PLOS ONE

RESEARCH ARTICLE

A New Comparative-Genomics Approach for Defining Phenotype-Specific Indicators Reveals Specific Genetic Markers in Predatory Bacteria

Zohar Pasternak[1]*, Tom Ben Sasson[2], Yossi Cohen[1], Elad Segev[2], Edouard Jurkevitch[1]

1 Department of Plant Pathology and Microbiology, the Robert H. Smith Faculty of Agriculture, Food and Environment, the Hebrew University of Jerusalem, Rehovot, Israel, **2** Department of Applied Mathematics, Holon Institute of Technology, Holon, Israel

* zpast@yahoo.com

Citation: Pasternak Z, Ben Sasson T, Cohen Y, Segev E, Jurkevitch E (2015) A New Comparative-Genomics Approach for Defining Phenotype-Specific Indicators Reveals Specific Genetic Markers in Predatory Bacteria. PLoS ONE 10(11): e0142933. doi:10.1371/journal.pone.0142933

Editor: Michael Platten, University Hospital of Heidelberg, GERMANY

Received: August 29, 2015

Accepted: October 28, 2015

Published: November 16, 2015

Data Availability Statement: Datasets were deposited at the MG-RAST database (http://metagenomics.anl.gov/linkin.cgi?project=13062) under accession numbers 4624348.3-4624351.3.

Funding: This work was supported by the Israel Science Foundation (grant 1583/12) and the German Research Foundation (grant CH 731/2-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Predatory bacteria seek and consume other live bacteria. Although belonging to taxonomically diverse groups, relatively few bacterial predator species are known. Consequently, it is difficult to assess the impact of predation within the bacterial realm. As no genetic signatures distinguishing them from non-predatory bacteria are known, genomic resources cannot be exploited to uncover novel predators. In order to identify genes specific to predatory bacteria, we developed a bioinformatic tool called DiffGene. This tool automatically identifies marker genes that are specific to phenotypic or taxonomic groups, by mapping the complete gene content of all available fully-sequenced genomes for the presence/absence of each gene in each genome. A putative 'predator region' of ~60 amino acids in the tryptophan 2,3-dioxygenase (TDO) protein was found to probably be a predator-specific marker. This region is found in all known obligate predator and a few facultative predator genomes, and is absent from most facultative predators and all non-predatory bacteria. We designed PCR primers that uniquely amplify a ~180bp-long sequence within the predators' TDO gene, and validated them in monocultures as well as in metagenetic analysis of environmental wastewater samples. This marker, in addition to its usage in predator identification and phylogenetics, may finally permit reliable enumeration and cataloguing of predatory bacteria from environmental samples, as well as uncovering novel predators.

Introduction

Predation between microorganisms affects major ecological processes on a global scale. Micropredation (as defined by the destruction of a viable microbial cell) is an important component of the marine microbial loop, through the consumption of bacteria and archaea by protists and their lysis by phages [1]. Protists and phages also have strong effects on freshwater microbial food webs by contributing to prokaryotic mortality [2,3]. In soil, protozoa enhance nitrogen

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

mineralization, leading to increased plant nitrogen uptake and plant growth [4]. Although much less is known on soil phages, they may be present at high densities, potentially contributing to microbial turnover [5,6]. In addition to protozoa and phages, bacteria can also perform predation on one another. Predatory bacteria include obligate and facultative predators, which together can prey on a large variety of other bacteria [7]. Moreover, they have the capacity to attack and consume a variety of multidrug-resistant clinical strains, maintaining their predation regardless of prey antimicrobial resistance [8]; hence, they might be used as therapeutic agents where antimicrobial drugs fail. Although predators are distributed between many of the higher phyla, their currently-known total diversity amounts to only about 20 genera [7]. This dearth stems from our inability to identify predatory interactions from microscopic observation of natural samples and by the limitation of culture-based characterization by the growth requirements of both prey and predator.

Recently we developed an approach by which predator-enriched or predator-depleted protein families were identified by comparing the proteomes of predatory vs. non predatory bacteria, enabling the detection of predatory capacities in full genomes [9,10]. Nevertheless, no genes have been found to be unique to predators. With such a tool at hand, it may become possible to further screen genomes for potential predatory abilities and also assess the abundance of bacterial predators in the environment, an essential step toward understanding the effects of bacterial predation in nature. Now, with the growing availability of whole genome data, new methods have emerged for systematically finding optimal genetic markers to distinguish between phylogenetic or phenotypic groups [11,12,13,14]. However, none of the existing methods enabled us to find enriched genes in thousands of genomes simultaneously in a user-friendly and efficient manner. To that end, we designed a novel bioinformatics tool, and used it to find, for the first time, a predation marker gene; we further investigated this gene to design PCR primers for identification and analysis of predatory organisms in monoculture. Finally, we showed that the gene specifically detects predatory bacteria in environmental samples.

Materials and Methods

Software development

A novel software, DiffGene, takes advantage of the orthologous gene cluster table created and maintained by the microbial genome database (MBGD) [15] and freely available at http://mbgd.genome.ad.jp/htbin/view_arch.cgi. This table is updated twice a year, and is arranged so that each row in the table is an orthologous cluster (i.e. the same gene) and each column is a genome. The ortholog identification and grouping procedure, called DomClust, is described in full elsewhere [16]; in short, it takes as input all-against-all protein BLAST similarity data and classifies genes based on subsequent hierarchical clustering with UPGMA [17]. During clustering, it detects domain fusion or fission events, splits clusters into domains (if required), and then splits the resulting trees such that intra-species paralogous genes are divided into different groups so as to create plausible orthologous groups. Next, a second procedure called DomRefine [15] improves domain-level clustering using multiple sequence alignment information, and a third (MergeTree) [18] adds new genomes to the table.

The raw MBGD data, as is often the case for such bulky datasets, is too large and complex to be handled on a personal computer. In DiffGene, these data are automatically cleaned, unneeded and redundant data are deleted, and all gene occurrences are transformed from gene names into binary data so that each datapoint in the table only contains either a one or a zero (the gene is present or absent in the genome, respectively). This reduces the file size by two orders of magnitude, and enables efficient algorithm usage. Genomes are then assigned into two groups, 'present' and 'absent'; for each gene, the proportion of genomes in each group

which contain that specific gene is calculated. Different fraction threshold levels for the two groups can be assigned so the output of the analysis contains only genes which appear in at least the indicated fraction of 'present' genomes and at most in the indicated fraction of 'absent' genomes. The software is freely available at http://departments.agri.huji.ac.il/plantpath/jurkevitch/ej-software.html.

Search for predation-specific marker

Predators are phylogenetically diverse but share some phenotypic or ecological traits (S1 Table). We employed DiffGene to find genes enriched in predatory compared to non-predatory genomes. DNA and protein sequences of the most-discriminating gene were taken from the NCBI RefSeq database, representing predatory and non-predatory bacteria, as well as eukaryotic organisms. Protein sequences were aligned using MUSCLE [19] and a maximum-likelihood phylogenetic tree was constructed in MEGA6 [20].

Experimental confirmation of specificity

DNA sequences of the candidate marker gene were also aligned and manually inspected to find potential primers for a PCR reaction. The most predator-specific primers were named TDO-F (5'-TAYGARYTVTGGTTYAARCARAT-3') and TDO-R (5'-GGMGTCATSSTYTCVA-3') (for nucleotide ambiguity codes, see [21]). DNA to be used as PCR template was extracted with PowerSoil isolation kit (MoBio laboratories, Carlsbad, CA) from pure cultures of three predators (*Bdellovibrio bacteriovorus* HD100, *Bdellovibrio exovorus* JSS, and *Peredibacter starrii* A3.12) and five non-predators from various phyla (*Escherichiacoli* ML35, *Pseudomonas* sp., *Flavobacterium* sp., *Burkholderia* sp. and *Photobacterium* sp.). PCR reactions were performed using 12.5 ul master mix (0.1 U/μl Taq Polymerase, 500 μM dNTP each, 20 mM Tris-HCl (pH 8.3), 100 mM KCl, 3 mM MgCl2), 8.5 ul double-distilled water, 2 ul of each primer and 2 ul of template DNA; amplification conditions were 95°C for 5 min, followed by 36 cycles of 95°C for 30 sec, 50°C for 30 sec, 72°C for 30 sec, and a final stage of 72°C for 7 min. The same primers were used to assess the predator communities in four environmental samples from a wastewater treatment plant. DNA extraction and PCR amplification were as before, and MiSeq next-generation sequencing (Illumina, USA) was performed as previously described [22]. Sequences were processed in MOTHUR v1.34 [23]: quality, length and adapter trimming were performed on the forward (non-paired) reads as previously described [24], resulting in >50,000 reads per sample with a uniform length of 184 nucleotides per read. Datasets were deposited at the MG-RAST database (http://metagenomics.anl.gov/linkin.cgi?project=13062) under accession numbers 4624348.3–4624351.3. Sequences sharing 97% identity were clustered into the same operational taxonomic unit (OTU) and the representative sequence from each OTU was phylotyped using BLAST. The representative sequences of the 100-most abundant OTUs, along with marker gene sequences from predators and non-predators, were used for creating a phylogenetic tree as above.

Results

Multi-locus typing of all predatory bacteria

Of the 2286 complete (non-draft) bacterial genomes available on the MBGD dataset at the time of analysis, 14 belonged to known predator species (S1 Table) and the other 2272 were considered non-predators. DiffGene analysis discovered several genes which were quite specific to either genomes of predators or of non-predators, but none of them was 100% specific to either group. Combining three of these genes–*kynA*, *waaL* and *gntR* (Table 1)–such that a genome would be considered 'predatory' if it contained the former two and lacked the latter, resulted in

Table 1. Abundance of marker genes in genomes of predatory and non-predatory bacteria.

Gene	Rep. accession	Predators	Non-predators
[*kynA*] Tryptophan 2,3-dioxygenase	NP_968676	14/14 = 100%	302/2272 = 13%
[*waaL*] O-antigen ligase	NP_968553	14/14 = 100%	434/2272 = 19%
[*gntR*] Transcription regulator	YP_004789625	0/14 = 0%	1221/2272 = 54%

Rep., representative.

doi:10.1371/journal.pone.0142933.t001

correct classification of 14/14 (100%) of the predators and 2255/2272 (99.3%) of the non-predators. It bears noting that since one bacterial class, namely delta-proteobacteria, was over-represented in the 'predatory' genomes, the three-gene set could in fact have been indicative not of predators but of delta-proteobacteria. However, this is unlikely because of the 50 'non-predatory' delta-proteobacterial genomes, five contained *kynA*, five *waaL*, and 39 *gntR*, thus all were correctly identified as non-predatory (except for *Anaeromyxobacter* which is a potential predator).

Marker gene for obligatory predatory bacteria

The top marker gene, *kynA* (encoding tryptophan 2,3-dioxygenase—TDO) was found in 100% (14/14) of predator and 13.3% (302/2272) of non-predators genomes. Of the 302 non-predatory genomes containing *kynA*, five were delta-proteobacteria (out of 50 delta-proteobacteria in the database). Protein sequence analysis of *kynA* representatives revealed that sequences from all obligate and two facultative predatory bacteria, as well as eukaryotic organisms, although phylogenetically extremely diverse, are longer than the non-predatory bacterial ones, as is reflected in their proximity in the TDO phylogenetic tree (Fig 1). The main feature distinguishing the sequences from the two groups is a segment ~60 amino acids long (S1 Fig), absent from all non-predatory bacteria (0%, 0/2272) and most facultative predatory bacteria. This entire segment is a long alpha-helix, which is hydrophilic (mean±SD Kyte-Doolittle hydropathy of -0.76±0.75) and charged (-1.84 at pH = 7). TDO, by itself in bacteria and together with indoleamine 2,3-dioxygenase (IDO) in the mammalian liver, catalyses the first and rate-limiting step in the kynurenine pathway, converting L-tryptophan to N-formyl-kynurenine [25]. The next step in this pathway is converting the N-formyl-kynurenine to formyl-anthranilate (by the enzyme kynureninase) or to L-kynurenine (by the enzyme kynurenine formamidase). Further investigation revealed that all eukaryotes, bacterial non-predators and facultative predators which possessed the *kynA* gene, also possessed the genes required for completing the kynurenine pathway, with the genes for kynureninase and/or kynurenine formamidase usually adjacent to the *kynA* gene. However, in all obligate bacterial predators, no other gene belonging to this pathway was found.

Validation of TDO sequence specificity in predators

The nucleotide sequences of the TDO gene from representative species were aligned, and PCR primers were manually designed to selectively amplify the specific sequence in the TDO genes of predators. PCR amplification was performed on DNA extracted from cultured representative predatory and non-predatory bacteria. A ~180 bp-long PCR product was obtained from the genomes of all predators, whereas no amplicon was obtained from any of the non-predators (S2 Fig). The amplicon was subsequently sequenced and validated to be part of the TDO gene. To further test the specificity of the *kynA* PCR primers, metagenetic next-generation sequencing was conducted on four environmental wastewater samples. Such samples include many

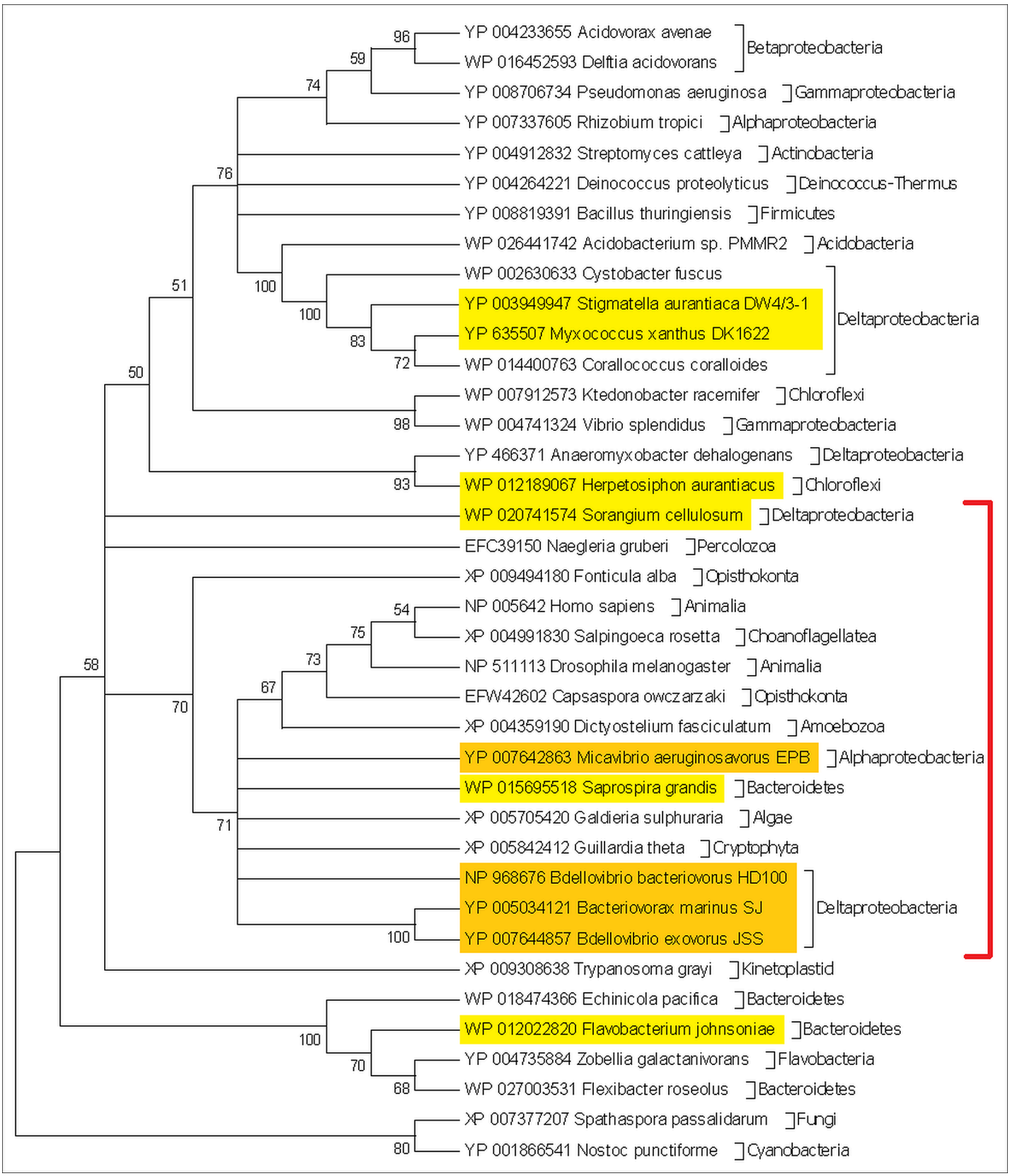
YP 004233655 Acidovorax avenae
WP 016452593 Delftia acidovorans
Betaproteobacteria
YP 008706734 Pseudomonas aeruginosa
Gammaproteobacteria
YP 007337605 Rhizobium tropici
Alphaproteobacteria
YP 004912832 Streptomyces cattleya
Actinobacteria
YP 004264221 Deinococcus proteolyticus
Deinococcus-Thermus
YP 008819391 Bacillus thuringiensis
Firmicutes
WP 026441742 Acidobacterium sp. PMMR2
Acidobacteria
WP 002630633 Cystobacter fuscus
YP 003949947 Stigmatella aurantiaca DW4/3-1
YP 635507 Myxococcus xanthus DK1622
WP 014400763 Corallococcus coralloides
Deltaproteobacteria
WP 007912573 Ktedonobacter racemifer
Chloroflexi
WP 004741324 Vibrio splendidus
Gammaproteobacteria
YP 466371 Anaeromyxobacter dehalogenans
Deltaproteobacteria
WP 012189067 Herpetosiphon aurantiacus
Chloroflexi
WP 020741574 Sorangium cellulosum
Deltaproteobacteria
EFC39150 Naegleria gruberi
Percolozoa
XP 009494180 Fonticula alba
Opisthokonta
NP 005642 Homo sapiens
Animalia
XP 004991830 Salpingoeca rosetta
Choanoflagellatea
NP 511113 Drosophila melanogaster
Animalia
EFW42602 Capsaspora owczarzaki
Opisthokonta
XP 004359190 Dictyostelium fasciculatum
Amoebozoa
YP 007642863 Micavibrio aeruginosavorus EPB
Alphaproteobacteria
WP 015695518 Saprospira grandis
Bacteroidetes
XP 005705420 Galdieria sulphuraria
Algae
XP 005842412 Guillardia theta
Cryptophyta
NP 968676 Bdellovibrio bacteriovorus HD100
YP 005034121 Bacteriovorax marinus SJ
YP 007644857 Bdellovibrio exovorus JSS
Deltaproteobacteria
XP 009308638 Trypanosoma grayi
Kinetoplastid
WP 018474366 Echinicola pacifica
Bacteroidetes
WP 012022820 Flavobacterium johnsoniae
Bacteroidetes
YP 004735884 Zobellia galactanivorans
Flavobacteria
WP 027003531 Flexibacter roseolus
Bacteroidetes
XP 007377207 Spathaspora passalidarum
Fungi
YP 001866541 Nostoc punctiforme
Cyanobacteria
96
59
74
76
51
100
100
83
72
50
98
93
58
54
75
73
67
70
71
100
100
70
68
80

Fig 1. Maximum-likelihood phylogenetic tree of the tryptophan 2,3-dioxygenase protein. The percentage of trees in which the associated taxa clustered together (out of 100 bootstraps) is shown next to the branches; branches with <50% were collapsed. Obligate bacterial predators are marked orange, facultative yellow. Red line indicates genomes with the ~60 amino acid-long insert.

doi:10.1371/journal.pone.0142933.g001

more genomes than the databases upon which in-silico analysis was performed, thus providing a more stringent test for specificity. The sequences in each wastewater sample were clustered into operational taxonomic units (OTUs), where all the sequences in a single OTU are at least 97% similar to each other. All samples contained between 67–91 OTUs, and all four rarefaction curves reached saturation after ~30,000 reads (S3 Fig), implying that all the diversity in the samples was detected. When comparing both known and wastewater TDO sequences in a phylogenetic tree, all known non-predators formed an outgroup; of the 100-most abundant OTUs in the wastewater samples, comprising >99% of all sequences, none were phylogenetically close to any known non-predator (Fig 2). This result was also observed when BLASTing each environmental TDO sequence against the NCBI database. All TDO wastewater OTUs belonged to two broad phylogenetic groups: the first, similar to many of the known predators, and the second, similar only to a single known bacterium, *Niastella koreensis* (Fig 2). The seven most highly-abundant environmental OTUs (named OTU1-OTU7 in Fig 2) encompassed 65% of all environmental sequences, and all of these belonged to the "*Niastella*-like" group.

Discussion

Predatory interactions between bacteria are difficult to detect and therefore it is difficult to address their effect in nature. Further complicating the assessment of their diversity, distribution and abundance is the lack of genetic markers so that metagenomic data are almost blind in ascertaining their status in environmental samples. In order to overcome these difficulties, we have developed DiffGene, a software that allows a quick and easy characterization of marker genes for microbial groups, as long as sequenced genomes representing the groups of study are available. Using all available fully-sequenced genomes, marker genes are optimally chosen so that their presence or absence (rather than their sequence or abundance) is an indication of the genomes' grouping. Finding marker genes for specific microbial organisms, grouped by either phenotype or taxonomy, can be a challenging task. In our previous work [9,10] proteome similarity matrices were used on a much smaller scale to detect genes unique to bacterial predators. The gene matrix, being based on gene abundance rather than absence/presence data, resulted in less-specific markers but revealed specific genomic properties of most predators. Most strikingly, several genes from the mevalonate pathway (isoprenoid biosynthesis) were highly enriched in most predators; however, while useful for screening cultured organisms, its absence from the genomes of a few predators may lead to false negative results. Its further presence in Archaea and in some non-predatory bacteria would make these genes poor markers for predators in environmental samples. Appropriately, the absence/presence matrix developed here hardly marked any mevalonate pathway genes as highly predator-specific.

We found the *kynA* gene, encoding tryptophan 2,3-dioxygenase (TDO), to be the most predator-specific gene. TDO is the first enzyme in the degradation of tryptophan pathway, and was previously found among the predator-enriched complement of genes [9]. Surprisingly, no other gene belonging to this pathway was found in the genomes of obligate predatory bacteria, whereas all other genomes containing *kynA* contained the genes necessary to complete the tryptophan degradation. This suggests that obligate predators either degrade L-tryptophan by another, unknown pathway, or that they do not catabolize L-tryptophan at all, instead using the N-formyl-kynurenine produced by the TDO for another, unknown purpose. The second most potent marker for predators, the *WaaL* protein, is also a metabolic gene. It catalyzes a

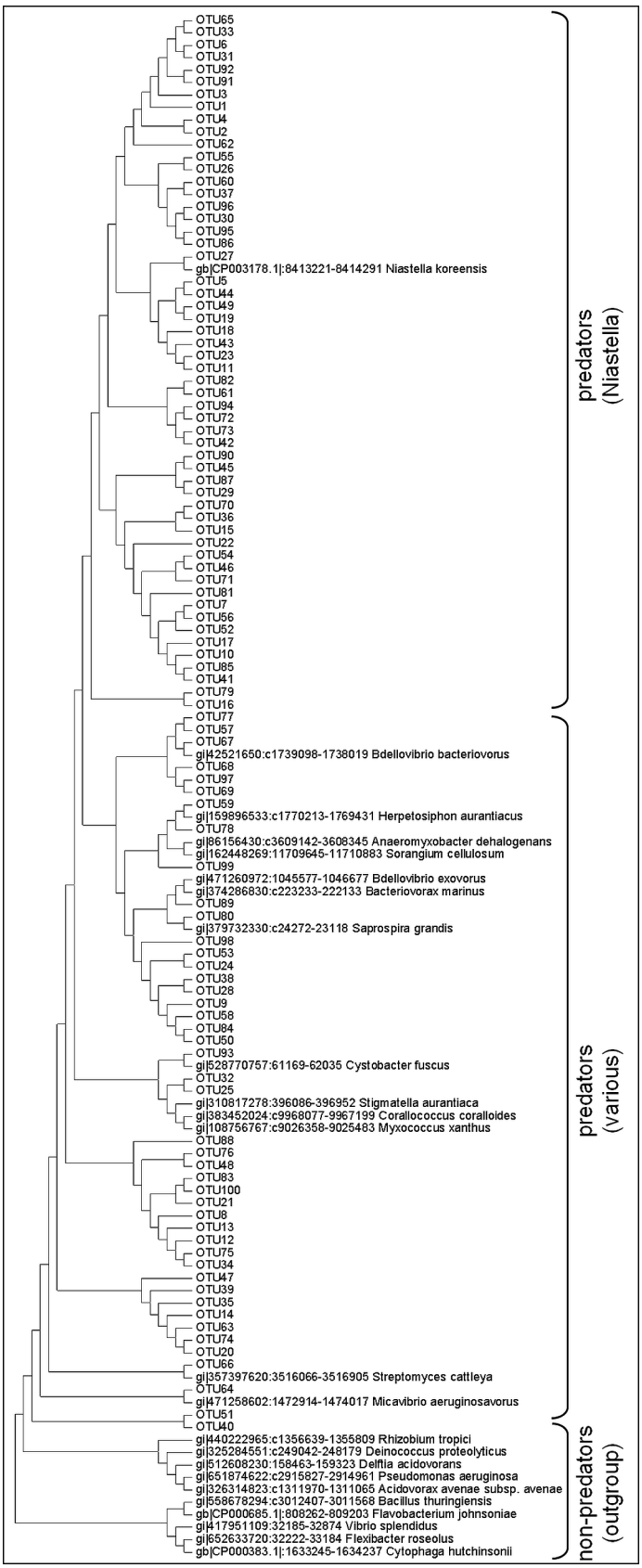
OTU65
OTU33
OTU6
OTU31
OTU92
OTU91
OTU3
OTU1
OTU4
OTU2
OTU62
OTU55
OTU26
OTU60
OTU37
OTU96
OTU30
OTU95
OTU86
OTU27
gb|CP003178.1|:8413221-8414291 Niastella koreensis
OTU5
OTU44
OTU49
OTU19
OTU18
OTU43
OTU23
OTU11
OTU82
OTU61
OTU94
OTU72
OTU73
OTU42
OTU90
OTU45
OTU87
OTU29
OTU70
OTU36
OTU15
OTU22
OTU54
OTU46
OTU71
OTU81
OTU7
OTU56
OTU52
OTU17
OTU10
OTU85
OTU41
OTU79
OTU16
OTU77
OTU57
OTU67
gi|42521650:c1739098-1738019 Bdellovibrio bacteriovorus
OTU68
OTU97
OTU69
OTU59
gi|159896533:c1770213-1769431 Herpetosiphon aurantiacus
OTU78
gi|86156430:c3609142-3608345 Anaeromyxobacter dehalogenans
gi|162448269:11709645-11710883 Sorangium cellulosum
OTU99
gi|471260972:1045577-1046677 Bdellovibrio exovorus
gi|374286830:c223233-222133 Bacteriovorax marinus
OTU89
OTU80
gi|379732330:c24272-23118 Saprospira grandis
OTU98
OTU53
OTU24
OTU38
OTU28
OTU9
OTU58
OTU84
OTU50
OTU93
gi|528770757:61169-62035 Cystobacter fuscus
OTU32
OTU25
gi|310817278:396086-396952 Stigmatella aurantiaca
gi|383452024:c9968077-9967199 Corallococcus coralloides
gi|108756767:c9026358-9025483 Myxococcus xanthus
OTU88
OTU76
OTU48
OTU83
OTU100
OTU21
OTU8
OTU13
OTU12
OTU75
OTU34
OTU47
OTU39
OTU35
OTU14
OTU63
OTU74
OTU20
OTU66
gi|357397620:3516066-3516905 Streptomyces cattleya
OTU64
gi|471258602:1472914-1474017 Micavibrio aeruginosavorus
OTU51
OTU40
gi|440222965:c1356639-1355809 Rhizobium tropici
gi|325284551:c249042-248179 Deinococcus proteolyticus
gi|512608230:158463-159323 Delftia acidovorans
gi|651874622:c2915827-2914961 Pseudomonas aeruginosa
gi|326314823:c1311970-1311065 Acidovorax avenae subsp. avenae
gi|558678294:c3012407-3011568 Bacillus thuringiensis
gb|CP000685.1|:808262-809203 Flavobacterium johnsoniae
gi|417951109:32185-32874 Vibrio splendidus
gi|652633720:32222-33184 Flexibacter roseolus
gb|CP000383.1|:1633245-1634237 Cytophaga hutchinsonii
predators (Niastella)
predators (various)
non-predators (outgroup)

Fig 2. Maximum-likelihood phylogenetic tree of the representative sequences of the 100-most abundant OTUs in the metagenetic analysis, including representative TDO sequences from known predatory and non-predatory bacteria. The bootstrap consensus tree inferred from 100 replicates is taken to represent the evolutionary history of the taxa analyzed. Wastewater OTU names are according to abundance, i.e. OTU1 is the most abundant OTU in the environmental samples, OTU2 is the second-most abundant, and so on. Known sequence names include GI accession, coordinates within the genome, and species name.

doi:10.1371/journal.pone.0142933.g002

critical step in lipopolysaccharide synthesis as it establishes the glycosidic bond of a sugar at the proximal end of the undecaprenyl-diphosphate (Und-PP)-O-antigen with a terminal sugar of the lipid A-core oligosaccharide [26]. The widely distributed regulatory protein *GntR*, absent from the predators, is involved in amino acids, sugars, fatty acids and alkylphosphonate metabolism and pyridoxal phosphate-dependent aminotransfer [27]. It has been proposed that this family goes back to the last common universal ancestor [27], implying that, for an unknown reason, it is selectively lost in predatory bacteria. Using these three genes as a multi-locus sequence typing scheme yielded near-perfect results, except for 17 out of 2272 'non-predators' which were classified as 'predators'; it is possible that at least some of these 17 are indeed previously-undetected predators. Most of them are suspected predators such as *Flexibacter* [28,29] and *Anaeromyxobacter* [30] or gliding heterotrophs (e.g. *Owenweeksia*, *Fluviicola*, *Echinicola*) that may yet prove to be predatory.

The finding of a predation-specific marker advances the research of predatory bacteria in three important aspects: first, eliminating—for the first time—the need for a laborious *in-vitro* predator and prey co-culture in order to ascertain whether a species is indeed predatory. This was confirmed using cultured strains. Second, shedding new light on predator phylogenetics, since the TDO gene obviously bears some significance for the predatory lifestyle and may thus prove more phylogenetically informative than the ubiquitous 16S rRNA gene. Third, applying the predator-specific TDO PCR primers in real-time PCR and next-generation metagenetic sequencing applications may finally permit reliable enumeration and cataloguing, respectively, of predators from environmental samples, as well as uncovering novel predatory bacteria. In the tested wastewater samples, none of the 100-most abundant OTUs (encompassing >99% of all sequences) could be traced to known non-predators; while these OTUs could theoretically hail from unknown non-predators, it is more likely that these findings confirm that the approach is indeed predator-specific. Interestingly, our preliminary analysis of wastewater samples revealed that a large portion of the predator OTUs and sequences had no known relatives except *Niastella koreensis*, a gliding bacterium of the phylum Bacteroidetes which is most closely related to the genera *Flexibacter*, *Cytophaga* and *Flavobacterium* [31]; as it happens, all three of these genera contain species which are known predators [7,29], leading us to assume that *Niastella koreensis*, as well as the entire *"Niastella*-like" group apparent in Fig 2, are indeed predators. The many unrecognized predator OTUs suggest that predatory bacteria are unrepresented in culture collections.

Surprisingly, the predator-specific TDO protein is most similar to the eukaryotic one. It has long been thought that, 1.5 billion years ago, the eukaryotic cell originated from a merger of two prokaryotes, an archaeal host and a bacterial endosymbiont [32]; then, during the evolutionary transition from an endosymbiont to an organelle, the bacterium transferred some of its DNA to the host chromosomes [33]. However, since prokaryotes are unable to perform phagocytosis, the means by which the endosymbiont originally entered its host is an enigma. Davidov and Jurkevitch [34], based on [35], suggested that this process was facilitated by a predatory bacterium which penetrated and replicated within the host periplasm, and later became the mitochondria. In a previous study [9], we found that the mevalonate pathway, which is the

isoprenoid synthesizing mechanism in eukaryotes (but not in most bacteria), is also strongly enriched in predators. Together, the mevalonate and TDO data add a tantalizing phylogenetic clue to a possible connection between predatory bacteria and eukaryotic evolution.

The genome-centered approach developed in DiffGene and the flexibility awarded to alter the required specificity of markers can provide strict specificity allowing for unambiguous identification of membership in a particular bacterial group at the cost of potential false negative. In contrast, more relaxed specificity settings enable the assignment of microorganisms to groups even if they may not all contain all the marker genes, thus discovering genomes enriched for particular functions and pathways. The approach developed here may be useful for additional purposes as well. Many microbial pathogens are characterized using multi-locus typing, where up to 16 genes are selected as molecular markers and compared between isolates, either by their presence/absence or sequence [36]. Nevertheless, this approach can lead to erroneous typing due to the genes being non-representative [37] and requires many marker genes per group in order to verify the isolates' membership. Furthermore, genetic markers are often selected *ad hoc*, using too few reference genomes and/or manual inspection of the results [38]. Applying DiffGene for marker search in this context may help overcome such limitations.

Ethics statement

This study did not involve humans, human data or animals; therefore, no ethics approval was required.

Consent

This article is not a prospective study involving human participants, and does not contain individual clinical data; therefore, no consent for publication was required.

Supporting Information

S1 Fig. Multiple alignment of the region specific to obligate bacterial predators and eukaryotes of the tryptophan 2,3-dioxygenase protein. Top, obligate bacterial predators and eukaryotes; bottom, non-predatory and facultative predatory bacteria. Residues are colored according to the ClustalX color scheme: blue = hydrophobic, green = polar, magenta = negatively charged, red = positively charged, pink = cysteine, orange = glycine, yellow = proline.
(TIF)

S2 Fig. PCR amplification using predator-specific TDO primers on DNA extracted from cultures of predatory and non-predatory bacteria. L, ladder (100bp); 1, *Bdellovibrio bacteriovorus* HD100; 2, *Bdellovibrio exovorus* JSS; 3, *Peredibacter starrii* A3.12; 4, *Escherichia coli* ML35; 5, *Pseudomonas* sp.; 6, *Flavobacterium* sp.; 7, *Burkholderia* sp.; 8, *Photobacterium* sp.; C, negative control.
(TIF)

S3 Fig. Rarefaction curves of bacterial communities at 97% sequence similarity level in the four samples from various areas of the wastewater treatment plant.
(TIF)

S1 Table. Predatory bacteria analyzed in this study.
(DOCX)

Acknowledgments

This work was supported by the Israel Science Foundation (grant 1583/12) and the **Deutsche Forschungsgemeinschaft** (DFG) (grant CH 731/2-1).

Author Contributions

Conceived and designed the experiments: ZP ES EJ. Performed the experiments: ZP ES YC TBS. Analyzed the data: ZP ES YC EJ TBS. Wrote the paper: ZP ES EJ.

References

1. Johnke J, Cohen Y, de Leeuw M, Kushmaro A, Jurkevitch E, Chatzinotas A. Multiple micro-predators controlling bacterial communities in the environment. Curr Opin Biotechnol. 2014; 27: 185–190. doi: 10.1016/j.copbio.2014.02.003 PMID: 24598212
2. Personnic S, Domaizon I, Dorigo U, Berdjeb L, Jacquet S. Seasonal and spatial variability of virio-, bacterio-, and picophytoplanktonic abundances in three peri-alpine lakes. Hydrobiologia. 2009; 627(1): 99–116.
3. Pirlot S, Unrein F, Descy JP, Servais P. Fate of heterotrophic bacteria in Lake Tanganyika (East Africa). FEMS Microbiol Ecol. 2007; 62(3): 354–364. PMID: 17983442
4. Ekelund F, Saj S, Vestergård M, Bertaux J, Mikola J. The "soil microbial loop"is not always needed to explain protozoan stimulation of plants. Soil Biol Biochem. 2009; 41: 2336e2342.
5. Buée M, De Boer W, Martin F, Van Overbeek L, Jurkevitch E. The rhizosphere zoo: an overview of plant-associated communities of microorganisms, including phages, bacteria, archaea, and fungi, and of some of their structuring factors. Plant Soil. 2009; 321(1–2): 189–212.
6. Williamson KE, Corzo KA, Drissi CL, Buckingham JM, Thompson CP, Helton RR. Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. Biol Fertil Soils. 2013; 49: 857–869.
7. Jurkevitch E, Davidov Y. Phylogenetic diversity and evolution of predatory prokaryotes. In: Predatory Prokaryotes—Biology, ecology, and evolution. 2007. Jurkevitch E. (ed.). Springer-Verlag, Heidelberg.
8. Kadouri DE, To K, Shanks RMQ, Doi Y. Predatory Bacteria: A Potential Ally against Multidrug-Resistant Gram-Negative Pathogens. PLoS ONE. 2013; 8(5): e63397. doi: 10.1371/journal.pone.0063397 PMID: 23650563
9. Pasternak Z, Pietrokovski S, Rotem O, Gophna U, Lurie-Weinberger MN, Jurkevitch E. By their genes ye shall know them: genomic signatures of predatory bacteria. ISME J. 2013; 7(4): 756–769. doi: 10.1038/ismej.2012.149 PMID: 23190728
10. Pasternak Z, Njagi M, Shani Y, Chanyi R, Rotem O, Lurie-Weinberger MN, et al. In and out: an analysis of epibiotic vs periplasmic bacterial predators. ISME J. 2014; doi: 10.1038/ismej.2013.164 PMID: 24088628
11. Cody AJ, Bennett JS, Maiden MCJ. Multi-locus sequence typing and the gene-by-gene approach to bacterial classification and analysis of population variation, In: Michael Goodfellow, Iain Sutcliffe and Jongsik Chun, Editor(s), Methods in Microbiology, Academic Press. 2014; 41: 201–219.
12. Bohle HM, Gabaldón T. Selection of marker genes using whole-genome DNA polymorphism analysis. Evol Bioinform Online. 2012; 8: 161–169. doi: 10.4137/EBO.S8989 PMID: 22474405
13. Yemin L, Morrison JC, Hershberg R, Ro GL. POGO-DB—a database of pairwise-comparisons of genomes and conserved orthologous genes. Nucleic Acids Res. 2013; doi: 10.1093/nar/gkt1094 PMID: 24198250
14. Huang K, Brady A, Mahurkar A, White O, Gevers D, Huttenhower C, et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. Nucl Acids Res. 2013; doi: 10.1093/nar/gkt1078 PMID: 24203705
15. Uchiyama I, Mihara M, Nishide H, Chiba H. MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucl Acids Res. 2015; 43 (D1): D270–D276.
16. Uchiyama I. Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. Nucl Acids Res. 2006; 34: 647–658. PMID: 16436801
17. Sneath PHA & Sokal RR. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: Freeman. 1973; 573 p.
18. Uchiyama I, Mihara M, Nishide H, Chiba H. MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. Nucl Acids Res. 2013; 41(D1): D631–D635.

19. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32(5): 1792–1797. PMID: 15034147

20. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol. 2013; 30: 2725–2729. doi: 10.1093/molbev/mst197 PMID: 24132122

21. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucl Acids Res. 1985; 13: 3021–3030. PMID: 2582368

22. Green SJ, Venkatramanan R, Naqib A. Deconstructing the Polymerase Chain Reaction: Understanding and Correcting Bias Associated with Primer Degeneracies and Primer-Template Mismatches. PLoS ONE. 2015; 10(5): e0128122. doi: 10.1371/journal.pone.0128122 PMID: 25996930

23. Schloss PD, Wescott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009; 75(23):7537–7541. doi: 10.1128/AEM.01541-09 PMID: 19801464

24. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeqIllumina sequencing platform. App Env Microbiol. 2013; 79(17): 5112–5120.

25. Littlejohn TK, Takikawa O, Truscott RJW, Walker MJ. Asp274 and his346 are essential for heme binding and catalytic function of human indoleamine 2,3-dioxygenase. J Biol Chem. 2003; 278: 29525–29531. PMID: 12766158

26. Han W, Wu B, Li L, Zhao G, Woodward R, Pettit N, et al. Defining function of lipopolysaccharide O-antigen ligase waaL using chemoenzymatically synthesized substrates. J Biol Chem. 2012; 287: 5357–5365. doi: 10.1074/jbc.M111.308486 PMID: 22158874

27. Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM. The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev. 2005; 29(2): 231–262. PMID: 15808743

28. Sallal AK. Lysis of cyanobacteria with *Flexibacter* spp isolated from domestic sewage. Microbios. 1994; 77: 57–67. PMID: 8159127

29. Chen H, Young S, Berhane T-K, Williams HN. Predatory Bacteriovorax Communities Ordered by Various Prey Species. PLoS ONE. 2012; 7(3): e34174. doi: 10.1371/journal.pone.0034174 PMID: 22461907

30. Thomas SH, Wagner RD, Arakaki AK, Skolnick J, Kirby JR, Shimkets LJ, et al. The mosaic genome of *Anaeromyxobacterdehalogenans* strain 2CP-C suggests an aerobic common ancestor to the delta-Proteobacteria. PLoS ONE. 2008; 3(5): e2103. doi: 10.1371/journal.pone.0002103 PMID: 18461135

31. Weon HY, Kim BY, Yoo SH, Lee SY, Kwon SW, Go SJ, et al. *Niastella koreensis* gen. nov., sp. nov.And *Niastella yeongjuensis* sp. nov., novel members of the phylum Bacteroidetes, isolated from soil cultivated with Korean ginseng. Int J Syst Evol Microbiol. 2006; 56: 1777–1782. PMID: 16902007

32. Sagan L. On the origin of mitosing cells. J Theor Biol. 1967; 14(3): 255–274. PMID: 11541392

33. Timmis JN, Ayliffe MA, Huang CY, Martin W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature Rev Genet. 2004; 5: 123–135. PMID: 14735123

34. Davidov Y and Jurkevitch E. Predation between prokaryotes and the origin of eukaryotes. BioEssay. 2009; 31: 748–757.

35. Guerrero R, Pedros-Alio C, Esteve I, Mas J, Chase D, Margulis L. Predatory prokaryotes: predation and primary consumptionevolved in bacteria. Proc Natl Acad Sci USA. 1986; 83: 2138–2142. PMID: 11542073

36. Pérez-Losada M, Cabezas P, Castro-Nallar E, Crandall KA. Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. Infect Genet Evol. 2013; 16: 38–53. doi: 10.1016/j.meegid.2013.01.009 PMID: 23357583

37. Leopold SR, Sawyer SA, Whittam TS, Tarr PI. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol Biol. 2011; 11: 183. doi: 10.1186/1471-2148-11-183 PMID: 21708031

38. Dutilh BE, Snel B, Ettema TJ, Huynen MA. Signature genes as a phylogenomic tool. Mol Biol Evol. 2008; 25(8): 1659–1667. doi: 10.1093/molbev/msn115 PMID: 18492663