

GETPrime 2.0: gene- and transcript-specific qPCR primers for 13 species including polymorphisms

Fabrice P.A. David^{1,2}, Jacques Rougemont^{1,2,*} and Bart Deplancke^{2,3,*}

¹Bioinformatics and Biostatistics Core Facility, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland, ²Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland and ³Laboratory of Systems Biology and Genetics, Institute of Bio-engineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

Received August 15, 2016; Accepted October 04, 2016

ABSTRACT

GETPrime (<http://bbcftools.epfl.ch/getprime>) is a database with a web frontend providing gene- and transcript-specific, pre-computed qPCR primer pairs. The primers have been optimized for genome-wide specificity and for allowing the selective amplification of one or several splice variants of most known genes. To ease selection, primers have also been ranked according to defined criteria such as genome-wide specificity (with BLAST), amplicon size, and isoform coverage. Here, we report a major upgrade (2.0) of the database: eight new species (yeast, chicken, macaque, chimpanzee, rat, platypus, pufferfish, and *Anolis carolinensis*) now complement the five already included in the previous version (human, mouse, zebrafish, fly, and worm). Furthermore, the genomic reference has been updated to Ensembl v81 (while keeping earlier versions for backward compatibility) as a result of re-designing the back-end database and automating the import of relevant sections of the Ensembl database in species-independent fashion. This also allowed us to map known polymorphisms to the primers (on average three per primer for human), with the aim of reducing experimental error when targeting specific strains or individuals. Another consequence is that the inclusion of future Ensembl releases and other species has now become a relatively straightforward task.

INTRODUCTION

Genome-scale experiments have accumulated massive information over recent years and have greatly contributed to our understanding of gene expression and its regulatory mechanisms. These experiments have clearly revealed the ubiquitous nature of alternative splicing and isoform

dosage effects (1,2). It is in this regard key to perform precise, quantitative measurements of selected genes and transcripts to assess specific expression patterns or functions. Such experiments typically involve the quantitative real-time polymerase chain reaction (qPCR), and the value of these qPCR assays depends in large part on the quality of the selected primer pair for the respective, targeted transcription unit (3).

We have therefore undertaken the systematic design of primer pairs for every known gene and transcript for organisms with well-annotated genome references with *in silico* verification of optimal specificity. The design of these primer pairs follows the pipeline described in (4), which we briefly recall here: for designing gene- or transcript-specific primers pairs, exon junctions that are included in respectively the largest or smallest number of isoforms for each gene are first identified after which the corresponding transcript is processed with PerlPrimer (5) for the best primer set that overlaps these junctions. Candidate primers are then filtered according to (i) genome-wide specificity (running BLAST with an *E*-value of 100) and (ii) not spanning 5' or 3' untranslated regions (UTR), as well as ranked according to the number of isoforms they cover, amplicon length, and other primer quality parameters that were previously discussed (3,4). The top three primer pairs are then retained and displayed in the database with a star-based quality flag corresponding to the rank in this list. If no pair passes the filters, then the original primer design constraints are progressively relaxed until a candidate pair emerges, hence the warnings associated with some primers (the 'warnings' column that can be observed in Figure 1).

Since its inception in 2011, the database has been used continuously and access statistics show a large user base. For example, the GETPrime web interface received nearly 1800 visits (by 1000 users) over the first 6 months of 2016 alone. Individual users also provided constructive feedback to further improve GETPrime, which in large part prompted the major update of the database (2.0) that is presented here.

*To whom correspondence should be addressed. Tel: +41 21 693 1821; Email: bart.deplancke@epfl.ch
Correspondence may also be addressed to Jacques Rougemont. Tel: +41 21 693 9573; Email: Jacques.rougemont@epfl.ch



GETPrime

Gene and Transcript-specific primer generator for real-time PCR

Search Downloads API documentation

NEWS: Last update finished on 2015 November 12 using Ensembl release 81. More species are available (13, see below).

Ensembl release: 81
 Organism: Homo sapiens
 Limit: 100
 Comma or space-separated list of identifiers: mdm1
 Search

30 primer pairs found. [Download](#)

Search all columns:

ID	Gene	Transcripts	Rank	# Transcripts	Amplicon length	Forward primer	Tm fwd	Reverse primer	Tm rev	Ensembl status	Warnings
2111374	ENSG00000111554 MDM1	ENST00000540418 ENST00000303145 ENST00000411698 ENST00000538454	★★★★ (1)	4/15	133	12:68321627-68323091 AGTGTCTCCTGAAAGGAAGAG 4 SNPs	59	12:68321436-68321532 AATTCACCTTCCCAAGCCT 5 SNPs	59	Known	Details View in UCSC
2111375	ENSG00000111554 MDM1	ENST00000540418 ENST00000303145 ENST00000411698 ENST00000538454	★★★☆☆ (2)	4/15	134	12:68321627-68323091 AGTGTCTCCTGAAAGGAAGAG 4 SNPs	59	12:68321435-68321531 GAATTCACCTTCCCAAGCC 5 SNPs	59	Known	Details View in UCSC
2111376	ENSG00000111554 MDM1	ENST00000540418 ENST00000303145 ENST00000411698 ENST00000538454	★★★☆☆ (3)	4/15	135	12:68321627-68323091 AGTGTCTCCTGAAAGGAAGAG 4 SNPs	59	12:68321434-68321530 GGAATTCACCTTCCCAAGC 6 SNPs	59	Known	Details View in UCSC
2111377	ENSG00000111554 MDM1	ENST00000303145 ENST00000411698 ENST00000541686 ENST00000430606 ENST00000545774	★★★★ (1)	5/15	80	12:68331156-68331177 TTTGTCCGAGTCTTGAATTC 4 SNPs	59	12:68327012-68331118 TTCGTGATGCCTAATGTCTG 5 SNPs	59	Known	Details View in UCSC

Figure 1. The GETPrime 2.0 search interface and tabular display. The figure shows several of the 30 primer pairs found for human gene *MDM1*. Results can be downloaded in tab-separated format through the 'Download' link. The search is restricted to an organism, Ensemble release, and a maximum number of lines (the smaller the number, the faster the query). Each result line corresponds to a single primer pair, and displays its unique ID, the gene, and transcript(s) it targets, its star-based rank (among the best three pairs found for the gene), the fraction of isoforms it covers, the amplicon length, the primer sequences and their respective melting temperatures, and the Ensembl annotation for the gene (KNOWN or NOVEL). The last two columns provide respectively warnings if the primer search did not work with standard parameters and a link to a primer pair-specific page shown in Figure 3.

Data integration

GETPrime 2.0 cross-references a number of data sources to document gene structures, transcript sequences, genome sequences, and annotated variants. The database now incorporates data from three versions of Ensembl (6): 50 (July 2008), 61 (February 2011), and 81 (July 2015). This is to keep backward compatibility with the first release of GETPrime, while updates will be performed on a regular basis. Relevant data from Ensembl is automatically imported into our PostgreSQL database (<https://www.postgresql.org>). Thanks to the uniform structure of the Ensembl database for various species, we can now easily select additional species and we currently host yeast, chicken, macaque, chimpanzee, rat, platypus, pufferfish, and *Anolis carolinensis* next to the previously established primers pairs for human, mouse, zebrafish, fly, and worm. Compared to version 1.0 (4), the database schema has been re-designed to improve the speed of queries via the web user interface and to provide two new interaction modes: a batch download capability and a programmatic interface (RESTful API).

User interface

The user interface of GETPrime 2.0 has been re-designed to make it faster, friendlier, and richer. It is based on a new 3-tier Ruby on Rails (RoR) (<http://rubyonrails.org>) application. Among many other features, this frame-

work improves the efficiency of database queries and simplifies the rendering on web pages. It also implements a RESTful API that allows programmers to access the data directly (see documentation at <http://bbcf-tools.epfl.ch/getprime/api-documentation>). A new search engine allows searching by gene name, Ensembl gene ID or transcript ID or directly by the internal primer pair ID (Figure 1). The search box accepts up to 10 identifiers per search. When only one identifier is provided and does not match perfectly, a regular expression search is performed. This search tool uses the *Jquery* (mostly the Ajax method) and *datatables.js* Javascript libraries. The Ajax technology is used to update portions of the web pages following user selections without reloading the whole page. This improves the responsiveness and flexibility of the display.

Primers are linked to a view in the UCSC genome browser (7) where they are displayed in their genomic context. In the UCSC view, primer pairs are identified by a unique numeric ID, by the gene and transcript they target, and by their rank in the list of candidates (Figure 2). This UCSC display is generated by uploading a single custom track (as a BED file) generated for each organism and Ensembl version. The BED file can be directly downloaded as well as the full database as TAB-separated files. Each primer pair is clickable and linked back to the GETPrime website, and more specifically to the page containing details about the primer. This page contains more information than the pre-

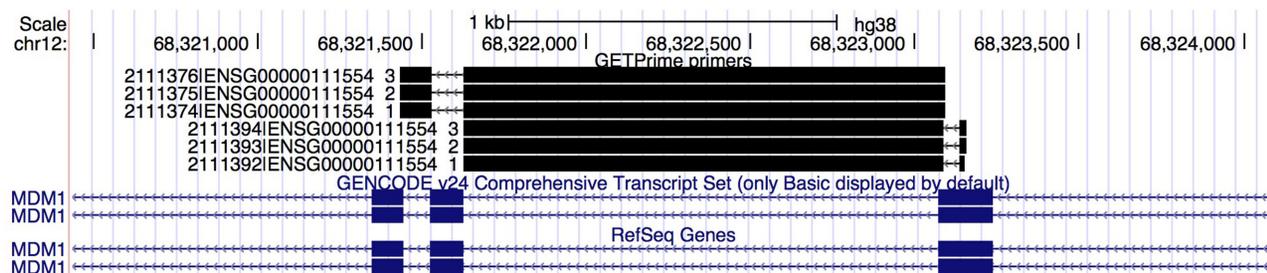


Figure 2. The UCSC view of GETPrime 2.0 primer pairs. The two primers (in black) of each pair are displayed as thick bars connected by thin arrows revealing on which strand the pair of primers will amplify DNA. They are also mapped to their genomic coordinates, including the intron(s) that each primer potentially spans. In this example, six primer pairs are displayed. For the first three, both forward and reverse primers span an intron, whereas for the three other pairs, only the reverse primer spans an intron. Note that the format of the displayed identifier is the following: GETPrimeID|Ensembl-gene-ID.GETPrime-rank (e.g. 2111376|ENSG00000111554.3) and that the other primer pairs for *MDM1* are not visible within this screenshot.

Details of primer pair 2111374

ID	2111374
Rank	1 ★★☆☆
Gene Ensembl ID	ENSG00000111554
Gene Name	MDM1
Gene Ensembl status	Known
Chromosome	12
Transcript IDs	ENST00000540418 ENST00000303145 ENST00000411698 ENST00000538454
Warnings	-
UCSC link	View in UCSC

Forward primer 68321627-68323091

Sequence	A G T G T C T C T G A A A G G A A G [Intron of 1444bp] A G	
Strand	Minus	
Tm	59	
Type	Junction	
SNPs	Position	rs number
	68323073	112751628 [Ensembl]
	68323080	545434178 [Ensembl]
	68323084	200011845 [Ensembl]
	68323086	374673461 [Ensembl]

Reverse primer 68321436-68321532

Sequence	A A T T C A C C T T C [Intron of 78bp] C C A A G C C T	
Strand	Plus	
Tm	59	
Type	Junction	
SNPs	Position	rs number
	68321439	777723960 [Ensembl]
	68321446	753853844 [Ensembl]
	68321447	757312225 [Ensembl]
	68321526	778938836 [Ensembl]
	68321528	200965515 [Ensembl]
	68321530	758897142 [Ensembl]

Figure 3. The GETPrime 2.0 primer details page. All information about one particular primer pair is summarized in this page: gene and transcript IDs, GETPrime warnings, and detailed information about each forward and reverse primer. Particularly relevant are the indication of SNP positions (in red) and whether a primer spans an intron as well as the UCSC display link.

vious version of GETPrime. For example, next to the position in the genome of the primer sequences, the position and the length of the introns are reported when applicable.

Sequence polymorphisms

Our knowledge of genomic variation within species and how such variants drive molecular and organismal diversity is rapidly increasing (8–12). One of the benefits of these advances is that we are now able to incorporate variant information (when available) in genomic experiments since such genetic variants may be an important source of experimen-

tal variability or even failure (13,14). Thus, to reduce experimental error, we decided to start displaying the presence of known SNPs within the GETPrime 2.0 primers to aid users in the design and interpretation of their experiments. So far, we were able to cover SNPs for human and mouse by importing them from dbSNP v145 (15) and to map these SNPs to the primers that overlap them. Corresponding positions in the primer sequences are then highlighted (Figure 3) and a link to the dbSNP-based evidence allows a more detailed evaluation of the nature and relevance of the polymorphism(s).

Table 1. Global statistics of GETPrime 2.0 for each of the 13 included species

Species	Number of genes in ensembl v81	Number of genes covered (% of total genes)	Number of primer pairs	Number of variants
<i>Anolis carolinensis</i>	19	19 (100%)	57	
<i>Caenorhabditis elegans</i>	20 447	20 412 (99.8%)	104 810	
<i>Danio rerio</i>	22 337	21 805 (97.6%)	121 576	
<i>Drosophila melanogaster</i>	13 918	13 911 (99.9%)	99 032	
<i>Gallus gallus</i>	5222	5204 (99.6%)	18 791	
<i>Homo sapiens</i>	22 017	21 653 (98.3%)	444 256	2 864 885
<i>Macaca mulatta</i>	8693	1154 (13.2%)	5345	
<i>Mus musculus</i>	22 155	21 835 (98.6%)	268 855	492 968
<i>Ornithothynchus anatinus</i>	170	149 (87.6%)	606	
<i>Pan troglodytes</i>	140	140 (100%)	474	
<i>Rattus norvegicus</i>	21 470	20 841 (97.0%)	88 311	
<i>Saccharomyces cerevisiae</i>	6692	6620 (98.9%)	19 923	
<i>Tetraodon nigroviridis</i>	1130	1125 (99.6%)	3838	

Database content

The GETPrime 2.0 database currently contains a total of 1 175 874 primer pairs (444 256 in human, 268 855 in mouse), corresponding to an average of six pairs per covered gene (across 13 species). In human, there are more than 20 pairs per gene and 12 in mouse. On average, 92% of Ensembl protein-coding genes are covered by our database, the remainder corresponding to non-unique sequences for which specific primers could not be designed (Table 1). Importantly, for human and mouse, this number exceeds 98%. However, some species are still only partially covered due to differences in the Ensembl annotation compared to the human database. In particular, for *A. carolinensis* or macaque, only a fraction of the annotated genes were processed in the pipeline (Table 1). Moreover, the incomplete status of the macaque assembly led to a high failure rate of the pipeline probably due to the repetitive nature of unassembled contigs (Table 1). We plan to resolve both issues in a next release. Regarding polymorphisms, a total of 2 864 885 variants were mapped to human primers (492 968 in mouse), indicating that more than 80% of human primers overlap a documented variant, with an average of about three SNPs per primer. This illustrates the importance of considering this information when designing or using primers.

CONCLUSION AND PERSPECTIVE

The steady access statistics of the GETPrime database are a testimony that the embedded primer information is useful and the release of GetPrime 2.0 responds to user feedback that we have received, namely: update the genomic data, extend to new species, and cross-reference new types of genomic data (polymorphisms). Our plan for the future is to maintain the availability of the database, keep it up-to-date and add new species when possible. In addition, we intend for GETPrime to closely follow and reflect the growth of genomic data resources at Ensembl and elsewhere. One additional important aspect would be a broader experimental validation of our *in silico*-designed primers. One way to do so would be to accommodate user feedback. We intend to implement a system that would allow the flagging of primers that have been successfully (or possibly even unsuccessfully) used in experiments, including links to the respective papers.

FUNDING

Swiss National Science Foundation Grant [#31003A_16273 5 to B.D.]; SyBIT project of SystemsX.ch (to J.R.); Swiss Federal Institute of Technology in Lausanne (EPFL). The open access publication charge for this paper has been waived by Oxford University Press—*NAR* Editorial Board members are entitled to one free paper per year in recognition of their work on behalf of the journal.

Conflict of interest statement. None declared.

REFERENCES

1. Pelechano, V., Wei, W. and Steinmetz, L.M. (2013) Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, **497**, 127–131.
2. Kornblihtt, A.R., Schor, I.E., Alló, M., Dujardin, G., Petrillo, E. and Muñoz, M.J. (2013) Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.*, **14**, 153–165.
3. Derveaux, S., Vandesompele, J. and Hellemans, J. (2010) How to do successful gene expression analysis using real-time PCR. *Methods*, **50**, 227–230.
4. Gubelmann, C., Gattiker, A., Massouras, A., Hens, K., David, F., Decouttere, F., Rougemont, J. and Deplancke, B. (2011) GETPrime: a gene- or transcript-specific primer database for quantitative real-time PCR. *Database*, **2011**, bar040.
5. Marshall, O.J. (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. *Bioinformatics*, **20**, 2471–2472.
6. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
7. Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
8. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
9. Deplancke, B., Alpern, D. and Gardeux, V. (2016) The genetics of transcription factor DNA binding variation. *Cell*, **166**, 538–554.
10. Albert, F.W. and Kruglyak, L. (2015) The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.*, **16**, 197–212.
11. Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M. *et al.* (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature*, **477**, 289–294.
12. Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F. *et al.* (2014) Natural variation in genome architecture among 205

- Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.*, **24**, 1193–1208.
13. Taxis, N., Lang, R.P. and Camara, M.D. (2008) Sequence polymorphism can produce serious artefacts in real-time PCR assays: hard lessons from Pacific oysters. *BMC Genomics*, **9**, 234.
 14. Boyle, B., Dallaire, N. and MacKay, J. (2009) Evaluation of the impact of single nucleotide polymorphisms and primer mismatches on quantitative PCR. *BMC Biotechnol.*, **9**, 75.
 15. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.