Evolutionary Bioinformatics

METHODOLOGY

Selection of Marker Genes Using Whole-Genome DNA Polymorphism Analysis

Harry M. Bohle[1,*] and Toni Gabaldón[1,2,*]

[1]Bioinformática, Universidad Internacional de Andalucía, Málaga, Spain. [2]Bioinformatics and Genomics, Centre for Genomic Regulation (CRG), and UPF, Barcelona, Spain.
*These authors contributed equally to this work.
Corresponding authors website: http://gabaldonlab.crg.eu

Abstract: Molecular markers serve to assign individual samples to specific groups. Such markers should be easily identified and have a high discrimination power, being highly conserved within groups while showing sufficient variability between the groups that are to be distinguished. The availability of a large number of complete genomic sequences now enables the informed selection of genes as molecular markers based on the observed patterns of variability. We derived a new scoring system based on observed DNA polymorphic differences, and which uses the Bayes theorem as adapted by Wilcox. For validation, we applied this system to the problem of identifying individual species within a prokaryotic (*Vibrio*) and a eukaryotic (*Diphyllobothrium*) genus for validation. Top-scoring candidates genes Chromosome segregation ATPase and ATPase-subunit 6 showed better discrimination power in *Vibrio* and *Diphyllobothrium*, respectively, as compared to standard molecular markers (*recA*, *dnaJ* and *atpA* for *Vibrio*, and 18s rRNA, ITS and COX1 for *Diphyllobothrium*).

Keywords: molecular marker, genome analysis, Bayes's theorem, DNA polymorphism

Evolutionary Bioinformatics 2012:8 161–169

doi: 10.4137/EBO.S8989

This article is available from http://www.la-press.com.

Background

Molecular methods to assign biological samples to specific groups (eg, taxonomic groups) have largely replaced morphological comparisons, allowing hundreds or even thousands of characters to be compared across samples.[1] Historically, numerous DNA-based approaches encompassing random whole-genomic analysis have been used to discriminate groups of organisms. These include methods like, among many others, restriction fragment length polymorphism (RFLP), or random amplification of polymorphic DNA (RAPD).[2,3] Alternatively, sequences from genes, usually selected by their conserved, housekeeping roles, can be used.[2] However, it is often the case that existing markers provide insufficient resolution or are confounded by homoplasy, homologous recombination and lateral gene transfer.[4,5] In recent years, thanks to great advances in sequencing technologies,[6,7] the number and diversity of completely sequenced genomes is growing exponentially. This provides the basis for optimizing the selection of marker genes based on the analysis of the whole genetic complement of a given set of organisms. Earlier attempts to use whole-genome information to select marker genes that could best serve as predictors of phylogenetic relatedness include the use of scores based on the level of sequence identities from whole-genome alignments,[8] or the selection of unique sequence signatures present in a few species.[9] These methods, however, do not exploit the information from sequence variability within a species. Here we propose and evaluate an alternative algorithm for the selection of optimal genetic markers, which is based on the comparison of complete genomes. In brief, the basis of our strategy is to rank different genes according to the level of DNA polymorphism within and between defined taxonomic groups. More specifically, DNA polymorphism is measured as the average number of nucleotide differences per site,[10] and a conditional probabilistic statistic based on Bayes's Theorem as adapted by Willcox[11] is used to prioritize genes, so that genes presenting higher levels of polymorphism between groups but lower variation within a group receive higher scores. In order to validate the methodology, we apply it to the problem of selecting marker genes for the identification of individual species within a prokaryotic (*Vibrio*) and a Eukaryotic (*Diphyllobothrium*) genus. Publicly available genomic sequences were analyzed to select high-scoring marker genes, which were subsequently amplified and sequenced in a set of additional, non-sequenced strains of these groups. The discrimination power (DP) of these newly obtained sequences was compared to that of traditional marker genes.

Methods

Sequence data

Complete genome sequences were downloaded from the National Center of Bioinformatics Information (NCBI) in Genbank (.GBK) format. These were: (i) chromosome I from the following *Vibrio* species and strains: *V. cholerae* (NC_002505), *V. vulnificus* (NC_004459), *V. parahaemolyticus* (NC_004603), *V. harveyi* (NC_009783), *V. fischeri* (NC_006840), *Alivibrio salmonicida* (NC_011312), *V. splendidus* (NC_011753), *V. cholerae* (NC_009457), *V. cholerae* (NC_012578), and *V. cholerae* (NC_012668); (ii) Whole mitochondrial genomes from Different *Diphyllobothrium* species and strains: *D. latum* (NC_008945), *D. nihonkaiense* (NC_009463), *D. latum* (AB269325) and *D. latum* (DQ985706).

Alignments, polymorphism analysis, and molecular marker score calculation

Genome sequences mentioned above were divided into four different groups: (1) VibrioDS, containing only one representative genome for each *Vibrio* species, using the *Vibrio cholerae* strain (NC_002505); (2) VibrioSS, comprising the four different *Vibrio cholerae* strains; (3) DiphyllobothriumDS containing one genome per *Diphyllobothrium* species using NC_008945 as *D. latum* representative; (4) DiphyllobothriumSS containing all *D. latum* strains. Each group was aligned using MAUVE v2.3.1 using the progressiveAligner option.[12] Output files were re-formatted to Variscan—extended multi-FASTA (XMFA) format with a custom PERL Script (XMFA.pl) and analyzed using Variscan v2.0.[13] The resulting files were used as an input for the molecular marker score calculation implemented in a custom PERL script (SCORE.pl), and using two different window sizes of 300pb and 500 pb, for *Vibrio* and *Diphyllobothrium*, respectively. The final output

consists of a plain text file listing the potential marker genes, sorted in a descending order of their scores.

Algorithm

The Bohle-Gabaldón (BG) score calculation is based on the level of DNA polymorphism in the Distinct Species (*DS*) group and Same Species (*SS*) groups, as inferred from the average of nucleotide differences per site ($\hat{\pi}$). Not more than one *SS* group may be considered. The Bayes's theorem as adapted by Willcox[11] is used as follows. If the number of genome sequences in *DS* group is lower than 4 and there is no length constraint for the marker, formula (1) is used.

If molecular marker with specific size is required (S_{ref}) formula (2) is used, S_i is the nucleotides length of gene *i*. Also, if the amount of whole-genomes for *DS* group is 4 or more, is possible include Tajima's *D* (D_i) without specific size requirement (3) or with (4), which better account for the possibility of rare haplotypes. Based on Willcox conditions, higher $\hat{\pi}$ in Different Species ($\pi_{i(DS)}$) and lower in Same Species ($\pi_{i(SS)}$) is better. For ($D_{i(DS)}$) in *DS* group more negative values are preferred and, finally, the size of molecular marker (S_{ref}) is arbitrary. In order to reduce sequencing costs we selected rather small sizes (300 pb–500 pb).

BG score using DNA polymorphism (less than 4 genomes):

$$Score_i = \hat{\pi}_{i(DS)}(1-\hat{\pi}_{i(SS)}) \quad (1)$$

Scoring using DNA polymorphism and Size (less than 4 genomes)

$$Score_i^{+Size} = \hat{\pi}_{i(DS)}(1-\hat{\pi}_{i(SS)})\left(\frac{S_i}{S_i+\left|S_{ref}-S_i\right|}\right) \quad (2)$$

Scoring using DNA polymorphism and Tajima's *D*[14] (4 genomes and more):

$$Score_i^{+Tajima\ -Size} = \hat{\pi}_{i(DS)}(1-\hat{\pi}_{i(SS)})\left(-\frac{\hat{D}_{i(DS)}}{2}\right) \quad (3)$$

Scoring using DNA polymorphism, Tajima's *D* and Size (4 genomes and more):

$$Score_i^{+Tajima\ -Size} = \hat{\pi}_{i(DS)}(1-\hat{\pi}_{i(SS)}) \times\left(-\frac{\hat{D}_{i(DS)}}{2}\right)\left(\frac{S_i}{S_i+\left|S_{ref}-S_i\right|}\right) \quad (4)$$

The maximum value for Score is 1 using $\pi_{i(DS)} = 1$, $\pi_{i(SS)} = 0$, Tajima's $D = -2$ and $S_i = S_{ref}$. The minimum value for Score is 0 considering $\pi_{i(DS)} = 0$, $\pi_{i(SS)} = 1$, Tajima's $D = +2$ and $S_i \neq S_{ref}$.

Experimental validation analysis

Additional *Vibrio* sequences for the candidate genes were obtained from biological samples stored in the Collection of Aquatic Important Microorganism (CAIM) at the Center of Research for Nutrition and Development (Mexico). Collected strains were: *V. ordalii* CAIM608, *V. aestuarianus* CAIM592, *V. orientalis* CAIM332, *V. tubiashii* CAIM313, *V. splendidus* CAIM319, *V. cyclitrophicus* CAIM 596, *V. fortis* CAIM629, *V. parahaemolyticus* CAIM320, *V. harveyi* CAIM513, *V. rotiferianus* CAIM577, *V. mytili* CAIM528, *V. navarrensis* CAIM609, *V. fluvialis* CAIM593, *V. agarivorans* CAIM615, *V. mimicus* CAIM602, *V. metschnikovii* CAIM317, *V. vulnificus* CAIM610, *V. aerogenes* CAIM906 and *V. neptunius* CAIM532. Similarly, additional sequences for candidate *Diphyllobothrium* marker genes were obtained from samples fixed in ethanol at the Parasitology Institute of Biology Center of the Czech Republic. These included the strains *D. latum* TS-07/17, *D. pacificum* TS-06/30a.b., *D. dendriticum* TS-04/39, *D. nihonkaiense* TS-06/236, *D. polyrugosum* TS-05/58 and *D. ditremum* TS-02/32.

DNA purification and amplification

Genomic DNA from *Vibrio* species was purified using E.Z.N.A. Bacterial DNA Kit (Omega Biotek, USA). *Diphyllobothrium* samples were diluted (1) in nuclease-free water, macerated with mortar, to subsequently purify DNA using E.Z.N.A. Tissue DNA Kit (Omega Biotek, USA), following manufacturer's instructions. The final volume for PCR were 50 µL with 5 µL Buffer 10x (20 nM Tris-HCl pH 8.0, 40 nM NaCl, 2 mM Sodium phosphate, 0.1 mM EDTA, 1 mM DTT, stabilizers, 50% (v/v) glycerol), 1 µL

Table 1. 10 top-scoring marker genes for *Vibrio* species discrimination using S_i = 300 pb.

Score$_i$	Locus tag	Size (pb)	$\pi_{i(DS)}$	$\pi_{i(SS)}$	Tarima's $D_{(DS)}$
0.00308	VC1988	0.98387	0.03469	0.00000	–0.09022
0.00252	VC1954	0.33667	0.05809	0.00000	–0.12885
0.00238	VC2163	0.78667	0.03703	0.00000	–0.08185
0.00237	VC2354	0.47667	0.04847	0.00000	–0.10258
0.00233	VC2665	0.96667	0.03374	0.00000	–0.07132
0.00222	VC2189	0.59667	0.04396	0.00000	–0.08477
0.00212	VC1986	0.60653	0.04145	0.00000	–0.08437
0.00208	VC2658	0.82189	0.03318	0.00000	–0.07621
0.00207	VC2652	0.56667	0.03689	0.00000	–0.09897
0.00207	VC1534	0.59817	0.04150	0.00000	–0.08352

dNTPs (10 mM), 6 µL $MgCl_2$ (50 mM), 1 µL primers (10 µM), 0.5 µL Platinum *Taq* DNA polymerase (2.5 U), 5 µL template DNA and 31.5 µL free nuclease water. Primers for target gene amplification were designed based on the level of observed sequence conservation. The primers used for *Vibrio* were forward 5′-ATG GTT TCA ATT AAN GGN TTR CCK CC-3′ and reverse 5′-TTA GAT GTA RAK ATC GAC MCC NA-3′ and for *Diphyllobothrium* target gene were forward 5′-ATG ATC TTT AGT GGT TAT TCA -3′ and reverse 5′-CTA ATG GTC CAC TGA AAA TGA TAA TAT-3′. The thermal profile used was the following: initial activation (2 min, 95 °C), followed by 35 cycles of denaturation (1 min, 95 °C), annealing (1 min, 55 °C) and extension (1 min, 72 °C), and a final extension (4 min, 72 °C). Electrophoresis agarose gel (1.5%) stained with Ethidium bromide was used to identify the PCR products from *Vibrio* (~300 pb) and *Diphyllobothrium* (~500 pb). PCR products were purified using Minielute gel extraction kit (QIAGEN, USA) and cloned using CloneJET PCR cloning kit (Fermentas, USA). This kit includes the positive selection cloning vector pJET1.2/blunt that contains a lethal gene which is disrupted by ligation of a DNA insert into the cloning site. As a result, only cells with recombinant plasmids are able to propagate. Finally, DNA from the *E. coli* top 10 colonies was purified using E.Z.N.A. bacterial DNA Kit (Omega Biotek, USA). Total DNA obtained from clones was amplified using primers pJET1.2 forward and reverse (CloneJET, Fermentas, USA) with BigDye Terminator v3.1 Cycle sequencing Kit (Applied Biosystem, USA) using manufacturer's instructions. The PCRs products were purified for Dyes using Dye Terminator Removal kit (Omega Biotek, USA) and sequenced using ABI PRISM 310 machine (Applied Biosystem, USA). The sequences obtained were edited, assembled, aligned and compared using CLC Genomics Workbench v3.5.5 (CLC Bio, Denmark).

Table 2. 10 top-scoring marker genes o *Diphyllobothrium* species discrimination using S_i = 500 pb.

Score	Gen	Size (pb)	$\pi_{i(DS)}$	$\pi_{i(SS)}$
0.01175	ATP6	509	0.01196	0.00013
0.01066	ND6	458	0.01156	0.00015
0.00733	ND3	356	0.00944	0.00019
0.00563	ND4L	260	0.00833	0.00028
0.00524	COX2	569	0.00596	0.00023
0.00479	ND2	878	0.00841	0.00015
0.00433	ND4	1250	0.01083	0.00017
0.00404	ND1	890	0.00719	0.00022
0.00355	ND5	1568	0.01115	0.00047
0.00230	COX1	1565	0.00720	0.00004

Molecular marker discrimination power analysis

To prioritize the markers, we developed a simple Discrimination Power (DP) score (5) based in Bayes's Theorem adapted by Willcox[11] which evaluates the maximum identity (ΔI_i^{max}) for each species in each molecular marker gene (x) analyzed.

$$DP_x = \prod_{i=1}^{n} (1 - \Delta I_i^{\max}) \quad (5)$$

The maximum value for DP is 1 (ie, perfect molecular marker), if maximum difference of identity for the closest species in each species for each molecular marker tends to 0. The minimum value for DP is 0

Table 3. Prokaryotic molecular markers genes comparison using Discrimination power scoring.

Species	Accession number	SC	*recA*			*dnaJ*			*atpA*			*Chromosome segregation ATPase*		
			CSC	Id	(1-Id)	CSC	Id	(1-Id)	CSC	Id	(1-Id)	CSC	Id	(1-Id)
V. aestuarianus	JN040521	**1**	*2*	0.999	0.001	*5*	0.801	0.199	*8*	0.899	0.101	*18*	0.708	0.292
V. alginolyticus	NZ_AAPS01000071	**2**	*1*	0.999	0.001	*15*	0.883	0.117	*14*	0.967	0.033	*14*	0.809	0.191
V. cholerae	NC_002505	**3**	*11*	0.921	0.079	*11*	0.932	0.068	*11*	0.958	0.042	*11*	0.876	0.124
V. coralliilyticus	NZ_ACZN01000015	**4**	*12*	0.971	0.029	*12*	0.904	0.096	*20, 12*	0.957	0.043	*13*	0.779	0.221
V. cyclitrophicus	JN040526	**5**	*18*	0.924	0.076	*18*	0.904	0.096	*18*	0.973	0.027	*18*	0.844	0.156
V. fischeri	NC_006840	**6**	*16*	0.876	0.124	*16*	0.852	0.148	*16*	0.918	0.082	*16*	0.761	0.239
V. fluvialis	JN040529	**7**	*3, 11*	0.861	0.139	*2*	0.848	0.152	*4*	0.889	0.111	*1*	0.673	0.327
V. fortis	JN040527	**8**	*5*	0.882	0.118	*19*	0.853	0.147	*15*	0.942	0.058	*18*	0.802	0.198
V. harveyi	JN040517	**9**	*15*	0.979	0.021	*15*	0.925	0.075	*15*	0.979	0.021	*14*	0.868	0.132
V. metschnikovii	JN040531	**10**	*15*	0.845	0.155	*7*	0.825	0.175	*20*	0.835	0.165	*7*	0.646	0.354
V. mimicus	JN040530	**11**	*3*	0.921	0.079	*3*	0.932	0.068	*3*	0.958	0.042	*3*	0.876	0.124
V. neptunius	JN040535	**12**	*4*	0.971	0.029	*4*	0.904	0.096	*4*	0.957	0.043	*19*	0.577	0.423
V. orientalis	JN040523	**13**	*17*	0.893	0.107	*13*	0.851	0.149	*19*	0.974	0.026	*19*	0.787	0.213
V. parahaemolyticus	JN040516	**14**	*9*	0.917	0.083	*2*	0.867	0.133	*15*	0.970	0.030	*9*	0.868	0.132
V. rotiferianus	JN040518	**15**	*9*	0.979	0.021	*9*	0.925	0.075	*9*	0.979	0.021	*14*	0.774	0.226
V. salmonicida	NC_011312	**16**	*6*	0.876	0.124	*6*	0.852	0.148	*6*	0.918	0.082	*6*	0.761	0.239
V. shilonii	NC_ABCH01000040	**17**	*13*	0.893	0.107	*13*	0.826	0.174	*15*	0.912	0.088	*1*	0.539	0.461
V. splendidus	JN040524	**18**	*5*	0.924	0.076	*5*	0.904	0.096	*5*	0.973	0.027	*5*	0.844	0.156
V. tubiashii	JN040522	**19**	*9*	0.886	0.114	*14, 8*	0.853	0.147	*13*	0.935	0.065	*13*	0.787	0.213
V. vulnificus	JN040533	**20**	*9, 11*	0.859	0.141	*9*	0.842	0.158	*17*	0.904	0.096	*9*	0.700	0.3
Discrimination power score			**7.980×10^{-27}**			**3.530×10^{-19}**			**1.070×10^{-26}**			**6.310×10^{-14}**		

Notes: Underline Score is highest. JN040516-JN040535: In this work.
Abbreviations: SC, Specie code; CSC, Closest specie code; Id, Identity (Match nucleotides/total nucleotides).

when the level of identity of that marker in the closest species tends to 1 for each species.

Results

Automated prioritization of marker genes

Publicly available genomes from *Vibrio* and *Diphyllobothrium* were downloaded and subjected to the selection of marker genes approach aforementioned. For each genus, a list of potential marker genes sorted in descending order of their BG scores was produced. For *Vibrio* species (Table 1), the best molecular marker is a protein-coding gene with locus tag VC1988 in chromosome 1 of the reference genome *V. cholerae* NC_002505. This gene encodes a chromosome segregation ATPase, a protein essential for cell division that forms part of a chromosomal segregation complex. In the case of *Diphyllobothrium*, the analysis of completely sequenced mitochondrial genomes revealed the gene encoding the subunit 6 of the ATPase complex as the best potential marker gene (Table 2). This enzyme is part of the mitochondrial oxidative phosphorylation and is essential for the generation of ATP.[15]

Experimental Validation

In order to validate the effectiveness of our approach we amplified these marker genes from additional strains of known taxonomic assignment but with no current genomic sequences available. The effectiveness of the markers, as measured by the Discrimination Power score (DP) described above, was compared to that of common markers used previously for these species. These were *atpA,*[16] *dnaJ*[17] and *recA*,[18] for *Vibrio* and 18*S r*RNA, COX1 and 18*s r*RNA + ITS + 5.8*s r*RNA[19,20] for *Diphyllobothrium*.

Twenty new sequences were obtained from the chromosome segregation ATPase gene in different *Vibrio* species. Remarkably, this gene showed the best Discrimination Power value (Table 3) with a *DP* score of 6.3×10^{-14}. Standard markers showed lower discrimination powers: *dnaJ* ($DP_{dnaJ} = 3.5 \times 10^{-19}$), *atpA* ($DP_{atpA} = 1.1 \times 10^{-26}$) finally *recA* ($DP_{recA} = 7.9 \times 10^{-27}$). In the case of *Diphyllobothrium,* seven new sequences were obtained from ATPase-subunit 6 (ATP6) gene in different species. Again, the marker gene selected by our approach presented the highest Discrimination

Table 4. Eukaryotic molecular markers genes comparison using Discrimination power scoring.

Species	Accession number	SC	18s rRNA			COX1			18s + ITS + 5.8s rRNA			ATPase6		
			CSC	Id	(1-Id)	CSC	Id	(1-Id)	CSC	Id	(1-Id)	CSC	Id	(1-Id)
D. dendriticum	JN040538	1	*2*	0.999	0.001	*4*	0.936	0.064	*2*	0.999	0.001	*3*	<u>0.935</u>	0.065
D. ditremum	JN040539	2	*1*	0.999	0.001	*3*	0.902	0.098	*1*	0.999	0.001	*1,3*	<u>0.892</u>	0.108
D. latum	JN040536	3	*1,2*	0.999	0.001	*4*	<u>0.905</u>	0.095	*1,2*	0.989	0.011	*1*	0.935	0.065
D. nihonkaiense	JN040540	4	*1,2,3*	0.996	0.004	*3*	0.935	0.065	*1,2*	0.973	0.027	*1*	<u>0.902</u>	0.098
D. pacificum	JN040541	5	*1,2,3,4*	0.964	0.036	*2*	0.849	0.151	*1,2*	0.853	0.147	*1,3*	<u>0.823</u>	0.177
Discrimination power value			**1.440×10^{-13}**			**5.848×10^{-6}**			**4.365×10^{-11}**			**<u>7.914×10^{-6}</u>**		

Notes: Underline Score is higher. JN040536-JN040541: In this work.
Abbreviations: SC, Specie code; CSC, Closest specie code; Id, Identity (Match nucleotides/total nucleotides).

power (DP_{ATP6} = 7.9 × 10^{-6}), followed by COX1 (D_{PCOX1} = 5.8 × 10^{-6}), ITS rRNA(DP_{ITS} = 4.4 × 10^{-11}) and 18s rRNA(DP_{18sRNA} = 1.4 × 10^{-13}) (Table 4).

Discussion

We have proposed and validated a novel approach for the informed selection of marker genes based on the observed levels of DNA polymorphism[10] among whole genomic sequences. Our results indicate that our approach effectively selects marker genes for species differentiation. Besides having greater discrimination powers than traditional markers, our markers also reduced the number of species that showed identical sequences for the marker. Nevertheless, in both genera studies, there are still some species that are too closely related to be differentiated with a single marker. The use of a combination of markers, or the selection of specific markers for that group of species within the genus would be required. Our approach has some minimal requirements. For instance, if the goal is to obtain marker genes for species differentiation in a given genus, a minimum of three different strain genomes belonging to two different species within the genus is required. Moreover, the design of primers may present problems if the sequences are too divergent, although this problem is shared with other approaches.

Our approach and scoring system method provides a new, powerful tool for the exploitation of available genome sequences to assist in the selection of marker genes. In both the eukaryotic and prokaryotic genera tested, the theoretical analyses showed excellent correlation with empirical results and showed a better performance than molecular markers previously proposed by different authors for the same species. The adaptation of Bayes theorem permitted the use of a conditioned statistic that prioritizes genes showing low DNA polymorphism inside the same species (different strains), while displaying high DNA polymorphism between different species.

Acknowledgements

We would like to thank Dr. Bruno Gomez-Gil for the donation of fixed biological material from different *Vibrio* species and Professor Dr. Tomáš Scholz for the donation of fixed biological material from different *Diphyllobothrium* species. We would like to thank for Dr. Patricio Bustos from ADL Diagnostic Chile Ltda for economic support in empirical analysis. TG research is supported in part by a grant from the Spanish Ministry of Science (BFU2009-09168).

Author Contributions

Conceived and designed the experiments: HB, TG. Analysed the data: HB, TG. Wrote the first draft of the manuscript: HB, TG. Contributed to the writing of the manuscript: HB, TG. Agree with manuscript results and conclusions: HB, TG. Jointly developed the structure and arguments for the paper: HB, TG. Made critical revisions and approved final version: HB, TG. All authors reviewed and approved of the final manuscript.

Disclosures and Ethics

As a requirement of publication author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Pearson T, Okinaka RT, Foster JT, Keim P. Phylogenetic understanding of clonal populations in an era of whole genome Sequencing. *Genetics and Evolution*. 2009;9:1010–9.
2. Gürtler V, Mayall BC. Genomic approaches to typing, taxonomy and evolution of bacterial isolates. *International Journal of Systematic and Evolutionary Microbiology*. 2001;51:3–16.
3. Thompson FL, Gevers D, Thompson CC, et al. Phylogeny and molecular identification of vibrios on the basis of multilocus sequence analysis. *Applied Environmental Microbiology*. 2005:5107–15.
4. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology*. 2008;6:431–40.
5. Bapteste E, Boucher Y, Leigh J, Doolittle WF. Phylogenetic reconstruction and lateral gene transfer. *TRENDS in Microbiology*. 2004;12(9):406–11.
6. Binnewies TT, Motro Y, Hallin PF, et al. Ten years of bacterial genome sequencing: Comparative-genomics-based discoveries. *Functional & Integrative Genomics*. 2006;6(3):165–85.
7. Hernandez D, François P, Farinelli L, Osterås M, Schrenzel J. De novo bacterial genome sequencing. Millions of very short reads assembled on a desktop computer. *Genome Research*. 2008;18:802–9.

8. Zeigler DR. Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International Journal of Systematic and Evolutionary Microbiology*. 2003;53:1893–900.
9. Phillippy AM, Ayanbule K, Edwards NJ, Salzberg SL. Insignia: A DNA signature search web server for diagnostic assay development. *Nucleic Acids Research*. 2009;37:229–34.
10. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceeding of the National Academic of Science of the United States of America*. 1979;76:5269–73.
11. Willcox WR, Lapage SP, Bascomb S, Curtis MA. Identification of Bacteria by Computer: Theory and Programming. *Journal of General Microbiology*. 1997;77:317–30
12. Darling AE, Mau B, Perna NT. ProgressiveMauve. Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. *PLoS One*. 2010;5(6):e11147.
13. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. *Bioinformatics*. 2005;21:2791–3.
14. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585–95.
15. Lee SH, Lee S, Jun HS, et al. Expression of the mitochondrial ATPase6 gene and T*fam* in Down Syndrome. *Molecules and Cells*. 2002;15(2):181–5.
16. Thompson CC, Thompson FL, Vicente AC, Swings J. Phylogenetic analysis of vibrios and related species by means of *atpA* gene sequences. *International Journal of Systematic and Evolutionary Microbiology*. 2007;57:2480–4.
17. Nhung PH, Shah MM, Ohkusu K, et al. The *dnaJ* gene as a novel phylogenetic marker for identification of *Vibrio* species. *Systematic and Applied Microbiology*. 2007;30:309–15.
18. Thompson CC, Thompson FL, Vandemeulebroecke K, Hoste B, Dawyndt P, Swings J. Use of *recA* as an alternative phylogenetic marker in the family *Vibrionaceae*. *International Journal of Systematic and Evolutionary Microbiology*. 2004;54:919–24.
19. Jeon HK, Kim KH, Huh S, et al. Morphologic and Genetic Identification of *Diphyllobothrium nihonkaiense* in Korea. *Korean Journal of Parasitology*. 2009;47(4):369–75.
20. Scholz T, Garcia HH, Kuchta R, Wicht B. Update on the human broad tapeworm (Genus *Diphyllobothrium*), including clinical relevance. *Clinical Microbiology Reviews*. 2009:146–60.

Supplementary Data

The scoring system and the necessary re-formatting scripts have been implemented in PERL. The PERL scripts (SCORE.pl and XMFA.pl) and a user manual for Windows, Linux and Mac are available at http://www.bioinformatics.cl.