Published online 19 April 2008

Nucleic Acids Research, 2008, Vol. 36, No. 10 ***e56***
doi:10.1093/nar/gkn191

UniPrime: a workflow-based platform for improved universal primer design

Michaël Bekaert* and Emma C. Teeling

School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland

Received December 18, 2007; Revised March 19, 2008; Accepted April 1, 2008

ABSTRACT

UniPrime is an open-source software (http://uniprime.batlab.eu), which automatically designs large sets of universal primers by simply inputting a gene ID reference. UniPrime automatically retrieves and aligns homologous sequences from GenBank, identifies regions of conservation within the alignment and generates suitable primers that can amplify variable genomic regions. UniPrime differs from previous automatic primer design programs in that all steps of primer design are automated, saved and are phylogenetically limited. We have experimentally verified the efficiency and success of this program by amplifying and sequencing four diverse genes (*AOF2*, *EFEMP1*, *LRP6* and *OAZ1*) across multiple Orders of mammals. UniPrime is an experimentally validated, fully automated program that generates successful cross-species primers that take into account the biological aspects of the PCR.

INTRODUCTION

Comparing the genomic structure and content of phylogenetic and ecologically diverse taxa will act as a 'Rosetta' stone allowing us to annotate and interpret our own genome (1–4). Evolutionary analyses of whole genomes and targeted genes sequenced in divergent species have advanced our understanding of the patterns of human disease mutations in many inherited diseases and cancers (5–7). Comparative genomics is a powerful and expanding field, which is evident from the exponential increase in the number of non-human sequence entries in GenBank and EMBL within the past decade. GenBank has doubled in size about every 18 months. It currently contains over 65 billion nucleotide bases from more than 61 million individual sequences, with 15 million new sequences added in the past year (8). Since the completion of the human sequence project in 2001 (9,10), the number of whole genomes sequenced or in the process of being sequenced is also increasing. Currently, over 617 genomes are completed, 531 are in assembly stage, 652 are in process and 421 have been approved for whole genome sequence (11). Researchers are faced with vast quantities of molecular data that can only be stored, analysed and mined with appropriate computer-based algorithms and programs. Subsequently, the bioinformatics software used to mine these data are also increasing exponentially (4).

Although many species will be sequenced at the whole genome level, this number still only represents a small fraction of the diversity of life: e.g. 107/5000 mammals either sequenced, in the process or accepted to be sequenced (11,12). Therefore, smaller comparative genomic projects that target key genes in key taxa (4) will still play a large role in future comparative studies. Good primer design is a crucial step in any comparative genomics project and ensures specificity and efficiency of target amplification, necessary to achieve reliable PCR results. Traditionally, universal primers were designed by first generating a multispecies alignment then, manually identifying conserved regions in that alignment, finally an algorithm was used to estimate the melting temperature of candidate primers sequences within the conserved regions. This is a laborious process with many defined user steps, including downloading and aligning sequences of phylogenetic interest.

Recent advances in bioinformatics primer design software have increased the speed of some steps within this process. Most programs will automatically design compatible forward and reverse PCR primers from an inputted sequence [e.g. Primer3 (13), Oligo 6 (http://www.oligo.net), AutoPrime (14), CODEHOP (15), ExonPrimer (http://ihg.gsf.de/ihg/ExonPrimer.html)], or an inputted user defined multiple alignment [e.g. PrimaClade (16), Primer Premier (http://www.premierbiosoft.com)]. A recent program QPRIMER (17) generates universal primers within exons by automatically creating multi-genome alignments of human, mouse, rat, chicken, dog, zebrafish and fruit fly. Although, all of these programs enhance the speed and accuracy of primer design, none automate all steps in the process for all regions within the genome. To our knowledge, no automatic primer design program uses phylogenetic information to retrieve and align all homologous

*To whom correspondence should be addressed. Tel: +353 1 716 2263; Fax: +353 1 716 1152; Email: michael@batlab.eu

sequences from GenBank. They do not automatically assess levels of variation across the alignment, design non-degenerate primers, nor estimate the possibility of false positive amplifications.

To address these problems, we have designed the UniPrime program, an open-source suite where users can automatically design sets of universal primers to amplify regions of suitable inter-specific variation across divergent taxa, by simply inputting a reference sequence or accession number. UniPrime uniquely allows all steps of the process to be saved, including initial data retrieval, multispecies alignment and primer design sites. We have experimentally verified the efficiency and success of this program by amplifying and sequencing four diverse genes (*AOF2*, *EFEMP1*, *LRP6* and *OAZ1*) across multiple Orders of mammals. The program is available at http://uniprime.batlab.eu and is licensed under the Creative Commons GNU GPL.

METHODS

Algorithm

We designed an automated method to: (i) design successful universal primers for any given locus; (ii) maximize the number of variable positions in the fragment amplified given a user defined similarity threshold; (iii) evaluate the 'mis-priming' potential of the primers generated. This process is defined in five steps and shown schematically in Figure 1. Each step can be reviewed by the companion web interface to assess potential challenges resulting from sequencing error, incorrect annotation or an artifactual duplication of the locus. The main module of UniPrime uses a command line interface. The commands are simple, they are detailed in the 'readme' file and the number of parameters requested is small. Each step is described subsequently.

Step 1—Initial sequence. Initially, the GenBank GeneID (unique identifiers for genes provided by Entrez Gene) of the target locus (protein coding gene, snRNA, etc.) is input by the user. This reference code is used to retrieve the sequences related to this target locus (the genomic DNA and the mRNA sequences). The program then selects a single nucleotide sequence (proto-type sequence), which is usually the reference mRNA sequence of the longest isoform of the gene. If no mRNA sequence is known for the gene, the reference genomic DNA sequence is then used. If required, the user can enforce the use of the DNA rather than mRNA sequence. In such cases, the quality of the alignment is less accurate with multiple insertions/deletions introduced due to introns, but the overall scheme of the primer design method is unchanged.

Step 2—Search for orthologues. The prototype sequence is used as a 'query' for a Blastn (18) search of the NCBI database to identify highly similar homologous sequences. The user can delimit the search at varying phylogenetic levels by incorporating the Entrez Query BLAST option. The NCBI 'Reference genomic sequence' (RefSeq) database (19) is used as the default search database. An *e*-value

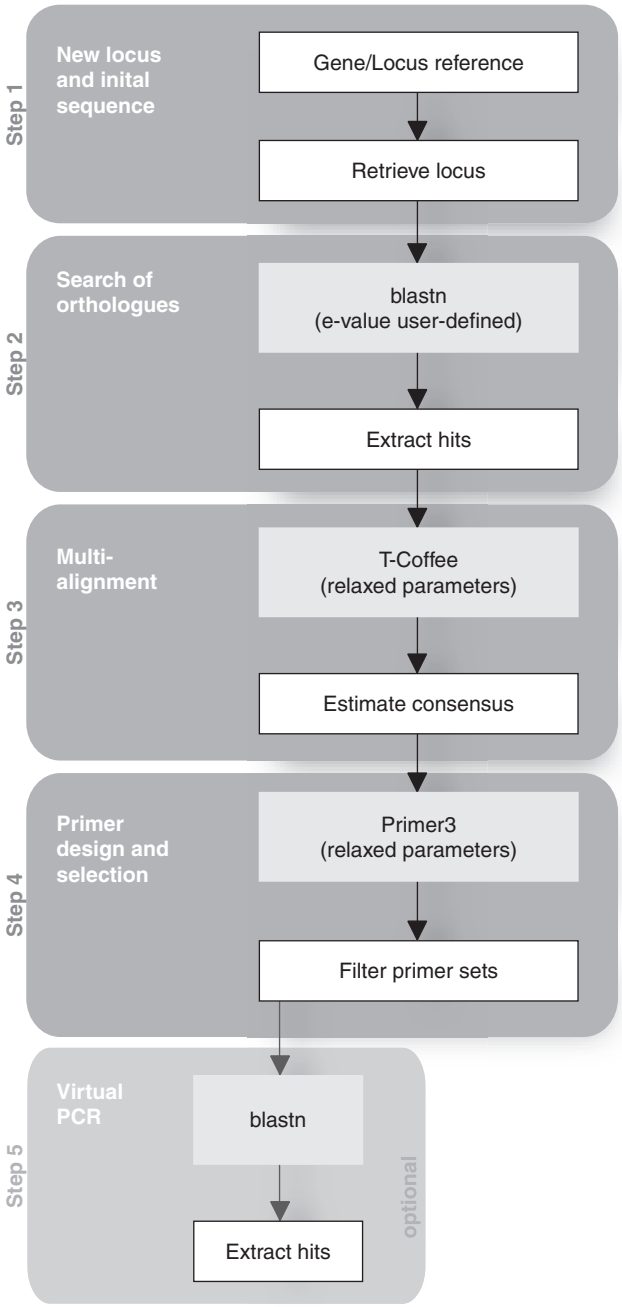

Figure 1. Schematic diagram of the UniPrime algorithm. Steps carried out by UniPrime are shown in white boxes. Steps performed by external programs, are shown in grey boxes. The five main functions of UniPrime are highlighted.

threshold of 10^{-100} is incorporated as the default cut-off point for valid hits. Both parameters are user-defined and can be modified. Due to problems with variable intron length and number within divergent taxa, this step preferentially uses an mRNA prototype sequence when available; otherwise, the DNA sequence of target locus is used. If the prototype sequence is mRNA, then only mRNA sequences will be retrieved. Likewise, if the prototype sequence is DNA then only DNA sequences will be retrieved. The sequences with the best *e*-value score

for each species within the database are retrieved, and stored in the database. Using this method, we hope to retrieve orthologous sequences; however, as our retrieval method is based only on a high blast score, this step may also gather closely conserved paralogous sequences. UniPrime allows the user to review the retrieved sequences and remove any sequences that are considered paralogous before primer design. Also, users can further modify the phylogenetic distance between the taxa they want incorporated in the primer design step by including or removing the required taxa.

Step 3—multi-species alignment. The stored sequences are concatenated into a single file, which is then imported into the alignment program T-Coffee (20). Users can restrict the taxa they want to be included in this input file. The sequences are aligned using the default parameters in T-Coffee (20). From this alignment, a consensus sequence is inferred with a conservative default threshold of 60% (i.e. only the nucleotide which occurs >60% at a single position in the alignment is represented in the consensus sequence, otherwise an N is reported). Only A, T, C, G and N are used in the consensus sequence. As the number of non-N-sites in the consensus sequence is limited by the threshold, increasing this threshold should lower the number of primers designed but increase their specificity. Likewise, decreasing this threshold should increase the number of primers designed but lower their specificity. If required the user can manually incorporate their own alignment.

Step 4—primer design. All possible primers along the consensus sequence are generated by Primer3 (13) using the following parameters: melting temperature (Tm) 55–65°C; primer length 20–25 bp; GC clamp (G or C at 3′ end); GC content 40–60%; and, an optimum product size of 600 bp. The user can vary the product size. To ensure that all possible primers are defined along the consensus sequence, they are generated using a sliding window approach. The sliding window of 250 bp moves along each nucleotide position in the consensus sequence and primers are designed where possible. By default, sites that are over 40% variable within the alignment are defined as Ns; therefore, primers cannot be generated within these regions. Users can change this threshold, thereby controlling the level of expected variability within the resulting amplicon. To select the optimal primer pairs each primer sequence is examined against all input sequences using the following filters: (i) the entire primer sequence must be found in all input sequences but can have a mis-match of 20% with each sequence; (ii) the last 5 bp at the 3′ end of each primer must be 100% conserved in all input sequences. This is an essential filter, as DNA polymerase cannot bind efficiently to a template if there are mismatches at the 3′ end of a primer, regardless of the specificity in 5′ end region. If the last two residues of the primer do not match the template, then no amplification can occur (21).

Step 5—virtual PCR. The possibility of amplifying non-target sequences using the proposed primer pairs is assessed in an optional last step by completing a 'virtual PCR'. The primer sequences are submitted for a Blastn search within the 'Whole-genome shotgun reads' (wgs) databases of GenBank to identify sequences that match the forward and reverse primer sequence within a compatible size range for PCR amplification (primers <10 kb apart).

All steps and results are stored in the UniPrime database and their details can be viewed using the web-based interface.

Computer implementation

We implemented this algorithm in a software suite called UniPrime. The basic algorithm requires Bioperl 1.4 (22), T-Coffee 3.2 (20), PostgreSQL 7 (http://www.postgresql.org) and Primer3 (13) version 1.0. Newer versions of these programs can also be used. Primer3 version 1.1.1 is used as it includes an advanced Tm calculation method (23).

Web interface

The intermediate results for each step of the UniPrime suite can also be viewed, accessed and modified through the web companion interface. This is implemented in PHP script using PHP 4.3 or above (http://www.php.net) and any PHP compliant web server.

Laboratory verification

We used UniPrime to generate primers from four diverse genes. We validated the primers designed by amplifying and sequencing five fragments from these genes in five divergent Orders across Class Mammalia. The sequences retrieved and used by UniPrime are available in the Supplementary data and belong to *Bos taurus* (cow), *Canis lupus familiaris* (dog), *Homo sapiens* (man), *Macaca mulatta* (*macaque*), *Monodelphis domestica* (opossum), *Mus musculus* (mouse), *Pan troglodytes* (chimp) and *Rattus norvegicus* (rat).

Taxa and genes

The genomic DNA from six divergent mammal species was used (Table 1). The four genes selected (Table 2) have diverse functions and show different variability levels as estimated by DnaSP 4.10.9 (24). The five selected primer sets are shown in Table 2.

Table 1. Mammalian species used

Species name	Abbrev.	Common name	Order
Myrmecophaga tridactyla	Mtri	Giant anteater	Xenarthra
Ornithorhynchus anatinus	Oana	Platypus	Monotremata
Rousettus lanosus	Rlan	Long-haired rousette	Chiroptera
Tragelaphus eurycerus	Teur	Bongo	Cetartiodactyla
Tupaia minor	Tmin	Pygmy tree shrew	Scandentia
Euphractus sexcinctus	Esex	Six-banded armadillo	Xenarthra

For each species the common name and the Order name are indicated (12).

Table 2. Primers used and estimated product length

Gene	GeneID		Primers (5′ → 3′)	CORE index	Size	Location (human)
AOF2	23028	F	ATGCAGTTCTCTGTACCCTTCC	41	550	Exon 15–Exon 16
		R	AACATGCCCNAACAAATTGAC			
EFEMP1	2202	F	GCATTGCAAAACTCTGTATGG	37	650	Intron 6–Exon 7
		R	TACCTTCACAGTTGAGCCTGTC			
LRP6 (1)	4040	F	ATCAGNTCCCTCAGTATCATGG	37	800	Exon 21–Exon 22
		R	TAATGTGATCGCTCTGTGG			
LRP6 (2)	4040	F	GAACTCAATTGTCCTGTNTGCTC	37	1200	Exon 18–Exon 19
		R	CAGTTCATCTGANTTGTCACTGC			
OAZ1	4946	R	TCCCTNCACTGCTGTAGTAACC	33	550	Exon 2–Exon 3
		F	CNGGGATCTCGATGTAGAGG			

For each gene/locus the forward (F) and reverse (R) primers are indicated. The 'GeneID' was the entry used for the initial step. The 'CORE index' has been provided by T-coffee, and is the reliability score of the alignments. The expected product length is an average value inferred from the multiple alignment but varies between species. Two sets of primers were used for the *LRP6* gene.

Table 3. Model and parameters

	Model	ti/tv ratio	α-shape	p-inv
AOF2	HKY + I + G	2.2047	4.1884	0.1939
EFEMP1	HKY + I + G	1.8285	1.6702	0
LRP6 (1)	HKY + I + G	1.6993	5.5588	0.0800
LRP6 (2)	HKY + I + G	1.3773	2.9999	0.0905
OAZ1	HKY + I + G	2.0732	0.5506	0.1105

Optimal model and parameters settings of sequence evolution estimated by Modeltest and used to establish the maximum likelihood bootstrap consensus trees. ti/tv, Transition/transversion ratio; α-shape, shape of the distribution; p-inv, proportion of invariable sites.

PCR and DNA sequencing

PCR was performed with 2 nM of each primer, 1.5 mM $MgCl_2$, 1 U of Platinum Taq DNA polymerase (Invitrogen Corporation, Carlsbad, California, USA), and 10 ng of genomic DNA. Touchdown conditions of amplification were used for all species, as follows: 10 cycles of denaturation at 95°C for 30 s, annealing at 65°C for 30 s −1°C per cycle, extension at 72°C for 60 s; followed by 35 cycles with 95°C for 30 s, annealing at 55°C for 30 s, extension at 72°C for 60 s. The initial denaturation step and the last extension step were 3 min each. The PCR products were separated and visualized in 1% agarose gel. Sequencing reactions were performed in both directions on PCR products, using the same primer set as for amplification.

Sequence validation

Newly generated sequences were concatenated and aligned, using T-Coffee (20), with the original sequences used to generate the primers. Maximum likelihood (ML) analyses were performed for each data set with PAUP 4.0b10 (25) using the parameters settings (Table 3) for the optimal model of sequence evolution as estimated by Modeltest (26). Starting trees were obtained via neighbor-joining (NJ). 100 ML bootstrap analyses were performed using tree-bisection and recombination branch swapping. *Ornithorhynchus anatinus* (platypus) was used as an outgroup in all analyses apart from *EFEMP1*, where *Monodelphis domestica* (opossum) was used.

Supporting information

The generated sequences were deposited in GenBank (Table 4).

RESULTS AND DISCUSSION

Empirical evaluation of the optimal number of input sequences required to successfully design universal primers

The design of optimal primers relies on the quality of the alignment and directly depends on the number of homologous sequences initially retrieved. We assessed how many optimal primer pairs UniPrime could design when the number of input sequences was varied. From 200 randomly selected mammalian genes, we identified and retrieved their homologous genes using UniPrime. The maximum number of species retrieved was 10 as only 10 mammalian genomes were annotated in the default RefSeq database at time of analyses. All possible combinations of retrieved sequences were generated for each gene and every combination was used to design primers: i.e. for each gene, regardless of phylogenetic affinity, we systematically jack-knifed the number of species present from 10 to one and each iteration was used to design primers. Figure 2A shows the correlation between the number of input sequences and the average number of primers identified per kilobase pair. The data fits with an inverse function ($R^2 = 0.97$): above five sequences, the number of primer sets reached an asymptotical value of 0.5. Therefore, only five to six initial sequences are necessary to design optimal primers using UniPrime.

This result has only been established for mammals [last shared a common ancestor ~220 MYA; (12)] and for the default consensus threshold of 60%. This result will vary at different Linnaean ranks and among phylogenetically diverse organisms. (Figure 2C). UniPrime will work with any data set regardless of genetic diversity as long as it is possible to create an alignment and thus a consensus sequence. The quality of the alignment will dictate the number of primers found, and this quality is dependent on both phylogenetic similarity and genetic diversity of the taxa or genes being used (Figure 2). UniPrime was consistently able to generate a similar amount of primers

Table 4. Genbank Accession numbers

	OAZ1	*LPR6* (1)	*LPR6* (2)	*EFEMP1*	*AOF2*
Euphractus sexcinctus	EF674548	EF674524	na	EF674539	EF674533
Rousettus lanosus	EF674547	EF674526	EF674529	EF674542	EF674535
Myrmecophaga tridactyla	EF674546	EF674525	EF674528	EF674541	EF674534
Tragelaphus eurycerus	EF674545	EF674522	EF674530	EF674543	EF674536
Ornithorhynchus anatinus	EF674544	EF674527	EF674531	na	EF674538
Tupaia minor	na	EF674523	EF674532	EF674540	EF674537

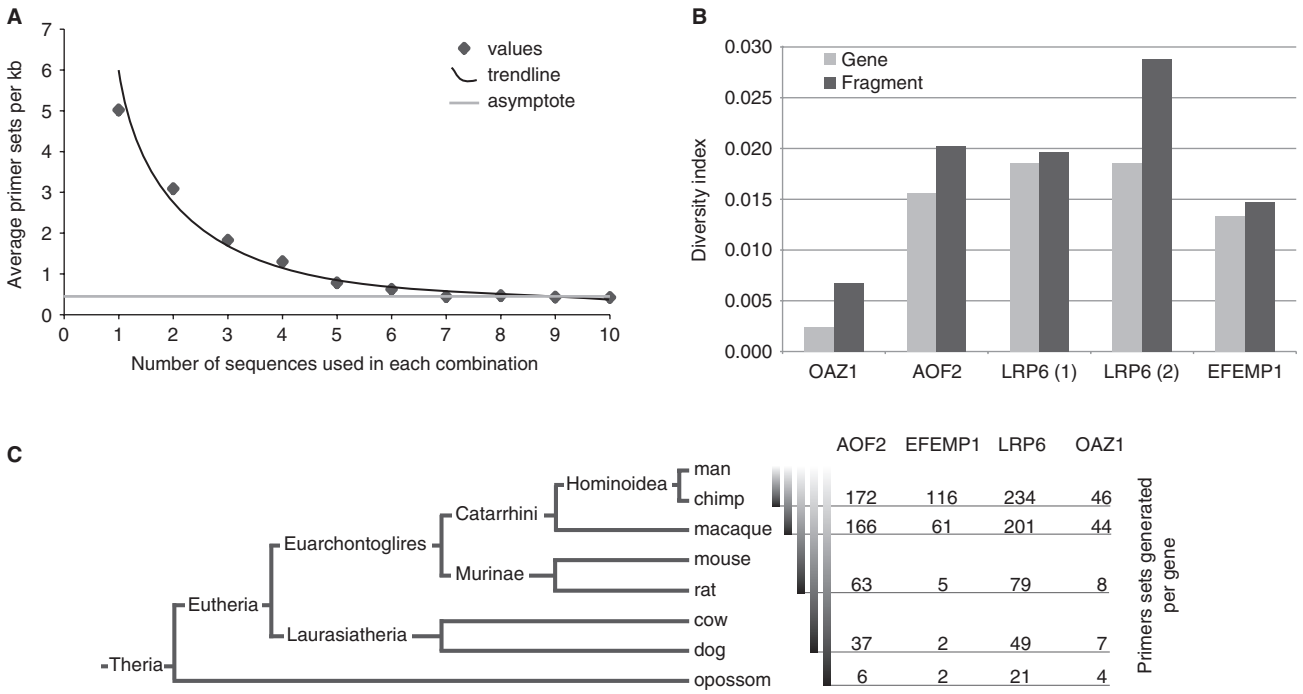

Figure 2. (**A**) Impact of the number of sequences used for the alignment on the primer selection. The number of sequence used for the multi-alignment step varied from 1 to 10. The number of primer sets selected is expressed in primer sets per kb. The 200 random mammalian genes were used from 1 to 300 kb-long. The curve is asymptotic to a value of 0.5 primer set per kilobase (grey line). (**B**) Sequence diversity. Diversity indices (average number of variation per site) across the alignment of the entire gene (light grey) and of the amplified fragment (dark grey). The amplified regions are more variable than the complete gene as is expected due to the primer design process. (**C**) Primers sets generated per gene. For each gene, the number of primer sets generated is assessed across phylogenetic distances.

per kilobase pair regardless of gene function and structure, indicating that this is a robust and reliable primer design method. Interestingly, there is no difference in the number of primers designed when more than five input sequences are used. As at least seven whole mammalian genomes are available, it appears that it is possible to create reliable mammalian primers for coding regions using this method. Therefore, this is an invaluable tool for future phylogenetic and comparative studies.

Laboratory benchmark

We randomly selected a total of four genes that are well studied, found throughout the genome and show a variable degree of diversity (Figure 2B) within the human population and among mammals: (i) *AOF2* (also called *LSD1*) is a component of several histone deacetylase complexes and silences genes by functioning as a histone demethylase (27); (ii) *EFEMP1*, an extracellular matrix protein expressed in retina (28); (iii) *LRP6*, a low density lipoprotein receptor protein and putative tumour suppressor of leukaemia (29) and (iv) *OAZ1*, the ornithine decarboxylase antizymes, that regulates polyamine synthesis (30). Among the Class Mammalia, we applied our algorithm to these four genes to design five primer pairs providing amplification products of about 600 bp (Table 2 and Supplementary material for the alignments used). Despite the high levels of evolutionary divergence (~220 MYA) among our input sequences, all genes were successfully amplified, sequenced and verified using phylogenetic analyses (Figure 3). The gene trees obtained for these sequences were congruent with the established species phylogeny (Figure 3).

Test design

For *OAZ1*, the sequences retrieved and primers generated as an output by UniPrime are shown in Figure 4.

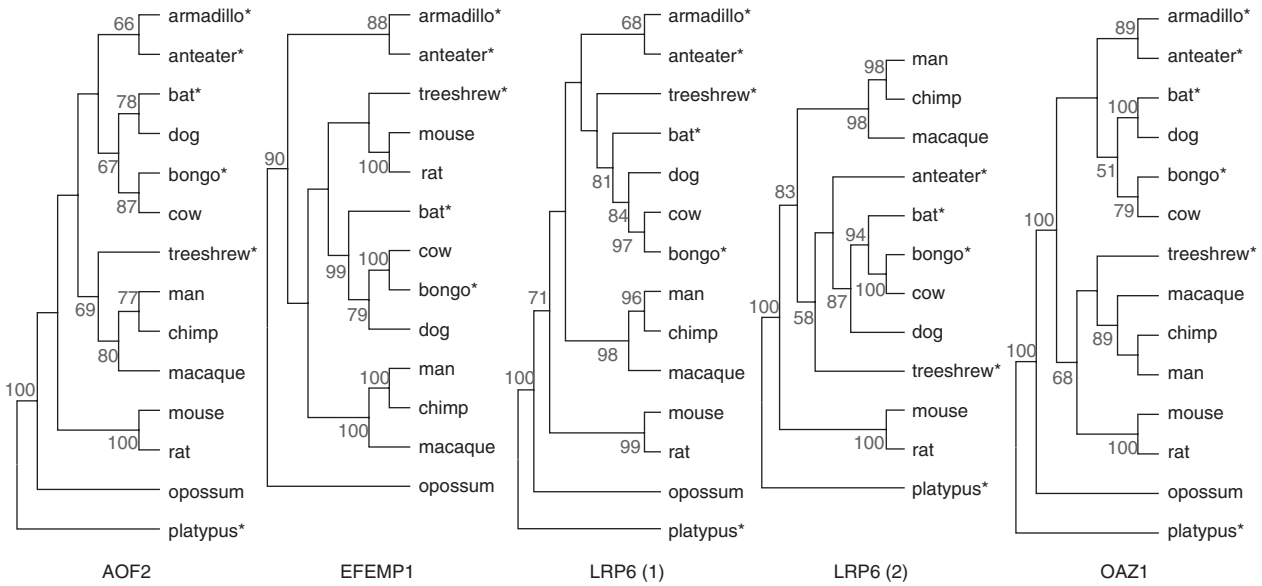

Figure 3. Benchmark ML bootstrap consensus trees based on our amplified fragments (indicated by *) and retrieved aligned sequences. Bootstrap support over 50% is shown.

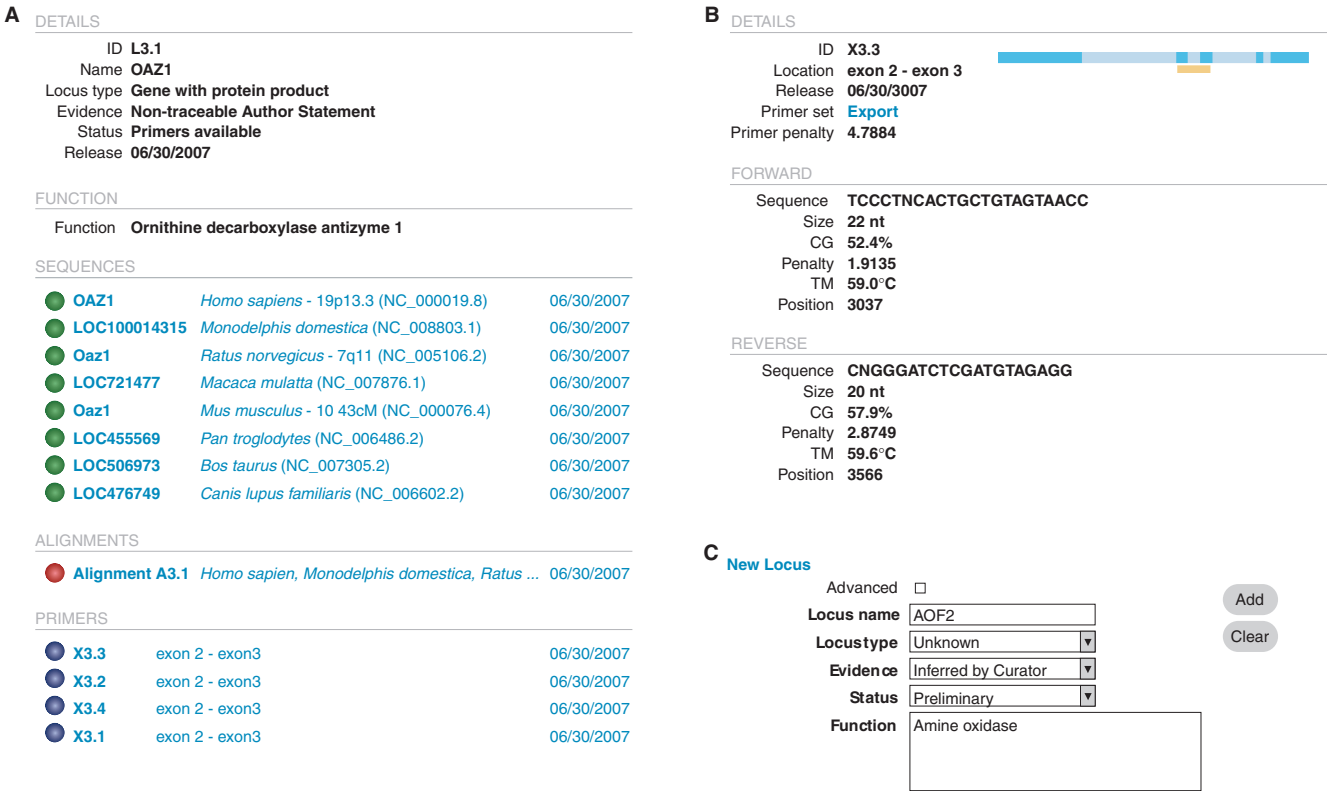

Figure 4. Screenshot of UniPrime web interface. (**A**) Detail of the output for the OAZ1 locus, including the sequences retrieved, the alignment and the primers available. (**B**) Details of one the primer sets generated. (**C**) Screenshot of the manual entry of a new locus (step 1).

Detailed information including related references for each sequence, alignment or primer pair are available, and can be reviewed through the web companion interface Figure 4A. Users can update, add or remove a locus, sequence alignment or primers through the web-based interface (Figure 4B and C). The response time of the web companion interface is nearly instantaneous, regardless of the quantity of information stored in the database or the number of target loci analysed. The primer generation time depends on the response time of the NCBI server but on average takes <10 min per locus (not including step 5—'virtual PCR').

To date, UniPrime has been implemented in four phylogenetically diverse projects at University College Dublin, Ireland. UniPrime is user friendly and has generated over 100 primer sets among which 90% have successfully amplified the target locus (data not shown). One of unique qualities of UniPrime is that only the Gene reference ID is required to enable the user generate a full suite of universal primers. UniPrime also stores and allows easy access to wealth of information via the web interface. UniPrime is an attractive alternative to the long and troublesome steps required for manual retrieval and alignment of homologous sequence from databases. UniPrime represents a new generation of primer design programs that builds on previous programs, automates all steps, enables great user versatility and efficiently mines the ever-expanding genomic databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Alisha Goodbla for her assistance with laboratory work, William J. Murphy and two anonymous reviewers for their constructive comments. This work was supported by a Science Foundation Ireland PIYRA 06/YI3/B932, award to ECT. Funding to pay the Open Access publication charges for this article was provided by University College Dublin, Seed Funding Scheme.

Conflict of interest statement. None declared.

REFERENCES

1. Margulies,E.H., Blanchette,M., Haussler,D. and Green,E.D. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res.*, **13**, 2507–2518.
2. Murphy,W.J., Pevzner,P.A. and O'Brien,S.J. (2004) Mammalian phylogenomics comes of age. *Trends Genet.*, **20**, 631–639.
3. O'Brien,S.J. and Fraser,C.M. (2005) Genomes and evolution: the power of comparative genomics. *Curr. Opin. Genet. Dev.*, **15**, 569–571.
4. Tuggle,C.K., Dekkers,J.C. and Reecy,J.M. (2006) Integration of structural and functional genomics. *Animal Genet.*, **37 (Suppl. 1)**, 1–6.
5. Fleming,M.A., Potter,J.D., Ramirez,C.J., Ostrander,G.K. and Ostrander,E.A. (2003) Understanding missense mutations in the BRCA1 gene: an evolutionary approach. *Proc. Natl Acad. Sci. USA*, **100**, 1151–1156.
6. Burk-Herrick,A., Scally,M., Amrine-Madsen,H., Stanhope,M.J. and Springer,M.S. (2006) Natural selection and mammalian BRCA1 sequences: elucidating functionally important sites relevant to breast cancer susceptibility in humans. *Mamm. Genome*, **17**, 257–270.
7. Thomas,J.W., Touchman,J.W., Blakesley,R.W., Bouffard,G.G., Beckstrom-Sternberg,S.M., Margulies,E.H., Blanchette,M., Siepel,A.C., Thomas,P.J., McDowell,J.C. *et al.* (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**, 788–793.
8. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res.*, **35**, D21–D25.
9. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
10. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
11. NCBI. (2007) Genome sequencing projects statistics. [http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html].
12. Springer,M.S. and Murphy,W.J. (2007) Mammalian evolution and biomedicine: new views from phylogeny. *Biol. Rev. Camb. Philos. Soc.*, **82**, 375–392.
13. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
14. Wrobel,G., Kokocinski,F. and Lichter,P. (2004) AutoPrime: selecting primers for expressed sequences. *Genome Biol.*, **5**, P11.
15. Rose,T.M., Henikoff,J.G. and Henikoff,S. (2003) CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. *Nucleic Acids Res.*, **31**, 3763–3766.
16. Gadberry,M.D., Malcomber,S.T., Doust,A.N. and Kellogg,E.A. (2005) Primaclade—a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21**, 1263–1264.
17. Kim,N. and Lee,C. (2007) QPRIMER: a quick web-based application for designing conserved PCR primers from multigenome alignments. *Bioinformatics*, **23**, 2331–2333.
18. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.
19. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
20. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
21. Drenkard,E., Richter,B.G., Rozen,S., Stutius,L.M., Angell,N.A., Mindrinos,M., Cho,R.J., Oefner,P.J., Davis,R.W. and Ausubel,F.M. (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. *Plant Physiol.*, **124**, 1483–1492.
22. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
23. Koressaar,T. and Remm,M. (2007) Enhancements and modifications of primer design program Primer3. *Bioinformatics*, **23**, 1289–1291.
24. Rozas,J., Sanchez-DelBarrio,J.C., Messeguer,X. and Rozas,R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, **19**, 2496–2497
25. Swofford,D.L. (2003) PAUP* 4.0. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
26. Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.
27. Huang,J., Sengupta,R., Espejo,A.B., Lee,M.G., Dorsey,J.A., Richter,M., Opravil,S., Shiekhattar,R., Bedford,M.T., Jenuwein,T. *et al.* (2007) p53 is regulated by the lysine demethylase LSD1. *Nature*, **449**, 105–108.
28. Downs,K., Zacks,D.N., Caruso,R., Karoukis,A.J., Branham,K., Yashar,B.M., Haimann,M.H., Trzupek,K., Meltzer,M., Blain,D. *et al.* (2007) Molecular testing for hereditary retinal disease as part of clinical care. *Arch. Ophthalmol.*, **125**, 252–258.
29. Mani,A., Radhakrishnan,J., Wang,H., Mani,A., Mani,M.A., Nelson-Williams,C., Carew,K.S., Mane,S., Najmabadi,H., Wu,D. *et al.* (2007) LRP6 mutation in a family with early coronary disease and metabolic risk factors. *Science*, **315**, 1278–1282.
30. Ivanov,I.P. and Atkins,J.F. (2007) Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res.*, **35**, 1842–1858.