMs were selective enough to distinguish type I from type II PKS members but some were not able to distinguish the type I PKS from the evolutionary closely related type I fatty acid synthases [49]. For these cases the trees were successfully applied to make the distinction between the type I PKS and type I fatty acid synthases. Due to the low length of metagenomic sequences we could not detect complete type I PKS proteins. Yet, our analysis has enabled us to estimate the potential of the environments to serve as resources for potentially new PKS members with pharmacological relevance.

The results of all three screenings show that the combination of HMM based sequence search (contributing sensitivity) followed by the construction of maximum-likelihood trees (to increase the total specificity) turned out to be a powerful tool to harvest the predicted protein from metagenomic data sets as well as from conventional sources (*UniRef* [31]).

3.3 Conclusions

The analyses presented here are case studies which demonstrate that environmental sequence collections can provide useful data to make statements about the influence of environments on their microbial inhabitants. We also presented how these data sets can contribute to the pool of biotechnologically and pharmacologically relevant proteins.

Additionally, it became clear that a comparative approach by taking samples from very different environments into account can help to understand the outstanding features of the single sample and is more meaningful than an isolated analysis.

On the other hand, our analyses reveal some of the difficulties that can occur during the analysis of this type of data and point out the need to adapt approaches and computational methods to deal with them. It is not only the

sample's intrinsic difficulties like short sequence length have to be overcome but also the problem of the absence of standards for such sequencing projects that makes the comparison of different environmental samples challenging.

Despite these difficulties, metagenomics is a powerful approach that increasingly helps to understand the microbial population of numerous biotopes. Environmental sequencing is still at its beginning. Upcoming techniques like non-Sanger sequencing methods [10, 11] and single cell sequencing [51] as well as the improvement of computation methods and the increase of computational resources will speed up the generation and handling of metagenomic data. These developments in different fields will help to tackle current challenges in the analysis of microbial communities and pave the way to further genetic, metabolic and ecological insights.

3.4 Perception of the studies

The publication *Environments shape the nucleotide composition of genomes* [1] (Appendix A) was well accepted by the scientific community (so far 23 citations). The review *Comparative analysis of environmental sequences: potential and challenges* [2] (Appendix B) that gave an early overview of this young research field was cited 10 times until now. The later review *Get the most out of your metagenome: computational analysis of environmental sequence data* [3] (Appendix C) was cited 4 times. The studies *A Molecular Study of Microbe Transfer between Distant Environments* [4] (Appendix D), *A nitrile hydratase in the eukaryote monosiga brevicollis.* [5] (Appendix F) and *A computational screen for type I polyketide synthases in metagenomics* [6] (Appendix E) are just recently published respectively submitted or in press.

Chapter 4

Acknowledgments

This work would not exist without the help of many other people. Research is usually team work and I had the luck to work in a great team and in an excellent environment at EMBL. I want to thank my supervisor Peer Bork for giving me the chance to do my PhD in his group and all my colleagues who supported me with competence and a pleasant, amicable atmosphere – especially my office mates Jeroen Raes, Takuji Yamada and Sean Hooper. Sincere thanks for discussions and very helpful feedback are given to the members of my thesis advisory committee – Toby Gibson, Lars Steinmetz and Thomas Dandekar. Thomas was so kind to accept me as an external PhD student and made it possible for me to have the Julius-Maximilians-Universität Würzburg as partner university.

The time of my PhD was not only from a scientific perspective a huge gain and inspiring for me but also from the personal one. I made many friends for life and would like to thank them for the great time we had and will have.

Last but not least many thanks go to my family for their love and constant support.

Bibliography

[1] Konrad U Foerstner, Christian von Mering, Sean D Hooper, and Peer Bork. Environments shape the nucleotide composition of genomes. *EMBO Rep*, 6(12):1208–1213, Dec 2005.

[2] Konrad U Foerstner, Christian von Mering, and Peer Bork. Comparative analysis of environmental sequences: potential and challenges. *Philos Trans R Soc Lond B Biol Sci*, 361(1467):519–523, Mar 2006.

[3] Jeroen Raes, Konrad U Foerstner, and Peer Bork. Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol*, Oct 2007.

[4] Sean D Hooper, Jeroen Raes, Konrad U Foerstner, Eoghan D Harrington, Daniel Dalevi, and Peer Bork. A molecular study of microbe transfer between distant environments. *PLoS ONE*, 3(7):e2607, 2008.

[5] Konrad U Foerstner, Tobias Doerks, Jean Muller, Jeroen Raes, and Peer Bork. A nitrile hydratase in the eukaryote Monosiga brevicollis. *PLoS ONE*, submitted.

[6] Konrad U Foerstner, Tobias Doerks, Anja Doerks, and Peer Bork. A computational screen for type I polyketide synthases in metagenomics shotgun data. *PloS ONE*, in press.

[7] Wikipedia contributors. Genome. *Wikipedia - The Free Encyclopedia*, oldid=233106637, 20 August, 13:11, 2008.

[8] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, and Joseph M Merrick. Whole-genome random sequencing and assembly of Haemophilus influenzae rd. *Science*, 269(5223):496–512, Jul 1995.

[9] Axel Bernal, Uy Ear, and Nikos Kyrpides. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acids Res*, 29(1):126–127, Jan 2001.

[10] Kelly Rae Chi. The year of sequencing. *Nat Methods*, 5(1):11–14, Jan 2008.

[11] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nat Methods*, 5(1):16–18, Jan 2008.

[12] Nathan Blow. DNA sequencing: generation next-next. *Nature Methods*, 5:267–274, 2008.

[13] Jo Handelsman. Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev*, 68(4):669–85, 2004.

[14] James T Staley and Allan Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, 39:321–346, 1985.

[15] Norman R. Pace. Opening the door onto the natural microbial world: molecular microbial ecology. *Harvey Lect*, 91:59–78, 1995.

[16] David. A Stahl, David J Lane, Gary J Olsen, and Norman R Pace. Characterization of a yellowstone hot spring microbial community by 5S rRNA sequences. *Appl Environ Microbiol*, 49(6):1379–1384, Jun 1985.

[17] Stephen J Giovannoni, Theresa B Britschgi, Craig L Moyer, and Katharine G Field. Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345(6270):60–63, May 1990.

[18] Thomas M Schmidt, Edward F DeLong, and Norman R Pace. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol*, 173(14):4371–4378, Jul 1991.

[19] Wikipedia contributors. Metagenomics. *Wikipedia - The Free Encyclopedia*, oldid=231481389, 12 August, 16:39,, 2008.

[20] Jo Handelsman, Michelle R Rondon, Sean. F Brady, Jon Clardy, and Robert M Goodman. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol*, 5(10):R245–9, 1998.

[21] J Craig Venter, Karin Remington, John F Heidelberg, Aaron L Halpern, Doug Rusch, Jonathan A Eisen, Dongying Wu, Ian Paulsen, Karen E Nelson, William Nelson, Derrick E Fouts, Samuel Levy, Anthony H Knap, Michael W Lomas, Ken Nealson, Owen White, Jeremy Peterson, Jo Hoffman, Rachel Parsons, Holly Baden-Tillson, Cynthia Pfannkoch, Yu-Hui Rogers, and Hamilton O Smith. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74, 2004. Enter text here.

[22] Gene W Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E Allen, Rachna J Ram, Paul M Richardson, Victor V Solovyev, Edward M Rubin, Daniel S Rokhsar, and Jillian F Banfield. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43, Mar 2004.

[23] Douglas B Rusch, Aaron L Halpern, Granger Sutton, Karla B Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, Jonathan A Eisen, Jeff M Hoffman, Karin Remington, Karen Beeson, Bao Tran, Hamilton Smith, Holly Baden-Tillson, Clare Stewart, Joyce Thorpe, Jason Freeman, Cynthia Andrews-Pfannkoch, Joseph E Venter, Kelvin Li, Saul Kravitz, John F Heidelberg, Terry Utterback, Yu-Hui Rogers, Luisa I Falcón, Valeria Souza, Germán Bonilla-Rosso, Luis E Eguiarte, David M Karl, Shubha Sathyendranath, Trevor Platt, Eldredge

Bermingham, Victor Gallardo, Giselle Tamayo-Castillo, Michael R Ferrari, Robert L Strausberg, Kenneth Nealson, Robert Friedman, Marvin Frazier, and J. Craig Venter. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol*, 5(3):e77, Mar 2007.

[24] Susannah Green Tringe, Christian von Mering, Arthur Kobayashi, Asaf A Salamov, Kevin Chen, Hwai W Chang, Mircea Podar, Jay M Short, Eric J Mathur, John C Detter, Peer Bork, Philip Hugenholtz, and Edward M Rubin. Comparative metagenomics of microbial communities. *Science*, 308(5721):554–557, Apr 2005.

[25] Alexander H Treusch, Arnurf Kletzin, Guenter Raddatz, Torsten Ochsenreiter, Achim Quaiser, Guido Meurer, Stephan C Schuster, and Christa Schleper. Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ Microbiol*, 6(9):970–80, 2004.

[26] Shibu Yooseph, Granger Sutton, Douglas B Rusch, Aaron L Halpern, Shannon J Williamson, Karin Remington, Jonathan A Eisen, Karla B Heidelberg, Gerard Manning, Weizhong Li, Lukasz Jaroszewski, Piotr Cieplak, Christopher S Miller, Huiying Li, Susan T Mashiyama, Marcin P Joachimiak, Christopher van Belle, John-Marc Chandonia, David A Soergel, Yufeng Zhai, Kannan Natarajan, Shaun Lee, Benjamin J Raphael, Vineet Bafna, Robert Friedman, Steven E Brenner, Adam Godzik, David Eisenberg, Jack E Dixon, Susan S Taylor, Robert L Strausberg, Marvin Frazier, and J. Craig Venter. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*, 5(3):e16, Mar 2007.

[27] Edward F DeLong, Christina M Preston, Tracy Mincer, Virginia Rich, Steven J Hallam, Niels-Ulrik Frigaard, Asuncion Martinez, Matthew B Sullivan, Robert Edwards, Beltran Rodriguez Brito, Sallie W Chisholm, and David M Karl. Community genomics among stratified microbial

assemblages in the ocean's interior. *Science*, 311(5760):496–503, Jan 2006.

[28] Steven R Gill, Mihai Pop, Robert T Deboy, Paul B Eckburg, Peter J Turnbaugh, Buck S Samuel, Jeffrey I Gordon, David A Relman, Claire M Fraser-Liggett, and Karen E Nelson. Metagenomic analysis of the human distal gut microbiome. *Science*, 312(5778):1355–1359, Jun 2006.

[29] Héctor García Martín, Natalia Ivanova, Victor Kunin, Falk Warnecke, Kerrie W Barry, Alice C McHardy, Christine Yeates, Shaomei He, Asaf A Salamov, Ernest Szeto, Eileen Dalin, Nik H Putnam, Harris J Shapiro, Jasmyn L Pangilinan, Isidore Rigoutsos, Nikos C Kyrpides, Linda Louise Blackall, Katherine D McMahon, and Philip Hugenholtz. Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol*, 24(10):1263–1269, Oct 2006.

[30] Christian von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Krüger, Berend Snel, and Peer Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–D362, Jan 2007.

[31] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, May 2007.

[32] Nicole King, M. Jody Westbrook, Susan L Young, Alan Kuo, Monika Abedin, Jarrod Chapman, Stephen Fairclough, Uffe Hellsten, Yoh Isogai, Ivica Letunic, Michael Marr, David Pincus, Nicholas Putnam, Antonis Rokas, Kevin J Wright, Richard Zuzow, William Dirks, Matthew Good, David Goodstein, Derek Lemons, Wanqing Li, Jessica B Lyons, Andrea Morris, Scott Nichols, Daniel J Richter, Asaf Salamov, J. G I Sequencing, Peer Bork, Wendell A Lim, Gerard Manning, W. Todd Miller, William McGinnis, Harris Shapiro, Robert Tjian, Igor V Grigoriev, and Daniel Rokhsar. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, 451(7180):783–788, Feb 2008.

[33] Stephen D Bentley and Julian Parkhill. Comparative genomic structure of prokaryotes. *Annu Rev Genet*, 38:771–91, 2004.

[34] Huai-Chun Wang, Edward Susko, and Andrew J Roger. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem Biophys Res Commun*, 342(3):681–684, Apr 2006.

[35] Hugo Naya, Héctor Romero, Alejandro Zavala, Beatriz Alvarez, and Héctor Musto. Aerobiosis increases the genomic guanine plus cytosine content (GC) in prokaryotes. *J Mol Evol*, 55(3):260–264, Sep 2002.

[36] C E McEwan, D Gatherer, and N R McEwan. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas*, 128(2):173–178, 1998.

[37] Eduardo P C Rocha and Antoine Danchin. Base composition bias might result from competition for metabolic resources. *Trends Genet*, 18(6):291–294, Jun 2002.

[38] Gerhard J Herndl, Gerald Müller-Niklas, and Jürgen Frick. Major role of ultraviolet-B in controlling bacterioplankton growth in the surface layer of the ocean. *Nature*, 361:717 – 719, 1993.

[39] Patrick Lorenz and Jürgen Eck. Metagenomics and industrial applications. *Nat Rev Microbiol*, 3(6):510–6, 2005.

[40] Taku Uchiyama, Takashi Abe, Toshimichi Ikemura, and Kazuya Watanabe. Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol*, 23(1):88–93, Jan 2005.

[41] Sean R Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[42] Stéphane Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, Oct 2003.

[43] Wikipedia contributors. Hidden markov model. *Wikipedia - The Free Encyclopedia*, oldid=232477305, 17 August, 11:41, 2008.

[44] Helen C Pace and Charles Brenner. The nitrilase superfamily: classification, structure and function. *Genome Biol*, 2(1):REVIEWS0001, 2001.

[45] Michihiko Kobayashi and Sakayu Shimizu. Nitrile hydrolases. *Curr Opin Chem Biol*, 4(1):95–102, Feb 2000.

[46] Toru Nagasawa and Hideaki Yamada. Microbial production of commodity chemicals. *Pure & Appl. Chern.*, 67(7):1241–1256, 1995.

[47] Hideaki Yamada, Sakayu Shimizu, and Michihiko Kobayashi. Hydratases involved in nitrile conversion: screening, characterization and application. *Chem Rec*, 1(2):152–161, 2001.

[48] James Staunton and Kira J Weissman. Polyketide biosynthesis: a millennium review. *Nat Prod Rep*, 18(4):380–416, Aug 2001.

[49] Holger Jenke-Kodama, Axel Sandmann, Rolf Müller, and Elke Dittmann. Evolutionary implications of bacterial polyketide synthases. *Mol Biol Evol*, 22(10):2027–2039, Oct 2005.

[50] Jeroen Raes, Jan O Korbel, Martin J Lercher, Christian von Mering, and Peer Bork. Prediction of effective genome size in metagenomic samples. *Genome Biol*, 8(1):R10, 2007.

[51] Thomas Ishoey, Tanja Woyke, Ramunas Stepanauskas, Mark Novotny, and Roger S Lasken. Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol*, 11(3):198–204, Jun 2008.

Appendix A

Environments shape the nucleotide composition of genomes

EMBO reports

scientific report

Environments shape the nucleotide composition of genomes

Konrad U. Foerstner[1], Christian von Mering[1], Sean D. Hooper[1] & Peer Bork[1,2+]

[1]European Molecular Biology Laboratory, Heidelberg, Germany, and [2]Max-Delbrück Centre for Molecular Medicine, Berlin, Germany

To test the impact of environments on genome evolution, we analysed the relative abundance of the nucleotides guanine and cytosine ('GC content') of large numbers of sequences from four distinct environmental samples (ocean surface water, farm soil, an acidophilic mine drainage biofilm and deep-sea whale carcasses). We show that the GC content of complex microbial communities seems to be globally and actively influenced by the environment. The observed nucleotide compositions cannot be easily explained by distinct phylogenetic origins of the species in the environments; the genomic GC content may change faster than was previously thought, and is also reflected in the amino-acid composition of the proteins in these habitats.

Keywords: ecology; environment; evolution; GC content; metagenomics

EMBO reports (2005) **6**, 1208–1213. doi:10.1038/sj.embor.7400538

INTRODUCTION

The relative abundance of the nucleotides guanine and cytosine ('GC content') varies widely between genomes of different species and even between entire phyla (Sueoka, 1962). However, it is unclear whether this is due to intrinsic, organism-specific mechanisms or external factors, and whether it is the result of neutral processes or selection. Several hypotheses have been put forward to explain variations in the GC content of organisms, some of which are controversial (discussed by Bentley & Parkhill, 2004). These hypotheses are often based on observed, simple correlations of GC content with another (intrinsic or extrinsic) measure. One of the intrinsic correlations is a tendency of large genomes to be GC rich and small genomes to be GC poor (Heddi *et al*, 1998; Moran, 2002; Rocha & Danchin, 2002). Because large genomes are presumably found in more complex, variable environments, there could be an indirect link between GC content and niche complexity. One possible reason for this is the higher cost of synthesis of ATP than of UDP (in complex environments, growth and ATP synthesis are presumed to be slower). The need for being able to quickly mobilize ATP may also have a role in the case of small genomes (Rocha & Danchin, 2002). As random mutations of DNA are mainly the conversion from C to T and from G to A, the lack of repair mechanisms in reduced genomes could also be a reason for small genomes being AT rich (Glass *et al*, 2000). Another factor could be the preferred growth temperature of an organism, which has been proposed to correlate with GC content (Musto *et al*, 2004), but this is under debate (Marashi & Ghalanbor, 2004; Musto *et al*, 2005). Growth temperature is known to correlate with polypurine (AG) tracts in messenger RNAs (Lobry & Chessel, 2003; Paz *et al*, 2004). Although this alone does not preclude a correlation with GC, it would disfavour extreme GC levels in thermophilic organisms. It has been observed that genomes of some nitrogen-fixing organisms contain a higher fraction of guanine and cytosine than the genomes of nonfixing species of the same genus (McEwan *et al*, 1998). Likewise, Naya *et al* (2002) put forward a connection between an aerobic lifestyle and an increased GC content.

However, most of the above correlations are not very strong, and could obviously be merely indirect consequences of other, as yet unknown, factors that influence genomic GC content more directly. Another complication is that, so far, the field has focused on available genome sequences, which are derived from single isolates from a wide variety of environments. This has precluded the analysis of community effects (in natural settings, microbes may live in large communities of hundreds or thousands of different species), and of global influences of the environment. In addition, it neglects the large fraction of environmental microbes that resist cultivation in the laboratory (Staley & Konopka, 1985). Only recently, random shotgun sequence data from environmental DNA preparations have become available, allowing an unbiased view on the genomic characteristics of an entire environmental community. Here we show, using the large-scale data from Sargasso Sea surface water (Venter *et al*, 2004), from a biofilm in an underground acid drainage mine (Tyson *et al*, 2004) as well as from farm soil and deep-sea whale carcasses (Tringe *et al*, 2005), that the environment indeed has a considerable impact on GC content and implicitly also on the amino-acid composition of the proteins in a habitat.

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
[2]Max-Delbrück Centre for Molecular Medicine, Robert-Rössle-Strasse 10, 13092 Berlin, Germany
[+]Corresponding author. Tel: +49 6221 387 8526; Fax: +49 6221 387 517; E-mail: bork@embl.de

Received 7 July 2005; revised 15 August 2005; accepted 19 August 2005; published online 30 September 2005

RESULTS

Unexpected GC-content distributions in environments

To obtain a representative, quantitative estimate of the environmental GC-content distribution, raw sequencing reads were analysed (not the assembled contigs). However, analysis of raw sequencing reads may generate some inaccuracies, as they can contain regions of poor sequencing quality. Therefore, consistency checks of increasing stringency were executed, invariably confirming the initial GC-content distributions, whether by limiting the analysis to open reading frames with clear homology, or even to a restricted set of translation-related marker genes (see Methods for details).

Owing to the large amounts of DNA (numerous independent reads totalling more than 100 Mbp for each of the four habitats), the GC-content patterns are very robust, and (sub)samples from similar environments tend to have similar GC-content patterns (Fig 1A). Surprisingly, the samples from farm soil and ocean surface water—both of which contain DNA from more than 1,000 diverse, non-abundant species (Venter *et al*, 2004; Tringe *et al*, 2005)—are very different, with the surface water sample having a GC-content median of around 34% and the soil sample around 61%. To test whether these differences are simply the result of distinct phylogenetic compositions of the samples, we estimated the GC-content distribution that the environments were expected to have, on the basis of the known abundances of the various phyla and the GC content of previously known genomes from these phyla. Both water and soil samples deviated strongly from expectations (Fig 1B; supplementary Fig 1 online; expected distributions were estimated by re-creating the communities from known genomes and matching the reported phylogenetic compositions). Strikingly, the GC content in these two complex environments is more narrowly distributed than that of most bacterial phyla, which is unexpected as the environments contain species from many phyla and should therefore have an even broader distribution than the 162 completely sequenced genomes known today (see bottom of Fig 1A for comparison). In addition, we observe that GC-content differences exist even for closely related sequences (Fig 2B), suggesting an active, continuing process.

The above trends are weaker for the acidic biofilm and the whale carcasses, but these environments are much younger (far less than 100 years old; Tyson *et al*, 2004; Tringe *et al*, 2005), and seem to contain only a few species.

Unconstrained nucleotides show the largest differences

To avoid possible biases due to habitat-specific, perhaps unusual, features of non-coding DNA and to measure functional constraints, we restricted the analysis to the open reading frames themselves (of length 150 codons or longer; Fickett, 1995), and analysed the GC-content distribution separately for each of the three codon positions. We found that the third codon position is even more extreme with respect to GC distribution than the average of all three positions (Fig 2, the median in farm soil is 74%, versus 24% in the ocean surface water). The third codon position is relatively free to evolve (owing to the degeneracy of the genetic code), and its extreme GC-content distribution suggests that the process that drives GC-content changes is (at least to some extent) kept in check by coding requirements.

Global differences in amino-acid usage in proteins

The overall frequencies of the various amino acids in encoded proteins are known to vary with changes in overall GC content in microbial genomes (Sueoka, 1961). To confirm and assess this dependency in the case of environmental communities, we globally counted amino acids in predicted proteins, and computed the relative fraction of each amino acid in the various samples (Fig 1C; supplementary Table 1 online). The following amino acids are encoded by AT-rich codons, and are thus expected to be over-represented in low-GC environments: F, Y, M, I, N and K. Conversely, the following amino acids are expected to be over-represented in high-GC environments: G, A, R and P. The abundance ratio of the two groups (the so-called 'FYMINK/GARP' index; Foster *et al*, 1997) correlates inversely with overall GC content, as expected (supplementary Table 1 online).

DISCUSSION

Environmental microbial communities seem to show distinct, and unexpectedly narrow, GC-content distributions. The observed GC patterns are not simply a result of differing species compositions in each environment, as simulations of these compositions using sequenced genomes with the same phylogenetic distribution results in distinct GC patterns (see Fig 1B for a striking example; also see supplementary Fig 1 online). Even closely related sequences, when they are from different environments, show marked differences in GC content, more so than when they are from the same environment (Fig 2B). We can exclude an impact of certain enriched gene families, because the differences remain when the analysis is restricted to a set of essential genes that occur only once per genome and are present in each environment (Fig 1B; supplementary Fig 1 online). However, we cannot completely rule out effects due to differences in experimental protocols (such as DNA preparation or cloning). A weak correlation between genome size and GC content (Moran, 2002; supplementary Fig 2 online) might reflect one possible environmental impact: genomes in ocean surface water are smaller than in soil (Venter *et al*, 2004; Tringe *et al*, 2005). In any case, the narrow distributions of the GC content in complex habitats indicate that mainly external environmental factors influence the GC nucleotide composition of a community, either selectively or by causing a directed, mechanistic mutational bias. These factors have to be more global than the previously suggested lifestyle influences (Bentley & Parkhill, 2004), such as the use of oxygen as an energy source (Naya *et al*, 2002), the ability to fix nitrogen (McEwan *et al*, 1998) or differences in effective population size (Moran, 1996; also see supplementary Fig 4 online). One possibility would be ultraviolet irradiation, which is particularly high in surface water, to the extent that it influences bacterioplankton productivity (Herndl *et al*, 1993). Whatever is causing the differences in GC content, it could either actively change the GC content of the existing organisms in an environment, or alternatively, it could limit the type of microbes that can successfully populate an environment in the first place. Genome-wide changes of GC content are thought to occur on relatively slow timescales—1% of change in CG content is projected to require about 3 Mio years (Haywood-Farmer & Otto, 2003). In contrast, microbial communities are presumably broken up and re-assembled on much shorter timescales (open oceans, for example, have strong water currents—with global ocean mixing occurring fast, in only a

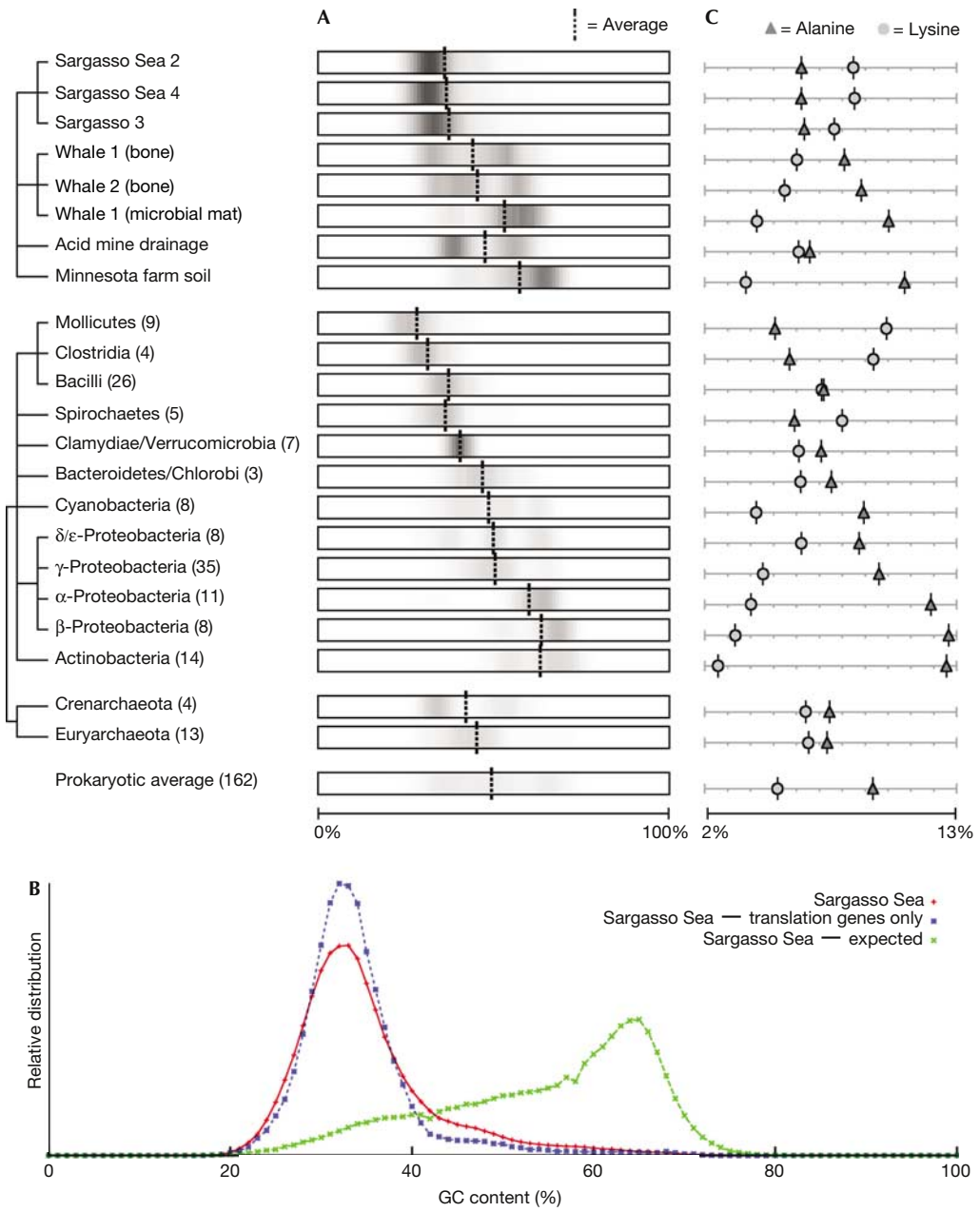


Fig 1 | Guanine and cytosine content of environmental sequences. Guanine and cytosine content distributions and predicted frequencies of amino acids in four environments (eight subsamples in total, all containing >90% prokaryotic species), compared with completely sequenced prokaryotic genomes grouped into phyla and subphyla. The trees depict the relationships between the samples (Tringe *et al*, 2005), and between phyla and subphyla to which the genomes belong. The number of sequenced genomes available for each taxonomic group is given in parentheses. Only phyla with at least three completely sequenced genomes have been included, and only those environmental sequence fragments that contain at least one predicted open reading frame with significant similarity to a known gene (60 bits or better) are shown. (**A**) Relative distributions of Guanine and cytosine (GC) content values, averaged over individual sequence reads. For comparability, virtual reads were generated for completely sequenced genomes. The darker the colour, the higher the number of reads with the respective GC content. Vertical dashed lines denote the average value of each sample/group. (**B**) Comparison of the GC distribution of Sargasso Sea reads (subsamples #2–#4) with (i) a subset that contains only translation genes occurring once per genome and (ii) with a simulated sample derived from completely sequenced genomes and selected to contain the same distribution of phyla. Translation genes show a distribution similar to the whole set, indicating that no bias is introduced by gene content (larger genomes may contain many genes with unusual GC content); the deviation from the simulated sample shows that GC content is apparently not always a simple function of the broad phylogenetic distribution of the species in an environment. (**C**) Frequencies of the amino acids lysine and alanine among encoded proteins. Notice the dependency on GC content (for other amino acids, as well as a compound index, see supplementary Table 1 online).

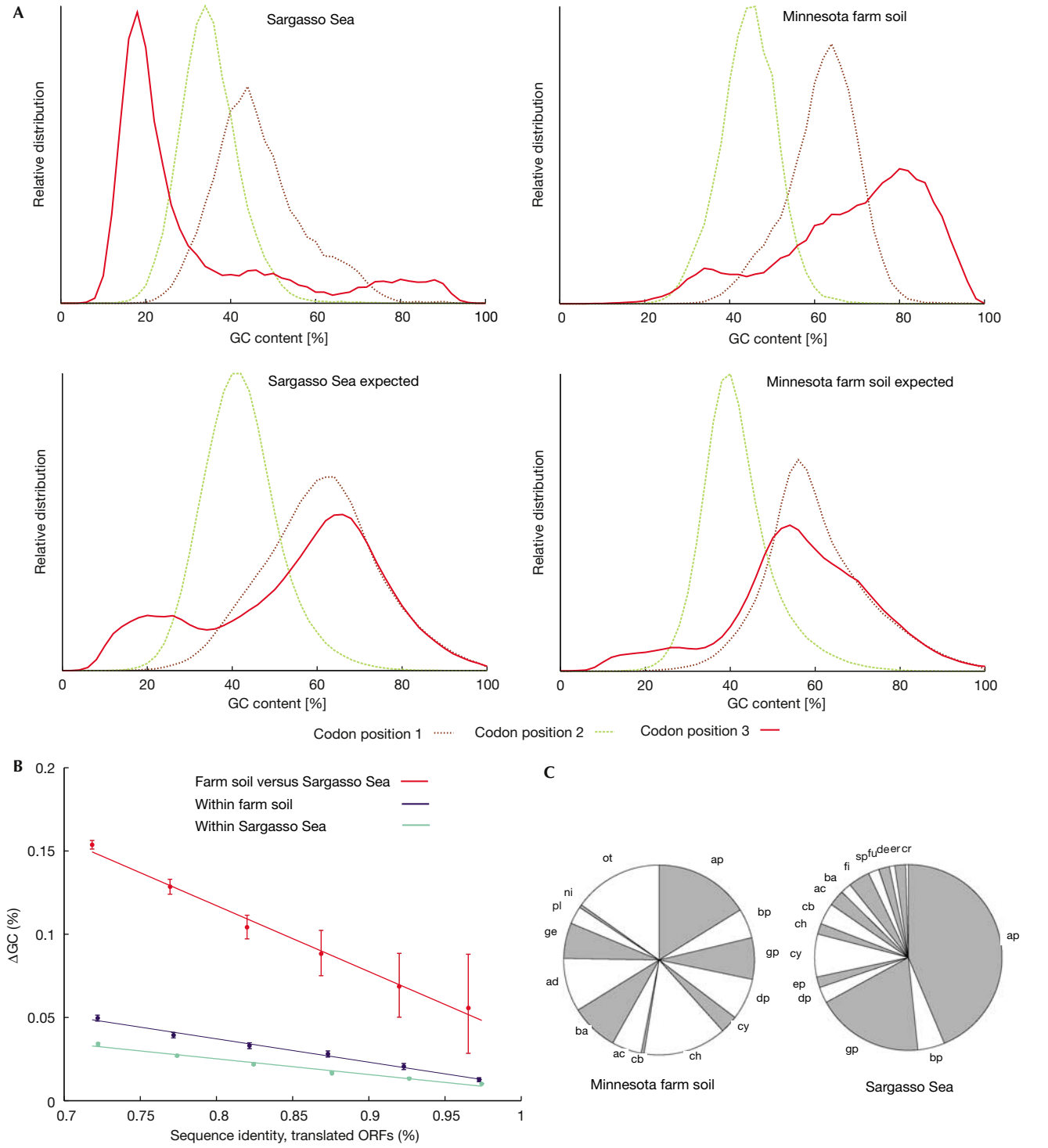


Fig 2 | Guanine and cytosine content analysis of open reading frames. (**A**) Deviation from expectation. Guanine and cytosine (GC) content distributions are shown for each environmental sample, separately for each codon position. The curves are compared with the expected distributions; the latter were derived from known genomes by sampling their DNA in amounts matching the overall phylogenetic compositions reported for the samples. (**B**) GC-content differences for paired open reading frames (ORFs) of high sequence similarity (that is, recent divergence). ORFs were paired on the basis of reciprocal best matches in BLAST searches (see supplementary Figure 3 online for more details). Error bars denote 90% confidence intervals of the mean. (**C**) Phylogenetic distributions of organisms, as reported from 16S ribosomal RNA analysis, for two principal samples. Note the wide range of phyla present. ac, Actinobacteria; ad, Acidobacteria; ap, α-Proteobacteria; ba, Bacteriodetes; bp, β-Proteobacteria; cb, Chlorobi; ch, Chloroflexi; cr, Crenarchaeota; cy, Cyanobacteria; de, Deinococcus-Thermus; dp, δ-Proteobacteria; ep, ε-Proteobacteria; er, Eryarchaeota; fi, Firmicutes; fu, Fusobacteria; ge, Gemmatimonadetes; gp, γ-Proteobacteria; ni, Nitrospira; ot, others; pl, Planctomycetes; sp, Spirochaetes.

few centuries; Stuiver *et al*, 1984). This would argue that community GC-content patterns originate at the time of community assembly, by selective pressures restricting the set of appropriate organisms from a larger pool of available organisms. Supporting this, we observe (in all environments tested) that the distribution of GC content is much more narrow than the GC content of a simple, unbiased mix of all prokaryotes known at present (Fig 1A).

The observed GC-content differences have a direct impact on the amino-acid composition of proteins in the respective environments (Fig 1C), a correlation (Sueoka, 1961) that is well established for individual genomes (Bharanidharan *et al*, 2004), and that can now be extended to the genetic material of whole communities. GC-rich communities contain more amino acids encoded by GC-rich codons, whereas the opposite is true for GC-poor communities (Fig 1C; supplementary Table 1 online). Considering the relatively young age of any given microbial community, it seems that the local amino-acid usage fluctuates rapidly, complementary to a drift at evolutionary timescales that has been observed recently (Jordan *et al*, 2005).

METHODS

Data. At the time of this study, four distinct environments had been analysed through cultivation-independent, large-scale DNA shotgun sequencing ('large scale' being arbitrarily defined as more than 100 Mbp of raw sequence): surface sea water from the Sargasso Sea (Venter *et al*, 2004); a pair of deep-sea whale carcasses ('whale fall') from distinct geographic locations (Tringe *et al*, 2005); an acidophilic biofilm from an underground mine drainage flow (Tyson *et al*, 2004); and agricultural surface soil from a farm in Minnesota (Tringe *et al*, 2005). Collectively, more than 2 Gbp of sequence data are available, and they provide the first opportunity for an unbiased assessment of the nucleotide composition of community DNA, because previous DNA collections (PCR based or cultivation dependent) can be assumed to have substantial experimental bias (Suzuki & Giovannoni, 1996). For all four environments, most of the sequences found (>90%) were from prokaryotic organisms, together with an unknown fraction of associated bacteriophages (but phage DNA did not influence the results; see below for specific tests).

Sargasso Sea surface ocean water. For this environment, a total of 1,986,782 raw sequencing reads are available (Venter *et al*, 2004) from seven different water samples (~2 Gbp of raw sequence). We chose to limit the analysis to samples #2–#4, constituting about 51% of the data, for two reasons: sample #1 is somewhat controversial (Delong, 2005), being the only sample that contains several dominating species—large fractions of their complete genomes could actually be assembled from the data. These dominating species showed a suspiciously low number of polymorphisms, and were not re-discovered in an independent sample from the same site. Therefore, it cannot be excluded that sample #1 has a certain fraction of clonally expanded, contaminating microbes—which is why it was omitted here. Samples #5–#7 were omitted because they had undergone various changes in filtering regimes (some selecting for large particle sizes only), and because they were not used for the assembly in the original publication.

Minnesota farm surface soil. This data set consists of 198,529 raw sequencing reads (220 Mbp). However, the library preparation procedure that was applied to this sample included an amplification step, resulting in several clones with identical inserts. After removal of this redundancy, 149,139 sequencing reads remained, which were used for the present analysis.

Acidic mine drainage biofilm. In all, 124,805 raw sequencing reads have been generated for this sample (Tyson *et al*, 2004), totalling about 124 Mbp of sequence. The original publication focused mainly on those reads that contributed to genome assembly, but for this study all reads were considered, independent of assembly.

Deep-sea whale carcasses ('whale fall'). Three subsamples have been analysed (Tringe *et al*, 2005), from two distinct carcasses, generating a total of 116,464 raw sequencing reads. The two carcasses are from distinct geographic locations, several thousand miles apart.

All four environments vary with respect to the relative abundance and diversity of the bacterial species they contain. This leads to marked differences in the extent to which the raw reads could be assembled into larger contigs. The most extensive assembly was reported for the acid mine drainage community—here, more than two-thirds of the sequencing reads could be assembled into contigs (enabling the almost complete assembly of five genomes). At the other extreme, less than 1% of the soil sequences could be assembled (arguing for a very large diversity of species in the soil). The other two environments were between these two extremes: assembly rates were about 60% and 45% for the Sargasso Sea data and the whale-fall data, respectively.

GC-content distributions. Generally, GC content was measured separately for each read, and all the values for an entire sample were then binned and plotted as a relative distribution of GC content. This indicates that the 'window size' of the GC-content measurement was equivalent to the average read length (between 900 and 1,100 bp, depending on sample). As a first consistency check, the analysis was limited to reads that showed an unequivocal homology to a known protein (scoring at least 60 bits in BLAST searches), or had been properly assembled into a longer contig that showed such homology (Fig 1A,B; supplementary Fig 1 online). This procedure filtered out reads of overall poor quality. As a second check, the analysis was further restricted to sequences that were clearly homologous to a set of 61 marker genes known to be present in all prokaryotic genomes studied so far, usually as single-copy genes (Fig 1B; supplementary Fig 1 online). This ensured that the result was not influenced by gene families of unknown or peripheral function that are potentially more amenable to horizontal transfer. The check also excluded any influence of bacteriophages, because the set of 61 marker genes—mainly ribosomal and translation-related genes—is usually absent from phages and viruses.

Expected GC-content distributions. For each environmental data set, the approximate phylogenetic distribution of organisms was known (from marker genes or ribosomal RNA sequences). This allowed the computation of an expected GC-content distribution on the basis of traditional genome sequences, as follows: expected distributions were generated by sampling—from the 163 complete prokaryotic genome sequences in the STRING database (von Mering *et al*, 2005)—DNA fragments of lengths comparable with raw sequencing reads (a further two recent genomes were included to cover phyla that are not yet represented in STRING). The various phyla to be sampled were weighted to match the phylum distribution of the environmental sample studied (within

each phylum, genomes were sampled evenly). From the sampled reads, the GC-content distributions were derived exactly in the same way as for the environments (Fig 1B; supplementary Fig 1 online). **Supplementary information** is available at *EMBO reports* online (http://www.emboreports.org).

ACKNOWLEDGEMENTS

This work was supported by the European Union, grant no. LSHG-CT-2003-503265. S.D.H. was supported by the Knut and Alice Wallenberg foundation.

REFERENCES

Bentley SD, Parkhill J (2004) Comparative genomic structure of prokaryotes. *Annu Rev Genet* **38:** 771–792

Bharanidharan D, Bhargavi GR, Uthanumallian K, Gautham N (2004) Correlations between nucleotide frequencies and amino acid composition in 115 bacterial species. *Biochem Biophys Res Commun* **315:** 1097–1103

Delong EF (2005) Microbial community genomics in the ocean. *Nat Rev Microbiol* **6:** 459–469

Fickett JW (1995) ORFs and genes: how strong a connection? *J Comput Biol* **2:** 117–123

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* **44:** 282–288

Glass JI, Lefkowitz EJ, Glass JS, Heiner CR, Chen EY, Cassell GH (2000) The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407:** 757–762

Haywood-Farmer E, Otto SP (2003) The evolution of genomic base composition in bacteria. *Evol Int J Org Evol* **57:** 1783–1792

Heddi A, Charles H, Khatchadourian C, Bonnot G, Nardon P (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *J Mol Evol* **47:** 52–61

Herndl GJ, Mullerniklaus G, Frick J (1993) Major role of ultraviolet-B in controlling bacterioplankton growth in the surface-layer of the ocean. *Nature* **361:** 717–719

Jordan IK, Kondrashov FA, Adzhubei IA, Wolf YI, Koonin EV, Kondrashov AS, Sunyaev S (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature* **433:** 633–638

Lobry JR, Chessel D (2003) Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. *J Appl Genet* **44:** 235–261

Marashi SA, Ghalanbor Z (2004) Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. *Biochem Biophys Res Commun* **325:** 381–383

McEwan CE, Gatherer D, McEwan NR (1998) Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* **128:** 173–178

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci USA* **93:** 2873–2878

Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108:** 583–586

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett* **573:** 73–77

Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G (2005) The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor. *Biochem Biophys Res Commun* **330:** 357–360

Naya H, Romero H, Zavala A, Alvarez B, Musto H (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* **55:** 260–264

Paz A, Mester D, Baca I, Nevo E, Korol A (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proc Natl Acad Sci USA* **101:** 2951–2956

Rocha EP, Danchin A (2002) Base composition bias might result from competition for metabolic resources. *Trends Genet* **18:** 291–294

Staley JT, Konopka A (1985) Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39:** 321–346

Stuiver M, Quay PD, Ostlund HG (1984) Abyssal water carbon-14 distribution and the age of the world oceans. *Science* **219:** 849–851

Sueoka N (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc Natl Acad Sci USA* **47:** 1141–1149

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc Natl Acad Sci USA* **48:** 582–592

Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62:** 625–630

Tringe SG *et al* (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37–43

Venter JC *et al* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33:** D433–D437

Appendix B

Comparative analysis of environmental sequences: potential and challenges

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B
FirstCite® e-publishing

Phil. Trans. R. Soc. B
doi:10.1098/rstb.2005.1809
Published online

Comparative analysis of environmental sequences: potential and challenges

Konrad U. Foerstner[1,†], Christian von Mering[1,†] and Peer Bork[1,2,*]

[1]*European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg 69117, Germany*
[2]*Max-Delbrück-Centre for Molecular Medicine, Robert-Rössle-Strasse 10, Berlin 13092, Germany*

Environmental sequencing, also dubbed metagenomics, is increasingly being used to obtain insights into organismal communities in diverse habitats, and has a variety of potential applications foreseeable in biotechnology and medicine. The first public large-scale data provide already a wealth of information hidden in vast amounts of fragmented pieces of DNA from unknown species residing in these environments. Comparative sequence analysis is essential for the interpretation of such data. However, different layers of complexity that are intrinsic to each sample require the establishment of some baselines for comparison: how to normalize for the differences in phylogenetic and functional diversity, how to avoid biases from incomplete data, and how to deal with differences in species dominance or genome sizes? Here we discuss a few of these items and delineate some simple discriminative sequence properties for four distinct habitats.

Keywords: comparison; diversity; environments; metagenomics

1. INTRODUCTION

After the delivery of the first completely sequenced bacterial genomes in 1995, environmental sequencing was already extensively discussed as a promising avenue (Stein *et al*. 1996), and the term 'metagenome' for the collective genomic information of a habitat appeared in the scientific literature as early as 1998 (Handelsman *et al*. 1998). Yet, until recently, it was mostly the sequencing of large amounts of 16/18S rRNA that gave the first insights into the species complexity within a number of different habitats (e.g. Rappe & Giovannoni 2003), whereby bacterial species seem by far the most abundant. All together, more than 120 000 sequences of 16S rRNAs from different prokaryotic species are currently captured in databases such as RDP (Cole *et al*. 2005). In contrast to the large numbers of species implied by their rRNA sequences, there are so far only a little more than 200 completely sequenced genomes published, and any in-depth analysis of building plans and functional repertoires is limited to those (mostly prokaryotic) species. Furthermore, the current genome sequences represent a biased view of living matter on earth, as they have been derived from a very few eukaryotic model organisms and from a variety of prokaryotes that can be cultivated and grown in a laboratory. However, cultivation is only possible (using standard conditions) for about 1% of all microbial species, and natural populations are greatly distorted under laboratory conditions (this is known as 'great plate anomaly' Staley & Konopka 1985). Only in 2004, the first large-scale metagenomics studies appeared (Tyson *et al*. 2004; Venter *et al*. 2004), which were cultivation-independent because they employed 'shotgun' approaches directly on environmental DNA preparations. To sub-clone the DNA, various strategies are being used, and especially the long-insert fosmid or BAC libraries are very promising for the future; they have already delivered the first results, either through random end-sequencing or through screening for and targeted sequencing of specific functional systems (Beja *et al*. 2002; Treusch *et al*. 2004).

Whatever technology will be driving the data generation a few years from now, it is already clear that massive environmental sequencing is feasible and that it will generate a wealth of data for basic science, but also for more direct applications in many disciplines. The first areas that come to mind are biotechnology and medicine, with surveys for pathogens (Schmeisser *et al*. 2003) or the discovery of novel antibiotics and specific degradation pathways to be utilized, but applications are likely to be much more far-reaching (see figure 1 for a few of the potentials and hopes).

Here, we will explore the first large-scale metagenomics datasets available, and discuss some of their properties and how they can be compared. In contrast to complete genomes, which are defined entities, all these data are incomplete so far to an almost unknown extent, perhaps analogous to the first EST data that were generated in the early 1990s, stimulating speculations on human gene numbers. More importantly, we will point to different layers of complexity that are imposed by differences in experimental and computational protocols and raise the question of how to compare the different datasets in a meaningful way. Despite these notes of caution, we claim that it is possible to extract specific information towards both the phylogenetic and functional characterization of

*Author for correspondence (bork@embl.de).
†These authors contributed equally to the study.

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

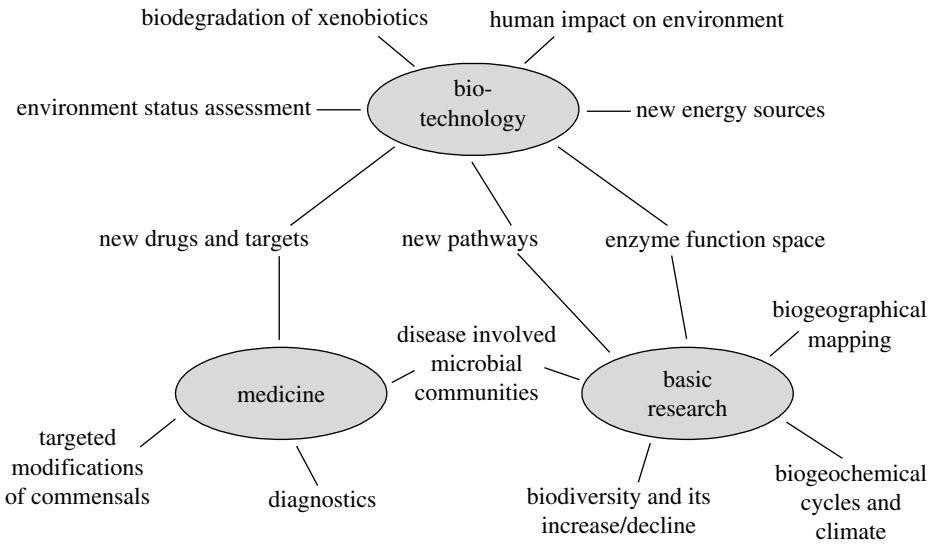


Figure 1. Potential applications of environmental sequencing approaches.

microbial communities if one is aware of possible biases and formulates the questions accordingly.

2. CHARACTERIZING THE FIRST LARGE-SCALE METAGENOMICS DATASETS: APPLES AND ORANGES

The first truly large-scale random shotgun sequencing data from an environment have been published only recently (Tyson *et al.* 2004), characterizing an underground biofilm under extremely acidic conditions (less than pH 1) in an iron mine drainage path. Just a month later, a much more complex environmental sample from surface water of the Sargasso sea has been reported (Venter *et al.* 2004), containing an order of magnitude more data (see table 1). This latter dataset alone comprises more predicted open reading frames (ORFs) than contained in all the completely sequenced genomes available at the time (although metagenomics ORFs are sometimes fragmented). Early in 2005, two more shotgun datasets have been released, from yet other, very different habitats, namely 116 Mbp from whalebone samples in more than 500 m water depth in two different oceans (hereafter whalefall), as well as 208 Mbp from surface soil on a Minnesota farm (Tringe *et al.* 2005; see table 1 for a summary). Several more datasets of up to 200 Mb are underway, as is a more data-rich and systematic sampling of ocean water.

Although the resulting sequences are hard data, the experimental sampling protocols can be quite different, leading to considerable biases. For example, size filters have been used in the Sargasso sea that are likely to select against small viruses as well as against larger eukaryotic cells. This is simplifying the analysis of prokaryotic diversity, but has to be taken into account when re-analysing and comparing the data to other samples. Furthermore, as the data come from different laboratories, the protocols for read quality filtering, assembly and gene prediction can vary considerably, making it difficult to compare basic properties between different habitats such as the number of annotated ORFs or the degree of assembly. This will also have an impact on downstream analyses, such as determining the phylogenetic or functional composition.

Unfortunately (for details see table 1), not only the habitats, sampling procedures and the data treatments vary considerably but also the nature of the data itself. In some environments, certain species dominate, as exemplified in the acid mine drainage sample where five prokaryotes contribute greater than 80% of all the sequences obtained (notably, one of them, *Leptospirillum*, was the first sequenced member of an entire phylum, that of *Nitrospira*, illustrating the bias in classical genome sequencing).

On the contrary, the assembly rate of the much more complex soil data (less than 1%) indicates that a single species is unlikely to be abundant in this sample. It has been estimated that at least 1 Gbp (Tringe *et al.* 2005) would have to be sequenced before the most abundant species could be reasonably covered by assembling the reads. Thus, while the amount sequenced might have been sufficient to capture the major trends and functional repertoires in the acid mine drainage data, the coverage of the soil might still not be fully representative despite consisting of more than 200 Mbp of raw sequence.

Another factor to consider is the diversity of species within an environment, which is presumably much higher in 0.5 g of soil than even in hundreds of litres of ocean water (e.g. Torsvik *et al.* 2002). This is also reflected in higher estimates of species numbers: more than 3000 in the soil sample versus 1800 in the Sargasso sea samples (Venter *et al.* 2004; Tringe *et al.* 2005). In addition, the heterogeneity of a sample (0.5 g of soil harbours various differently populated subhabitats) and the number of individuals can only be estimated, yet will impact the data. The different constraints imposed by the environments are reflected in the genome sizes (estimates range from 2 to 6 Mbp in water and soil, respectively; Venter *et al.* 2004; Tringe *et al.* 2005). This all makes it difficult to extrapolate from individual ORFs to entire species in a sample and leaves a considerable uncertainty in ORF-based estimates. However, the elucidation of the

Table 1. Large-scale environmental sequencing projects: properties and scope.

	acid mine drainage	Sargasso sea[a]	farm soil	whale falls
particle size filtering	none	>0.1 μm; <0.8 μm	none	none
number of subsamples	1	4[a]	1	3
total amount sequenced–raw	124 Mbp	1687 Mbp	208 Mbp	116 Mbp
total amount sequenced–quality filtered	76 Mbp	1350 Mbp	104 Mbp[b]	78 Mbp
read average size–raw	996 bp	1015 bp	1046 bp	993 bp
read average size–quality filtered	737 bp	818 bp	696 bp	673 bp
fraction of reads failing any assembly	∼20%	∼40%	>99%	∼55%
genomes reported as largely assembled	5	3	none	none
number of ORFs annotated	>12 000	>1 000 000	>180 000	>120 000
minimum number of species found	5	1000	847[c]	17[c,d]
estimated total number of species	n.r.	>1800	>3000	25–150[d]
reference	(Tyson *et al.* 2004)	(Venter *et al.* 2004)	(Tringe *et al.* 2005)	(Tringe *et al.* 2005)

[a] not including data from the Sorcerer II expedition–these data (samples 5–7) were not considered in the original publication (Venter *et al.* 2004) for the pooled assembly; in addition, they were generated using a variety of different filtering protocols.
[b] filtering here included removal of redundant reads generated by library amplification prior to cloning.
[c] 'ribotypes'; species defined as having 97% identical rRNA sequences.
[d] depending on sub-sample studied.

phylogenetic composition of the communities in each sample remains one of the big scientific challenges in metagenomics. Is the current overrepresentation of proteobacteria in the set of completely sequenced genomes a result of their general abundance, or of a sampling bias? They certainly seem to dominate in the more complex samples of soil and surface water, but this might be a chicken-and-egg problem as we can possibly identify them better than other phyla, knowing more about them already.

3. PHYLOGENETIC VERSUS FUNCTIONAL DIFFERENCES BETWEEN METAGENOMES

While several metagenome properties are obvious, or easy to obtain (e.g. table 1), other features such as the phylogenetic spectrum or the functional repertoire of a sample are more difficult to compute due to the different nature of the samples. A simplifying, best-hit similarity analysis of the ORFs should nevertheless give some rough trends (table 2), although even there major biases could have been introduced. For example, virus genes tend to evolve quickly and their homologues will be easily overlooked, and the size filter used for the Sargasso sea data introduces an extra bias against viruses in this particular sample. Furthermore, many of the predicted ORFs do not have any obvious homologue in the public databases so far. For the most complex soil data, as many as 47% of the reads do not show any obvious hit and even in the sample for which most ORFs have an homology assignment, that of the Sargasso sea, more than a quarter of all ORFs seem entirely novel. This fraction could easily be enriched in viruses, or hitherto undescribed archaea, making the estimates in table 2 even less reliable. What the data do confirm is that the bacterial domain contributes by far the most ORFs in complex environments, and also that extreme habitats can indeed differ. Given the diverse phylogenetic backgrounds, another hope is that the metagenomics data can reveal the adaptation of the communities to their environments; some of this has already been characterized by looking at individual samples (Tyson *et al.* 2004; Venter *et al.* 2004) and indeed the first comparative study revealed different features of the environments that impose constraints on the genomes, e.g. the dominant energy sources available in different environments, or different concentrations of ions (Tringe *et al.* in press).

4. BASE COMPOSITION AS A PROPERTY THAT DISCRIMINATES METAGENOMES FROM DIFFERENT HABITATS

As we still know very little about metagenomes, there might be many other basic community properties that can differ substantially, imposing further challenges for comparative analyses. For example, in the absence of any phylogenetic information, the base composition of DNA fragments should be an indicator of unexpected distortions or differences. It has long been known that organisms and phyla differ in their overall base composition (for review see Karlin *et al.* 1998; Bentley & Parkhill 2004). This has been studied at several levels of detail—ranging from simple compositional measures such as GC content or dinucleotide frequencies, to codon usage, and higher order measures such as hexanucleotide frequencies (White *et al.* 1993; Elhai 2001).

The distributions of GC content values for all four environmental genomics datasets were expected to cover a wide range of values because they each consist of a complex mixture of many species. However, both the soil DNA and the surface water seem to have relatively narrow ranges of GC content values (Foerstner *et al.* 2005). While it certainly cannot be excluded that this narrow distribution of GC content values is due to sampling or cloning biases, the datasets do contain sequences from hundreds of species from a wide variety of bacterial phyla, and so no major biases are immediately obvious. It is not yet fully understood what drives the evolution of GC content, although a number of correlations with environmental parameters have been reported (and sometimes disputed). These include temperature, oxygen availability and other rather indirect factors such as the average genome size (which correlates weakly with GC content and is

Table 2. Summary of BLAST similarity searches, showing the distribution of best hits across the three domains of life (and viruses/phages). (Only open reading frames of at least 300 bp were considered. Database searched: UniREF (08/2004). ORFs generating no hits or hits below 80 bits were counted under 'no homology'. Assembly depth correction: ORFs from highly covered parts of the assembly were given proportionally more weight, because they represent more abundant species in the environment. The analysis was repeated with other parameters, and for longer, more reliable ORFs (greater than or equal to 450 nt), similar results were obtained. When lowering the threshold for accepting homologies from 80 to 60 bits in the BLAST scoring scheme, *ca* 20% more assignments were possible, but they are likely to include a considerable number of false positives.)

	best hit prokaryotic (%)	best hit archaeal (%)	best hit eukaryotic (%)	best hit phage/virus (%)	no homology (%)
farm soil	48.7	2.3	1.1	0.2	47.7
Sargasso sea	69.5	2.0	2.4	0.3	25.8
whale falls	61.4	1.3	1.2	0.2	35.9
acid mine drainage	26.6	42.5	0.5	0.1	30.3

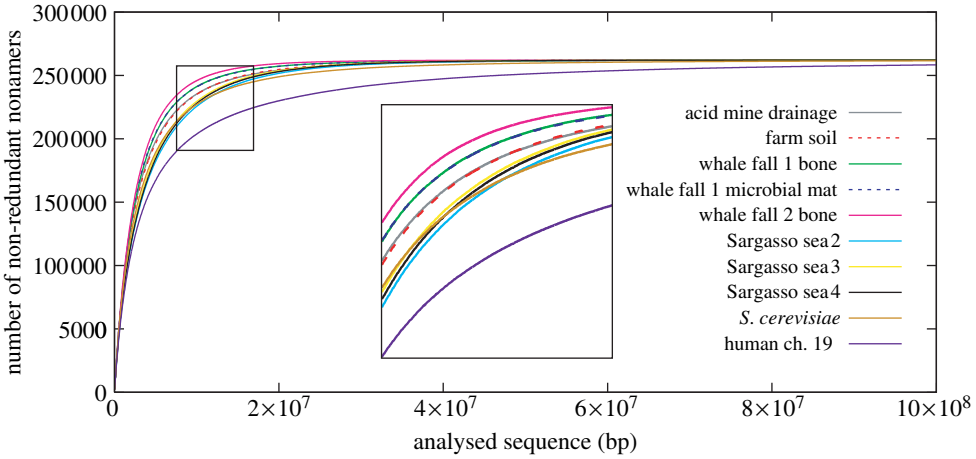


Figure 2. DNA complexity analysis. The curves show the simulated accumulation of nonamer occurrences (each distinct nonamer is counted only once), generated by random sampling of nonamers from the environmental sequences. As controls, the genome of *Saccharomyces cerevisiae*, and the human chromosome 19 were similarly sampled. The maximum number of 262 144 (4^9) distinct nonamers was reached in each environmental sample after analysing a total sequence length in the order of 10^8 bp.

itself probably related to environmental factors; see the following references for discussions on these and other factors: McEwan *et al.* 1998; Hurst & Merchant 2001; Moran 2002; Naya *et al.* 2002; Rocha & Danchin 2002; Bentley & Parkhill 2004; Musto *et al.* 2004). The validity and relative contribution of the above factors remain largely unclear and leave room for other, yet unknown, selective pressures that may force the GC content within a community to be more similar than expected. The GC content does have an impact on codon usage and thus on the proteins encoded in the metagenomes, as exemplified by the differences in amino acid compositions of the predicted proteins. The interplay of these compositional differences and environment-specific functional constraints remains to be elucidated.

While the above theories provide ways to discuss and interpret the observed distinct GC patterns in the samples, for other compositional features we have fewer explanations. For example, a complexity analysis using nucleotide nonamer frequencies (the largest oligomers for which the majority of permutations are still present in large genomes and samples) revealed some unexpected similarities between samples. We simulated the accumulation of distinct nonamers for each of eight environmental (sub)-samples by selecting the sequencing reads in random order, and repeated the procedure with bakers' yeast and human chromosome 19 as controls (figure 2). Sequences with low complexity (i.e. high repeat density) should show a flatter accumulation curve, as is observed for the human chromosome. The data implicitly indicate a slightly higher gene density in environmental samples than in *Saccharomyces cerevisiae* (where it is 72%), confirming the high prokaryotic gene content of the samples. The detailed behaviour of the samples in this simulation cannot be easily explained. Although the subsamples tend to cluster together, whalefall DNA seems to be more complex than soil, although the latter has the highest species diversity. It is tempting to link the nonamer occurrence simply to GC content and claim that the numbers of non-redundant nonamers is limited by unbalanced AT–GC distributions. Yet many other factors might contribute as we are only now starting to understand the metagenomes and the biases of the approach for deciphering them.

5. CONCLUSIONS

It is clear that environmental genomics approaches represent an entirely new quality of sequencing projects in terms of scope and complexity. This comes along with unique features and pitfalls, and poses various new challenges for the analysis and interpretation of the data. Simple technical differences in the sample

preparation and subsequent analysis might have a much larger impact on the resulting data than is the case for current genome projects, where the assembly of only a single entity (a genome) and external information such as physical maps can give some feedback on the original quality. In metagenome assemblies, shared phages or recently horizontally transferred fragments of DNA might cause species to merge artificially. Thus, as with the deposition of raw sequencing traces in genome projects, resources that allow for the deposition of intermediate steps of the data treatment (such as details on quality filtering and assembly, e.g. Salzberg *et al*. 2004) become important. This enables other scientists to follow the treatment of the raw data, as various different questions in the promising avenue of metagenomics probably each require different approaches to the data. The maintenance and extension of such data resources should not be underestimated when applying for funds, as only a comparative analysis of many different habitats under many conditions will provide the context sufficient for understanding each individual sample. All these technical hurdles and problems are clearly outweighed by the enormous potentials of the metagenomics approach. Despite the early struggle to understand and dissect the different layers of complexity, comparative metagenome analysis is well suited to tackle many new, exciting questions, from finding a surprising new gene variant to estimating the total number of genes and species on earth. The practical impacts are equally promising and the application areas summarized in table 1 can be easily extended.

REFERENCES

Beja, O., Suzuki, M. T., Heidelberg, J. F., Nelson, W. C., Preston, C. M., Hamada, T., Eisen, J. A., Fraser, C. M. & DeLong, E. F. 2002 Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* **415**, 630–633. (doi:10.1038/415630a)

Bentley, S. D. & Parkhill, J. 2004 Comparative genomic structure of prokaryotes. *Annu. Rev. Genet.* **38**, 771–792. (doi:10.1146/annurev.genet.38.072902.094318)

Cole, J. R., Chai, B., Farris, R. J., Wang, Q., Kulam, S. A., McGarrell, D. M., Garrity, G. M. & Tiedje, J. M. 2005 The ribosomal database project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res.* **33**, D294–D296. (doi:10.1093/nar/gki038)

Elhai, J. 2001 Determination of bias in the relative abundance of oligonucleotides in DNA sequences. *J. Comput. Biol.* **8**, 151–175. (doi:10.1089/106652701300312922)

Foerstner, K. U., Von Mering, C., Hooper, S. D. & Bork, P. 2005 Environments shape the nucleotide composition of genomes. *EMBO Rep.* **6**, 1208–1213.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. 1998 Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, R245–R249. (doi:10.1016/S1074-5521(98)90108-9)

Hurst, L. D. & Merchant, A. R. 2001 High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. *Proc. R. Soc. B* **268**, 493–497. (doi:10.1098/rspb.2000.1397)

Karlin, S., Campbell, A. M. & Mrazek, J. 1998 Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.* **32**, 185–225. (doi:10.1146/annurev.genet.32.1.185)

McEwan, C. E., Gatherer, D. & McEwan, N. R. 1998 Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas* **128**, 173–178. (doi:10.1111/j.1601-5223.1998.00173.x)

Moran, N. A. 2002 Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **108**, 583–586. (doi:10.1016/S0092-8674(02)00665-7)

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F. & Bernardi, G. 2004 Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* **573**, 73–77. (doi:10.1016/j.febslet.2004.07.056)

Naya, H., Romero, H., Zavala, A., Alvarez, B. & Musto, H. 2002 Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.* **55**, 260–264. (doi:10.1007/s00239-002-2323-3)

Rappe, M. S. & Giovannoni, S. J. 2003 The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369–394. (doi:10.1146/annurev.micro.57.030502.090759)

Rocha, E. P. & Danchin, A. 2002 Base composition bias might result from competition for metabolic resources. *Trends Genet.* **18**, 291–294. (doi:10.1016/S0168-9525(02)02690-2)

Salzberg, S. L., Church, D., DiCuccio, M., Yaschenko, E. & Ostell, J. 2004 The genome assembly archive: a new public resource. *PLoS Biol.* **2**, E285. (doi:10.1371/journal.pbio.0020285)

Schmeisser, C. *et al*. 2003 Metagenome survey of biofilms in drinking-water networks. *Appl. Environ. Microbiol.* **69**, 7298–7309. (doi:10.1128/AEM.69.12.7298-7309.2003)

Staley, J. T. & Konopka, A. 1985 Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39**, 321–346. (doi:10.1146/annurev.mi.39.100185.001541)

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. 1996 Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* **178**, 591–599.

Torsvik, V., Ovreas, L. & Thingstad, T. F. 2002 Prokaryotic diversity—magnitude, dynamics, and controlling factors. *Science* **296**, 1064–1066. (doi:10.1126/science.1071698)

Treusch, A. H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G., Schuster, S. C. & Schleper, C. 2004 Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea. *Environ. Microbiol.* **6**, 970–980. (doi:10.1111/j.1462-2920.2004.00663.x)

Tringe, S. G. *et al*. 2005 Comparative metagenomics of microbial communities. *Science* **308**, 554–557.

Tyson, G. W. *et al*. 2004 Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43. (doi:10.1038/nature02340)

Venter, J. C. *et al*. 2004 Environmental genome shotgun sequencing of the Sargasso sea. *Science* **304**, 66–74. (doi:10.1126/science.1093857)

White, O., Dunning, T., Sutton, G., Adams, M., Venter, J. C. & Fields, C. 1993 A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res.* **21**, 3829–3838.

Appendix C

Get the most out of your metagenome: computational analysis of environmental sequence data

ELSEVIER

Get the most out of your metagenome: computational analysis of environmental sequence data

Jeroen Raes, Konrad Ulrich Foerstner and Peer Bork

New advances in sequencing technologies bring random shotgun sequencing of ecosystems within reach of smaller labs, but the complexity of metagenomics data can be overwhelming. Recently, many novel computational tools have been developed to unravel ecosystem properties starting from fragmented sequences. In addition, the so-called 'comparative metagenomics' approaches have allowed the discovery of specific genomic and community adaptations to environmental factors. However, many of the parameters extracted from these data to describe the environment at hand (e.g. genomic features, functional complement, phylogenetic composition) are interdependent and influenced by technical aspects of sample preparation and data treatment, leading to various pitfalls during analysis. To avoid this and complement existing initiatives in data standards, we propose a minimal standard for metagenomics data analysis ('MINIMESS') to be able to take full advantage of the power of comparative metagenomics in understanding microbial life on earth.

Addresses
European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Corresponding author: Bork, Peer (bork@embl.de)

Current Opinion in Microbiology 2007, **10**:490–498

This review comes from a themed issue on
Genomics
Edited by Claire M. Fraser-Liggett and Jean Weissenbach

Available online 23rd October 2007

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2007.09.001

Introduction

Since the first publications of large-scale environmental shotgun sequencing projects [1–3], we witness exponentially increasing efforts to investigate the genetic basis of environmental diversity using this technique [4]. The combined 'metagenome' of a community complements traditional (16S) phylotyping approaches and genome sequencing of culturable ecosystem members [5]. Given recent advances in sequencing technologies, this approach promises to uncover the identity as well as functionality of the 'unculturable majority' of microbial species on earth, and might lay a firm basis for our understanding of ecosystem functioning. Today, about four times as many genes have been generated in five years of metagenomic sequencing than in over a decade of complete genome sequencing [4], and with dramatically dropping sequencing costs, metagenome projects will be initiated almost everywhere on earth. However, because of the great complexity of generated data, their analysis, an essential step in each project, is far from easy and requires accessible and user-friendly tools that are mostly not available yet. Consequently, we witness an emerging field within computational biology that aims not only at the development of those tools but also at an understanding of the ecosystem imprinted in the sequence data. Recently, several computational methods have been developed and applied to analyze the functional and phylogenetic composition of individual samples (environments) and to derive various properties of the inhabiting microbial communities. In addition, the comparison of results between different (sub)environments ('comparative metagenomics' [6••]) allows to draw more general conclusions about the relationship between metagenome properties and the habitat they were derived from. Here, following the workflow of a typical metagenomics data analysis, we review the state-of-the-art in methodologies that address each step and describe the discoveries they have led to. In addition, we point to possible interdependencies of different metagenome properties and various pitfalls in the analysis of metagenomics data. Finally, we suggest a minimal set of computational analyses to be performed to be able to properly describe and compare an environment.

A typical metagenomics data analysis workflow

Starting from raw reads, assembly is usually the first step to increase fragment length and gain insights in population structure. After that, gene calling is performed, as most (but not all) functional analyses are done at the protein level. Next, higher level metagenome descriptors are derived: basic properties (e.g. sequence composition), species composition, functional composition and population properties. These descriptors provide first insights in the communities but become even more powerful when compared with environments using comparative metagenomics approaches.

From reads to proteins: assembly and gene calling

In bacterial genome sequencing projects, the protocols to go from raw reads to complete and high-quality proteomes have become well established [7]. Metagenomes, however, are sending bioinformaticians back to the drawing board. Already the assembly process can pose a great challenge. One of the main reasons for this is the

phylogenetic complexity of samples: while it is usually possible to assemble most of the genomes from environments containing a small number of (dominant) species (e.g. [2,8•]), samples with large species richness, such as soil [6••], can hardly be assembled given a sequencing depth of up to 100 Mb per (sub)sample. A recent study based on simulated metagenomes confirmed this trend [9•]. On top of this complexity come the added complications regarding high frequency of polymorphisms and genome variations that have been reported even up to the subspecies level [10–12]. Also, the presence of viruses and/or inserted phages might hamper the assembly by increasing the chance of chimeric contigs [13]. Novel short-read sequencing technologies are imposing further complications. Although strategies are being developed to assemble these novel datatypes [14,15], so far, no specific metagenome assembly software has been published or can be simply downloaded. Two strategies to alleviate the assembly problem were identified, namely (i) the use of reference sequences [11••] and (ii) the pre-binning of reads into phylogenetic groups based on sequence composition [2,8•,16••]. Although both can provide improvements, for the former approach, the number of reference genomes still is insufficient for complex metagenome assembly, while for the latter, the binning process only seems to be satisfyingly work in very simple communities (see below). Another alternative are 'greedy' assembly approaches in which multiple metagenomic samples are combined into one superassembly [11••,17••], though with the associated risk of cross-assembling different strains and species. In any case, while currently assemblers such as phrap [18], Arachne [19], JAZZ, Forge and Celera Assembler [20] are being adapted and used to assemble metagenomes, this research area needs further active developments to increase assembly quality for complex metagenomes, sometimes generated using a combination of different sequence technologies.

Owing to the generally limited assembly of these data, also gene prediction methods have to be adapted to deal with (i) massive amounts of fragmented genes on short sequences, (ii) the phylogenetic diversity in samples that hampers the usage of species-specific training sets and (iii) lower end quality of sequences leading to within-frame stop codons and frameshifts. Recent developments use heuristically estimated codon models based on the GC content of the small fragments to overcome this [21]. Alternatively, extrinsic strategies to find coding regions based on (a) their similarity to other regions in a reference database (e.g. known (meta)genomes and/or the sequence set under investigation) and (b) their synonymous versus non-synonymous substitution rate (indicating evolutionary constraints) can be applied [22,23]. To avoid gene prediction problems altogether, some studies (e.g. [24•,25]) base their whole downstream analysis on blastx annotation of reads, limiting themselves to the 'known fraction' of their dataset. The development of metagenome gene prediction software is still in its infancy, and a rigorous evaluation of new and existing methods is needed. Recently, two 'classical' gene predictors (Fgenesb and Critica/Glimmer) were evaluated for this purpose, but unfortunately not compared with the abovementioned heuristic/extrinsic approaches [9•]. Fgenesb performed markedly better (especially on non-assembled sequences), but about 20% of genes were missed and another 10% wrongly predicted, leaving ample room for improvement, judged by the current performance of these methods on full genomes. The considerable differences between these two tools are somewhat worrying given the wide range of methodologies that were used to analyze the metagenomes published so far, suggesting artificial differences in the resulting gene sets which hamper comparative analysis ([4]; see below).

Towards community understanding: delineation of metagenome descriptors

To go from 'a bag of genes' towards a proper understanding of an ecosystem, its inhabitants and its functioning, a range of techniques are being developed to derive parameters that are helpful in this process. These can be subdivided into the following categories: (i) basic descriptors; (ii) phylogenetic composition; (iii) functional composition and (iv) population properties, though they are interdependent (see below). When combined, they can give a first glimpse into biodiversity and ecosystem functioning. Here, we will describe some of the first published methods towards this goal.

While more technical descriptors such as average read length, contig size or assembly rate (as, for example reported by references [3,16••]) make the nature of the dataset more transparent, intrinsic, **basic metagenome descriptors**, such as sequence composition reflect already some environmental constraints. The descriptors range from the GC content [26] over codon usage [21] to oligonucleotide composition [27,28]. These measurements are currently being used for phylogenetic read classification and/or gene prediction (for the latter see reference [21]). Another basic parameter that can be derived from metagenomic shotgun reads is the effective genome size (EGS; a measure of average genome size that takes associated plasmids, inserted elements and phages into account [29]). It can be used to normalize for genome-size effects in comparative metagenomics analyses (see below), and, when combined with assembly information, allows estimation of species richness [3,17••]. It further can provide guidance on coverage issues, for example how much more sequencing is needed to capture most of unique sequences in a sample [3] or to complete the most dominant genome [6••].

To understand the contribution of the different inhabitants to the community, the deduction of the **species**

composition from metagenome data is of crucial importance, but far from trivial. Two distinct concepts with different aims should be discerned: (i) the classification of each read/contig to species (or at least some level of phylogenetic grouping), to possibly link functions of genes encoded by the reads to the community members exerting them and (ii) the quantitative determination of general species composition of the environment at hand. Several approaches for these two concepts have been recently proposed.

Assigning contigs and genes to taxonomic groups is, for complex samples, currently mostly done by 'best-BLAST-hit' mapping (e.g. [17••,24•,30•,31,32•,33]) because of the low computational efforts and the usage of the full spectrum of known genes as reference. On the down side, it is generally not very accurate, cannot deal with horizontal gene transfer (HGT) and does not allow mapping reads to internal nodes of the tree of life (leading to misleading mappings if the best-hit is from a phylogenetic group that is underrepresented in sequence databases) [9•,34•]. Though methods are being developed to deal with these problems, they still only allowed the correct assignment of 25% of reads of a missing species in simulations [35].

Alternatively, sequence composition based 'binning' approaches might be less influenced by biases in sequence databases. Various techniques based on oligonucleotide (2–8 mer) frequency signatures are being applied and developed ([2,9•,16••,27,28,36]; see McHardy and Rigoutsos in this issue). In a recent analysis on simulated metagenomes, the phylopythia binning tool outperforms a 'best-BLAST hit' and a basic oligonucleotide frequency method [9•] but was unfortunately not compared with any of the other recently developed methodologies. In addition, only results on larger contigs (>8 kb or >10 reads) were presented, while the big challenge in this field lies in assigning <1 kb-sized reads that dominate unassembled, complex samples. Given that currently only 60% of larger contigs can be correctly assigned using the best approach [9•], this problem is still far from being solved.

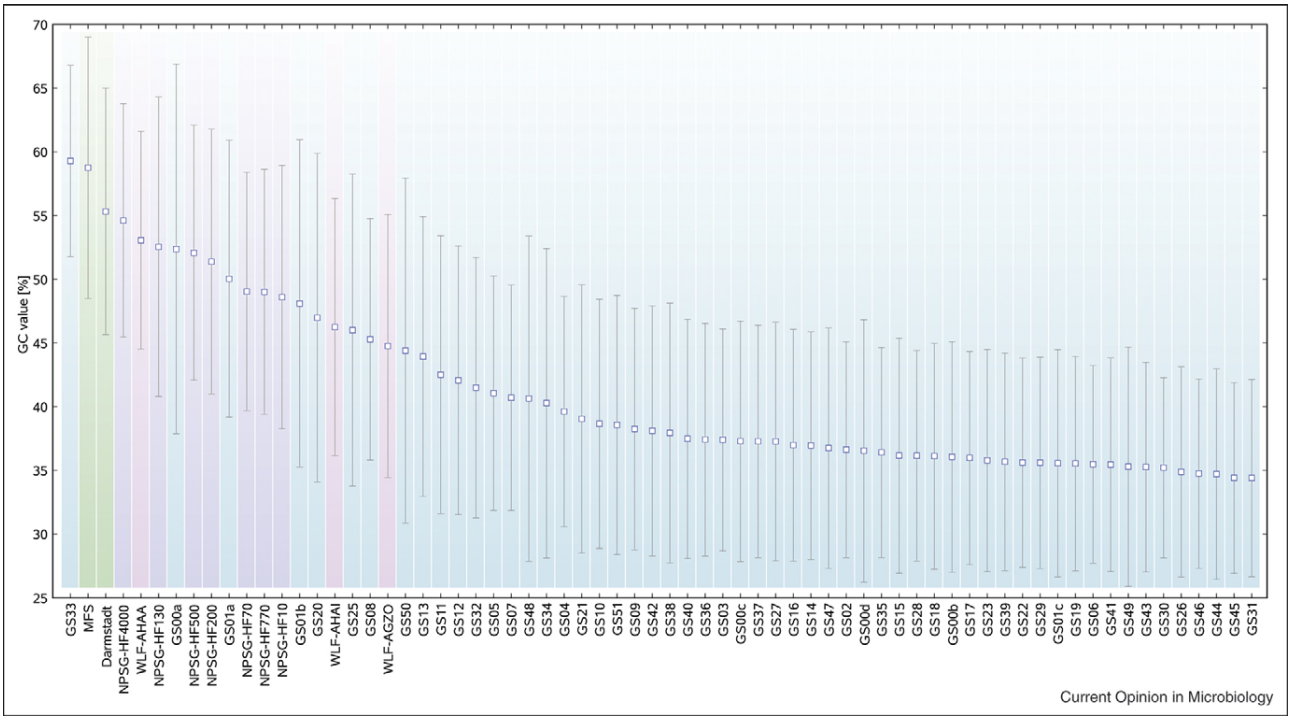
To estimate the species composition quantitatively ('who is there and how many of them are present?'), approaches based on single-copy or equal-copy marker genes, whose counts linearly scale with the number of individuals present (and not e.g. with genome size), are used. After early applications of this principle (using one marker gene at the time; e.g. [3]), a first large-scale phylogenetic approach was developed to map marker gene containing reads to (internal and external) nodes of a reference tree [34•]. This approach should also be more quantitative than classic 16S rRNA PCR based approaches, as it does not suffer from amplification bias or from quantification problems due to varying 16S copy numbers [34•]. However, as 16S methods have the advantage of being able to map to a much larger sampling of genera, both techniques should be regarded as complementary [34•].

To elucidate the **functional composition** of environments, first the predicted ORFs need functional annotation. Strategies analogous to genome sequencing projects have generally been used so far (see reference [4] for an overview). The most common approach is BLAST-based annotation by comparing ORFs against higher order databases such as NCBI Clusters of Orthologous Groups (COGs), TIGR funccats, STRING extended COGs, SEED, KEGG, and so on ([4] and refs therein). However, these techniques only allowed to functionally annotate ~25–50% of proteins per published metagenome [4], which might be due to the limited blast sensitivity for highly fragmented genes ([9•,37]; see below). In genome annotation, two additional methodologies are being employed to improve on this: profile-based homology searches [38] and gene context approaches [39]. The former are easy to implement by searching for protein modules using domain databases (e.g. [40–42]) while the latter offer more potential but need adaptation to the data. In particular, gene neighborhood analysis, a powerful function prediction concept for prokaryotic genomes that does not require homology for the ORFs to be annotated, can be used for function prediction in shotgun sequence data [4,23,43,44]. Recently, Harrington *et al.* published an improved methodology yielding high sensitivity on short contigs (including unassembled reads), allowing function prediction (using a combination of blast-based and profile-based homology and neighborhood approaches) for ~50–80% of proteins in four metagenomics samples [44]. While homology-based approaches will be useful to trace new functionally distinct (sub)families within known superfamilies, neighborhood-based approaches are particularly useful to discover and annotate completely novel proteins associated to known processes, especially in the light of biomining for novel industrially relevant enzymes and catalysts (see Figure 1).

Recently, the first methods have been developed to derive **population properties** from metagenomics data. For example, Johnson and Slatkin [10] estimated growth rates and mutation rates on the basis of site-frequency mutation spectra derived from the raw reads (while incorporating Phred quality scores) and von Mering *et al.* [34•] measured relative evolutionary rate on the basis of phylogenetic mapping of metagenomic marker genes. Rusch *et al.* [11••] and Gill *et al.* [45••] used an elegant technique dubbed 'fragment recruitment' to investigate the amount of genome rearrangements and gain first insights in population structure by aligning mated sequence reads to reference genomes.

While, in complex samples, all the descriptors discussed above (and probably more) are needed to move towards a better understanding of ecosystem functioning, in less

Figure 1



GC content of the metagenomic soil and sea samples. The highest GC content is observed for both the soil samples while the ocean surface water samples have the lowest. The only exception is from a Global Ocean Sampling (GOS), sample taken in a hypersaline lagoon with an exceptionally high GC content. GOS samples taken from coastal (GS13) or fresh water, from a mangrove (GS32), embayment (GS5) or reef (GS25) also show a higher value—possibly because of mixing with soil. Contaminations (GS00a/Sargasso1 [55]) or a higher fraction of eukaryotes in the sample due to filter differences (GS01a, GS01 [3]) apparently also increase the GC content. Sample abbreviations: GSX, Global Ocean Sampling [11••]; MFS, Minnesota farm soil [6••]; Darmstadt, Darmstadt soil [33]; NPSG, North pacific subtropical gyre [24•]; WLF, Whale fall [6••].

complex ecosystems dominated by a few species, a combination of sequence composition binning and assembly seems already sufficient to (almost) completely sequence the community members. This, in turn, allows the assignment of some metabolic activities to individual ecosystem members. This allowed the reconstruction of metagenomic metabolic pathways and indicates cooperation between species/phylogenetic groups within one environment (e.g. [1,2,8•]) or assign specific roles for distinct species/phylogenetic groups in relation with their host (e.g. [16••]). As it has only been possible to start exploring 'who does what' in these simple ecosystems, a great challenge lies in the improvement of tools to address these issues in increasingly complex environments to understand the interrelationships of organisms living in soil, the ocean or in the human body.

Metagenomes in context: comparative analysis

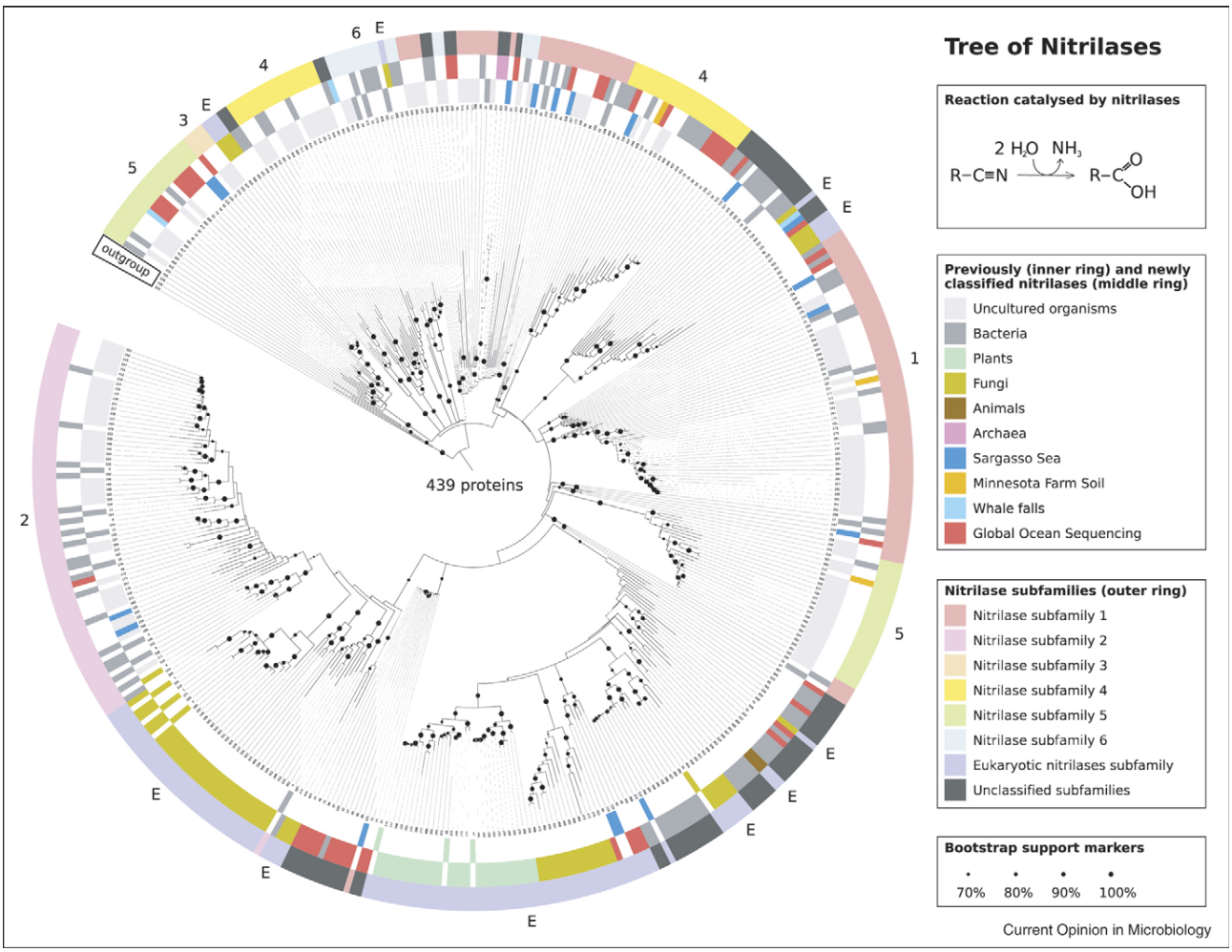
Although the analysis of individual metagenomes has greatly increased our understanding about microbial communities, genome sequence analysis has shown that great additive power comes from comparative approaches [46] as they provide context to the individual samples.

Comparison of different samples from *the same or similar* environments can reveal the influence of particular environmental factors on microbial communities. For example, in a gradual sampling of sea water from the surface to 4000 m depth, the increasing pressure and reduction of light was shown to influence the functional repertoire of organisms living at various depths [24•]. Similarly, comparisons of symbiont communities living in distinct murine intestines allowed linking disease (obesity) to the functional repertoire of the gut inhabitants (in this case the authors showed increased energy harvesting capacities [32•]).

Conversely, the comparison of *diverse* habitats' metagenomes allows the discovery of general trends that link metagenome and community properties with phenotypic features of environments. For instance, one of the most **basic properties**, the GC content, was shown to differ significantly between environments, ranging from high values in Minnesota farm soil to very low ones in surface sea water [26]. Although the original study was based on limited data, this observation is confirmed by the recent Global Ocean Survey (GOS) data [11••], where all open sea surface water sequences show low

GC content, while those from samples taken close to the land (e.g. lagoons and beaches) where water can mix with soil, show markedly higher values (see Figure 2). Likewise, a correlation between microbial genome size and environmental complexity was shown [29] along with the differences of evolutionary rates between environments [34•]. All these outcomes derived from the same environmental datasets seem somehow related: the Sargasso Sea surface water samples harbor the fastest-evolving, smallest genomes with the lowest GC content, while the Minnesota farm soil genomes appear to be the largest, have the highest GC content and evolve the slowest. Although individual links between GC content, genome size and replication/evolutionary rate have been

Figure 2



Biomining metagenomics data case study: nitrilases. Natural habitats are likely to harbor many novel enzymes with biotechnological potential. To detect them, functional, PCR-based or hybridization-based screening methods [56] can be complemented by computational mining of metagenomics datasets. To illustrate a typical computational screen, we extend here the study by Podar *et al.* that detected 17 new nitrilases in the Sargasso Sea (SGS) dataset [57] on an updated environmental dataset collection and UniRef (see Supplementary data for details and references [3,58–61] for similar studies). The tree (drawn using iTOL [62]) contains the 27 bait sequences and hits in UniRef (341), GOS (47), SGS (18; one more than reference [57]), MFS (3), WLF (3), and none in Acid Mine Drainage (AMD). All previously described nitrilases of uncultured species [63] were detected. The colored rings label genes described by [57] (inner ring), nitrilases found in this study (middle ring) and the previously proposed classification into subfamilies (outer ring [57]). The results imply a linear scaling of the identified enzymes with dataset size (at least for ocean surface water) as the nitrilases-per-screened-protein ratio is similar in Sargasso Sea and the much larger Global Ocean Sampling data. The many newly added nitrilases seem to challenge the proposed classification of this family into six bacterial and one eukaryotic subfamily indicating a need for systematic updates. Although most of the proposed subfamilies could be extended by new members (including from archaea!), some became uncertain because of low bootstrap values or even fell apart. Furthermore, novel subfamilies with probably distinct substrate specificities became apparent. The eukaryotes now clearly comprise several distinct subfamilies (e.g. at least two distinct fungal groups exist, only one being related to the bacterial subfamily 2), and several novel bacterial subfamilies can be assumed with confidence (black in our ring), some of which are deeply branching indicating a diverse substrate spectrum.

hypothesized before [47], the precise reasons for this observation remain unknown.

Not only is the breadth of the metagenomic **functional complement** linked to environmental factors (as measured from genome size), but also its composition. When one considers a habitat as one big intercommunicating 'soup' of organisms that carry their genes to maintain this interplay, the combined metagenome should reveal properties of the community and the environment as a whole. Indeed, this gene-centric approach has shown that the more similar the inferred functional composition of metagenomes is, the more related are the respective environmental phenotypes [6••,11••,32•]. Beyond general trends, comparative approaches can also pinpoint particular proteins, protein families and cellular processes that are likely to be responsible for the specific adaptations to particular environments, as could be shown in the first comparative study that used normalized overrepresentions and underrepresentions of such functional units [6••].

Comparative metagenomics can be also used to learn about differences in the **phylogenetic composition** of environments. It could be shown, for instance that the detectable taxonomic groups of microbes have distinct habitat preferences up to the subphylum level, which are remarkably stable in time [34•]. Likewise, a metagenomics study revealed a clear non-random distribution of phages in four ocean sampling sites, with a linear correlation between genetic and geographic distance [17••]. Finally, an analysis of the GOS data for aerobic anoxygenic photosynthesisers (AAnPs) showed great variation in diversity, abundance and composition of AAnP assemblages in different oceanic regions [48].

Interdependencies and pitfalls in comparative metagenomics

Despite the great potential of comparative metagenomics approaches, one should apply them cautiously. Various environment-specific biological factors (see above) and many (usually sample-specific) technical issues hamper the direct comparison of environments, as they influence each other and most results derived from these data.

As for the **basic descriptors**, differences in average genome size of samples (e.g. measurable by EGS [29]) will implicitly lead to differences in the relative functional composition of samples. For example, the sample with the smallest EGS should always have a significant overrepresentation of housekeeping genes as they are a constant fraction while other functional categories grow with genome size; no further biological conclusions should be drawn for differences here without proper normalization. Also, the observed differences in GC content [26] have an impact on homology searches, phylogenetic analyses and binning. However, the extent of such effects still needs to be determined.

The **phylogenetic complexity** of a sample strongly influences the (feasibility of) downstream analyses. For example, the lower the species complexity, the higher the coverage of each individual leading to better assembly and consequently better gene prediction. Longer contigs also improve the efficiency of neighborhood techniques for function prediction and increase the chance of correct phylogenetic assignment of fragments. Therefore, in these samples, the 'who does what' question will be easier to address.

Different **functional constraints** in environments can result in different evolutionary rates [34•] and thus can lead to skews in the gene and function detection rates (e.g. faster evolving genes are more difficult to capture by Blast and orthology assignment methods).

Limited sample coverage and phylogenetic diversity might hamper the direct comparison of **population genetic parameters** as robust estimates based on few data points are difficult, and abundant species might hide the real population structure in samples.

Besides the various biological factors, many **technical issues** related to sampling, sequencing and annotation influence downstream analyses. For instance, the frequent usage of filters or other selection methods for sampling directly influences the phylogenetic and functional composition of the sample. For example, Johnston *et al.* [49] described a surprising paucity of particular nitrogen-fixing genes in the first Sargasso Sea dataset [3], which was later criticized because of its failure to take into account that the main contributors to these genes (cyanobacteria) were probably not in the dataset because of the filtering [50•]. Likewise, in their comparison of the phylogenetic composition within several metagenomics datasets, von Mering *et al.* noticed a conspicuous lack of endospore-forming organisms, which could be linked with their ability to withstand DNA extraction protocols [34•].

Another effect comes from the sequencing technology and protocols. It is first reflected in the read length that depends on the technology (capillary (Sanger) sequencing versus 454 pyrosequencing) but also on parameter choices and other protocols (e.g. Sanger sequencing reads from the whale fall and soil datasets average at ~700 bp after quality clipping, while the sargasso sea ones are ~850 bp [29]). Together with coverage (partly also depending on the sequencing technology) this directly influences the amount of assembly. The resulting differences in contig length influence the success and quality of gene predictions, and the subsequent assignment of gene functions. Short reads as produced by the first generation

of 454 GS20 sequencers are especially little informative as they often are insignificant in the BLAST statistics (e.g. in the mouse gut dataset ~95% of 454 reads were unassignable to known genes/COGs/KEGGs, while for Sanger reads this was ~20–30% [32•]).

In addition to the factors described above, the selected assembly, gene calling and annotation protocols themselves are yet another factor that complicates a direct comparison of samples, for example regarding functional composition. So far, a plethora of methodologies was used in the different projects [4], necessitating a uniform, standardized way of treating metagenomic data in order to be able to compare results from different projects (see Box 1). Only then and in conjunction with good coverage estimates, presence and absence as well as overrepresentation and underrepresentation of genes can be interpreted more confidently. (Given the estimated diversity, was coverage high enough to expect the absence of a gene by chance?) To measure functional and/or phylogenetic coverage, several techniques have been used, ranging from the analysis of single-copy, non-linked genes (mostly used when some full genomes can be almost assembled, e.g. sludge [8•]), via theoretical calculations based on the Lander–Waterman equation (e.g. [3,8•]) to rarefaction approaches (e.g. [6••]).

Taken together, despite recent progress in method development to derive individual parameters for metagenomics samples, considerable effort has to go into the analysis of their interdependencies and the normalization of data from different production lines. Standards for some of the steps would be very helpful to make data comparable and thus add enormous value to them for little cost.

Box 1 Suggestion for a minimal metagenome sequence analysis standard (MINIMESS) to derive indicators for a dataset including annotation protocol, coverage estimates and community descriptors

The previously proposed Minimum Information about a Metagenomic Sequence (MIMS) standard covers detailed information about primary information such as sampling location and procedure, DNA isolation and sequencing and is an indispensable tool to interpret metagenomic datasets [51•,52]. While this is essential towards comparative metagenomics, the many interdependencies and pitfalls (see text) call for an additional, complimentary layer of reporting that provides a standardized description of the metagenome and its inferred community properties. A first prerequisite is the transparent and complete description of data treatment (e.g. assembly, gene calling and functional annotation protocol including the parameter settings). In addition, we propose the reporting of a basic set of metagenome descriptors, resulting from a standardized list of analyses to be performed on each published dataset. This set of descriptors provides an indispensable tool for the proper interpretation and post-analysis of the data and the comparison of metagenomes from independent samples.

(1) Basic sequence analysis: reporting of detailed assembly statistics (including contig composition), gene density, average gene length and fraction of predicted proteins with functional assignment.

(2) Species composition: quantitative description of species composition (marker gene approach, ideally complemented by 16S PCR based method) and species richness estimate.

(3) Functional composition: higher level functional content distribution (e.g. COG/KEGG/SEED, see main text).

(4) Species and gene coverage estimates.

(5) Linking of species and function: although phylomapping tools are just emerging (see dedicated section in this review), it would be favorable to provide a list of gene-species linkages, coming from phylogenetic assignment of reads/contigs based on homology, sequence composition and marker gene presence.

(6) Putative interfering *biological* factors: reporting of, for example GC content and average genome size (EGS).

(7) Putative interfering *technical* factors: reporting of read length and contig length distributions in relation to community complexity (see also coverage estimates).

Minimal standards for annotation and analysis

The more sampling conditions for metagenomics datasets are reported, the more detailed can be the inferences of environmental constraints. Not only exact sample site location and sampling methodology should be mentioned but also broad measurements on the (also non-obvious) physical and chemical properties of the environment, as well as detailed descriptions of the habitat should be made. However, often this is a considerable effort beyond the scope of the individual project and sometimes it is not even known what would be needed to record. Hence, the development of a 'Minimum Information about a Metagenomic Sequence (MIMS)' standard has been proposed in the community to enforce some essential measurements when submitting data to public databases [51•,52]. Beyond the proposed sample information, we believe that a complementing set of standard analyses is also necessary for proper interpretation of metagenomic data (MINIMESS, see Box 1) and should enable normalization of the heterogeneous data for various comparative analyses (see previous paragraph). A first step would be a transparent analysis protocol with all the parameters of the various methods properly reported as they can have a considerable impact on the results.

Over time, metagenomics data generation and analyses protocols will diversify further and there is a need for an accepted infrastructure [53,54] that can cope not only with the heterogeneous data but also with agreements on how to annotate and compare them. It is early days in environmental genomics and important findings will require robust and accurate tools and approaches—what we have reviewed here is just the beginning of a new exciting field emerging in computational biology.

Acknowledgements

The authors would like to thank Christian von Mering and members of the Bork group for stimulating discussions. This work was supported by the EU 6th Framework Programme (Contract No: LSHG-CT-2004-503567).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.mib.2007.09.001.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Hallam SJ, Putnam N, Preston CM, Detter JC, Rokhsar D, Richardson PM, DeLong EF: **Reverse methanogenesis: testing the hypothesis with environmental genomics**. *Science* 2004, **305**:1457-1462.

2. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.

3. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W *et al.*: **Environmental genome shotgun sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.

4. Raes J, Harrington ED, Singh AH, Bork P: **Protein function space: viewing the limits or limited by our view?** *Curr Opin Struct Biol* 2007, **17**:362-369.

5. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products**. *Chem Biol* 1998, **5**:R245-R249.

6. •• Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC *et al.*: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**:554-557.
The first comparative metagenomics study introducing the notion of a gene-centric view of environments.

7. Stothard P, Wishart DS: **Automated bacterial genome analysis and annotation**. *Curr Opin Microbiol* 2006, **9**:505-510.

8. • Garcia Martin H, Ivanova N, Kunin V, Warnecke F, Barry KW, McHardy AC, Yeates C, He S, Salamov AA, Szeto E *et al.*: **Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities**. *Nat Biotechnol* 2006, **24**:1263-1269.
An elegant linkage of functional and phylogenetic information in a simple system to reconstruct sludge metabolic pathways.

9. • Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M *et al.*: **Use of simulated data sets to evaluate the fidelity of metagenomic processing methods**. *Nat Methods* 2007, **4**:495-500.
An important pioneering effort to compare metagenomic data analysis tools, which should be expanded to include the recent wave developments.

10. Johnson PL, Slatkin M: **Inference of population genetic parameters in metagenomics: a clean look at messy data**. *Genome Res* 2006, **16**:1320-1327.

11. •• Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K *et al.*: **The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific**. *PLoS Biol* 2007, **5**:e77.
An impressive amount of data that will keep researchers busy for several years to come

12. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.

13. Salzberg SL, Yorke JA: **Beware of mis-assembled genomes**. *Bioinformatics* 2005, **21**:4320-4321.

14. Sundquist A, Ronaghi M, Tang H, Pevzner P, Batzoglou S: **Whole-genome sequencing and assembly with high-throughput, short-read technologies**. *PLoS ONE* 2007, **2**:e484.

15. Warren RL, Sutton GG, Jones SJ, Holt RA: **Assembling millions of short DNA sequences using SSAKE**. *Bioinformatics* 2007, **23**:500-501.

16. •• Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, Boffelli D, Anderson IJ, Barry KW, Shapiro HJ *et al.*: **Symbiosis insights through metagenomic analysis of a microbial consortium**. *Nature* 2006, **443**:950-955.
See annotation to reference [45••].

17. •• Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H *et al.*: **The marine viromes of four oceanic regions**. *PLoS Biol* 2006, **4**:e368.
See annotation to reference [30•].

18. Green P: Phrap (www.phrap.org).

19. Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC, Lander ES: **Whole-genome sequence assembly for mammalian genomes: Arachne 2**. *Genome Res* 2003, **13**:91-96.

20. Celera assembler (http://sourceforge.net/projects/wgs-assembler/).

21. Noguchi H, Park J, Takagi T: **MetaGene: prokaryotic gene finding from environmental genome shotgun sequences**. *Nucleic Acids Res* 2006, **34**:5623-5630.

22. Krause L, Diaz NN, Bartels D, Edwards RA, Puhler A, Rohwer F, Meyer F, Stoye J: **Finding novel genes in bacterial communities isolated from the environment**. *Bioinformatics* 2006, **22**:e281-e289.

23. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W *et al.*: **The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**:e16.

24. • DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR *et al.*: **Community genomics among stratified microbial assemblages in the ocean's interior**. *Science* 2006, **311**:496-503.
Well-designed comparative metagenomics study along an oceanic depth gradient

25. Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC Jr, Rohwer F: **Using pyrosequencing to shed light on deep mine microbial ecology**. *BMC Genomics* 2006, **7**:57.

26. Foerstner KU, von Mering C, Hooper SD, Bork P: **Environments shape the nucleotide composition of genomes**. *EMBO Rep* 2005, **6**:1208-1213.

27. McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I: **Accurate phylogenetic classification of variable-length DNA fragments**. *Nat Methods* 2007, **4**:63-72.

28. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO: **TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences**. *BMC Bioinformatics* 2004, **5**:163.

29. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: **Prediction of effective genome size in metagenomic samples**. *Genome Biol* 2007, **8**:R10.

30. • Culley AI, Lang AS, Suttle CA: **Metagenomic analysis of coastal RNA virus communities**. *Science* 2006, **312**:1795-1798.
Two studies that chart the vastly understudied world of virus diversity.

31. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A *et al.*: **Metagenomics**

to paleogenomics: large-scale sequencing of mammoth DNA. *Science* 2006, **311**:392-394.

32. • Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest**. *Nature* 2006, **444**:1027-1031.
See annotation to reference [45••].

33. Treusch AH, Kletzin A, Raddatz G, Ochsenreiter T, Quaiser A, Meurer G, Schuster SC, Schleper C: **Characterization of large-insert DNA libraries from soil for environmental genomic studies of Archaea**. *Environ Microbiol* 2004, **6**:970-980.

34. • von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N, Bork P: **Quantitative phylogenetic assessment of microbial communities in diverse environments**. *Science* 2007, **315**:1126-1130.
Study providing evidence for evolutionary rate difference between environments and a clear habitat preference of micro-organisms, which seems to be remarkably stable in time.

35. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Res* 2007, **17**:377-386.

36. Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T: **Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples**. *DNA Res* 2005, **12**:281-290.

37. Tress ML, Cozzetto D, Tramontano A, Valencia A: **An analysis of the Sargasso Sea resource and the consequences for database composition**. *BMC Bioinformatics* 2006, **7**:213.

38. Koonin EV, Tatusov RL, Galperin MY: **Beyond complete genomes: from sequence to structure and function**. *Curr Opin Struct Biol* 1998, **8**:355-363.

39. Huynen MA, Snel B, von Mering C, Bork P: **Function prediction and protein networks**. *Curr Opin Cell Biol* 2003, **15**:191-198.

40. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL *et al.*: **The Pfam protein families database**. *Nucleic Acids Res* 2004, **32**:D138-D141.

41. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P: **SMART 5: domains in the context of genomes and networks**. *Nucleic Acids Res* 2006, **34**:D257-D260.

42. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R *et al.*: **New developments in the InterPro database**. *Nucleic Acids Res* 2007, **35**:D224-D228.

43. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **The use of gene clusters to infer functional coupling**. *Proc Natl Acad Sci USA* 1999, **96**:2896-2901.

44. Harrington ED, Singh AH, Doerks T, Letunic I, Von Mering C, Jensen LJ, Raes J, Bork P: **Quantitative assesment of protein function prediction from metagenomics shotgun sequences**. *Proc Natl Acad Sci* 2007, **104**:13913-13918.

45. •• Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI, Relman DA, Fraser-Liggett CM, Nelson KE: **Metagenomic analysis of the human distal gut microbiome**. *Science* 2006, **312**:1355-1359.
Three studies allowing the first large-scale metagenomic insights into the intriguing world of symbiotic microbial communities.

46. Fraser CM, Eisen J, Fleischmann RD, Ketchum KA, Peterson S: **Comparative genomics and understanding of microbial biology**. *Emerg Infect Dis* 2000, **6**:505-512.

47. Bentley SD, Parkhill J: **Comparative genomic structure of prokaryotes**. *Annu Rev Genet* 2004, **38**:771-792.

48. Yutin N, Suzuki MT, Teeling H, Weber M, Venter JC, Rusch DB, Beja O: **Assessing diversity and biogeography of aerobic anoxygenic phototrophic bacteria in surface waters of the Atlantic and Pacific Oceans using the Global Ocean Sampling expedition metagenomes**. *Environ Microbiol* 2007, **9**:1464-1475.

49. Johnston AW, Li Y, Ogilvie L: **Metagenomic marine nitrogen fixation—feast or famine?** *Trends Microbiol* 2005, **13**:416-420.

50. • Remington KA, Heidelberg K, Venter JC: **Taking metagenomic studies in context**. *Trends Microbiol* 2005, **13**:404.
Comment that highlights the importance of taking technical factors (such as filtering) into account

51. • Field D, *et al.*: **Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification.** *Nat Biotechnol* (In community review).
Crucial proposal to provide a minimal amount of information on sampling, and so on. when submitting/publishing (meta)genome sequences.

52. Field D, Morrison N, Selengut J, Sterk P: **Meeting report: eGenomics: cataloguing our complete genome collection II**. *Omics* 2006, **10**:100-104.

53. Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavromatis K, Kunin V, Garcia Martin H *et al.*: **An experimental metagenome data management and analysis system**. *Bioinformatics* 2006, **22**:e359-e367.

54. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: a community resource for metagenomics**. *PLoS Biol* 2007, **5**:e75.

55. Mahenthiralingam E, Baldwin A, Drevinek P, Vanlaere E, Vandamme P, Lipuma JJ, Dowson CG: **Multilocus sequence typing breathes life into a microbial metagenome**. *PLoS ONE* 2006, **1**:e17.

56. Handelsman J: **Metagenomics: application of genomics to uncultured microorganisms**. *Microbiol Mol Biol Rev* 2004, **68**:669-685.

57. Podar M, Eads JR, Richardson TH: **Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study**. *BMC Evol Biol* 2005, **5**:42.

58. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G: **Structural and functional diversity of the microbial kinome**. *PLoS Biol* 2007, **5**:e17.

59. Zhu Y, Pulukkunat DK, Li Y: **Deciphering RNA structural diversity and systematic phylogeny from microbial metagenomes**. *Nucleic Acids Res* 2007, **35**:2283-2294.

60. van Loo B, Kingma J, Arand M, Wubbolts MG, Janssen DB: **Diversity and biocatalytic potential of epoxide hydrolases identified by genome analysis**. *Appl Environ Microbiol* 2006, **72**:2905-2917.

61. Rhee JK, Ahn DG, Kim YG, Oh JW: **New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library**. *Appl Environ Microbiol* 2005, **71**:817-825.

62. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. *Bioinformatics* 2007, **23**:127-128.

63. Robertson DE, Chaplin JA, DeSantis G, Podar M, Madden M, Chi E, Richardson T, Milan A, Miller M, Weiner DP *et al.*: **Exploring nitrilase sequence space for enantioselective catalysis**. *Appl Environ Microbiol* 2004, **70**:2429-2436.

Appendix D

A Molecular Study of Microbe Transfer between Distant Environments

OPEN ACCESS Freely available online

PLoS one

A Molecular Study of Microbe Transfer between Distant Environments

Sean D. Hooper[1], Jeroen Raes[2], Konrad U. Foerstner[2], Eoghan D. Harrington[2], Daniel Dalevi[3], Peer Bork[2]*

1 Department of Energy Joint Genome Institute (DOE-JGI), Walnut Creek, California, United States of America, **2** EMBL, Heidelberg, Germany, **3** Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America

Abstract

Background: Environments and their organic content are generally not static and isolated, but in a constant state of exchange and interaction with each other. Through physical or biological processes, organisms, especially microbes, may be transferred between environments whose characteristics may be quite different. The transferred microbes may not survive in their new environment, but their DNA will be deposited. In this study, we compare two environmental sequencing projects to find molecular evidence of transfer of microbes over vast geographical distances.

Methodology: By studying synonymous nucleotide composition, oligomer frequency and orthology between predicted genes in metagenomics data from two environments, terrestrial and aquatic, and by correlating with phylogenetic mappings, we find that both environments are likely to contain trace amounts of microbes which have been far removed from their original habitat. We also suggest a bias in direction from soil to sea, which is consistent with the cycles of planetary wind and water.

Conclusions: Our findings support the Baas-Becking hypothesis formulated in 1934, which states that due to dispersion and population sizes, microbes are likely to be found in widely disparate environments. Furthermore, the availability of genetic material from distant environments is a possible font of novel gene functions for lateral gene transfer.

Citation: Hooper SD, Raes J, Foerstner KU, Harrington ED, Dalevi D, et al. (2008) A Molecular Study of Microbe Transfer between Distant Environments. PLoS ONE 3(7): e2607. doi:10.1371/journal.pone.0002607

Editor: Dawn Field, NERC Centre for Ecology and Hydrology, United Kingdom

Received March 19, 2008; **Accepted** May 28, 2008; **Published** July 9, 2008

Funding: This work was supported by the EU 6th Framework Programme, Contract Nrs LSHG-CT-2004-503567 (GeneFun) and LSHG-CT-2003-503265 (BioSapiens). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: bork@embl.de

Introduction

The advances of environmental sequencing projects, or *metagenomes*, have brought methods and concepts from molecular biology and comparative genomics to the field of microbial ecology. Many of the same tools that are used in the analysis of isolate genomes can now be applied to whole communities of organisms [1]. In this work, we perform what can be described as comparative metagenomics, where we attempt to identify genetic material that originated from outside the environment, possibly transported by physical processes such as wind or water. For instance, dust clouds may carry microbes over vast distances [2], and carrier organisms such as birds and humans [3] are potential vehicles for transporting microbes. In other cases, drainage from cultured soils may pollute water [4], and it is conceivable that microbes may be transferred in the process.

The motility and sheer numbers of microbes form the basis for the Baas-Becking hypothesis formulated in 1934 [5]. It can be summed up as follows; everything is everywhere and the environment selects. For instance, the hypothesis implies that there is a good chance of finding trace amounts of a wide range of bacterial species wherever we look, but this does not mean that the species will grow or even survive in its new environment. Even if the transported microbe is inert, it would still contribute its genome (and DNA) to the new environment. Thus, the transported DNA may remain packaged within an inert host, within a surviving host, or may be free as the result of a ruptured or digested cell. Free-form DNA has been observed in for instance ocean sediments [6], where it comprises up to 90% of all DNA.

Regardless of the fate of the specific microbe, its DNA can be captured and detected at the time of a metagenomic sampling. Depending on the frequency of the DNA, reads will assemble into contigs or appear as single-reads, and can then be analysed computationally. In this work, we examine two such metagenomes: the Minnesota farm soil [7] data set and the Sargasso sea [8] data set, and attempt to evaluate the interchange, if any, of microbes between them using DNA sequences as proxies. Thus, we will evaluate the Baas-Becking hypothesis by examining the proportion of sequences that *i)* appear very different from other reads in its set and *ii)* appear more similar to reads in the other set. We will then study those sequences to which are potentially results of a microbe transfer across environments.

This comparative process is conceptually very similar to the study of lateral transfer genes (LGT) in isolate genomes. There is an extensive literature describing this approach, from early but seminal studies using atypical nucleotide composition as indicators

of LGT [9–11] to extensive phylogenetic studies covering hundreds of genomes [12]. All approaches require a careful choice of characteristics to use as discriminators of whether a sequence appears to be typical or not for its genome. We will substitute genomes for metagenomes in this study, so special attention must be given to the choice of discriminators.

We chose three distinct characteristics as discriminators; two nucleotide composition measures and one protein orthology measure. The first measure is based on the guanine/cytosine (GC) content of the sequence. GC content has been found to vary not only between species but also between environments [13]. For the farm soil and Sargasso sea data sets (hereafter referred to as *soil* and *sea*), we observe clear differences in the overall GC content. Soil has a high GC content at 61%, compared to only 34% in sea [13]. This difference is even more pronounced when comparing only the synonymous third codon position of genes (hereafter GC3s%) which avoids selection on the protein level. The more pronounced differences in GC3s% than GC suggest a mutational pressure on the choice of base exerted by exogenic factors, as previously described [13].

The second measure is based on oligomer frequency patterns (OFPs; [14–16]). For instance, the OFP of the oligomer TTATA, relative to the occurrences of T and A respectively, differs widely between organisms.One of the first systematic studies reported showed that the composition of dimers is conserved within genomes but different between genomes [17]. Since then many different methods have been developed to capture the genomic signature of bacteria and they have been use widely for either binning of metagenomic data [18] or the identification of lateral gene transfer [19].

Superficially, it could be assumed that GC3s% could be included in this measure, but the level of information is distinctly different in three aspects. Whereas GC3s% directly measures the mutational pressure, the OFP measures the effect of mutational context biases. Since OFPs are also normalized by nucleotide content, this measure is largely independent of GC3s%. Finally, since we study more than one base, OFPs are a more sensitive discriminator.

The third measure is based on protein similarity between translated open reading frames in both data sets. The rationale is that if a gene in e.g. soil has a substantially higher level of orthology to proteins in sea, compared to the rest of the proteins in soil, then it is less likely to be a common fixture of soil. If the two environments never interchange material, then we would expect high levels of orthology only for genes coding for highly conserved and ubiquitous functions, such as cell machinery. However, if a transfer of microbes occasionally occurs between soil and sea, we would expect to find non-ubiquitous yet highly orthologous genes.

For each of these discriminators individually, criticisms can be raised. For instance, bacteria which are parasites within soil eukaryotic cells may essentially live in a mini-environment similar to that of sea microbes, possibly resulting in similarities in GC3s%. Furthermore, organisms that are only distantly related but have similar DNA repair mechanisms could appear similar in OFPs. Orthology may also be spurious due to strict conservation of amino acid sequences of proteins, or by random chance.

Despite individual concerns such as those listed above, it becomes increasingly difficult to regard these open reading frames as false positives when all three discriminators are fulfilled.

In this work, we apply the three discriminators to predicted genes in the soil and sea sets in order to find genes that are consistent with an interchange of microbes between environments. This transfer of microbes did not specifically occur from the Minnesota farm soil to the Sargasso sea or vice versa, but from environments which share features with either the farm soil or Sargasso sea data. As both GC content and protein composition correlate with the similarity of environments [7,13], it is reasonable to assume that our three discriminators also account for transfers from environments that are at least geographically close to the sampling points or are of similar consistency [20,21].

Results

Starting with 184,000 genes in the soil set [7] and 700,000 genes in samples 2–4 from the sea set [8], we identified 1,216 genes that have a closer hit in the foreign environment than their own. These genes, together with their match in the foreign environment, formed pairs which allowed us to compare their features. To classify whether the GC content of these candidate gene pairs is endogenous in one but atypical in the other environment, we used the average of the two environmental GC3s% averages (48%) as a breakpoint (Fig. 1).

Of the 1216 ORF pairs, 284 sea genes had atypical GC3s% values (>48%); a strong over-representation both in absolute terms and in significance ($p<10^{-13}$), when compared to the expected number of 109 (based on the proportion of all sea genes with atypical GC3%s values). Conversely, the over-representation of soil genes with GC3s%<48% is not as strong, yet significant: 221 compared to an expected 174 ($p<10^{-3}$).

Quadrant A (Fig. 1) thus represents gene pairs where the soil genes have typical GC3s% values and the sea genes have GC3s% values much higher than the sea average. Accordingly, quadrant B has lower than average soil GC3s% and typical sea GC3s%. Overall, quadrant A has 170% as many pairs as expected and B 127%. However, since sample sizes are unequal, we subsampled the sea set 10 times into random subsamples of a size roughly equal to soil (Table 1). The degrees of over-representation remained at 165% and 123% respectively. Details of quadrants A–D are provided as supplementary Tables S1–S4.

At this point, we have created three classes of genes using orthology and GC3s% as discriminators. These classes represent genes that may have been transferred into a new environment (quadrants A and B) or simply conserved genes (quadrant C). If all three classes are actually false positives, we would expect OFPs to be distributed according to random expectation, i.e. soil genes would have OFPs similar to the soil set in large, and analogously for sea genes. Out of 16,450 random soil genes, 14,040 map to soil (85.3%). For sea, 14,257 of 16,476 map correctly to sea (86.5%). Thus, we expect that of the 284 soil genes in A, 242 should map to soil. However, we observe that only 53 soil genes map better to soil than sea. This is a strong under-representation ($p<10^{-137}$). In quadrant B, we expect that 191 of the 221 sea genes would map to sea, but observe only 13 ($p<10^{-161}$). The OFPs of quadrants A and B, compared to soil and sea sets are visualized as chaos game representations ([22]; see methods) in Fig. 2.

Finally, as an additional control, we study genes in quadrant C, which we do not believe to be transferred. Here, we find 674 of 678 unique sea genes and 579 of 663 unique soil genes mapped to their own environments. The sea mappings are actually significantly over-represented ($p<10^{-14}$), and soil genes map to soil about as often as expected.

The results strongly imply that genes in quadrants A and B are not only atypical in their current environments, but also highly similar to the external and internal mutational pressures in the other environment.

Amino acid identity is a measure of the similarity between the amino acid translations of genes, and as such focuses on the non-synonymous bases. A further comparison would be to measure the synonymous substitution rate *Ks* [23] of a gene pair, since this rate quickly becomes saturated over time. A *Ks* value of over 2.0 suggests that each base has on average been substituted at least

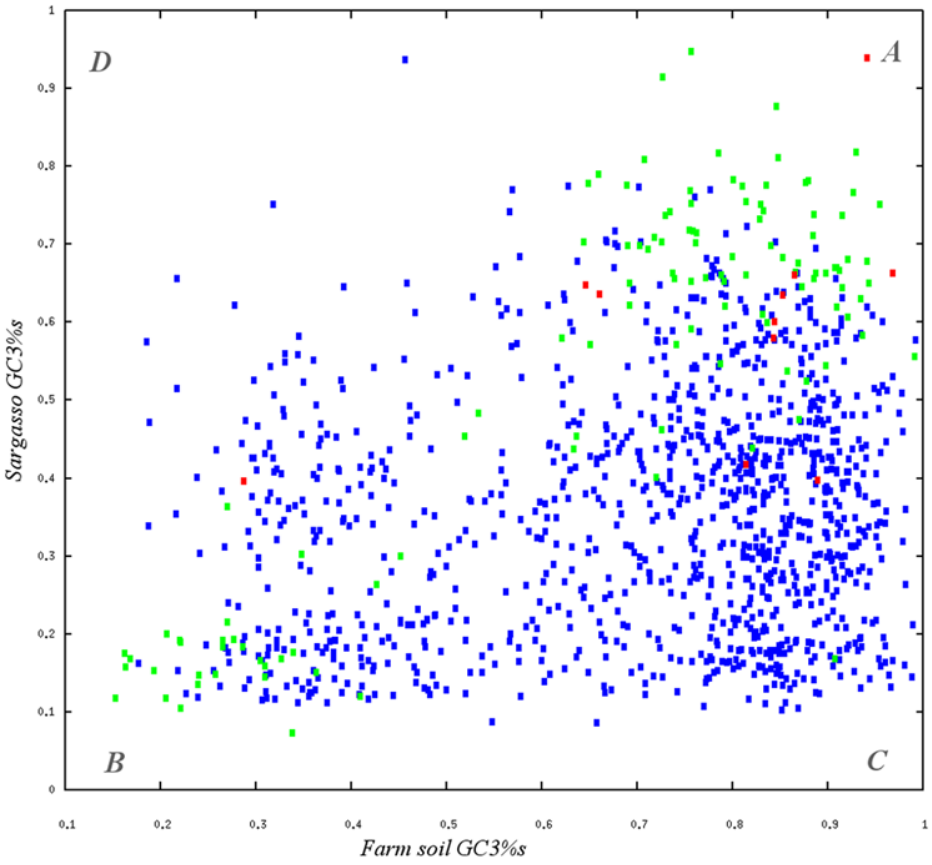


Figure 1. GC3s distribution of orthologous Genes. Distributions of GC3s for each of 1216 ORF pairs with closer similarity in the foreign environment. Using GC3s% = 48% as a separator (dotted lines), the ORF pairs are classified based on the GC content of its two members. Category A (upper right) is the quadrant where we expect to find possible transfer events from soil to sea, since these pairs have high GC3%s values for both members. Pairs in category B (lower left) have low GC3%s scores for both Genes, which could suggest a transfer from a sea-like environment to soil. Category C (lower right) has typical GC3s% values for both members of ORF pairs. These pairs are likely to be ancient conserved sequences. Finally, Category D (upper left) has atypical values for both Genes, close to the expected given the shape of the GC3% distribution (28 observed, 24 expected). Unsaturated Ks values are green, and pairs with Kn/Ks>1 are red.
doi:10.1371/journal.pone.0002607.g001

once; changes are therefore saturated. However, values of less than 2.0 suggest a higher level of similarity than evident from amino acid identities. Finding such unsaturated pairs in the quadrants would further suggest a shared recent history. In quadrant A we find 87 such pairs, and in B 31. Again, this is not what we expect if we consider the gene pairs to simply be ancient homologs. In quadrant C for instance, which we consider to be composed mainly of ancient homologs, we find only 8 unsaturated pairs out of 667. None score lower than 1.2. Thus, unsaturated gene pairs are significantly overrepresented in quadrants A and B (both at $p<10^{-14}$). Furthermore, the ratios (87 to 31) again suggest a bias in directionality from soil to sea. This also holds when the datasets are resampled (Table 1), suggesting that it is not an effect of sample sizes.

If genes in quadrants A and B are the results of microbes, alive or not, traversing large distances between soil and sea, then it would be interesting to know which species they come from. Determining which taxa are included in a metagenome is referred to as *binning*, and is not a straight-forward task. Since the focus of this paper is not on binning Minnesota farm soil and Sargasso, we employ a simple best hit approach and record the species for each gene. The results were then mapped onto the Interactive Tree of Life [24] and are available as supplementary figures (Fig. S1–S3).

Table 1. Resampling of sea set.

S	nA (74)	nB (117)	aA	aB	sA	sB	sA/sB
1	123	139	5	0	34	17	2.00
2	123	133	2	1	26	15	1.73
3	120	143	4	1	32	16	2.00
4	137	140	3	1	33	20	1.65
5	126	147	4	0	27	19	1.42
6	107	151	2	1	23	19	1.21
7	111	147	4	2	35	23	1.52
8	114	140	3	0	39	21	1.86
9	131	142	7	0	29	13	2.23
10	130	162	4	0	38	21	1.81
Full set	284	221	8	1	87	31	2.81

Distributions of genes in quadrants A and B. Key: **S**: sample number, **nA**; number of gene pairs in A, with the average expected number in parenthesis, **nB**; number of gene pairs in B, also with expected in parenthesis, **aA**; number of gene pairs in A with *Kn/Ks*>1, **aB**; number of gene pairs in B with *Kn/Ks*>1, **sA**; number of gene pairs in A with *Ks*<2, **sB**; number of gene pairs in B with *Ks*<2.
doi:10.1371/journal.pone.0002607.t001

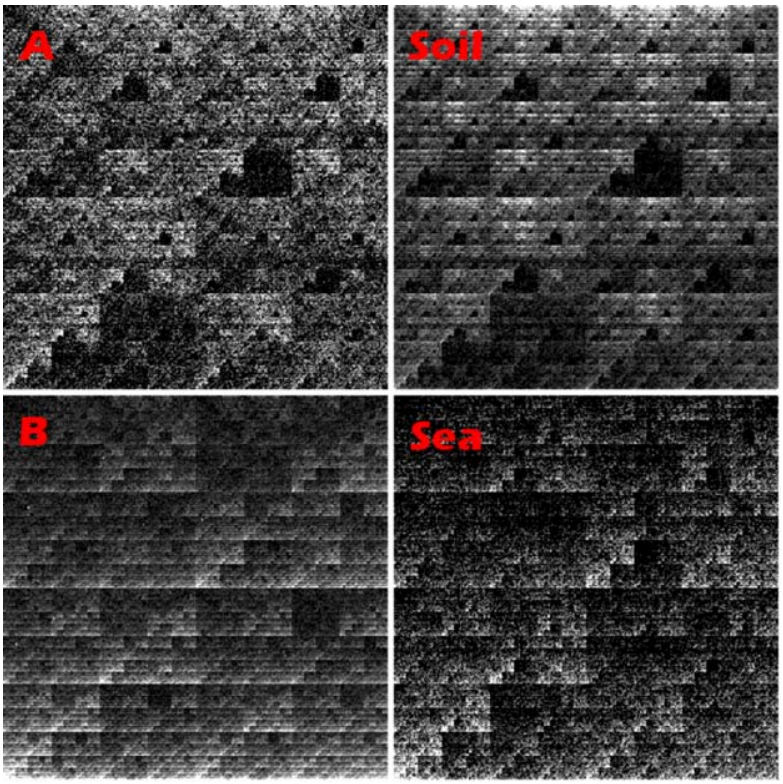


Figure 2. Chaos Game Representation (CGR) plot of oligomer frequencies of A and B vs soil and sea patterns. Note the similarities between A and soil, and B and sea respectively. Figure intensities have been normalized for clarity. CGR plots are a way of visualizing chain processes, such as oligomer patterns. See methods for details.
doi:10.1371/journal.pone.0002607.g002

For quadrant A, which should represent transfer from soil to sea, we find a relatively even contribution from a wide range of phyla and a considerable (~25%) contribution from the Rhizobium/Bradyrhizobium clades. Both of these families are predominantly terrestial bacteria, which is consistent with our findings.

Similarly, in quadrant B, we also observe an even contribution from a wide range, with a stronger representation from the bacteroides genus. Bacteroides are not predominantly soil bacteria, but can be found in the guts of farm animals. It is hence not subject to the mutational pressures of soil but is readily and consistently transferred from animals to soil. It is therefore not inconceivable that the contribution of bacteroides may be via animal waste to soil, and then to sea. This does not weaken our results, but rather strengthens the conclusion that transfer is predominantly from soil to sea rather than vice versa, and underlines the interaction of diverse environments other than the two we have studied.

For comparison, we include best hit binnings for the whole soil and sea sets (Fig. S4–S5). Noteworthy is the huge dominance of *Candidatus pelagibacter* (e.g. [25]) in the full sea set, but which is largely absent in quadrant A. Furthermore, our simplistic binning approach suggests that there is no strong contribution from any potential lab contaminant.

Protein function

The null hypothesis is that the transferred DNA is selected randomly, and therefore codes for random products. Supplementary Table S5 illustrates the distribution of protein functions by Cluster of Orthologous Gene categories [26]. Generally, quadrants A, B and C are consistent with a random selection of functions drawn from the distributions of the whole metagenome sets, but some differences nonetheless stand out. Quadrant B has lower numbers of ORFs coding for energy production and conversion and general function prediction (COG category C and R) than quadrants A and C, but higher numbers of ORFs coding for translation, replication and repair.

Fate of transfers

Based on our studies, quadrants A and B are consistent with a transfer effect. But what of the fate of these transported genes or DNA fragments? Most likely they will simply be degraded, but there is also a possibility of incorporation into indigenous genomes, constituting a true LGT. We can first study the GC content of the flanking DNA or neighboring gene, if any, and see if it is different or similar. Of the 505 candidates, the majority of flanking DNA has similar GC values. This suggests that large regions (likely entire genomes, plasmids or chromosomes) have been transferred but not assimilated. In quadrant A however, 31 of 284 sea genes have one or more neighbors with a GC3s%<48%, which may suggest that some genes are occasionally integrated into indigenous genomes. In this case, the ability to assemble contigs with several genes also suggests that these genes may have been incorporated into abundant species. Furthermore, we studied transfer candidates that seem to be under positive selection, as this would indicate an adaptation to the new environment and therefore LGT. We find 8 ORF pairs that suggest an accelerated evolution ($Kn/Ks>1$): 7 in quadrant A and 1 in B. The annotated functions [26] of these ORFs in the process of adaptation are diverse (Tables S1–S2). However, the function that is under selection is not necessarily the same as the annotated function [27] so we cannot exclude a

common functional theme due to the process of adaptive radiation [28]. Unfortunately, only 2 of the 7 ORFs have neighbors – both with similar GCs% values. This would suggest that these adapting ORFs may have been incorporated along with other genes which are not under selection in the new environment. Moreover, given the high rate of amelioration at the synonymous base, it is likely that many such ORFs would have $Kn/Ks<1$ despite adaptation. These 8 ORFs are therefore a conservative estimate.

Discussion

Microbe transfer

Through several different comparisons, we have found a set of genes, however small, for which the simplest explanation is microbe transfer. Specifically, we seem to detect a transfer of genetic material in the Sargasso samples from an environment very similar to Minnesota soil and vice versa. Given the prodigious population sizes and motility of e.g. bacteria, it should not be a surprising conclusion. However, detecting it is not as intuitive, and we believe we are the first to address this question using computational methods.

Our data is for natural reasons limited; other oceanic and soil samples may contain other transfer candidates, and the total transfer to the Atlantic Ocean from environments similar to Minnesota farm soil must be considerably larger. However, we believe it is an informative snapshot given the current data. With more large-scale sequencing projects, the picture will undoubtedly improve.

We also suggest a bias in transfer from soil to sea, in line with the generally accepted flow of water from land to oceans. Furthermore, the presence of bacteriodiales in quadrant B further tilts the scale in favor of transfer from soil to sea, since they are likely to have originated in a third environment – animal gut. Assessing the total proportion of foreign DNA in a metagenome is a difficult task at best. In this study, we focused on sets with quite different nucleotide compositions, which simplify detection of foreign DNA. Other sources of transfer may be more similar to the receiving environment, and detection is therefore more complicated. In addition, rare foreign DNA may be present in low numbers and is likely to evade detection by normal shotgun sequencing. Thus, in the case of transfers between soil and sea, the amount of transferred DNA seems to be abundant enough to be detectable by shotgun sequencing, even though it is only a fraction of the amount of indigenous DNA.

This study therefore suggests that the species abundance distributions of metagenomes which are not physically isolated may have exceedingly long 'tails' composed of rare organisms. It is therefore unlikely that sequencing projects of this type will reach full coverage in the near future.

Consequences for LGT

While little data is available on genes which have been incorporated into new hosts, our findings suggest that it is possible. Furthermore, it has been found that the extent of LGT in metagenome samples is comparable to that of isolate genomes [29], suggesting that LGT is an active process also within the soil and sea microbiomes. Combined with our findings, we suggest that the impact of LGT could be more far-reaching than previously thought, since functions need not be acquired from the immediate vicinity but from entirely different environments. This would also include non-microbial donors, such as genetically modified plants.

Materials and Methods

Our approach employs three basic discriminators to assess microbe transfer and is based on the study of lateral gene transfer. First, we test genes for their orthology against the other environment. If a gene in either set has a higher (20% better) homology score to an gene in the foreign environment than to its own, we select that gene pair for further investigation. Furthermore, all orthology must fulfil at least 80% protein similarity over at least 90% of the shortest gene. Genes under 100 base pairs in length were ignored. As a second measure, we calculated the GC content at the synonymous base (GC3s%). Using the GC3s% values of each member of a pair, we then classified pairs into three major categories depending on if one or no member had GC3s% values atypical for their environment. GC3s% was calculated using *codonw* (http://codonw.sourceforge.net/). As a third measure, oligomer frequencies were calculated using *softPSTk-Classifier* [15]. To further stress that the genomic signature of oligomers is not simply a result of the difference in GC between the two environments, we decided to also show the visualization using chaos game representations (CGR) plots [22]. The points in these graphs can easily be calculated recursively using the relation,

$$\begin{cases} r_i = \frac{1}{2}(r_{i+1} + u_i) \\ r_0 = (0.5, 0.5) \end{cases}$$

where

$$u_i \begin{cases} (0,0) | \text{if i : th position is A} \\ (1,0) | \text{if i : th position is C} \\ (1,1) | \text{if i : th position is G} \\ (0,1) | \text{if i : th position is T} \end{cases}$$

DNA with different composition will end up with different coordinates in the plot depending on the symbols. All points are bound to the unit square. Plot intensities have been normalized for clarity. Note that no conclusions have been drawn directly on the figure itself, rather from significance tests of the distributions in the quadrants. The figure is included for visualization purposes only.

Environmental data

We used the same data from Sargasso and Minnesota as were used previously by Tringe and co-workers [7]. Note that this data set does not include sample 1 from Sargasso, due to recent criticism [30].

Gene predictions were performed by the original authors, resulting in roughly 700 000 and 184 000 ORFs respectively.

Estimation of synonymous (Ks) and nonsynonymous substitution rates (Kn)

Nucleotide sequences were pairwise aligned by ClustalW [31] using the corresponding protein sequences as an alignment guide. Gaps and adjacent divergent positions in the alignments were removed. K_S estimates were obtained with the Codeml [23] algorithm in the PAML package (F3x4 model, gamma shape parameter and transition-transversion ratio estimated from the data [32]). Calculations were repeated five times to avoid incorrect Ks estimations due to suboptimal local maxima.

Supporting Information

Table S1 ORF pairs belonging to category A. Sea: Sargasso ORF. Soil: Minnesota ORF. ID: protein identity. Pos: protein positive similarity. Length: Overlap length of overlapping sequence. GC_sea: GC3s% of Sargasso ORF. GC_soil: GC3s% of Minnesota ORF.

Diff: GC3s% difference. KaKs: substitution ratio of synonymous to non-synonymous base. Ka: synonymous base substitution rate. Ks: non-synonymous substitution rate. COG: COG assignment. Func: COG functional category. Annotation: predicted function.
Found at: doi:10.1371/journal.pone.0002607.s001 (0.04 MB CSV)

Table S2 ORF pairs belonging to category B. Sea: Sargasso ORF. Soil: Minnesota ORF. ID: protein identity. Pos: protein positive similarity. Length: Overlap length of overlapping sequence. GC_sea: GC3s% of Sargasso ORF. GC_soil: GC3s% of Minnesota ORF. Diff: GC3s% difference. KaKs: substitution ratio of synonymous to non-synonymous base. Ka: synonymous base substitution rate. Ks: non-synonymous substitution rate. COG: COG assignment. Func: COG functional category. Annotation: predicted function.
Found at: doi:10.1371/journal.pone.0002607.s002 (0.03 MB CSV)

Table S3 ORF pairs belonging to category C. Sea: Sargasso ORF. Soil: Minnesota ORF. ID: protein identity. Pos: protein positive similarity. Length: Overlap length of overlapping sequence. GC_sea: GC3s% of Sargasso ORF. GC_soil: GC3s% of Minnesota ORF. Diff: GC3s% difference. KaKs: substitution ratio of synonymous to non-synonymous base. Ka: synonymous base substitution rate. Ks: non-synonymous substitution rate. COG: COG assignment. Func: COG functional category. Annotation: predicted function.
Found at: doi:10.1371/journal.pone.0002607.s003 (0.10 MB CSV)

Table S4 ORF pairs belonging to category D. Sea: Sargasso ORF. Soil: Minnesota ORF. ID: protein identity. Pos: protein positive similarity. Length: Overlap length of overlapping sequence. GC_sea: GC3s% of Sargasso ORF. GC_soil: GC3s% of Minnesota ORF. Diff: GC3s% difference. KaKs: substitution ratio of synonymous to non-synonymous base. Ka: synonymous base substitution rate. Ks: non-synonymous substitution rate. COG: COG assignment. Func: COG functional category. Annotation: predicted function.
Found at: doi:10.1371/journal.pone.0002607.s004 (0.00 MB CSV)

Table S5 A breakdown of COG categories by quadrant. The 'expected' occurence is based on the classification of the full soil and sea sets Cat: COG category. A,B,C: quadrants. Expected: expected number given random occurrence.
Found at: doi:10.1371/journal.pone.0002607.s005 (0.00 MB CSV)

Figure S1 Phylogenic distribution of category A.
Found at: doi:10.1371/journal.pone.0002607.s006 (18.08 MB TIF)

Figure S2 Phylogenic distribution of category B.
Found at: doi:10.1371/journal.pone.0002607.s007 (18.08 MB TIF)

Figure S3 Phylogenic distribution of category C.
Found at: doi:10.1371/journal.pone.0002607.s008 (18.08 MB TIF)

Figure S4 Full phylogenic distribution of the Sargasso set.
Found at: doi:10.1371/journal.pone.0002607.s009 (18.08 MB TIF)

Figure S5 Full phylogenic distribution of the soil set.
Found at: doi:10.1371/journal.pone.0002607.s010 (18.08 MB TIF)

Author Contributions

Conceived and designed the experiments: PB SH. Performed the experiments: SH. Analyzed the data: SH DD. Contributed reagents/materials/analysis tools: EH JR KF DD. Wrote the paper: PB JR SH.

References

1. Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10: 490–498.
2. Shinn EA, Griffin DW, Seba DB (2003) Atmospheric transport of mold spores in clouds of desert dust. Arch Environ Health 58: 498–504.
3. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, et al. (2003) Traces of human migrations in Helicobacter pylori populations. Science 299: 1582–1585.
4. Powell B, Martens M (2005) A review of acid sulfate soil impacts, actions and policies that impact on water quality in Great Barrier Reef catchments, including a case study on remediation at East Trinity. Mar Pollut Bull 51: 149–164.
5. Baas-Becking L (1934) Geobiologie of Inleiding Tot de Milieukunde. The Hague: Van Stockkum & Zoon.
6. Dell'Anno A, Danovaro R (2005) Extracellular DNA plays a key role in deep-sea ecosystem functioning. Science 309: 2179.
7. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
8. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
9. Mrazek J, Karlin S (1999) Detecting alien genes in bacterial genomes. Ann N Y Acad Sci 870: 314–329.
10. Lawrence JG, Ochman H (1997) Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 44: 383–397.
11. Hooper SD, Berg OG (2002) Detection of genes with atypical nucleotide sequence in microbial genomes. J Mol Evol 54: 365–375.
12. Choi IG, Kim SH (2007) Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A 104: 4489–4494.
13. Foerstner KU, von Mering C, Hooper SD, Bork P (2005) Environments shape the nucleotide composition of genomes. EMBO Rep 6: 1208–1213.
14. Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11: 283–290.
15. Dalevi D, Dubhashi D, Hermansson M (2006) Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. Bioinformatics 22: 517–522.
16. Ohno S (1988) Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. Proc Natl Acad Sci U S A 85: 9630–9634.
17. Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci U S A 89: 1358–1362.
18. Teeling H, Waldmann J, Lombardot T, Bauer M, Glockner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5: 163.
19. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P (2005) Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res 33: e6.
20. Green JL, Holmes AJ, Westoby M, Oliver I, Briscoe D, et al. (2004) Spatial scaling of microbial eukaryote diversity. Nature 432: 747–750.
21. Horner-Devine MC, Lage M, Hughes JB, Bohannan BJ (2004) A taxa-area relationship for bacteria. Nature 432: 750–753.
22. Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Res 18: 2163–2170.
23. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 725–736.
24. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127–128.
25. Tripp HJ, Kitner JB, Schwalbach MS, Dacey JW, Wilhelm LJ, et al. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature 452: 741–744.
26. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
27. Hooper SD, Berg OG (2003) On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol 20: 945–954.
28. Bergthorsson U, Andersson DI, Roth JR (2007) Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A 104: 17004–17009.
29. Tamames J, Moya A (2008) Estimating the extent of horizontal gene transfer in metagenomic sequences. BMC Genomics 9: 136.
30. DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3: 459–469.
31. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
32. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555–556.

Appendix E

A computational screen for type I polyketide synthases in metagenomics shotgun data

A computational screen for type I polyketide synthases in metagenomics shotgun data

Authors

Konrad U. Foerstner, Tobias Doerks, Christopher J. Creevey, Anja Doerks, Peer Bork

Affiliations

European Molecular Biology Laboratory, Heidelberg, Germany

Abstract

Background: Polyketides are a diverse group of biotechnologically important secondary metabolites that are produced by multi domain enzymes called polyketide synthases (PKS).
Methodology/Principle findings: We have estimated frequencies of type I PKS (PKS I) – a PKS subgroup – in natural environments by using Hidden-Markov-Models of eight domains to screen predicted proteins from six metagenomic shotgun data sets. As the complex PKS I have similarities to other multi-domain enzymes (like those for the fatty acid biosynthesis) we increased the reliability and resolution of the dataset by maximum-likelihood trees. The combined information of these trees was then used to discriminate true PKS I domains from evolutionary related but functionally different ones. We were able to identify numerous novel PKS I proteins, the highest density of which was found in Minnesota farm soil with 136 proteins out of 183,536 predicted genes. We also applied the protocol to UniRef database to improve the annotation of proteins with so far unknown function and identified some new instances of horizontal gene transfer.
Conclusions/Significance: The screening approach proved powerful in identifying PKS I sequences in large sequence data sets and is applicable to many other protein families.

Introduction

The majority of the microorganisms on earth cannot be cultured under standard laboratory conditions [1]. Therefore, uncultured organisms from environmental samples are promising sources of new enzymes and chemical compounds with biotechnological and pharmaceutical applications. Currently, three screening techniques are commonly applied for exploring protein functions in environmental samples: the function-based, the sequence-based and the substrate-induced gene-expression screening (SIGEX) [2]. Here we present a framework for sequence-based computational screens in environmental shotgun sequences, i.e. metagenomics data. It involves both homology-based and phylogenetic classification. While there has been some success in identifying important subfamilies in metagenomics data [3-6], there are also immense challenges ahead as tools and computational infrastructure often do not scale with the increase in metagenomics data and as many protein families have complicated evolutionary histories.

In order to explore a difficult and also important protein family in the context of diverse metagenomics data sets, we have chosen type I polyketide synthases (PKS I) as target proteins for screening. They synthesize a highly diverse group of secondary metabolites that covers many biological functions and have considerable medical relevance. Polyketides in general can act among other functions as antibiotics, immunosuppressants, pigments but also as toxins or carcinogens [7] via different mechanisms. Antibiotics like Erythromycin, Rifamycin and Oleandomycin are only a few examples with medical relevance. Polyketides are usually large chemical compounds that are synthesized in a series of repetitive steps. Similar to the synthesis of fatty acids short acyl-units are added to the growing molecule and are modified. All of these steps are catalyzed by a combination of domains, namely a acyltransferase domain (AT – transfers the acyl unit to the acyl carrier protein), a ketoacyl synthase domain (KS - performs the decarboxylative condensation), and an acyl carrier protein (PP - contains the phosphopantetheinyl arm) domain. Additionally the ketoreductase (KR), the dehydratase (DH), the enoyl reductase (ER) and the methyltransferase (MT) domain can modify the acyl unit after the condensation. The thioesterase domain (TE) releases the finished polyketide.. PKS members have been found in bacteria, fungi, plants, slime mold [8], Alveolata [9] and animals [10, 11]. Like the fatty acid synthases (FAS), PKS are classified according to the arrangement of their domains: type I with multiple domains per protein and type II in which each single domain represents an independent protein. Bacterial type I PKS are usually modular where each module is responsible for a single fusion step [12] while fungal type I PKS proteins usually occur as “iteratively” acting enzymes in which the domain combinations catalyze several steps. In plants a third class - PKS type III (chalcone synthases) – was discovered and later also described in bacteria [13]. It is common to classify the PKS into these three types although many exceptions of this classification are known [14, 15] as the evolution of PKS is rather complex [10, 12, 16-18].

There have been numerous attempts to identify PKS in environmental samples using non-computational methods (e.g. [19]). Here, we present a computational approach based on Hidden-Markov-Model (HMM) sequence searches (as done in other PKS focused studies

like [20]) followed by the construction of maximum-likelihood trees. This allows us to screen for multi-domain proteins and to estimate the potential of the different environments to serve as a source of PKS I sequences. Although the discrimination of type I PKS from type II PKS and type II FAS is simple, due to the large evolutionary distance [12] and PKS III are also a clearly separable group, a unique PKS I identification remains challenging. Reasons among others are the paralogy of type I PKS with type I fatty acid synthases [12] and with other enzymes and the fast evolution of PKS I. As PKS I proteins can be very large, it is unlikely that complete proteins are found in the highly fragmented shotgun metagenomic sequences. However, their multi-domain, repeated structure provides multiple instances of evidence to find real PKS I orthologs when searching independently with HMM of each of the eight domains introduced above.

Our approach included the creation and use of domain specific HMMs to find members of the type I PKS domain in six published metagenomic data sets - Minnesota farm soil (MSF) [21], Sargasso Sea (SGS) [22], human gut (HGUT) [23], acid mine drainage (AMD) [24], enhanced biological phosphorus removal sludges (EBPRS) [25] and whale falls (bones from sunken whales) (WLF) [21]. We used the UniRef database [26] as an reference set by treating it as another sample to be able to identify biases and the status of PKS I annotation. In contrast to most other studies that cover computational PKS analysis we did not only focus on AT and KS domains but took all eight domains into account. The results of the searches were the basis for the construction of maximum-likelihood trees which allowed the more precise classification of the HMM hits into type I PKS and non-PKS I members.

Results

Extracting PKS I candidate sequences using Hidden Markov Models

From 926 annotated type I PKS domain sequences in the PKSDB dataset [27], we generated multiple alignments and constructed eight Hidden Markov Models (one for each domain) that were searched against 6,613,204 predicted proteins in six metagenomics samples and UniRef (for details see methods).

In total 22,106 candidate sequences of the eight PKS I domains were retrieved and analyzed. They range from 45 MT domain sequences to 4355 sequences of the KS domain type (for individual datasets see Table S1). For most of the domains the UniRef set has the highest total and relative (compared to the total number of analyzed proteins) number of candidate type I PKS.

Refining potential PKS I sequences using maximum likelihood trees

Although we did not find type II PKS sequences, due to the similarity of PKS I to FAS I and other enzymes, HMMs alone were not sufficient to discriminate PKS I proteins and related enzymes. Therefore, we applied a phylogenetic approach [28] which allowed the subsequent characterization of type I PKS subgroups.

In agreement with previous knowledge the trees of the AT, DH, ER, KR, KS and PP domains show in general a consistent phylogenetic profile and contain PKS I and non-PKS I taxa (see Fig. 1 as an example, all other trees can be found in Methods S1 and S2). The main fraction of leaves in the PKS I branches is contributed by the Actinobacteria and clusters mostly together (see Table S2). Members of the Proteobacteria and other bacteria phyla occur in mixed groups. The fungal sequences form in most of the trees one or two groups within the PKS I branch and are closely located to sequences of other eukaryotes like *Dictyostelium* and animals. It was previously described that most of these animal proteins are FAS I members which are phylogenetically related to the fungal type I PKS [12, 16] and also the occurrence of PKS-like sequences in animal genomes (e.g. in sea urchin for the production of pigments) has been reported [10].

Not all domains perform equally in identifying PKS I members. For example, in the TE domain tree two clades are dominated by PKS I sequences but a clear discrimination between PKS I and non-PKS I members cannot be made for the rest of the tree. For example, the MT domain tree contains only a few members as the domain occurs quite rarely in type PKS I; also due to the short length of the PP domain the results in this tree are less resolved than those of the other seven domains.

The non-PKS I branches are large in some trees. In particular, in AT, KS and TE domain trees many unspecified acyltransferases, ketoacyl synthases and thioesterases respectively, were apparently not filtered out by the HMM searches. In the DH domain tree the non-PKS I sequences are predominately annotated as FAS members while ER and KR domain HMM searches seems to attract non-specified dehydrogenases and other oxidoreductases. The non-PKS I PP domain members were mainly adenylate amino acids or nonribosomal peptide synthetases (NRPS).

Quality analysis of the tree-based approach and HMM searches

The enormous computational requirements of the tree reconstructions made bootstrap analyses infeasible. However, the fragmented environmental sequences could strongly influence the quality and significance of the branches. We thus compared the trees with reference trees without metagenomic sequences and randomly created trees with the same amount of taxa. The Robinson-Foulds distances [29] between the test trees and the references trees were in general much smaller than the distances to random trees (see Fig. S2, Table S3 and Methods S3). Also, the log likelihood of the reference tress and trees with metagenomics samples show a much better fit to the sequence alignments and are much more similar to each other than to trees with random topologies (see Methods S3 and S4). This implies that the trees are a good representation of the phylogenetic signal in the dataset and that their topologies are not overly influenced by the inclusion of the metagenomic sequences.

To support the tree-based annotation of the metagenomics sequences, the placements of all manually annotated PKS I from PKSDB were checked. They should only be found in branches of the trees that are marked as PKS I containing branches. With exception of the TE domain set which has three PKSDB sequences that are located in non-PKS I branches (see Methods S5) all sequences are placed as expected in PKS I branches.

Using the trees for classification, it became apparent that the HMM bit score values are not a sufficient criterion for discriminating the type I PKS from the non-PKS I sequences. To quantify this, sequences of the HMM searches were grouped by their tree based annotation (implying that this is close to the true function). The bit score distributions of these groups were compared domain-wise and plotted as box plot (Fig. 2 for the AT domain, Fig. S1 for all domains). All domains have a higher median value for the PKS I than the non-PKS I. But for most of the domains there is a large overlap of the bit score value between these groups. Especially the many outliers with low bit scores in the type I PKS group coming from metagenomic proteins fall in the inter-quartile range of the non-PKS I group.

Taken together, these quality measurements indicate that the tree approach can properly classify the candidate sequences retrieved by HMMs into PKS and non-PKS I members.

PKS I domain densities in various environments

The number of domains that fall in branches which are classified as type I PKS members as they contain known PKS I sequences are visualized in Fig 3. In nearly all seven data sets the KS domain is found most frequently (with the exception of enhanced biological phosphorus removal sludge data sets) followed by the AT, PP or KR domains. ER and TE sequences occur generally in much lower counts. In agreement with previous studies the MT domain appears very rarely and could only be found in UniRef, the Minnesota farm soil sample and the phosphorus removal sludge. The discrepancy between the AT and KS domain occurrences might indicate different, domain specific HMM sensitivities as they tend to occur at equal copies, but it could have also biological reasons as the number of AT domains in PKS I proteins might differ from the number of KS domains if a trans-acting AT domain is involved [30].

The density of PKS I domains has the highest value in UniRef when the number of tree-refined PKS I sequences is normalized by the total number of proteins in each of the data sets (Fig 4A). It is around three times higher than that of Minnesota farm soil sample which has the highest in all environments.

In UniRef, many different PKS I domains are found in the same protein while the metagenomic sequences mostly encode protein fragments with a single domain due to the shotgun approach taken during data generation. Assuming that each of these metagenomic domain sequences represent a full type I PKS protein we normalized the number of single and multi domain hit proteins by the number of screened proteins (Fig. 4B). We found that only the farm soil has a higher PKS I density than UniRef , and PKS I seem most rare in the gut sample where only a single domain occurrence could be detected.

The identified PKS I proteins were also normalized by the number of genome equivalents for the Minnesota farm soil, Sargasso Sea, whale falls and acid drainage mine data sets as for these environments average effective genome sizes have been estimated [31]. With nearly seven type I PKS per genome equivalent, the farm soil has the highest density of these proteins (Fig. 5). This is in the range of fully sequenced genomes of organisms from soil habitats [12].

In UniRef, the largest proportion of potential PKS I proteins identified originated from Actinobacteria (5642 sequences), followed by Proteobacteria (3625 sequences). This is similar to statements of previous studies and may be biased by the number of sequenced genomes of these phylogenetic groups [12]. The counting of all taxonomic groups can be found in Table S2. We did not find potential type I PKS members in archaeal proteins. A possible reason for this is the lack of an FAS AT domain in archaea [12] and the low likelihood of horizontal transfer of PKS I genes. As the source organisms of proteins from

environmental samples are unknown, a detailed analysis of the taxonomic distribution is currently impossible.

As expected, the majority of the environmental sequences are located in clades dominated by bacterial PKS I domains, but there are metagenomic sequences that seem to have a closer relationship to eukaryotic type I PKS members. For example six Sargasso Sea sequences can be found close to *C. elegans* and Alveolata proteins in the AT domain tree. The originating species of these sequences is unfortunately unclear.

Despite the fragmentation of the metagenomic sequences we were able to find proteins with multiple domains in some of the six environments. In the Sargasso sea sample, 15 of these with a maximum number of seven domains were detected. The farm soil collection hosted nine multidomain proteins but none extended beyond two domains. The phosphorus removal sludge set contained six (up to three domains) and the whale fall one (two domains) of such sequences. The small number of multi domain hits found reflects the low coverage of the samples. But the fact that at least some are found give high confidence that we have detected real PKS I members and that these communities might be useful as sourced for further and more focused sequencing and screenings.

Distribution of potential type I PKS members in the different Sargasso Sea samples

The Sargasso Sea data set is composed of seven samples. It has been suggested that sample 1 of the Sargasso Sea data set was contaminated with *Burkholderia* and *Shewanella* species [32]. To exclude the possibility that this contamination biased the identification of PKS I proteins, the sample of origin of each of protein identified was examined. Additionally, their closest relatives in UniRef were determined by using BLAST. We found that seven of the 15 proteins with multiple domain hits were encoded by contigs mainly built from sample 1 reads, four from *Burkholderia* and two from *Shewanella* . Of the 171 single-domain hit proteins in the seven Sargasso samples, only 27 are found in contigs with contributions of sample 1 and none of these seems to be close related to *Burkholderia* proteins or *Shewanella* proteins. The high number of multi-domain protein hits coming from potential contaminations may be a result of the better coverage of these genomes in the first sample. However, the remaining single-domain hit proteins provide enough evidence that type I PKS proteins are not solely due to the contaminating species but that the uncontaminated ocean sample also hosts type I PKS producing organisms.

Detection of non-annotated PKS I members in UniRef

The screening and tree based refinement of UniRef proteins revealed type I PKS members that were so far not annotated as PKS I or PKS at all. This includes 971 proteins with multiple PKS I domain HMM hits and 760 proteins (mostly short, fragmented ones) with only one such hit. Additionally we could confirm the proposed annotation of further proteins, 197 proteins with multiple domain hits and 146 proteins with single domain hits,

that were marked as hypothetical, putative, probable or predicted PKS or PKS I.

The classification and functionality of PKS proteins in animals is still unclear. Based on the analysis of AT and KS domains Jenke-Kodama et al. [12] placed the animal FAS into the type I PKS family which makes them a subfamily of PKS. Castoe et al. [10] showed that sea urchins (*Strongylocentrotus purpuratus* and *Lytechinus variegatus*), birds (*Gallus gallus*), and fish (*Danio rerio* and *Tetraodon nigroviridis*) harbour PKS-like proteins with uncertain functionality, which are closely related to PKS members of *Dictyostelium*. In our study, the Metazoa contributed proteins with AT and KS domains (in some cases also the ER domain) that were placed in the PKS I branches of the trees while the remaining domains were found in non-PKS I branches. This distribution was the case for some insects, amphibia fish, echinodermata and mammals. In contrast all detected six domains of a protein in *Caenorhabditis briggsae* and eleven domains (except one DH domain) in *Caenorhabditis elegans* seem to be true type I PKS domains.

The proteins in the Alveolata *Cryptosporidium hominis*, *Cryptosporidium parvum*, *Toxoplasma gondii* are very large and contain only PKS I annotated domains. It confirms the described occurrence of PKS I in the protozoan pathogen *Cryptosporidium parvum* [9]. The detection of type I PKS members in *Ostreococcus tauri* and *Ostreococcus lucimarinus* sequences in UniRef supports a study that reported type I PKS proteins in unicellular green algae based on a KS domain tree [33]. The PKS I of these protists are described to be different from the currently known PKS proteins and might have a long separated evolution. The different domains detected were found to be placed close to disparate taxonomic groups (within bacteria and eukaryotes) in the trees generated.

Indication of horizontal gene transfer

The constructed phylogenetic trees also revealed some cases of potential horizontal gene transfers. An example is a small group of 3 fungal protein taxa in the AT domain tree that is placed in the Actinobacteria. In the DH domain tree, four *Danio rerio* (zebra fish) sequences are nested in a small group of fungal sequences that is surrounded by sequences from Actinobacteria. All proteins have the same domain structure including a KS, AT and KR domain in addition to the DH domain. It cannot be excluded that the detected protein originated from a genome contamination though. Protein identifiers of the described cases are listed in the Metods S3.

Discussion

Because of their size, modular structure, complicated evolution and similarity to type I FAS and other enzymes, PKS members are a challenging group of enzymes to identify and to classify. We were able to detect type I PKS proteins – one subgroup of the PKS group - in almost all the samples studied (Fig. 3). The Minnesota farm soil sample shows the highest density of PKS I which is not surprising as this environment has the highest species density which leads to strong competition and an “arms race” between species. The enormous potential for soil as source of useful secondary metabolites was already discussed earlier [34] and our results support these statements.

For both the human gut (145 Mb of reads, 46503 predicted genes) and acid mine drainage samples (140 Mb, 46862 predicted genes), the HMM searches identified only one candidate PKS I, albeit with high similarity to known PKS I sequences. This implies a low PKS I density in these environments and it has to be proven whether the respective species are members of the microbial communities or just temporal bystanders that came in via food or air. At least for AMD, one of the two detected PKS I proteins was found in one of the major community members, the *Leptospirillum* group III. This implies that even in an inhospitable environment like AMD, which contains only a small number of species, the community forces its inhabitants to arm themselves with expensive secondary metabolites. These kinds of environments have so far not been considered as sources of PKS proteins but our study indicates that novel attempts to search for antibiotics and other metabolites in them may reap rich rewards.

In addition to a quantification of PKS I in diverse environments, our study has also helped to classify unknown proteins in UniRef and improved their annotation. The usage of phylogenetic trees to discriminate between PKS I and non-PKS I sequences seems to be a feasible approach which also partially overcomes the problem of low bit score values and fragmentation of environmental proteins using traditional sequence similarity searches. Depending on the target sequences this method can be successfully applied to search in Sanger sequencing data sets and new generation 454 pyrosequencing data sets with read lengths starting from 450 bp (see Methods S4). The approach also shows the limits of current annotation schemes: If HMM searches had been the only approach used, this would have resulted in many false positives and false negative PKS I being identified. Despite this, the HMMs used here have been carefully designed, appear PKS I specific and are much more discriminative than those currently available (e.g. in PFAM [35] or TIGRFAM [36]). The HMMs have been deposited in SMART [37]. The combination of the information of all eight domain searches was shown to be a powerful detection method.

The approach outlined here can be applied to search further proteins of interest in environmental shotgun sequences and has been already successfully used to screen for the much smaller family of Nitrilases [6]. The rapidly increasing amount of metagenomic data that will be publicly released requires methods such as the one presented here to quickly and cheaply screen for proteins of interest.

Materials and Methods

Metagenomic and reference data sets

Sets of predicted proteins from the following metagenomics samples were analyzed in this study: Minnesota farm soil [21], Sargasso Sea [22], human gut [23], acid mine drainage [24], enhanced biological phosphorus removal sludges [25] and whale falls (sunken whale bones) [21]. Additional to the metagenomic samples proteins sequences from UniRef100 database [26] were used as reference set.

Hidden-Markov-Model creation and search

Due to the fact that neither Pfam [35] nor other resources offer Hidden-Markov-Models (HMM) of all the the eight PKS I domains, they were constructed based on a manually curated set of PKS I protein sequence hosted at PKSDB [27]. For each domain the sequences were aligned with *muscle* [38]. Based on these alignments HMMs were created and calibrated by *hmmbuild* and *hmmcalibrate* HMMER-package [39]. The UniRef protein sequences were screened with these HMMs. Alignments (by *muscle*) of extracted proteins were used to calculate maximum likelihood trees. The trees helped to manually select real PKS I members that were afterward aligned again. After a manual cleaning of these alignments they were used to generated HMMs (with the above described tools). Searches for type I PKS domains in the metagenomic sequences and UniRef were performed with these PKS I domain specific HMMs. A non-HMM based searching approach can be found in Methods S4 and Table S4, S5 and S6.

Tree construction

For each domain the sequences detected by the HMM were filtered by their e-values (see Methods S4). The selected sequences from UniRef and the metagenomic datasets were aligned by *hmmalign* (included in the HMMER-package [39]). For the KS and PP domain the UniRef sequence collection was shrunk to a set of representatives by making use of *blastclust* (from the NCBI BLAST package [40]) and a Python script [http:python.org]: Clusters based on a similarity cut-off of 90% were created and the annotation strings checked if all members were either PKS I or non-PKS I sequences. Without the resizing these two datasets would have been too large for further processing by *phyml*. Based on the alignments maximum likelihood trees were constructed using a slightly modified (removing limitation for memory usage - see Methods S6 for the patch file) version of *phyml* [28].

Data base construction and querying

Information like fasta file headers, HMM result quality, tree position and manual, tree

based classification of the sequences were combined in a *sqlite* database [http://www.sqlite.org] that was queried to created result statistics (see Methods S5).

Comparison of the tree topologies with reference trees

To test if the noise from the fragmented metagenomic samples overwhelms the phylogenetic signal of the reference set sequence from UniRef, a reference tree based on the alignments for the HMMs was built for each domain. The tree containing the environmental sequences and the reference trees were then pruned to their set of common taxa using *clann* [41]. For each domain 500 random trees containing the same leaf set as these common taxa trees were generated by the program *random_tree* (see sumplementary material) using a markovian approach. The pairwise Robinson-Foulds distances [29] of all combinations of these 502 trees were calculated with the *rfdist* function of *clann*. Supported by a python script box plots were created using *R* (http://www.r-project.org/).

Visualization and manual annotation

We used iTOL [42] for manual rerooting and visualizing of the trees. Tree nodes of proteins derived from UniRef or PKSDB were colorized by the taxonomic classification of the hosting species (different levels based on NCBI Taxonomy [43]). In addition automated, keyword based analysis of the annotation strings lead to a second color ring of the UniRef taxa. Further a source classifying color code was applied to environmental protein nodes. Both, UniRef and environmental proteins were marked by a color ring that reflects a value that we dubbed “global protein hit score” (GPHS). It is the difference of the number of domains in protein that are placed in PKS I branches and number of domains that are place in non-PKS I branches, divides by the total number of found domains (n_{PKS} - n_{Non_PKS}/ n_{PKS} + n_{Non_PKS}). Proteins with a GPHS higher than 0 are more likely to be PKS, Proteins with a GPHS lower than 0 are more likely to be non-PKS I. The GPHS can only be calculated for multi domain hit protein.

For a visualization of the results, a program is provided that creates graphical overviews of the proteins and the detected domains based on the database content.

Code and data availability

All python and C programs (Methods S6) that were created for this study are open source and available under the ISC license (http://www.opensource.org/licenses/isc-license.txt). The data base files and all other files are free availability under the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0/).

The generated detailed results are available in the supplementary material. This includes the resulting sequences of the HMM searches (Methods S7), the alignments (Methods

S7), the trees in Newick format (Methods S7), visualization of the trees (Methods S1 and S2) as well as the database that hold the integrated data (Methods S5). Also a text file of selected parts of the database is included (Methods S5). The created Hidden-Markov-Models are incorporate into domain search web service SMART [37].

Acknowledgments

We would like to thank all members of our group for support and feedback, especially Jeroen Reas, and the reviewers for their constructive and stimulating comments.

References

1 Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734-740.

2 Yun J, Ryu S (2005) Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. Microb Cell Fact 4: 8.

3 Beja O, Aravind L, Koonin E, Suzuki M., Hadd A., et al. (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science 289: 1902-6.

4 Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G (2007) Structural and Functional Diversity of the Microbial Kinome. PLoS Biol 5: e17.

5 Podar M, Eads J, Richardson T (2005) Evolution of a microbial nitrilase gene family: a comparative and environmental genomics study. BMC Evol Biol 5, 42.

6 Raes J, Foerstner KU, Bork P (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. Curr Opin Microbiol 10: 490-498.

7 Staunton J, Weissman KJ (2001) Polyketide biosynthesis: a millennium review. Nat Prod Rep 18: 380-416

8 Zucko J, Skunca N, Curk T, Zupan B, Long PF, et al. (2007) Polyketide synthase genes and the natural products potential of Dictyostelium discoideum. Bioinformatics 23: 2543-2549.

9 Zhu G, LaGier MJ, Stejskal F, Millership JJ, Cai X, et al. (2002) Cryptosporidium parvum: the first protist known to encode a putative polyketide synthase. Gene 298: 79-89.

10 Castoe, TA., Stephens T, Noonan BP, Calestani C (2007) A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. Gene 392: 47-58.

11 Calestani C, Rast JP, Davidson EH (2003) Isolation of pigment cell specific genes in the sea urchin embryo by differential macroarray screening. Development 130: 4587-4596.

12 Jenke-Kodama H, Sandmann A, Müller R, Dittmann E (2005) Evolutionary implications of bacterial polyketide synthases. Mol Biol Evol 22: 2027-2039.

13 Moore BS, Hopke JN (2001) Discovery of a new bacterial polyketide biosynthetic pathway. Chembiochem 2: 35-38.

14 Müller R (2004) Don't classify polyketide synthases. Chem Biol 11: 4-6.

15 Shen B (2003) Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. Curr Opin Chem Biol 7: 285-295.

16 Kroken S, Glass NL, Taylor JW, Yoder OC, Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci U S A 100: 15670-15675.

17 Jenke-Kodama H, Börner T, Dittmann E (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium Streptomyces avermitilis. PLoS Comput Biol 2: e132.

18 Ridley CP, Lee HY, Khosla C (2008) Evolution of polyketide synthases in bacteria. Proc Natl Acad Sci U S A 105: 4595-4600.

19 Schirmer A, Gadkari R, Reeves CD, Ibrahim F, DeLong EF, et al. (2005) Metagenomic analysis reveals diverse polyketide synthase gene clusters in microorganisms associated with the marine sponge Discodermia dissoluta. Appl Environ Microbiol 71: 4840-4849.

20 Minowa Y, Araki M, Kanehisa M (2004) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol. 368: 1500-17

21 Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.

22 Venter JC, Remington K, Heidelberg JF., Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66-74.

23 Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

24 Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.

25 Martín, HG, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR)

sludge communities. Nat Biotechnol. 24: 1263-1269.

26 Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282-1288.

27 Yadav G, Gokhale RS, Mohanty D (2003) SEARCHPKS: A program for detection and analysis of polyketide synthase domains. Nucleic Acids Res 31: 3654-3658.

28 Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.

29 Robinson DF, Foulds LR (1981) Comparison of Phylogenetic Trees. Mathematical Biosciences, 53: 131-147.

30 Cheng Y, Tang G, Shen B (2003) Type I polyketide synthase requiring a discrete acyltransferase for polyketide biosynthesis. Proc Natl Acad Sci U S A 100: 3149-3154.

31 Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. Genome Biol 8: R10.

32 DeLong EF (2005) Microbial community genomics in the ocean. Nat Rev Microbiol 3: 459-469.

33 John U, Beszteri B, Derelle E, de Peer YV,. Read B, et al. (2008) Novel insights into evolution of protistan polyketide synthases through phylogenomic analysis. Protist 159: 21-30.

34 Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-R249.

35 Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-D251.

36 Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371-373.

37 Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34: D257-D260.

38 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.

39 Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.

40 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

41 Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21: 390-392.

42 Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.

43 Wheeler DL, Chappey C, Lash AE, Leipe DD, Madden TL, et al. (2000) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 28: 10-14.

Figures

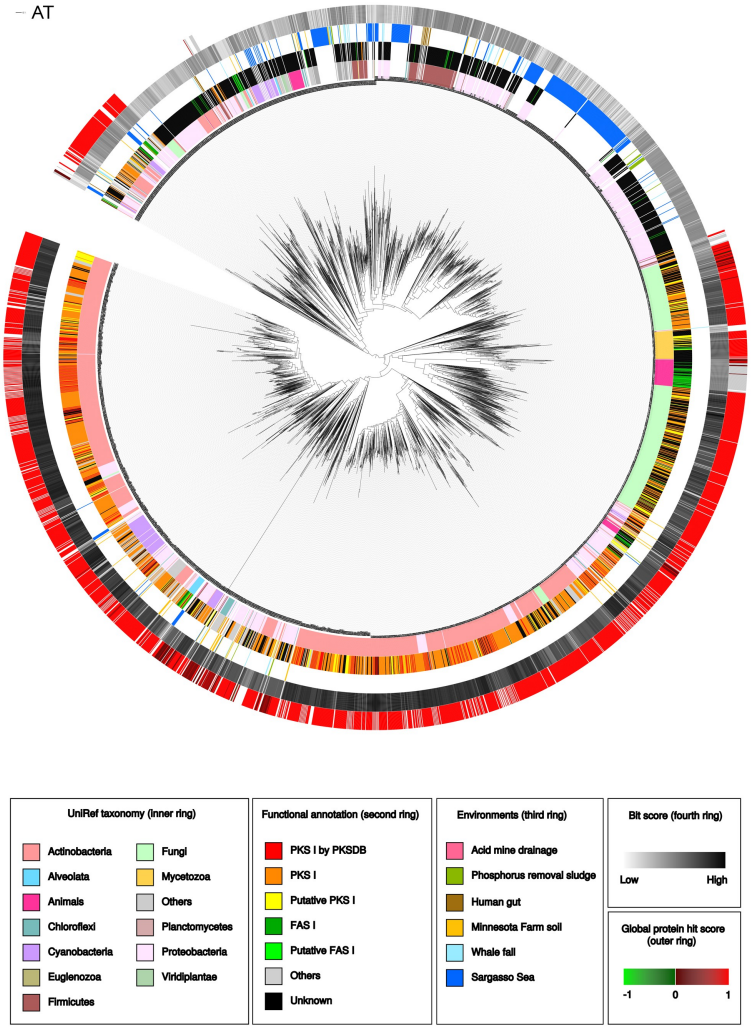


Figure 1: Maximum likelihood-tree of the AT domain.

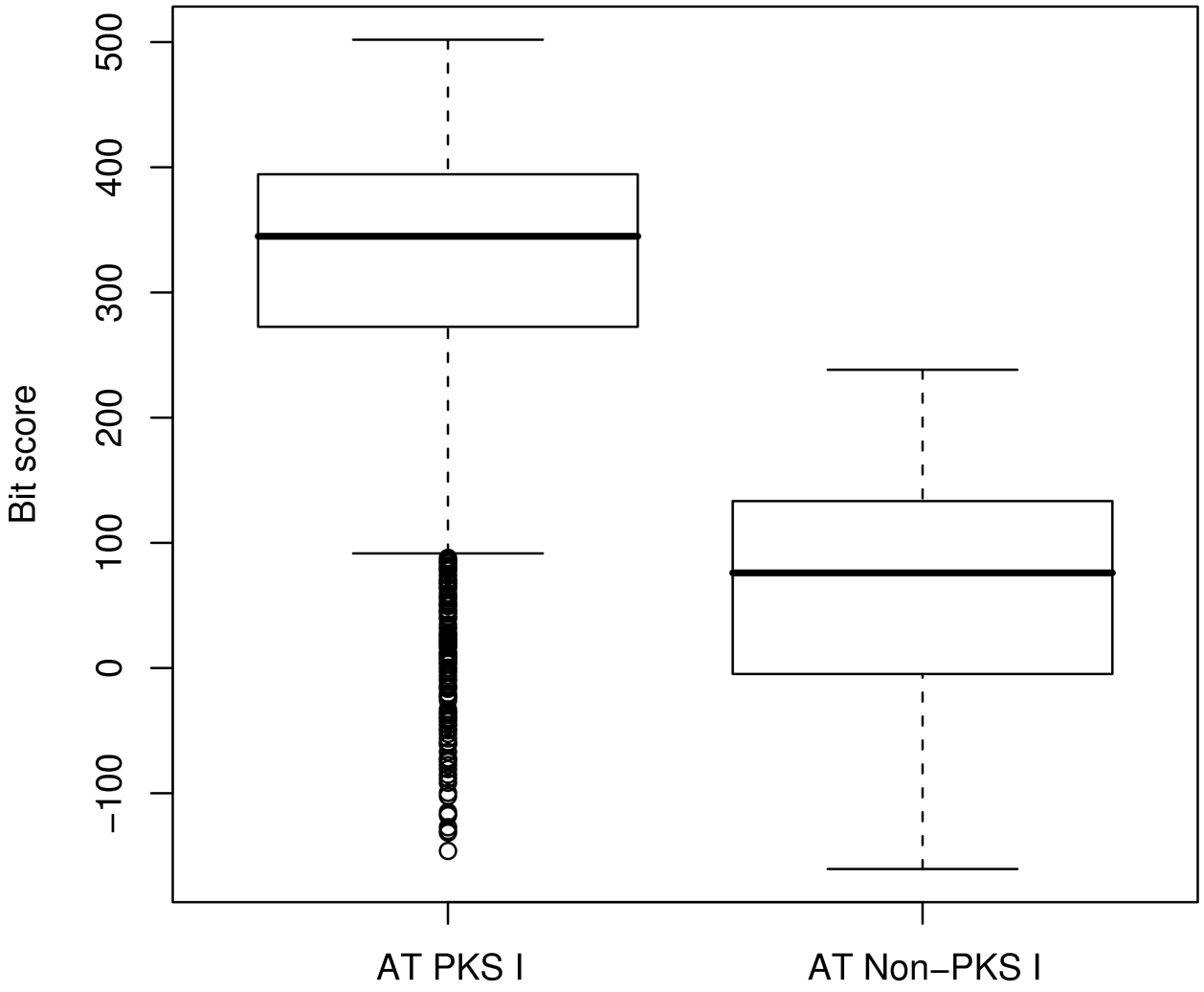


Figure 2: Box plots of the bit score distribution of HMM search result sequences for the AT domain classified as PKS I or as non-PKS I using the tree.

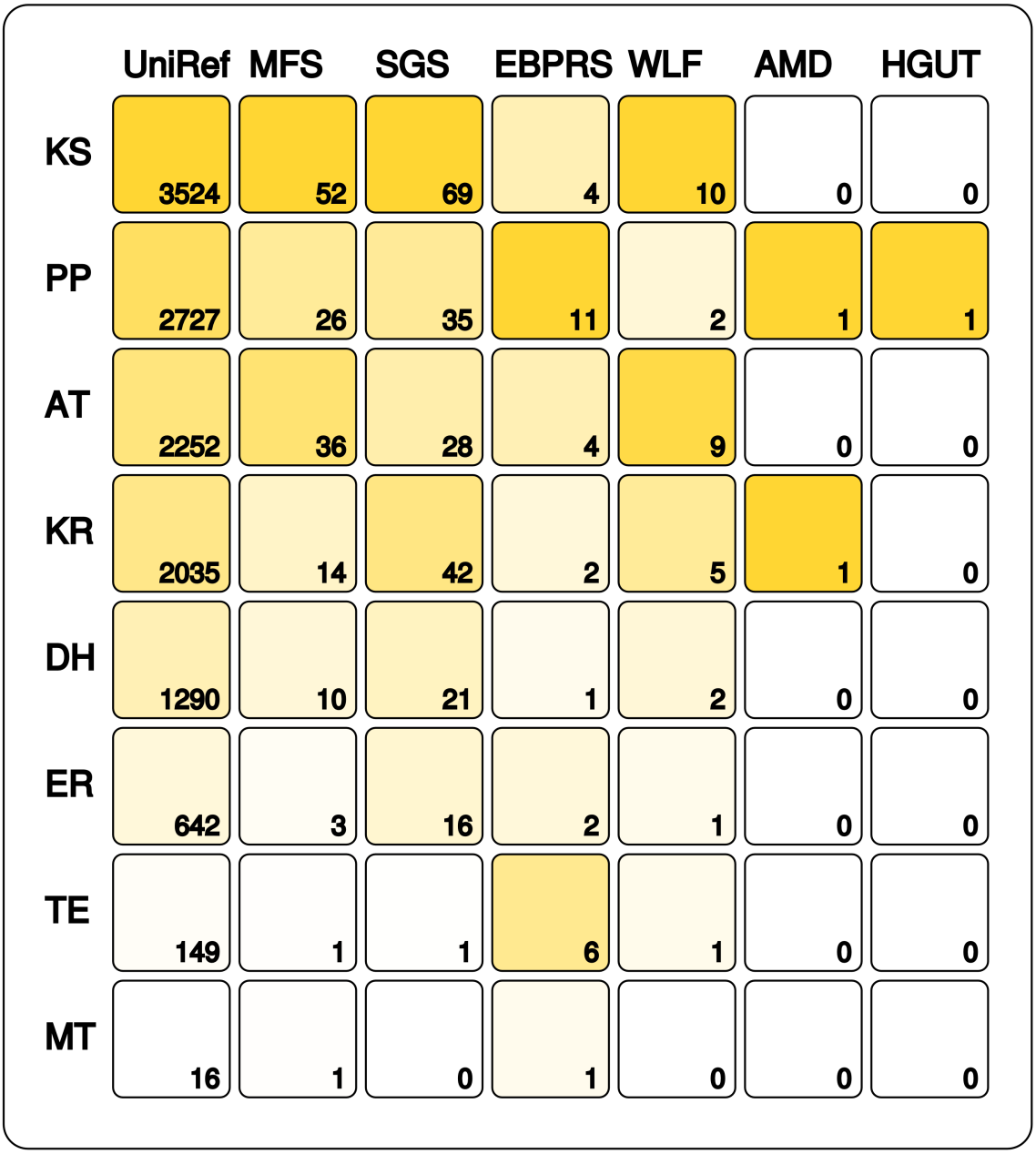


Figure 3: Number of sequences in the data sets that are annotated as type I PKS domains based on the maximum-likelihood tree. The intensity of the color is equivalent to the relative number of sequences inside a data set. The KS domain has in the larger data sets the highest number of hits and the ratio of the AT, KS and PP domain is mostly similar.

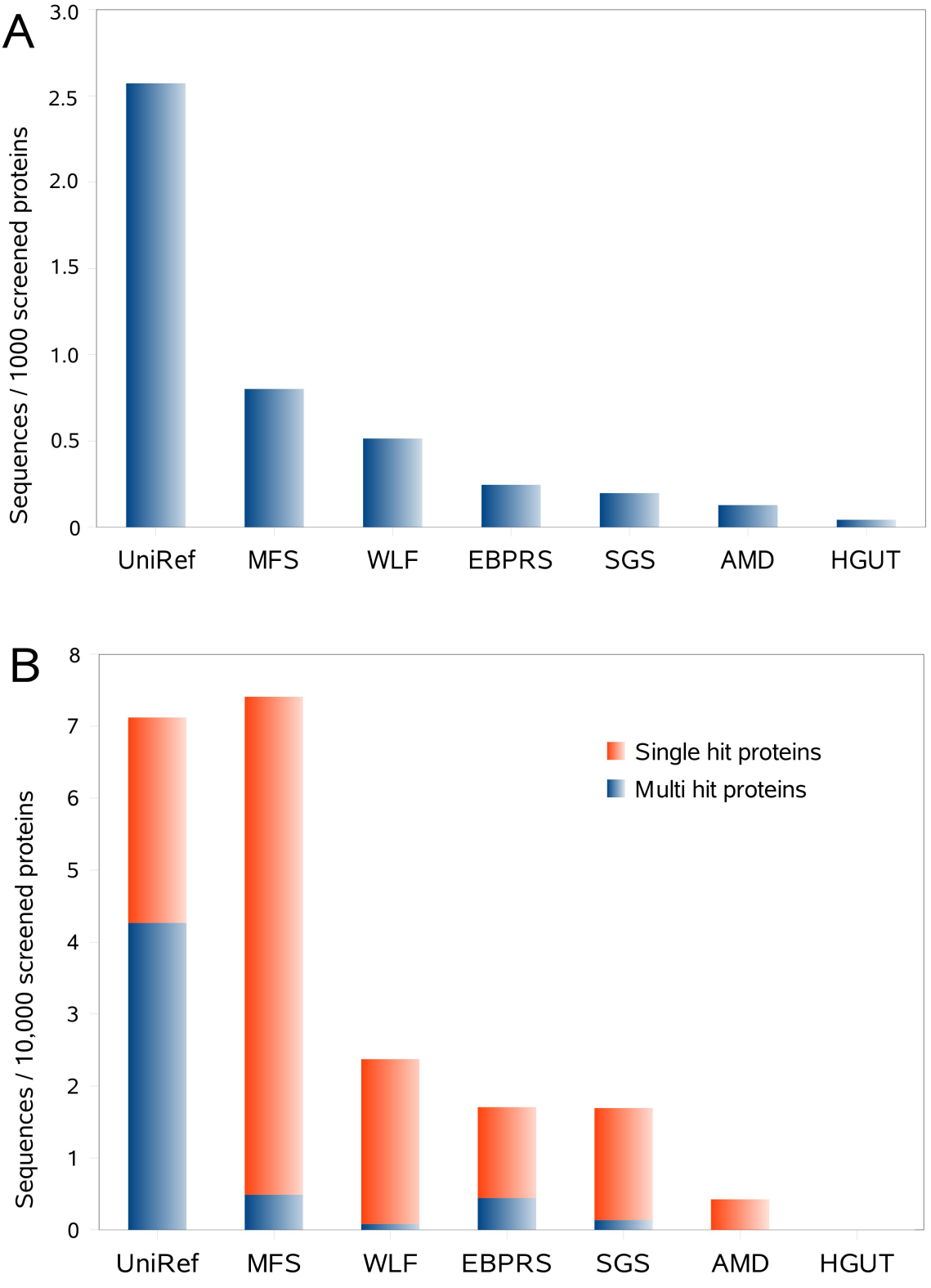


Figure 4: A – PKS I classified sequences normalized by total number of screened proteins. B – PKS I classified sequences normalized proteins-wise (all domains of one protein are counted together as one entity) by total number of screened proteins.

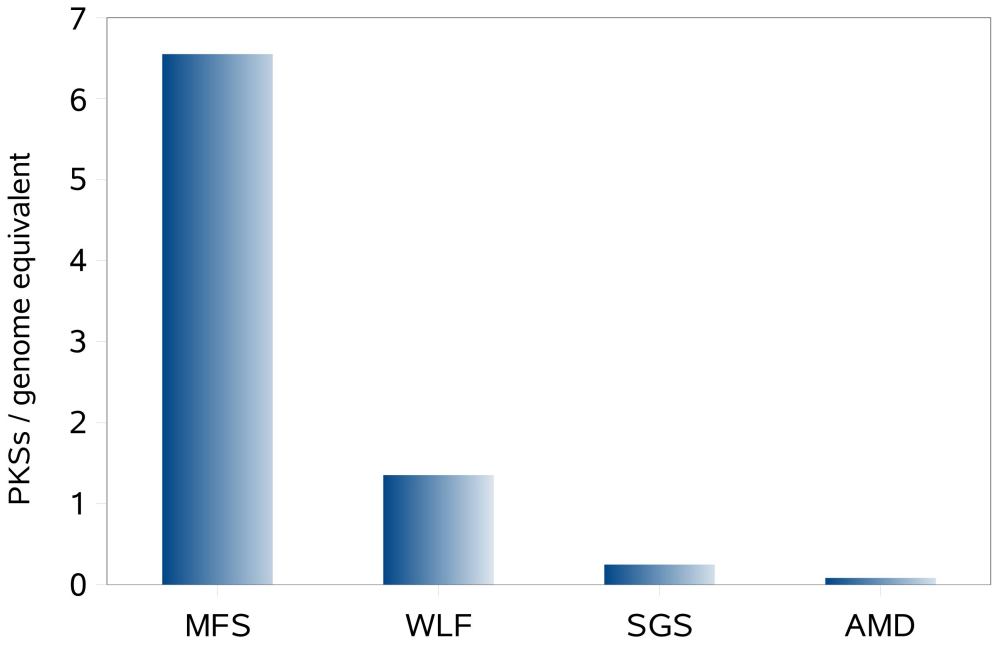


Figure 5: Type I PKS members per genome equivalent for the Minnesota farm soil, whale falls, Sargasso Sea and acid mine drainage sample estimated by Raes et al. [31]. The soil sample has the highest density of type I PKS per genome.

Appendix F

A nitrile hydratase in the eukaryote *Monosiga brevicollis*

A nitrile hydratase in the eukaryote *Monosiga brevicollis*

Abstract and Affiliations

Authors: Konrad U. Foerstner, Tobias Doerks, Jean Muller, Jeroen Raes, Peer Bork

Affiliations*: European Molecular Biology Laboratory, Heidelberg, Germany*

Abstract

Bacterial nitrile hydratase (NHases) are important industrial catalysts and waste water remediation tools. In a global computational screening of conventional and metagenomic sequence data for NHases, we detected the two usually separated NHase subunits fused in one protein of the choanoflagellate *Monosiga brevicollis*, a recently sequenced unicellular model organism from the closest sister group of Metazoa. This is the first time that an NHase is found in eukaryotes and the first time it is observed as a fusion protein. The presence of an intron, subunit fusion and expressed sequence tags covering parts of the gene exclude contamination and suggest a functional gene. Phylogenetic analyses and genomic context imply a probable ancient horizontal gene transfer (HGT) from proteobacteria. The newly discovered NHase might open biotechnological routes due to its unconventional structure, its new type of host and its apparent integration into eukaryotic protein networks.

Introduction

Nitril hydratases (NHases, E.C. 4.2.1.84) catalyze the hydrolysis of nitriles to their corresponding amids [1]. Often, this reaction is part of a two-step degradation pathway and is followed by an amidase catalyzed step. The respective amidase converts the amid into the corresponding carboxylic acids and ammonia. The structure [2,3] and reaction mechanism [4] of representative NHases have been extensively studied: The hetero-dimer or hetero-tetramer [2,3] consists of two kinds of subunits - α and β - and occurs as metalloenzyme that contains either iron (non-heme Fe(III)) or cobalt (non-corrin Co(III)) ions [5-8]. The biological function of the NHases is unknown so far but it was shown that they enable the respective organism to utilize aliphatic, aromatic and hetero-aromatic nitriles as sole nitrogen source under laboratory conditions e.g. [9,10]. Due to their ability to selectively and efficiently hydrolyze cyano groups, NHases are heavily used in biotechnological industry e.g. for the synthesis of the essential chemicals acrylamide (30,000 tons/year [11]) and nicotinamide (> 3500 tons/year [12]). In addition, their enzymatic activities are used to remove toxic nitriles (e.g. nitrile herbicides) during waste water treatment [13].

So far, NHases are described to occur in species belonging to the phyla Proteobacteria, Actionobacteria, Cyanobacteria and Firmicutes, in habitats ranging form soil [14], via costal

marine sediments [15] and deep sea sediments [10,16] to geothermal environments [17,18]. Here, using a large scale screen for NHases in public sequence databases and metagenomic datasets, we describe the identification of the first eukaryotic NHase and investigate its origin.

Results

In order to get an overview about the phylogenetic and habitat distribution of NHases, we created HMMs (Hidden-Markov-Model) for each of the two subunits based on 42 α and 48 β subunit sequences and screened 12,126,382 proteins (or protein fragments) from UniRef and seven metagenomic data sets from diverse environments. In total, 324 α (including 14 of thiocyanate hydrases (SCNases) [19]) and 265 β (including 4 SCNases) subunit members were found in this homology search step. The α subunit HMM seems to be more sensitive – the ratio of α to β sequences is not 1:1 as expected. Yet, the HMMs identify both subunits in most of the species in UniRef and also in some of the metagenomic scaffolds.

To confirm the NHases membership of the identified sequences, to study the taxonomic distribution of the originating organisms and to possibly define new subgroups we constructed maximum likelihood trees of both subunits. These trees (Figure 1) confirmed that the detected sequences are NHases and show taxonomic clustering. They illustrate that all sequences – also the metagenomic ones - seem to originate from bacterial species, with a large fraction of proteobacterial NHases found in the Global Ocean Sampling Expedition dataset (see supplementary material). There is one notable and surprising exception to this observation: both subunits are contained in a single hypothetical open reading frame (UniProt identifier A9V2C1) of the recently sequenced choanoflagellate *Monosiga brevicollis* [20], as deposited in the UniRef database.

The unicellular *Monosiga brevicollis* is one of more than 125 known choanoflagellates which represent the closest known relatives of metazoans (i.e. are closer to animals than plants and fungi). They can form simple multicellular colonies and are found in marine, brackish and freshwater habitats in which they use their apical flagellum to prey bacteria [21].

As *Monosiga* would be the first eukaryote that harbors an NHase, we analyzed the respective gene and encoding protein in detail.

The putative NHase is 496 amino acids long and contains the usually separately encoded subunits fused into one protein connected by a Histidin-rich stretch (Figure 2). Both subunits seem complete and the putative ion binding active site in the α subunit (single letter code: CXXCSC) that is necessary for NHase functioning [1] appears conserved. The orientation of the two subunits in the coding region of the genome of *Monosiga brevicollis* is similar to the operon structure of the subunits in proteobacteria; the β subunit is located 5'-terminal, the α subunit 3'-terminal. This observation is in line with the results of the phylogenetic analysis (Figure 1), where the protein clusters together with NHases of

proteobacterial origin and with BLAST-based analysis, which clearly indicates proteobacteria as the most similar homologs (see supplementary material).

In order to exclude contamination and check for likely functionality, we analyzed genomic features and EST (expressed sequence tag) data. The expression of the gene is strongly supported by the existence of two ESTs covering a large portion of the gene (Figure 2). Furthermore, one EST (accession number JGI_XYM3899.rev) implies that the gene contains a 96 bp long intron. The GC value of the corresponding transcripts (59.4 %) differs only slightly from the median GC value of all *Monosiga* transcripts (56.9%) which strengthen the assumption that it is a gene of *Monosiga* and not bacterial contamination of the genome sequence.

Putative amidases could be detected with HMMs in *Monosiga*'s protein set (as in other eukaryotes) but their genes are distantly located to the NHase in the genome and show only low similarity to the NHase-connected amidases in bacteria. Despite the fact that the identified amidases do not seem to be transferred from a proteobacterial donor together with the NHase, it is possible that an existing *Monosiga* amidase took over this functionality but we cannot exclude that the NHase products are processed differently in this choanoflagellate.

Discussion

The discovery of an NHase in an eukaryote, i.e. *Monosiga brevicollis,* from a sister group of animals, indicates a wider phylogenetic spread of NHases than currently believed. The presence of an intact domain structure, an (EST supported) intron and the similarity between the GC content of the gene and the surrounding genomic sequence makes a bacterial contamination extremely unlikely. As the eukaryotic NHase has a phylogenetic position within diverse bacterial NHases (Figure 1), the currently most parsimony explanation is that it resulted from an ancient horizontal gene transfer from bacteria into the choanoflagellate or a more ancient eukaryotic lineage. As it has been sustained for a considerable time to allow for GC amelioration, NHase functionality must have provided a selective advantage. The HGT hypothesis is corroborated by structural similarity to the proteobacterial NHase operon and its absence in any sequenced lower eukaryote so far, as well as the presence of highly repetitive stretches less than 10 bp upstream (5') of the gene which could have served as a site for homologous recombination and insertion of this gene. Although the alternative explanation (its presence at the root of all eukaryotes combined with multiple, independent losses in various eukaryotic lineages) cannot be excluded, all the above evidence makes it very unlikely.

Unfortunately, we are unable to predict the natural substrate of *Monosiga's* NHase and the low concentrations of nitriles expected in its habitats will likely hamper the determination of the precise role of the NHase in the physiology and ecology of this organism. For some aquatic bacteria, nitriles were previously reported to serve as nutritional sources [15,16,22]. We observe NHases in all samples of the Global Ocean Sampling Expedition and most samples of the North Pacific Subtropical Gyre implying a general ecological and nutritional

importance of this enzyme. Here we hypothesize that Monosiga has acquired the functionality to utilize nitriles for nutritional purposes.

From the biotechnological perspective, this newly discovered nitrile hydratase might be of relevance, too. The enzyme with fused subunits and a different type of host might have beneficial features like higher activity, higher stability or new substrate specificities.

Material and Methods

Data sets used

In this study sequences from the UniRef100 database [23] and the full set of proteins of *Monosiga brevicollis* [20] (downloaded from the JGI web site www.jgi.doe.gov) were analyzed. Additionally, we screened predicted proteins from the following metagenomics samples: Minnesota farm soil [24], Global Ocean Sampling Expedition [25], human gut flora [26], acid mine drainage [27], enhanced biological phosphorus removal sludges [28], North Pacific Subtropical Gyre [29] and whale falls (sunken whale bones) [24].

HMM creation

To create highly selective and specific Hidden-Markov-Models (HMM) of the two NHase subunits, available HMMs were retrieved from Pfam [30] (accession PF02979.7 and PF02211.6) and used for searches with *hmmsearch* (part of the HMMER package [31]) against the UniRef100 protein set. The extracted sequences were aligned with the program *muscle* [32]. Based on these manually cleaned alignments, we constructed and calibrated HMMs.

HMM search, tree construction and visualization

The UniRef and metagenomics protein data sets were screened by *hmmsearch* with the two NHase HMMs. After that the detected sequences were aligned with *hmmalign* (also included in the HMMER package). We manually added outgroup sequences to the alignments. The programs phyml [33], *clann* [34] and *seqboot* (PHYLIP packages [35]) constructed two trees (with 100 bootstrap repetitions) based on these alignments. After that Python scripts (www.python.org) integrated the sequence and taxomic information, annotation strings, trees and HMM search data into a database and created coloring files for iTOL [36] to visualize the trees.

Species mapping of environmental sequences

To map sequences from *Monosiga brevicollis* and metagenomic data sets to species a BLAST-based placing method was applied (see supplementary material).

Manual analysis

The manual analysis of the genomic region was performed with the tools *Artemis* [37] and *Clustal X* [38].

Acknowledgments

This work was supported by MetaHit (HEALTH-F4-2007-201052). We would like to thank Michihiko Kobayashi from the University of Tsukuba for providing us with help and all members of our group in particular Sean Powell for support and feedback.

References

1. Kobayashi M, Shimizu S (2000) Nitrile hydrolases. Curr Opin Chem Biol 4: 95-102.
2. Huang W, Jia J, Cummings J, Nelson M, Schneider G, et al. (1997) Crystal structure of nitrile hydratase reveals a novel iron centre in a novel fold. Structure 5: 691-699.
3. Nakasako M, Odaka M, Yohda M, Dohmae N, Takio K, et al. (1999) Tertiary and quaternary structures of photoreactive Fe-type nitrile hydratase from Rhodococcus sp. N-771: roles of hydration water molecules in stabilizing the structures and the structural origin of the substrate specificity of the enzyme. Biochemistry 38: 9887-9898.
4. Mitra S, Holz RC (2007) Unraveling the catalytic mechanism of nitrile hydratases. J Biol Chem 282: 7397-7404.
5. Banerjee A, Sharma R, Banerjee UC (2002) The nitrile-degrading enzymes: current status and future prospects. Appl Microbiol Biotechnol 60: 33-44.
6. Endo I, Nojiri M, Tsujimura M, Nakasako M, Nagashima S, et al. (2001) Fe-type nitrile hydratase. J Inorg Biochem 83: 247-253.
7. Harrop TC, Mascharak PK (2004) Fe(III) and Co(III) centers with carboxamido nitrogen and modified sulfur coordination: lessons learned from nitrile hydratase. Acc Chem Res 37: 253-260.
8. Kovacs JA (2004) Synthetic analogues of cysteinate-ligated non-heme iron and non-corrinoid cobalt enzymes. Chem Rev 104: 825-848.
9. Blakeya AJ, Colby J, Williams E, O'Reilly C (1995) Regio- and stereo-specific nitrile hydrolysis by the nitrile hydratase from Rhodococcus AJ270. FEMS Microbiology Letters 129: 57-61.
10. Layh N, Stolz A, Böhme J, Effenberger F, Knackmuss HJ (1994) Enantioselective hydrolysis of racemic naproxen nitrile and naproxen amide to S-naproxen by new bacterial isolates. J Biotechnol 33: 175-182.
11. Nagasawa T, Yamada H (1995) Microbial production of commodity chemicals. Pure andApplied Chemistry 67: 1241-1256.
12. Shaw NM, Robins KT, Kiener A (2003) Lonza: 20 Years of Biotransformations. Adv Synth Catal 345: 425-435.
13. Narayanasamy K, Shukla S, Parekh LJ (1990) Utilization of acrylonitrile by bacteria isolated from petrochemical waste waters. Indian J Exp Biol 28: 968-971.
14. DiGeronimo MJ, Antoine AD (1976) Metabolism of acetonitrile and propionitrile by Nocardia rhodochrous LL100-21. Appl Environ Microbiol 31: 900-906.
15. Langdahl BR, BISP P, Invorsen K (1996) Nitrile hydrolysis by Rhodococcus erythropolis BL1, an acetonitrile-tolerant strain isolated from a marine sediment. Microbiology 142 (1): 145-154.
16. Brandao PFB, Bull AT (2003) Nitrile hydrolysing activities of deep-sea and terrestrial mycolate actinomycetes. Antonie Van Leeuwenhoek 84: 89-98.
17. Pereira RA, Graham D, Rainey FA, Cowan DA (1998) A novel thermostable nitrile hydratase. Extremophiles 2: 347-357.
18. Toshifumi Y, Toshihiro O, Kiyoshi I, Takeshi N (1997) Cloning and Sequencing of a Nitrile Hydratase Gene from Pseudonocardia thermophila JCM3095. Journal of fermentation and bioengineering 83(5): 474-477.
19. Arakawa T, Kawano Y, Kataoka S, Katayama Y, Kamiya N, et al. (2007) Structure of thiocyanate hydrolase: a new nitrile hydratase family protein with a novel five-coordinate cobalt(III) center. J Mol Biol 366: 1497-1509.

20. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451: 783-788.
21. Buck KR, Garrison DL (1988) Distribution and abundance of choanoflagellates (Acanthoecidae) across the ice-edge zone in the Weddell Sea, Antarctica. Mar Biol 98: 263-269.
22. Colquhoun JA, Heald SC, Li L, Tamaoka J, Kato C, et al. (1998) Taxonomy and biotransformation activities of some deep-sea actinomycetes. Extremophiles 2: 269-277.
23. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23: 1282-1288.
24. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554-557.
25. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5: e77.
26. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.
27. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37-43.
28. Martin HG, Ivanova N, Kunin V, Warnecke F, Barry KW, et al. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. Nat Biotechnol 24: 1263-1269.
29. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.
30. Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: clans, web tools and services. Nucleic Acids Res 34: D247-D251.
31. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
32. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
33. Guindon Sp, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696-704.
34. Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21: 390-392.
35. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.
36. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.
37. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944-945.
38. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947-2948.

Figures

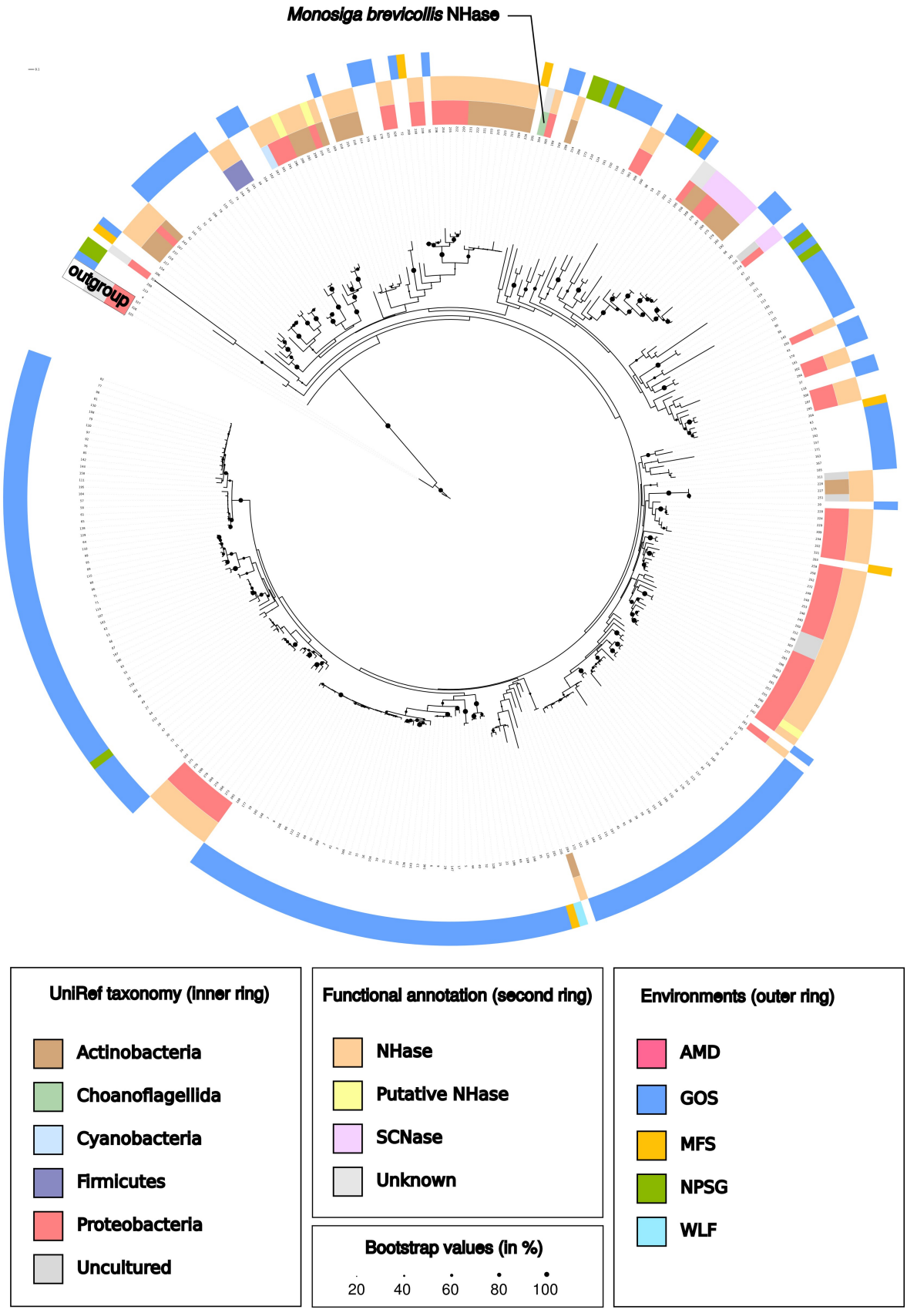

Monosiga brevicollis NHase
outgroup
UniRef taxonomy (inner ring)
Actinobacteria
Choanoflagellida
Cyanobacteria
Firmicutes
Proteobacteria
Uncultured
Functional annotation (second ring)
NHase
Putative NHase
SCNase
Unknown
Bootstrap values (in %)
20
40
60
80
100
Environments (outer ring)
AMD
GOS
MFS
NPSG
WLF

Figure 1: Maximum-likelihood tree of the NHase α subunit sequences. (AMD – acid mine drainage, MFS – Minnesota farm soil, GOS - Global Ocean Sampling Expedition, NPSG - North Pacific Subtropical Gyre, WLF – whale falls). The *Monosiga* sequence clusters together with sequences from GOS, MFS, NPSG and Actinobacteria and Proteobacteria from UniRef. A large fraction of GOS sequences form a separated branch (weak bootstrap support) with different subgroups. All these sequences seem to originated from Proteobacteria as our BLAST-based analysis indicate (see supplementary material). The β subunit shows a similar trend (see supplementary material).

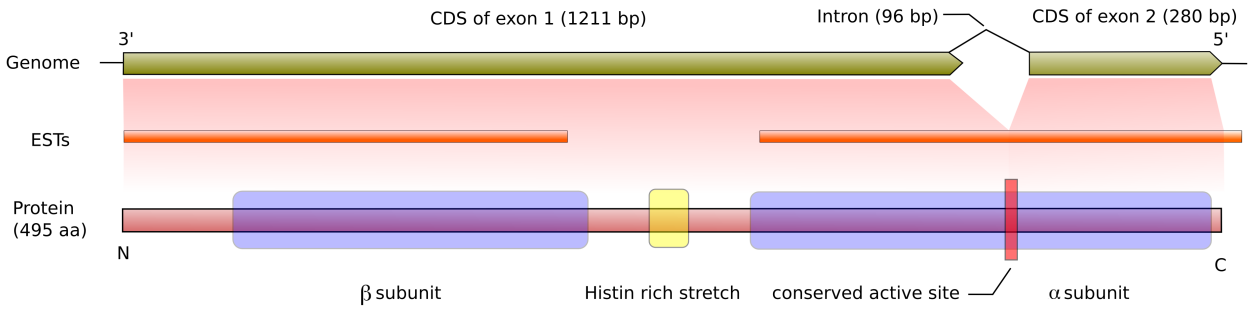


Figure 2: Scheme of the genomic region, ESTs and the protein of the NHases in *Monosiga brevicollis.* The β subunit and the Histidin-rich stretch are located in the protein part coded by the CDS of exon 1 while the α subunit consist of coding parts of exon 1 and exon 2. The putative active site is pinpointed in the β subunit and its coding sequence contains an intron. The two ESTs confirm the expression of both subunits and prove the splicing of the intron.

Appendix G

Splicing factors stimulate polyadenylation via USEs at non-canonical 30 end formation signals

The EMBO Journal (2007), 1–12 | © 2007 European Molecular Biology Organization | All Rights Reserved 0261-4189/07
www.embojournal.org

THE
EMBO
JOURNAL

Splicing factors stimulate polyadenylation via USEs at non-canonical 3′ end formation signals

Sven Danckwardt[1,2], Isabelle Kaufmann[3,5], Marc Gentzel[4], Konrad U Foerstner[4], Anne-Susan Gantzert[1,2], Niels H Gehring[1,2], Gabriele Neu-Yilik[1,2], Peer Bork[4], Walter Keller[3], Matthias Wilm[4], Matthias W Hentze[1,4,*] and Andreas E Kulozik[1,2,*]

[1]Department of Pediatric Oncology, Hematology and Immunology, University of Heidelberg, Germany, [2]Molecular Medicine Partnership Unit, EMBL and University of Heidelberg, Heidelberg, Germany, [3]Biozentrum, University of Basel, Basel, Switzerland and [4]European Molecular Biology Laboratory, Heidelberg, Germany

The prothrombin (F2) 3′ end formation signal is highly susceptible to thrombophilia-associated gain-of-function mutations. In its unusual architecture, the F2 3′ UTR contains an upstream sequence element (USE) that compensates for weak activities of the non-canonical cleavage site and the downstream U-rich element. Here, we address the mechanism of USE function. We show that the F2 USE contains a highly conserved nonameric core sequence, which promotes 3′ end formation in a position- and sequence-dependent manner. We identify proteins that specifically interact with the USE, and demonstrate their function as *trans*-acting factors that promote 3′ end formation. Interestingly, these include the splicing factors U2AF35, U2AF65 and hnRNPI. We show that these splicing factors not only modulate 3′ end formation via the USEs contained in the F2 and the complement C2 mRNAs, but also in the biocomputationally identified BCL2L2, IVNS and ACTR mRNAs, suggesting a broader functional role. These data uncover a novel mechanism that functionally links the splicing and 3′ end formation machineries of multiple cellular mRNAs in an USE-dependent manner.

The EMBO Journal advance online publication, 26 April 2007; doi:10.1038/sj.emboj.7601699
Subject Categories: RNA
Keywords: F2; mRNA processing; polyadenylation; splicing factor; upstream sequence element (USE)

*Corresponding authors. AE Kulozik, Department of Pediatric Oncology, Hematology and Immunology, University of Heidelberg, Im Neuenheimer Feld 156, 69120 Heidelberg, Germany. Tel.: +49 6221 564555; Fax: +49 6221 564559; E-mail: andreas.kulozik@med.uni-heidelberg.de or MW Hentze, European Molecular Biology Laboratory, Meyerhof str. 1, 69117 Heidelberg, Germany. Tel.: +49 6221 387501; Fax: +49 6221 387518; E-mail: Matthias.Hentze@embl.de
[5]Present address: Sir William Dunn School of Pathology, University of Oxford, UK

Received: 17 November 2006; accepted: 2 April 2007

Introduction

With the exception of some histone mRNAs, all eukaryotic mRNAs possess poly(A)-tails at their 3′ end, which are produced by a two-step reaction involving endonucleolytic cleavage and subsequent poly(A) tail addition (Colgan and Manley, 1997; Keller and Minvielle-Sebastia, 1997; Zhao *et al*, 1999; Gilmartin, 2005). The specificity and efficiency of 3′ end processing is determined by the binding of a multiprotein complex to the 3′ end processing signal. Most cellular pre-mRNAs contain two core elements. The canonical polyadenylation signal AAUAAA upstream of the cleavage site is recognized by the multimeric cleavage and polyadenylation specificity factor (CPSF). This RNA–protein interaction determines the site of cleavage 10–30 nt downstream, preferentially immediately 3′ of a CA dinucleotide. The second canonical sequence element is characterized by a high density of G/U or U residues and is located up to 30 nt downstream of the cleavage site. This downstream sequence element (DSE) is bound by the 64 kDa subunit of the heterotrimeric cleavage stimulating factor (CstF) that promotes the efficiency of 3′ end processing. Additional proteins, cleavage factors I and II (CF I and CF II), associate and the pre-mRNA is cleaved by CPSF 73 (Ryan *et al*, 2004; Dominski *et al*, 2005; Mandel *et al*, 2006). Subsequently, poly(A) polymerase (PAP) adds ~250 A-nucleotides to the 3′ end in a template-independent manner. Finally, several molecules of the poly(A)-binding protein II (PABPN1) bind to the growing poly(A) tail and determine its length. These proteins remain attached to the poly(A) tail during nuclear export and enhance both, the stability and the translation of the mRNA (von der Haar *et al*, 2004). Therefore, defects of mRNA 3′ end formation can profoundly alter cell viability, growth and development by interfering with essential and well-coordinated cellular processes.

Although almost all pre-mRNAs are constitutively polyadenylated, alternative and regulated poly(A) site selection represents an important regulatory mechanism for spatial and temporal control of gene expression (Zhao and Manley, 1996; Edwalds-Gilbert *et al*, 1997; Barabino and Keller, 1999; Zhao *et al*, 1999). Some 49% of human mRNAs contain more than one polyadenylation site (Yan and Marr, 2005). Alternative and regulated 3′ end processing serves to direct important cellular processes such as immunoglobulin class switch (Takagaki *et al*, 1996) or the regulated expression of the transcription factor NF-ATc during T-cell differentiation (Chuvpilo *et al*, 1999).

The medical consequences of errors of 3′ end processing are exemplified by the molecular sequelae of a common prothrombotic mutation in the prothrombin (F2) mRNA (F2 20210G→A). This mutation affects the most 3′ nucleotide of the mature mRNA, where the pre-mRNA is endonucleolytically cleaved and polyadenylated; it reverts the physiologically inefficient F2 cleavage site into the most favorable CA dinucleotide context, increasing cleavage site recognition and

resulting in the accumulation of correctly 3′ end processed F2 mRNA in the cytoplasm. From these studies, enhanced mRNA 3′ end formation efficiency emerged as a novel molecular principle underlying pathological gene expression and explaining the role of F2 20210G→A in the pathogenesis of thrombophilia (Gehring *et al*, 2001).

Subsequent analyses of the F2 mRNA 3′ end revealed an unusual architecture of non-canonical 3′ end processing signals that explain the susceptibility of the F2 3′ UTR and 3′ flanking sequence to additional, clinically relevant gain-of-function mutations (Danckwardt *et al*, 2004, 2006a, b). The presence of a sequence element that is located upstream (upstream sequence element (USE)) of the cleavage site within the 3′ UTR stimulates F2 3′ end processing. Moreover, this 15-nucleotide spanning element is both necessary and sufficient to enhance 3′ end processing when inserted into a heterologous β-globin mRNA 3′ UTR (Danckwardt *et al*, 2004).

Unlike (retro-)viral RNAs (Gilmartin *et al*, 1995; Graveley *et al*, 1996), stimulatory USEs have been experimentally documented in only a few mammalian mRNAs such as the human complement C2 (Moreira *et al*, 1998), lamin B2 (Brackenridge and Proudfoot, 2000), cyclooxygenase-2 (Hall-Pogar *et al*, 2005) and the collagen genes (Natalizio *et al*, 2002). Biocomputational analyses now predict that USEs may represent a common and evolutionarily conserved feature of mammalian 3′ end formation signals (Legendre and Gautheret, 2003; Hu *et al*, 2005), suggesting a broad role of USEs in cellular 3′ end mRNA processing.

We systematically analyzed the F2 USE and determined its mechanism of function. We show that several splicing factors, CPSF and CstF components specifically bind to the highly conserved USE. The functional characterization of these RNA-binding proteins by RNAi reveals a specific stimulatory effect of known splicing factors on the 3′ end processing of the F2 and C2 USE-containing pre-mRNAs as well as the biocomputationally predicted targets BCL2L2, IVNS and ACTR mRNAs. We propose a model of USE-directed 3′ end processing that involves a novel mRNP that integrates different nuclear pre-mRNA processing steps. Our data also implicate USE-dependent RNP complex formation in the physiology of important cellular processes such as hemostasis (and other thrombin-dependent processes) and the regulation of C2 gene expression as a component of innate immunity.

Results

The F2 USE increases mRNA 3′ end processing efficiency in a position- and sequence-dependent manner

To systematically define the F2 USE and study its mechanism of function, we established an internally controlled *in vivo* 3′ end processing assay (Danckwardt *et al*, 2004) and generated constructs that contain a tandem array of 3′ end formation signals, with modifications of the F2 USE within the 5′ site (Figure 1A). In contrast, the unmodified downstream site consists of sequences originating from the wild-type F2 3′ UTR and its 3′ flanking sequences. Thus, the smaller mRNA isoform detected in the poly(A) test (PAT) analysis has been cleaved and polyadenylated at the 5′ site, whereas the longer isoform has been processed at the 3′ site. This experimental setting enabled us to directly compare the processing efficiency of the (manipulated) 5′ site in relation to the control 3′ site, providing an internal control for other potential variables such as transcription or splicing efficiency, which could influence the abundance of the mRNA encoded by the transfected constructs.

The *in vivo* assay carried out in transiently transfected HeLa cells (Figure 1B) indicates that the replacement of the entire USE (Unrel., lane 2) almost completely abolishes 3′ end formation at the affected 5′site, when compared to F2 WT (USE, lane 1). In contrast, partial replacement of the first, second or third nucleotide quintett of the USE motif by an unrelated sequence reduces the 3′ end formation capacity at the respective site by ∼2-fold (Figure 1B, lanes 3–5), although significant 3′ end formation was still observed.

Because of the critical spatial relationship of canonical 3′ end formation signals to each other, we next analyzed the positional requirements of the USE on mRNA expression and 3′ end formation. For this purpose, it is important to note that the 15-nucleotide spanning USE *per se* is sufficient to enhance 3′ end processing even in a heterologous 3′ UTR in a context-independent manner (Danckwardt *et al*, 2004). Displacing the USE, therefore, assays the positional requirements of USE function and is not expected to be compounded by a potential disruption of the surrounding mRNA architecture. The successive shift of the USE downstream towards the polyadenylation signal enhances 3′ end processing (Figure 1B, compare lane 1 with lanes 6 and 7). In contrast, shifting the USE further upstream (by 10, 20 and 30 nucleotides, respectively) resulted in a successive down-modulation of mRNA expression through loss of function of 3′ end processing (Figure 1B, compare lane 1 with lanes 8, 9 and 10). Furthermore, the (relative) changes of the efficiency of the 5′ poly(A) site upon modification (in the context of the tandem construct) were also reflected on the level of absolute mRNA abundance (Supplementary Figure S1), which indicates that the results of the PAT analysis as shown here are not compounded by the fact that the normal F2 3′ end processing site is $<100\%$ efficient. Thus, the position of the USE with respect to the canonical polyadenylation signals seems to be a quantitative determinant of its function in 3′ end processing.

Previously published data suggest that USEs might stimulate 3′ end processing, at least in part, by recruiting components of the canonical CstF complex (Moreira *et al*, 1998), which, under normal circumstances, critically depends on the density of U residues. We therefore analyzed whether the F2 USE activity depends on its uridine (U) content or on a more specific sequence context. To this end, we tested constructs with increasing number of U residues within the USE core region (Figure 1C). While decreasing the number of U residues caused a gradual reduction of the 3′ end processing efficiency (Figure 1C, lanes 1–7 and lane 9), increasing the number of U residues also reduced the 3′ end formation efficiency (Figure 1C, lanes 10 and 11), eventually even ablating 3′ end maturation completely (lane 12). Furthermore, the USE of the L3 mRNA that is bound by hFip1 (Kaufmann *et al*, 2004) was less efficient as the wild-type F2 USE (Figure 1C, lanes 8 and 9). However, duplicating the wild-type USE motif had a stimulatory impact on 3′ end formation by ∼2-fold (Figure 1C, lane 13). These effects seem to be independent of a specific cell type, as similar results were obtained both in transfected HUH-7 and HeLa cells (not shown).

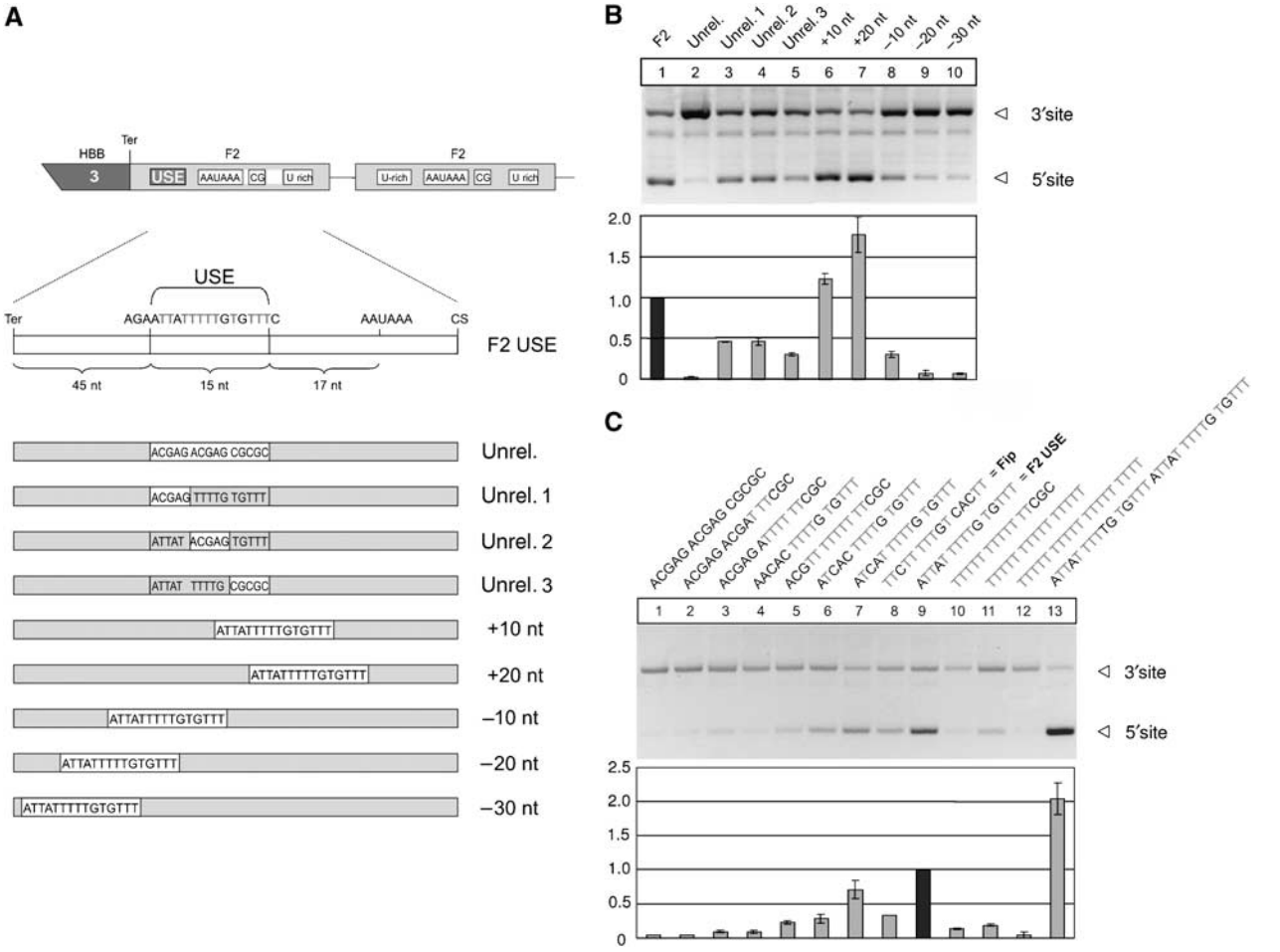


Figure 1 The F2 USE stimulates mRNA expression and 3′ end formation in a position- and sequence-dependent manner. (**A**) Schematic representation of the HBB (human β-globin)—F2 hybrid gene construct with a tandem array of two F2 3′ end formation signals used in transient transfection experiments (Ter: stop codon). The F2 USE was either completely or partially replaced by an unrelated nucleotide sequence, or displaced downstream and upstream of its original position as depicted. (**B**) *In vivo* assay carried out by transient transfection of a HBB-F2 hybrid gene construct with modifications of the USE as depicted in (A). The bar diagram shows the fold difference of the mRNA ratio processed at the 5′ or the 3′ site (5′/3′) relative to the F2 WT construct (highlighted) ± s.e. (at least four independent experiments). (**C**) *In vivo* assay carried out by transient transfection of a HBB-F2 hybrid gene construct with modifications of the USE as depicted. The bar diagram shows the fold difference of the mRNA ratio processed at the 5′ or the 3′ site (5′/3′) relative to the F2 WT construct (highlighted) ± s.e. (at least four independent experiments).

These results show that USE function is sequence and position sensitive, and that its potency is not simply determined by its U content. Because CstF binding at U-rich DSEs depends on the density of U-residues, these findings suggest that the F2 USE plays a specific role and does not simply compensate for the absent DSE in the F2 pre-mRNA. In this respect, the F2 mRNA appears to differ from the otherwise similar C2 mRNA (Moreira *et al*, 1998).

Finally, a sequence alignment revealed that the F2 USE is highly conserved among higher eukaryotes and is located at similar positions 17–22 nucleotides upstream of the poly(A) signal (Figure 2A). It comprises two highly conserved overlapping 3′ UTR motifs (UAUUUUU and UUUUGU) belonging to the top 10 out of 106 highly conserved 3′ UTR motifs, with a cross-species conservation rate of 30 and 24%, respectively (Xie *et al*, 2005). Interestingly, the disruption of either motif individually and/or the presence of only one motif highly correlated with loss of function (Figure 1B and C; Supplementary Figure S2, and data not shown), which emphasizes the importance of both sequence elements. It seems likely, therefore, that the F2 USE has evolved as an optimal sequence context that includes a nonameric core sequence (Figure 2A) in a functionally important region up to 40 nucleotides upstream of the poly(A) site (previously designated as core upstream element (CUE); Hu *et al*, 2005) to promote 3′ end processing. It is noteworthy that the USE as identified here does not include the tetramers UGUA and UAUA that have recently been shown to account for 3′ end formation at another non-canonical poly(A) site by recruiting the human 3′ processing factor CFIm (Venkataraman *et al*, 2005).

We next analyzed if other mRNAs that contain the nonameric USE core sequence can be identified. By using a sequence search algorithm that takes into consideration both the strand specificity and the typical length distribution for 3′ UTR motifs (peak >8-mers after exclusion of miRNAs target sites; Xie *et al*, 2005), we identified more than 1500 human transcripts that contain the nonameric USE core sequence (Figure 2B). Remarkably, a considerable amount of positive hits were identified in human transcripts with

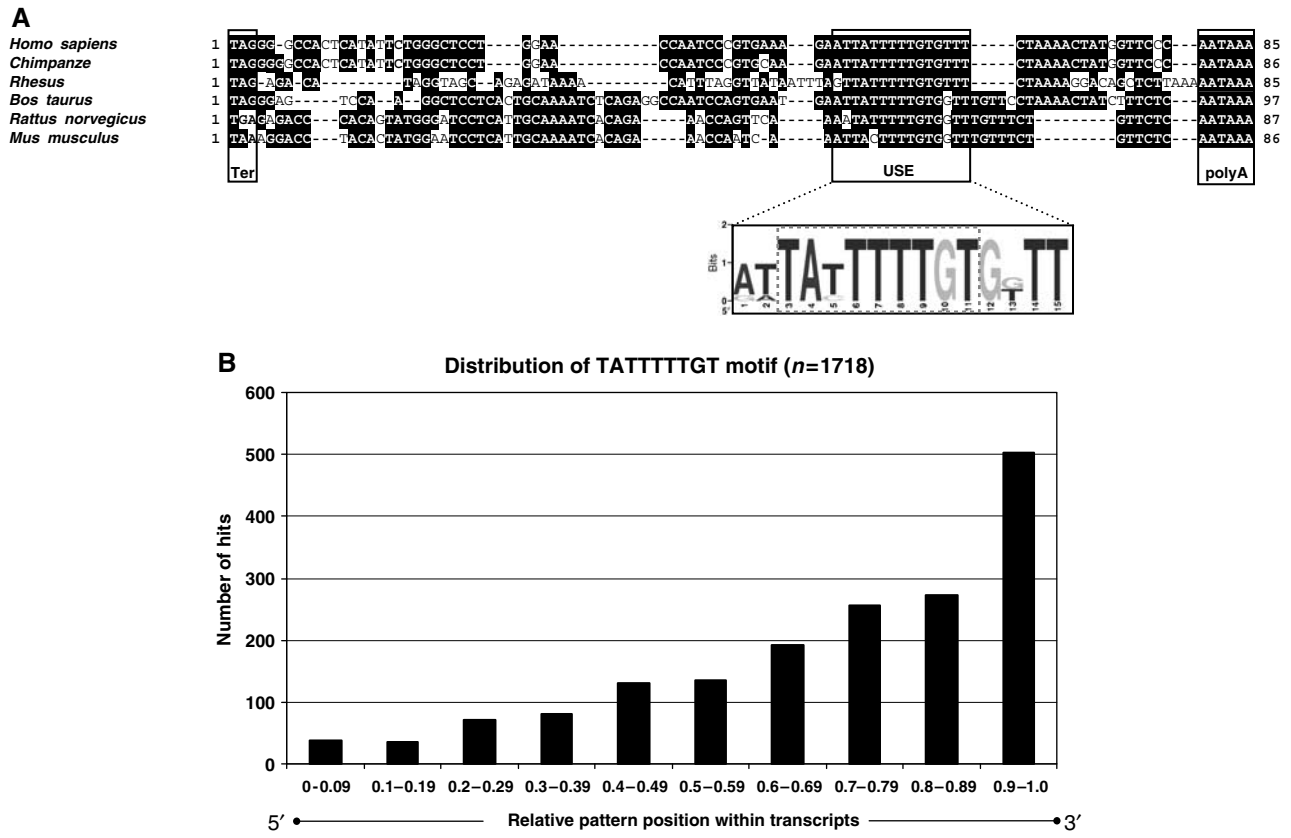


Figure 2 The F2 USE is highly conserved among higher eukaryotes. (**A**) Sequence comparison of the 3′ ends of vertebrate F2 genes (encompassing the entire 3′ UTR until the poly(A) signal). Shaded sequences denote identity. The graphical representation of the nucleic acid multiple sequence alignment (shown below) highlights the sequence conservation of the F2 USE according to the WebLogo 3 algorithm (Materials and methods), which contains a composite of two highly conserved sequence motifs (TATTTTTGT, highlighted; Xie *et al*, 2005). (**B**) Applying a sequence search algorithm that takes into consideration both the strand specificity and the typical length distribution for 3′ UTR motifs (peak >8-mers after exclusion of miRNAs target sites; Xie *et al*, 2005), more than 1700 human transcripts were identified to contain the nonameric USE core sequence motif. Number of hits are shown according to the localization of the sequence element within the mRNAs (bar diagram, *x*-axis, 5′ to 3′, left to right). Positive hits were filtered according to their relative location with respect to the poly(A) signal (AATAAA and ATTAAA, respectively). The identity of transcripts that contained the USE core sequence in close proximity to the poly(A) signal (<30 nucleotides upstream of the poly(A) signal (90 and 61 transcripts upstream of the AATAAA or ATTAAA, respectively)) is depicted in Supplementary Tables I and II.

unusually long 3′ UTRs (>1000 nucleotides, not shown). Filtering hits according to the localization of the sequence element within transcripts showed a polar distribution toward their 3′ end (Figure 2B), with more than 500 transcripts that contained the USE motif in the ultimate (tenth) part (0.9–1.0) in a 5′ to 3′ direction. Considering the critical spatial relationship of this sequence element for its function, we identified more than 150 human transcripts that contained the USE core sequence in close proximity to the poly(A) signal (less than 30 nucleotides upstream of the AATAAA and ATTAAA, respectively; see Supplementary Tables I and II). Interestingly, with the exception of four transcripts, all of them contained the USE core sequence motif in their 3′ UTRs. These finding suggest, therefore, that USE-dependent 3′ end processing plays a more general role in many transcripts.

Identification of specific nuclear USE-binding proteins

To identify *trans*-acting factors that specifically interact with the F2 USE to promote 3′ end processing, we next performed electromobility shift assays (EMSA) and UV crosslinking experiments. We used a ^{32}P-5′ end-labeled 21-mer RNA oligonucleotide probe including the 15 nucleotide USE core sequence that is both necessary and sufficient to promote 3′ end processing when inserted into a heterologous β-globin gene context (Danckwardt *et al*, 2004). Incubation of the USE probe with nuclear extract elicits a specific shift (Figure 3A, lanes 2–5 and 6–9). A 21-mer RNA oligonucleotide in which the USE core was replaced by a non-functional unrelated sequence fails to revert the observed shift (Unrel. comp cold, lane 10), whereas an RNA oligonucleotide containing the hFip1-binding site of the L3 mRNA (see above) competes for the formation of the shifted complex (Fip comp cold, lane 11), indicating that the hFip1-binding site interacts with at least one protein that is essential for the F2 USE gel shift. However, recombinant hFip1 failed to result in a shift of the USE oligonucleotide under physiological conditions, nor could it be identified as an interacting protein on the (entire) F2 3′ UTR by RNAse H protection analysis (not shown). No significant shift was observed when the USE probe was incubated with equal amounts of cytoplasmic extract (S100, lanes 12 and 13), indicating that at least one essential protein bound by the USE is nuclear.

We next investigated the USE-binding proteins by UV crosslinking (Figure 3B). The USE-specific 21-mer RNA was

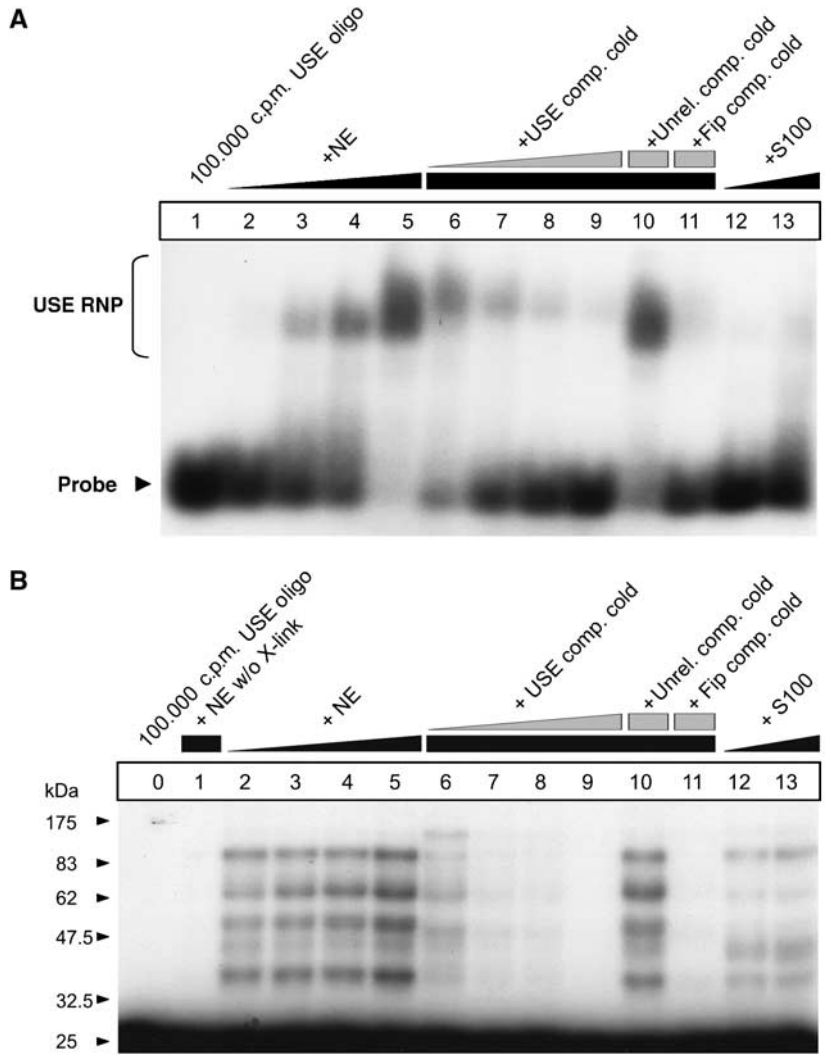


Figure 3 The F2 USE specifically interacts with nuclear proteins. (**A**) EMSA carried out with a F2 USE containing 21-mer RNA oligonucleotide (lane 1, free probe) after incubation in HeLa nuclear extract (NE, lanes 2–11) or cytoplasmatic extract (S100, lanes 12 and 13), respectively. Specificity of the RNA–protein interaction is shown by coincubation of an unlabeled F2 USE-containing 21-mer RNA oligonucleotide as cold competitor (lanes 6–9), of an unrelated competitor (lane 10) and a competitor, including the hFip1 binding site of the L3 RNA (lane 11). (**B**) UV crosslinking study carried out with the same USE-containing or competitor RNA oligonucleotides (lane 0, free probe) after incubation in HeLa nuclear extract (NE, lanes 1–11 (lane 1 without UV light exposure)) or cytoplasmatic extract (S100, lanes 12 and 13), respectively.

specifically UV crosslinked to at least five proteins of ~30–100 kDa. These crosslinks can be competed by cold USE and hFip1-binding site-specific 21-mers (lanes 6–9 and lane 11). Crosslinking studies with cytoplasmic extracts (lanes 12 and 13) showed that some of the USE-binding proteins also appear to be present in the cytoplasm, but the overall pattern of crosslinks is distinct from that generated with nuclear extracts. The affinity of the interaction between the USE and the crosslinking cytoplasmic proteins, however, does not appear to be sufficient to cause a shift in the EMSA. These results demonstrate that the F2 USE directly interacts with at least five different proteins that are predominantly located in the nuclear compartment. Furthermore, the functional significance of this interaction is highlighted by RNA–protein interaction studies using a template with a triple point mutation within the 15-nt USE core affecting the highly conserved nonamer (USEmut). This manipulation results in loss of protein binding (Supplementary Figure S2D and E), which highly correlates with loss of function of the USE (Supplementary Figure S2B, USE and USEmut; lanes 3 and 5).

Splicing factors and 3′ end processing proteins bind to the USE

We next aimed to identify the F2 USE-binding proteins by affinity purification followed by mass spectrometry. For this purpose, we first ascertained that the 3′biotin-TEG (triethylenglycol)-linker-modification used for immobilization of the RNA bait does not interfere with protein binding to the short 21-mer RNA oligonucleotides (Supplementary Figure S2A). As a specificity control, we used a template with a triple point mutation within the 15-nt USE core (USEmut), which results in similar loss of function as the replacement of the entire USE does (Supplementary Figure S2B, USE, USEmut and Unrel.; lanes 2, 3 and 5). RNA–protein interaction studies based on EMSA and UV crosslinking revealed that this loss of function correlates highly with loss of protein binding (Supplementary Figure S2D and E). The point mutated 21-mer USE sequence (USEmut) thus qualifies as an excellent specificity control for non-functional RNA–protein interactions during the affinity purification procedure. As an additional control for nonspecific RNA–protein interactions, an

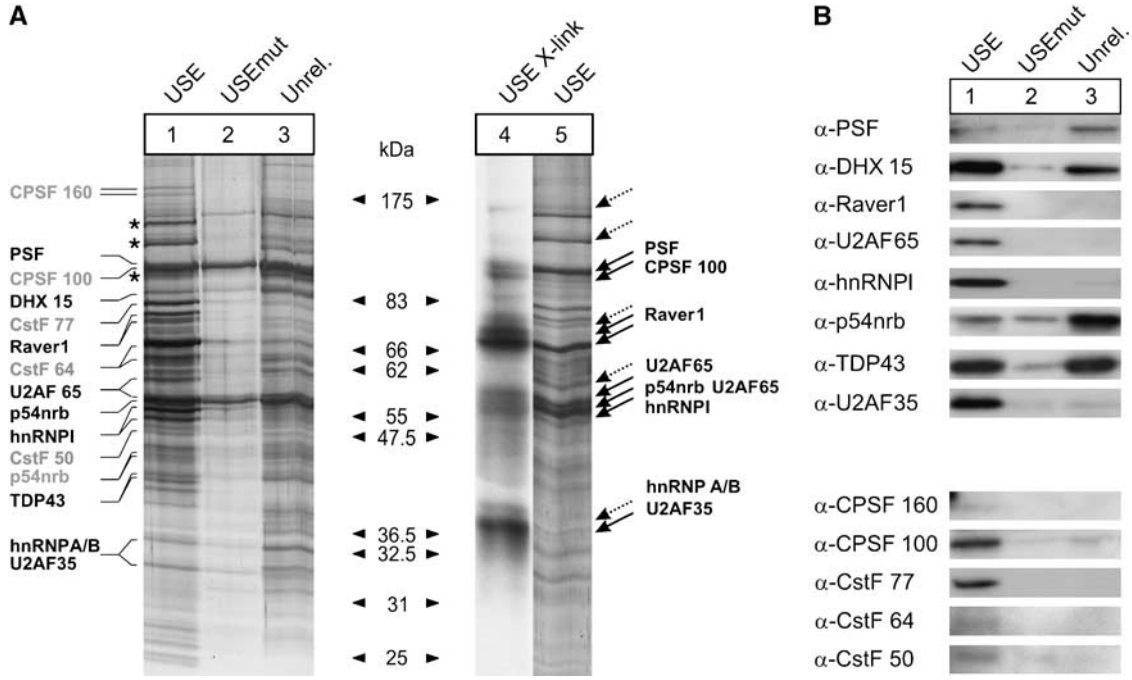


Figure 4 Affinity purification followed by mass spectroscopy identifies proteins that specifically interact with the F2 USE. (**A**) Silver-stained SDS–PAGE polyacrylamide gel of protein samples derived from affinity purification with immobilized 3′ biotinylated 21-mer RNA oligonucleotides with the F2 USE motif (USE, lanes 1 and 5), with a mutated F2 USE motif (USEmut, lane 2) or with an unrelated sequence context (Unrel., lane 3). Lanes 1–3 show protein samples eluted with up to 2000 mM NaCl (starting concentration 150 mM NaCl). Lane 4 shows the autoradiograph of a UV crosslink that highlights the size of direct interaction partners of the F2 USE (dotted arrows indicate putative direct USE-binding proteins that could not be unequivocally assigned by the mass spectrometry data) in comparison to the band pattern of directly and indirectly interacting proteins derived from affinity purification (lane 1). Silver-stained bands in lane 1 were cut out and subjected to mass spectrometry (protein names are only indicated at the respective size where the respective peptide score was maximal; canonical 3′ end processing factors are highlighted in yellow; p54nrb peptides of unexpected small size are highlighted in light gray; for mass spectrometry data, see also Table I). The analysis also revealed the presence of ATP citrate synthase, LRP 130, HSP 90- and tubulin (asterisks), which were judged to represent likely contaminants and were not included in Table I. (**B**) Immunoblots of eluates from affinity purifications for proteins identified by mass spectrometry (Table I). Lanes 1–3 correspond to samples as indicated in Figure 4A. Additional information on controls for unspecific RNA–protein interaction and equal loading is available in Supplementary Figure S2.

immobilized 21-mer RNA oligonucleotide with an unrelated sequence was used (Supplementary Figure S2C).

Affinity purification yielded several bands that were specific for the USE bait compared with the controls (Figure 4A, lanes 1–3). Comparison of the USE affinity purification-specific band pattern with the patterns of UV crosslinking experiments revealed that the size of some of the affinity-purified proteins corresponds to the size of the proteins identified by UV crosslinking; these proteins thus likely interact with the USE motif directly (Figure 4A, lanes 4 and 5, that is, PSF, p54nrb, U2AF65, hnRNPI, UFAF35). This analysis revealed that the F2 USE-binding proteins include factors known to be involved in 3′ end processing and, surprisingly, in splicing (Table I).

We next confirmed the identity of the proteins found by mass spectrometry. Immunoblots of eluates derived from affinity purification with the USE bait and the respective controls (Figure 4B; Table I) show that six out of 13 proteins identified by mass spectrometry were specifically present in the eluates of the USE columns. For three proteins (CPSF 160, CstF 64 and CstF 50) the signal is weak, which may indicate the low abundance of these proteins in the eluates or reflect a lower affinity of the antibodies. Four of the proteins (PSF, DHX 15, p54nrb, TDP43) are also present in the eluates of the USEunrel and/or USEmut controls, which indicates nonspecific RNA-binding and/or indirect RNA-interaction properties via yet unidentified proteins bound to the RNA baits in lanes 1, 2 and 3. Taken together, these results reveal that the USE interacts specifically with two predominant classes of proteins with known roles in splicing and 3′ end processing (Table I).

RNAi demonstrates a role of splicing factors in USE-dependent F2 3′ end processing

We next analyzed the functional importance of the identified proteins on USE-dependent 3′ end processing. We first established the siRNA-mediated, target-specific depletion of the splicing factors shown in the upper panel of Figure 4B, and subsequently performed an *in vivo* 3′ end formation assay by transfecting suitable constructs that contain a tandem array of 3′ end formation signals, with and without an F2 USE within the 5′ site (Figure 5B).

Functional RNAi resulted in efficient protein depletion of each target protein to below 25% of control levels (Figure 5A, each panel, lanes 1 and 2–5). Importantly, the protein abundance of the other seven proteins was unaffected by the target-specific depletions (not shown).

Expectedly, the analysis of transfected F2 mRNA reporter abundance revealed a significant upmodulation of 3′ end processing at the 5′ site in the presence of a functional F2 USE of ~7.6-fold, when compared with the respective mRNA counterpart derived from constructs without an USE

Table I The F2 USE specifically interacts with splicing factors, 3′ end processing proteins and other RNA binding proteins

	Characteristics of USE-binding proteins	Accession number	Score max	Peptides	Immunoblot
Splicing factors					
PSF/SFPQ	Splicing factor, proline and glutamine rich	P23246	547	15	USE = Unrel.
p54nrb/NonO	Non-POU domain-containing octamer-binding protein	Q15233	414	15	USE < Unrel.
hnRNPI/PTB	Polypyrimidine tract-binding protein 1	P26599	262	8	USE
DHX 15	Putative pre-mRNA splicing factor RNA helicase (DEAH box protein 15)	O43143	190	7	USE = Unrel.
TDP43	TAR DNA-binding protein 43	Q13148	161	3	USE = Unrel.
DDX5 p68	Probable RNA-dependent helicase p68 (DEAD/H box 5)	P17844	84	2	Very weak
SF1	Splicing factor 1 (zinc-finger protein 162) (transcription factor ZFM1)	Q15637	78	5	Not detectable
U2AF65	Splicing factor U2AF 65-kDa subunit	P26368	74	2	USE
U2AF35	Splicing factor U2AF 35-kDa subunit	Q01081	50	1	USE
SF3B	Splicing factor 3B subunit 3	Q15393	36	2	Not determined
3′ End processing					
CstF 50	Cleavage stimulation factor, 50-kDa subunit	Q05048	182	8	USE
CPSF 160	Cleavage and polyadenylation specificity factor, 160-kDa subunit	Q10570	168	8	USE
CstF 77	Cleavage stimulation factor subunit 3	gi|4557495	134	6	USE
CstF 64	Cleavage stimulation factor, 64-kDa subunit	P33240	102	3	USE
CPSF 100	Cleavage and polyadenylation specificity factor, 100-kDa subunit	Q9P2I0	89	3	USE
Other					
Similar to Raver1?	Unknown (protein for IMAGE:5113697)	gi|22902182	91	2	USE
hnRNP C1/C2	Heterogeneous nuclear ribonucleoproteins C1/C2	P07910	83	2	USE = USEmut = Unrel.
hnRNP A2/B1	Heterogeneous nuclear ribonucleoproteins A2/B1	P22626	50	2	Not detectable
Nuclear DNA helicase II	ATP-dependent RNA helicase A	Q08211	43	2	Not determined

USE-binding proteins identified by affinity purification and subsequent mass spectrometry of USE-specific bands shown in Figure 4A. The analysis also revealed the presence of ATP citrate synthase, LRP 130, HSP 90-alpha and tubulin (Figure 4A, lane 1, black asterisks), which were judged to represent likely contaminants and were not included in the table. Data interpretation was performed with the MASCOT V2.1 search engine for mass spectrometric data (see Materials and methods section). The maximal scores obtained for each individual protein and the number of aligned peptides are shown. The immunoblot data shown in Figure 4B are summarized.

(Figure 5C, lanes 17 and 18). This USE-dependent stimulatory effect on 3′ end processing was slightly reduced in cells upon depletion of PSF and p54nrb (to 3.3- and 2.1-fold, lanes 7 and 8, and 9 and 10), and almost completely abolished after depletion of the splicing factors hnRNPI, U2AF35 and U2AF65 (lanes 11 and 12, 13 and 14, 15 and 16). In contrast, depletion of Raver1, DHX 15 and TDP did not reduce USE-mediated 3′ end processing (lanes 1–6). These data indicate that USE function on F2 3′ end processing critically depends on the splicing factors hnRNPI, U2AF35 and U2AF65. It should be noted that we did not observe a significant proportion of unspliced reporter mRNAs upon depletion of these splicing factors in the PAT assay. Efficient depletion of PSF, U2AF35 and U2AF65, however, strongly affected cell morphology and plasmid transfection efficiencies.

hnRNPI and U2AF65 interact with the F2 USE directly

To identify the functionally relevant splicing factors that directly interact with the F2 mRNA *in vivo*, we next performed an RNP immunoprecipitation (IP) assay and monitored the endogenous F2 mRNA contained in the IPs. For this purpose, IPs were carried out with cell lysates after UV or formaldehyde (FA) crosslinking to specifically assay for direct RNA–protein interactions (Niranjanakumari *et al*, 2002), with antibodies directed against hnRNPI, U2AF35, p54nrb, U2AF65 and PSF. Nonspecific association of mRNAs with IP reagents was controlled by parallel incubations with anti-mouse antibodies (Figure 6A).

In IPs carried out with antibodies directed against hnRNPI and U2AF65, the endogenous F2 mRNA was specifically enriched in samples derived from cells after UV and FA crosslinking (Figure 6A, lanes 2 and 5, 8 and 11), whereas the F2 mRNA could not be detected in IPs with other antibodies or in IPs with cell lysates that were not cross-linked, respectively. These results thus indicate that hnRNPI and U2AF65 interact with the F2 mRNA directly. In contrast, U2AF35, p54nrb and PSF either do not directly interact with the F2 mRNA or a direct interaction is masked or otherwise undetectable.

We next investigated whether hnRNPI and U2AF65 interact with the F2 mRNA in a USE-dependent manner (Figure 6B). For this purpose, we extended the *in vivo* RNA–protein interaction study to cells transfected with reporter constructs either with or without a functional F2 USE (compare Supplementary Figure S2B, USE, USEmut and Unrel.), followed by assaying the FA-crosslinked reporter-specific mRNA in the IP material. In order to compensate for the ~5-fold difference of mRNA expression that depends on the functionality/presence of the USE (Figure 1B and data not shown), the amount of transfected reporter plasmid DNA was adjusted accordingly.

As shown in the left panel in Figure 6B, the USE-containing reporter mRNA was specifically detected in IPs carried out with antibodies directed against hnRNPI and U2AF65 (lanes 2 and 5). In contrast, reporter mRNAs with a non-functional USE (USEmut, middle panel) or without a USE (Unrel., right panel) were not detectable in IPs carried out with lysates of cells transfected with the USEmut or Unrel. constructs, respectively. The intact pyrimidine-rich F2 USE thus represents a direct binding site for hnRNPI and U2AF65 (Singh *et al*, 1995).

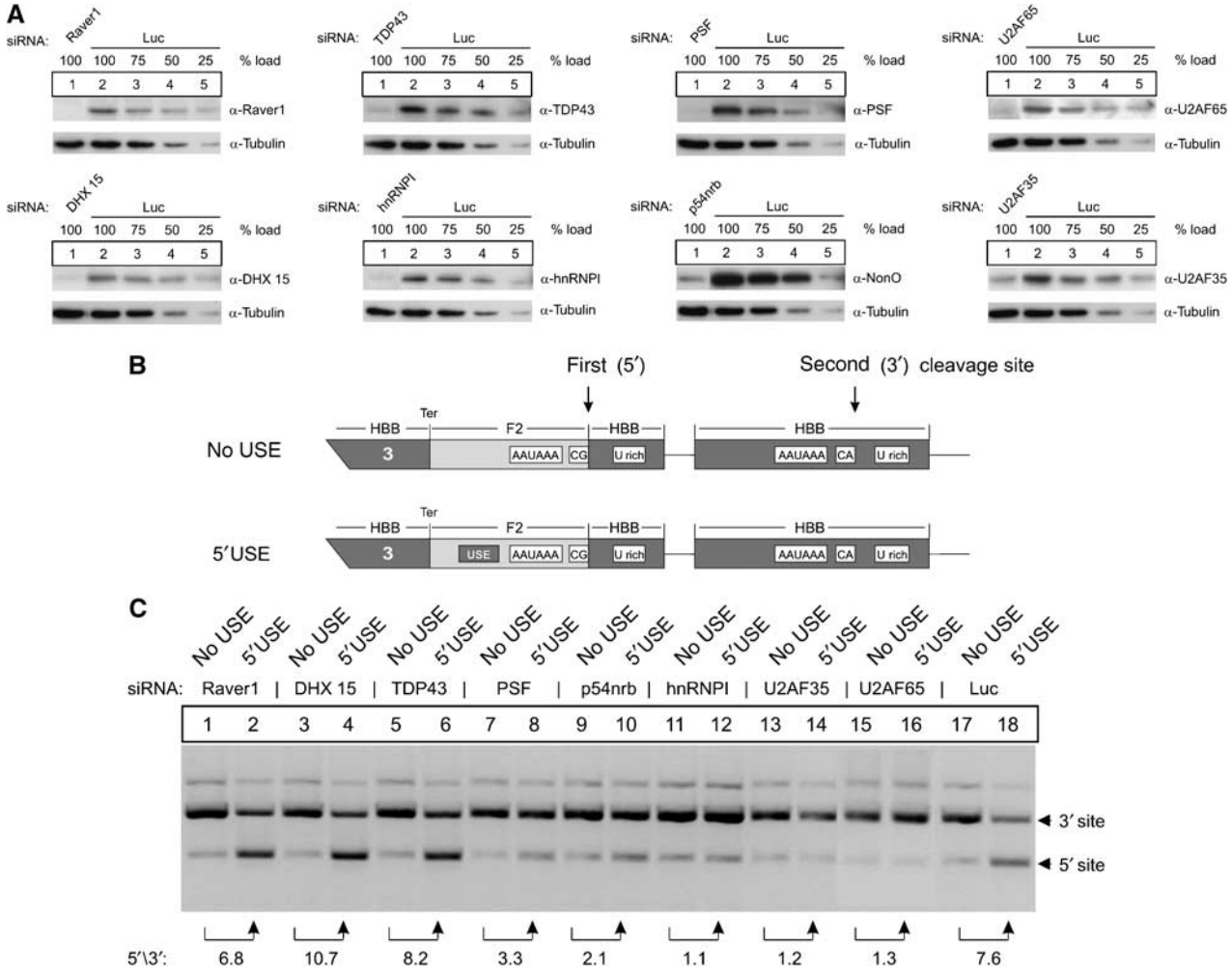


Figure 5 USE-binding splicing factors specifically promote the expression of USE-containing mRNAs. (**A**) Representative immunoblots of protein lysates obtained from cells treated with siRNAs directed against USE-binding proteins (lane 1, each panel). Lanes 2–5 in each panel show a serial reduction of the load of protein lysates obtained from cells treated with a luciferase siRNAs as control for nonspecific RNAi effects and for quantification of the RNAi efficiency. (**B**) Schematic representation of the HBB-F2 hybrid gene construct with a tandem array of two 3′ end formation signals. The F2 USE of the 5′ located 3′ end processing signal was either maintained ('5′USE' construct) or completely replaced by an unrelated nucleotide sequence ('no USE' construct). (**C**) *In vivo* assay carried out by transient transfection of a HBB-F2 hybrid gene construct, with or without a USE (B), after depletion of USE-binding proteins (A). Each number represents the fold difference of the mRNA ratio processed at the 5′ or the 3′ site (5′/3′) relative to the respective ratio in the odd numbered lanes (representative figure of three independent experiments).

RNAi demonstrates a role of splicing factors in USE-dependent mRNA expression of several endogenous transcripts

We next analyzed the functional importance of the identified proteins in USE-dependent mRNA expression of endogenous transcripts. For this purpose, we first established the siRNA-mediated, target-specific depletion of the splicing factors in HUH-7 cells and monitored endogenous mRNA abundance by RT–PCR (Figure 6C). The quantification of endogenous F2 mRNA abundance revealed a significant down-modulation to approximately 80% upon depletion of TDP43 and PSF. The quantitatively most profound reduction (to below 60%) resulted from depletion of the splicing factors p54nrb, U2AF35, hnRNPI and U2AF65 (Figure 6C, quantified after normalization against the ACTB mRNA abundance, which lacks the nonameric USE core sequence motif). Finally, we analyzed whether the functional effects on F2 mRNA abundance could be extended to other USE-containing mRNAs. For this analysis, we selected the USE core sequence-containing BCL2L2, IVNS1ABP and ACTR3B mRNAs (see Supplementary Tables I and II), and the C2 mRNA that has previously been shown to be processed in an USE- and hnRNPI-dependent manner (Moreira *et al*, 1998). Whereas depletion of U2AF35 resulted in a down-modulation of the F2, IVNS1ABP and C2 mRNAs, depletion of hnRNPI and U2AF65 strikingly affected all tested USE-containing mRNAs (Figure 6C, compare green and yellow bars).

In order to demonstrate that the functional effects of RNAi of the USE-binding proteins was specific for the USE core sequence-containing genes, we also monitored the expression of actin (ACTG1), the hypoxanthine guanine phosphoribosyltransferase 1 (HPRT1) and the polyomavirus enhancer-binding protein 2 (CBFB) mRNAs as representative examples of spliced mRNAs with a conventional 3′ end formation signal. Furthermore, we analyzed the expression of the mitogen-activated protein kinase kinase kinase 1 (MAP3K1) that contains the nonameric USE core sequence within the ORF

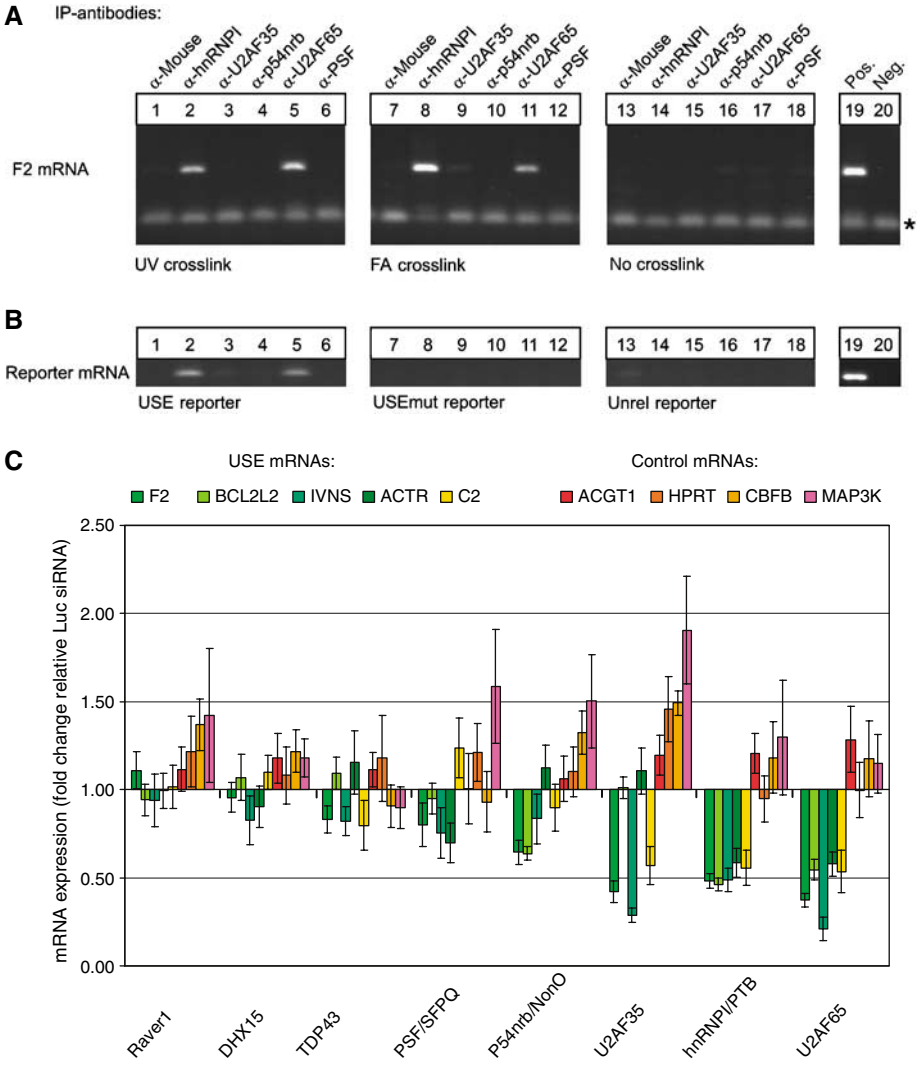


Figure 6 Depletion of the USE-binding protein hnRNPI and U2AF65 specifically reduces the mRNA expression of USE-containing endogenous transcripts. (**A**) *In vivo* RNA–protein interaction assay based on mRNA analyses in RNP IPs carried out with antibodies as indicated. IPs were carried out with HUH-7 cell lysates after UV crosslinking (lanes 1–6), FA crosslinking (lanes 7–12) or without crosslinking (lanes 13–18; for further information see Materials and methods). The F2 mRNA was analyzed by RT–PCR (primer dimers are indicated by asterisks); lanes 19 and 20 represent positive and negative controls, respectively. (**B**) *In vivo* RNA–protein interaction assay based on reporter mRNA analysis in IP samples derived from FA crosslinked HeLa cells transfected with reporter constructs containing either a F2 USE (USE reporter), a mutated F2 USE (USEmut reporter) or an unrelated sequence context (Unrel. reporter, see Supplementary Figure S2). (**C**) Endogenous F2, BCL2L2, IVNS, ACTR and C2 mRNA, and ACTG1, HPRT, CBFB and MAP3K mRNA abundance of HUH-7 cells after RNAi directed against USE-binding proteins as indicated (*x*-axis). The fold change of mRNA expression upon candidate siRNA treatment (*y*-axis) is quantified relative to the mRNA expression of cells treated with luciferase siRNAs after normalization against endogenous ACTB mRNA. Each bar represents values of at least five independent RNAi experiments determined by quantitative RT–PCR in duplicates (±s.e.).

far upstream of a potential downstream poly(A) signal: The quantification of the ACTG1, HPRT1, CBFB and MAP3K1 mRNAs shows that these controls are not down-modulated upon depletion of the USE-binding proteins (Figure 6C, red bars). Thus, the depletion of the USE-binding splicing factors hnRNPI, U2AF65 and—in part also U2AF35—reduces the expression of the nonameric USE core sequence-containing F2, BCL2L2, IVNS1ABP and ACTR3B mRNAs, whereas the USE core sequence-containing MAP3K1 mRNA that lacks a downstream poly(A) signal in close proximity was unaffected by these manipulations.

These results thus recapitulate the positional effect of USE function (Figure 1) and highlight the specific stimulatory effect of these splicing factors on USE-mediated mRNA expression for several endogenous transcripts.

Discussion

USE-dependent 3′ end processing has been experimentally documented for a number of genes that are involved in important physiological processes such as hemostasis (prothrombin; Danckwardt *et al*, 2004), innate immunity (complement C2; Moreira *et al*, 1998), inflammation (cyclooxygenase-2; Hall-Pogar *et al*, 2005) and in the maintenance of cell structure (lamin B2; Brackenridge and Proudfoot, 2000) and collagen (Natalizio *et al*, 2002). Biocomputational analyses predict that USE-dependent 3′ end processing may be quite common among cellular mRNAs (Legendre and Gautheret, 2003; Hu *et al*, 2005). USEs thus represent one of the important regulatory sequence elements contained in 3′ UTRs.

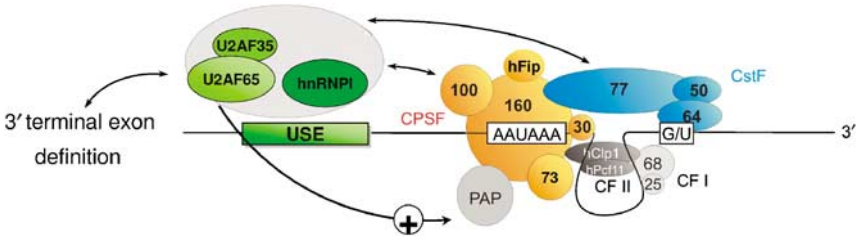


Figure 7 Model for USE-dependent RNA processing at (non-canonical) 3′ end formation signals. 3′ End processing of USE-containing mRNAs is proposed to depend on the formation of USE-dependent RNP complexes that participate in an extensive molecular network coordinating gene expression. USE-binding splicing factors are depicted as positive effectors (green complex), potentially involving a direct stimulation of PAP via U2AF65. The USE-binding protein complex (Table I) is indicated by gray shading. The USE-binding proteins are proposed to bridge and promote splicing and 3′ end processing. Specifically, the 65 kDa subunit of the U2AF dimer may link 3′terminal exon definition with efficient 3′ end processing. On the other hand, the USE-dependent RNP complexes establish the link to the 3′ end processing machinery by interaction with CPSF 100 and CstF 77. The polypyrimidine tract-binding protein PTB/hnRNPI and U2AF65 likely represent direct USE-interacting proteins and may thus nucleate the USE-dependent RNP complexes for efficient stimulation of 3′ end formation.

Sequence comparisons of the entire USE did not reveal a clear consensus in other genes, including those with functionally defined USEs. However, when recently identified conserved 3′ UTR sequence motifs (Hu *et al*, 2005; Xie *et al*, 2005) were considered, it became apparent that the F2 USE includes two overlapping motifs (UAUUUUU and UUUUGU) belonging to the top 10 out of 106 highly conserved 3′ UTR motifs with a cross-species conservation rate of 30 and 24%, respectively (Xie *et al*, 2005). Of note, the conservation rate of the highly conserved polyadenylation signal is 46%, whereas control random motifs show a conservation rate of only 10%. Importantly, the UAUUUUU motif is destroyed in the non-functional USE motif (USEmut), which highly correlates with loss of protein binding (Supplementary Figure S2). However, the presence of the UAUUUUU motif alone yielded only moderate 3′ end processing efficiencies, whereas full activity was observed in the presence of both elements (Figure 1). This indicates that the F2 USE has evolved as an optimal sequence context consisting of a composite of two highly conserved 3′ UTR motifs.

Interestingly, the complement C2 mRNA contains a suboptimal match of another conserved top 10 hexamers (UGUUUU; Hu *et al*, 2005), which can also be found at the 3′ end of the F2 USE. The biocomputational predictions and the functional analyses reported here thus implicate that different, highly conserved U-rich (Legendre and Gautheret, 2003) 3′ UTR sequence motifs can function as USEs and enhance 3′ end processing in a position-dependent manner, showing highest activities when located in a region recently designated as core upstream element (CUE; Hu *et al*, 2005). In that respect, the sequence element identified here differs from tetrameric sequence elements that have recently been identified to account for 3′ end processing at the non-canonical poly(A) site of the PAPOLA and PAPOLG mRNA by recruiting CFIm (Venkataraman *et al*, 2005).

Different steps of gene expression pathways are thought to be coupled (Hirose and Manley, 2000; Proudfoot *et al*, 2002). In this context, the important finding reported here is that two different classes of RNA processing proteins bind to the F2 USE (Table I) and thus provide further evidence for an extensive molecular network that effectively coordinates gene expression (Maniatis and Reed, 2002). In line with the notion that processing factors involved in pre-mRNA splicing and 3′ end formation can influence each other positively (Niwa *et al*, 1990; Wassarman and Steitz, 1993; Gunderson *et al*, 1994; Lutz *et al*, 1996; Vagner *et al*, 2000; Li *et al*, 2001; McCracken *et al*, 2002, 2003; Millevoi *et al*, 2002, 2006; Awasthi and Alwine, 2003; Kyburz *et al*, 2006), we identify here that the splicing factors hnRNPI, U2AF65 and, in part, U2AF35, represent components of a functionally relevant USE-dependent RNP. The requirement of these splicing factors for 3′ end processing seemed to be USE specific, because the depletion of these factors did not influence the expression of those tested mRNAs that do not contain the nonameric USE core sequence motif. Furthermore, the USE stimulates polyadenylation (Supplementary Figure S3) and its function critically depends on a tight spatial relationship to the canonical 3′ end formation signals. Although it is formally possible that the USE and the 3′ terminal splice site may potentiate polyadenylation through common means, our findings indicate that the functional effect of the F2 USE does not result from a nonspecific coupling of splicing or 3′ end terminal exon definition and 3′ end processing.

Previously, p54nrb and PSF have been shown to interact with the carboxy-terminal domain (CTD) of RNA polymerase II to link transcriptional activities with splicing (Emili *et al*, 2002; Kameoka *et al*, 2004), whereas 3′ end processing has been suggested to be more indirectly affected by these proteins (Rosonina *et al*, 2005). p54nrb has recently been shown to be a component of the snRNP-free U1A (SF-A) complex that appears to directly promote pre-mRNA cleavage (Liang and Lutz, 2006). Moreover, hnRNPI, U2AF35, U2AF65 and PSF have previously been identified in CF II preparations (de Vries *et al*, 2000). With respect to our findings presented here, it is particularly interesting that U2AF65 has previously been shown to directly interact with the CTD of the PAP (Vagner *et al*, 2000) and cleavage factor CF I (Millevoi *et al*, 2006). U2AF65 has also been reported to stimulate the 3′ end cleavage reaction when tethered more than 150 nucleotides upstream of the AAUAAA hexanucleotide (Millevoi *et al*, 2002). However, this finding is unlikely to be related to the USE effect, because we show that the USE is virtually nonfunctional when shifted more than 50 nucleotides upstream of the AAUAAA (Figure 1).

The second class of proteins interacting with the USE are canonical 3′ end processing factors (Table I). Therefore, the USE-dependent RNP complex may promote 3′ end processing by serving as an additional anchor for the (canonical) 3′ end

processing machinery or by stabilizing the RNA interaction of both CPSF and CstF components (Brackenridge and Proudfoot, 2000). Importantly, cooperative binding has also been implicated to account for CstF binding to the DSE, which greatly enhances the affinity of CPSF to the AAUAAA hexamer and vice versa (Colgan and Manley, 1997; Zhao *et al*, 1999). We therefore propose the existence of a novel and complex RNP, most likely consisting of at least two components of two distinct complexes (U2AF65 of the heterodimeric complex U2AF35/U2AF65, and hnRNPI known to interact with p54nrb and PSF) that cooperatively promote 3′ end processing in a USE-dependent manner (Figure 7). However, it should be noted that USE-dependent 3′ end processing of some of the USE core sequence-containing transcripts analyzed here seems to require different cofactors (Figure 6C), opening the perspective of transcript specific regulation of 3′ end processing. Despite previous reports implying a more general function of p54nrb and PSF in 3′ end processing (Lutz *et al*, 1998; Liang and Lutz, 2006), these splicing factors are likely dispensable for a USE-specific function in 3′ end processing (Figures 4B and 6C).

Taken together, the data presented here functionally link the splicing and 3′ end processing machineries in a USE-dependent manner. It will be interesting to dissect the different cofactor requirements and to analyze whether this novel mechanism is subject to specific regulatory steps that may respond to external stimuli.

Materials and methods

Detailed information on Materials and methods is available in Supplementary data.

Supplementary data

Supplementary data are available at *The EMBO Journal* Online (http://www.embojournal.org).

Acknowledgements

We thank Margit Happich for excellent technical assistance, Pavel Ivanov, Stephen Breit, Marcelo Viegas and other members of the Molecular Medicine Partnership Unit for advice and helpful discussions. We also acknowledge Brigitte Jockusch for kindly providing the anti-Raver1 antibody. This work was funded by grants from the Deutsche Forschungsgemeinschaft (KU563/7-1 and KU563/8-1 to AEK), the Fritz-Thyssen Stiftung (grant 1999-1076 to AEK) and by the 'Young Investigator Award' fellowship from the University of Heidelberg (to SD). This work was supported by the DFG Forschergruppe (FOR 426): Complex RNA–protein interactions in the maturation and function of eukaryotic mRNA. Work in the laboratory of WK was supported by the University of Basel and the Swiss National Science Foundation.

References

Awasthi S, Alwine JC (2003) Association of polyadenylation cleavage factor I with U1 snRNP. *RNA* **9:** 1400–1409

Barabino SM, Keller W (1999) Last but not least: regulated poly(A) tail formation. *Cell* **99:** 9–11

Brackenridge S, Proudfoot NJ (2000) Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. *Mol Cell Biol* **20:** 2660–2669

Chuvpilo S, Zimmer M, Kerstan A, Glockner J, Avots A, Escher C, Fischer C, Inashkina I, Jankevics E, Berberich-Siebelt F, Schmitt E, Serfling E (1999) Alternative polyadenylation events contribute to the induction of NF-ATc in effector T cells. *Immunity* **10:** 261–269

Colgan DF, Manley JL (1997) Mechanism and regulation of mRNA polyadenylation. *Genes Dev* **11:** 2755–2766

Danckwardt S, Gehring NH, Neu-Yilik G, Hundsdoerfer P, Pforsich M, Frede U, Hentze MW, Kulozik AE (2004) The prothrombin 3′end formation signal reveals a unique architecture that is sensitive to thrombophilic gain-of-function mutations. *Blood* **104:** 428–435

Danckwardt S, Hartmann K, Gehring NH, Hentze MW, Kulozik AE (2006a) 3′ End processing of the prothrombin mRNA in thrombophilia. *Acta Haematol* **115:** 192–197

Danckwardt S, Hartmann K, Katz B, Hentze M, Levy Y, Eichele R, Deutsch V, Kulozik A, Ben-Tal O (2006b) The prothrombin 20209 C>T mutation in Jewish–Moroccan Caucasians: molecular analysis of gain-of-function of 3′end processing. *J Thromb Haemost* **4:** 1078–1085

de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W (2000) Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J* **19:** 5895–5904

Dominski Z, Yang XC, Marzluff WF (2005) The polyadenylation factor CPSF 73 is involved in histone-pre-mRNA processing. *Cell* **123:** 37–48

Edwalds-Gilbert G, Veraldi KL, Milcarek C (1997) Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Res* **25:** 2547–2561

Emili A, Shales M, McCracken S, Xie W, Tucker PW, Kobayashi R, Blencowe BJ, Ingles CJ (2002) Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA* **8:** 1102–1111

Gehring NH, Frede U, Neu-Yilik G, Hundsdoerfer P, Vetter B, Hentze MW, Kulozik AE (2001) Increased efficiency of mRNA 3′ end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nat Genet* **28:** 389–392

Gilmartin GM (2005) Eukaryotic mRNA 3′ processing: a common means to different ends. *Genes Dev* **19:** 2517–2521

Gilmartin GM, Fleming ES, Oetjen J, Graveley BR (1995) CPSF recognition of an HIV-1 mRNA 3′-processing enhancer: multiple sequence contacts involved in poly(A) site definition. *Genes Dev* **9:** 72–83

Graveley BR, Fleming ES, Gilmartin GM (1996) RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. *Mol Cell Biol* **16:** 4942–4951

Gunderson SI, Beyer K, Martin G, Keller W, Boelens WC, Mattaj LW (1994) The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase. *Cell* **76:** 531–541

Hall-Pogar T, Zhang H, Tian B, Lutz CS (2005) Alternative polyadenylation of cyclooxygenase-2. *Nucleic Acids Res* **33:** 2565–2579

Hirose Y, Manley JL (2000) RNA polymerase II and the integration of nuclear events. *Genes Dev* **14:** 1415–1429

Hu J, Lutz CS, Wilusz J, Tian B (2005) Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA* **11:** 1485–1493

Kameoka S, Duque P, Konarska MM (2004) p54(nrb) associates with the 5′ splice site within large transcription/splicing complexes. *EMBO J* **23:** 1782–1791

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W (2004) Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23:** 616–626

Keller W, Minvielle-Sebastia L (1997) A comparison of mammalian and yeast pre-mRNA 3′-end processing. *Curr Opin Cell Biol* **9:** 329–336

Kyburz A, Friedlein A, Langen H, Keller W (2006) Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3′ end processing and splicing. *Mol Cell* **23:** 195–205

Legendre M, Gautheret D (2003) Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4:** 7

Li Y, Chen ZY, Wang W, Baker CC, Krug RM (2001) The 3′-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAs *in vivo*. *RNA* **7:** 920–931

Liang S, Lutz CS (2006) p54nrb is a component of the snRNP-free U1A (SF-A) complex that promotes pre-mRNA cleavage during polyadenylation. *RNA* **12:** 111–121

Lutz CS, Cooke C, O'Connor JP, Kobayashi R, Alwine JC (1998) The snRNP-free U1A (SF-A) complex(es): identification of the largest subunit as PSF, the polypyrimidine-tract binding protein-associated splicing factor. *RNA* **4:** 1493–1499

Lutz CS, Murthy KG, Schek N, O'Connor JP, Manley JL, Alwine JC (1996) Interaction between the U1 snRNP-A protein and the 160-kDa subunit of cleavage-polyadenylation specificity factor increases polyadenylation efficiency *in vitro*. *Genes Dev* **10:** 325–337

Mandel C, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley J, Tong L (2006) Polyadenylation factor CPSF 73 is the pre-mRNA 3′-end-processing endonuclease. *Nature* **444:** 953–956

Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. *Nature* **416:** 499–506

McCracken S, Lambermon M, Blencowe BJ (2002) SRm160 splicing coactivator promotes transcript 3′-end cleavage. *Mol Cell Biol* **22:** 148–160

McCracken S, Longman D, Johnstone IL, Caceres JF, Blencowe BJ (2003) An evolutionarily conserved role for SRm160 in 3′-end processing that functions independently of exon junction complex formation. *J Biol Chem* **278:** 44153–44160

Millevoi S, Geraghty F, Idowu B, Tam JL, Antoniou M, Vagner S (2002) A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3′-end processing. *EMBO Rep* **3:** 869–874

Millevoi S, Loulergue C, Dettwiler S, Karaa SZ, Keller W, Antoniou M, Vagner S (2006) An interaction between U2AF 65 and CF I(m) links the splicing and 3′ end processing machineries. *EMBO J* **25:** 4854–4864

Moreira A, Takagaki Y, Brackenridge S, Wollerton M, Manley JL, Proudfoot NJ (1998) The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3′ end formation by two distinct mechanisms. *Genes Dev* **12:** 2522–2534

Natalizio BJ, Muniz LC, Arhin GK, Wilusz J, Lutz CS (2002) Upstream elements present in the 3′-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *J Biol Chem* **277:** 42733–42740

Niranjanakumari S, Lasda E, Brazas R, Garcia-Blanco MA (2002) Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions *in vivo*. *Methods* **26:** 182–190

Niwa M, Rose SD, Berget SM (1990) *In vitro* polyadenylation is stimulated by the presence of an upstream intron. *Genes Dev* **4:** 1552–1559

Proudfoot NJ, Furger A, Dye MJ (2002) Integrating mRNA Processing with transcription. *Cell* **108:** 501–512

Rosonina E, Ip JYY, Calarco JA, Bakowski MA, Emili A, McCracken S, Tucker P, Ingles CJ, Blencowe BJ (2005) Role for PSF in mediating transcriptional activator-dependent stimulation of pre-mRNA processing *in vivo*. *Mol Cell Biol* **25:** 6734–6746

Ryan K, Calvo O, Manley JL (2004) Evidence that polyadenylation factor CPSF 73 is the mRNA 3′ processing endonuclease. *RNA* **10:** 565–573

Singh R, Valcarcel J, Green MR (1995) Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. *Science* **268:** 1173–1176

Takagaki Y, Seipelt RL, Peterson ML, Manley JL (1996) The polyadenylation factor CstF 64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* **87:** 941–952

Vagner S, Vagner C, Mattaj IW (2000) The carboxyl terminus of vertebrate poly(A) polymerase interacts with U2AF 65 to couple 3′-end processing and splicing. *Genes Dev* **14:** 403–413

Venkataraman K, Brown KM, Gilmartin GM (2005) Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev* **19:** 1315–1327

von der Haar T, Gross JD, Wagner G, McCarthy JE (2004) The mRNA cap-binding protein eIF4E in post-transcriptional gene expression. *Nat Struct Mol Biol* **11:** 503–511

Wassarman KM, Steitz JA (1993) Association with terminal exons in pre-mRNAs: a new role for the U1 snRNP? *Genes Dev* **7:** 647–659

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345

Yan J, Marr TG (2005) Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Res* **15:** 369–375

Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63:** 405–445

Zhao W, Manley JL (1996) Complex alternative RNA processing generates an unexpected diversity of poly(A) polymerase isoforms. *Mol Cell Biol* **16:** 2378–2386