ExonSurfer: User Guide

Elena Cristina Rusu*
Pablo Monfort Lanzas†

March 15, 2023

*Email: ecrisru@alumni.uv.es
†Email: pablo.monfort@i-med.ac.at

Contents

1 Overview of ExonSurfer and its purpose

ExonSurfer is a web-based tool designed to assist researchers in designing primers for transcript-specific detection and avoiding genomic DNA or hnRNA contamination during reverse transcription uniplex PCR experiments. The tool combines various steps of primer design processes, including target region selection, avoiding of common polymorphisms, and checks for specificity and self-complementarity. ExonSurfer selects primer pairs spanning exon-exon junctions and automatically verifies the primers' specificity using BLAST. The user can customize the primer length, thermal parameters, and maximum amplicon length to suit their amplification procedure. ExonSurfer selects the most specific exon-exon junction for the input target(s), enabling highly accurate and efficient primer design for a wide range of applications. Its purpose is to provide researchers with a reliable and user-friendly tool that streamlines the primer design process and enhances the accuracy and specificity of PCR experiments.

ExonSurfer supports primer design for *Homo sapiens*, *Mus musculus* and *Rattus norvegicus*, for Ensembl's release 108. The command line version for ExonSurfer can be accessed at https://github.com/CrisRu95/ExonSurfer, and this web tool can be inspected at https://github.com/pamonlan/ExonSurferWeb.

2 User Interface

2.1 How to navigate through the interface

2.1.1 Input information

In order to start primer design, the user needs to select the species and write a gene name.

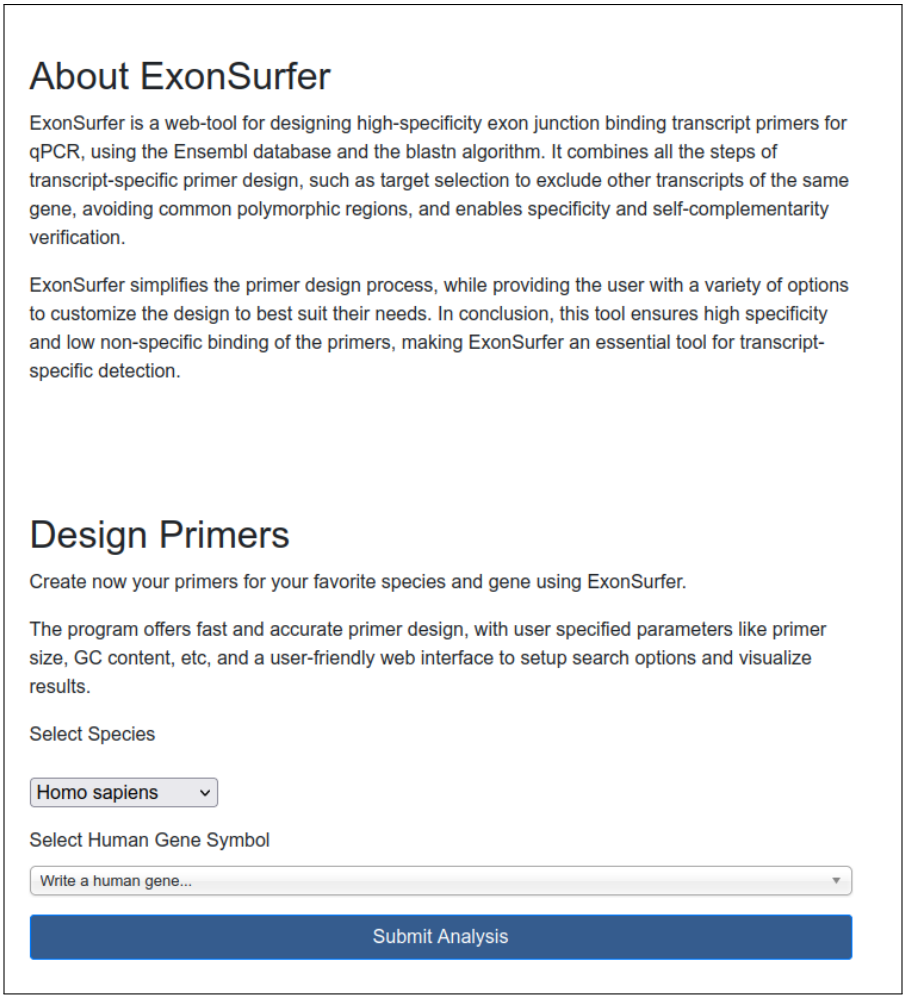

Figure 1: User input. Species and gene selection.

After pressing ”Submit Analysis”, the user needs to select the targeted transcripts and click again on the ”Submit Analysis” button.

Figure 2: Transcript selection. Transcript table and view.

2.1.2 Results

The main results page shows the ”Query Summary”, with information on the input provided by the user, the ”Top Primer Pairs”, which are the best primer pairs designed for that region and the complete ”Primer Results”, with a table of all the primer pairs designed.

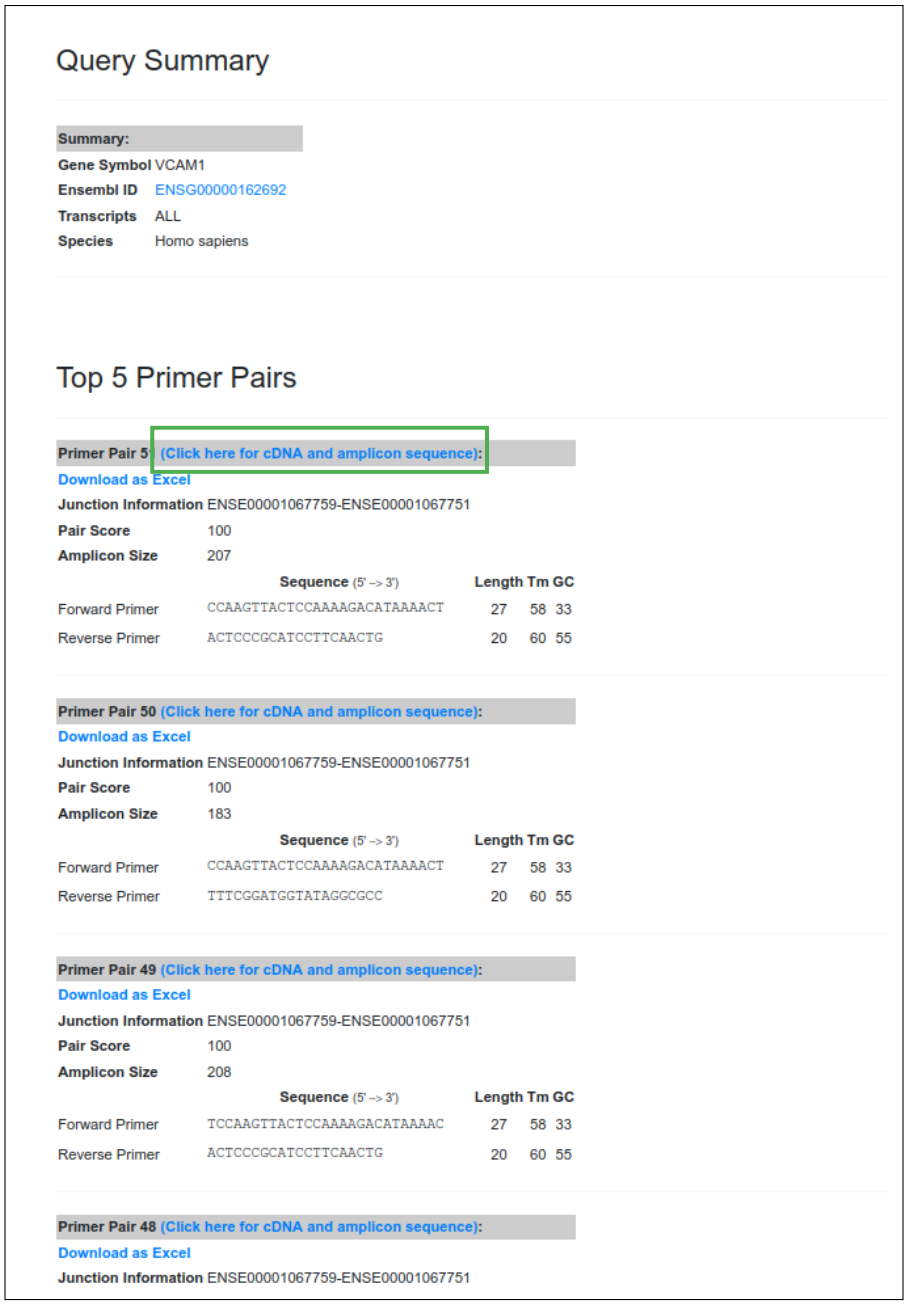

Figure 3: Results display. Query Summary and Top Primer Pairs.

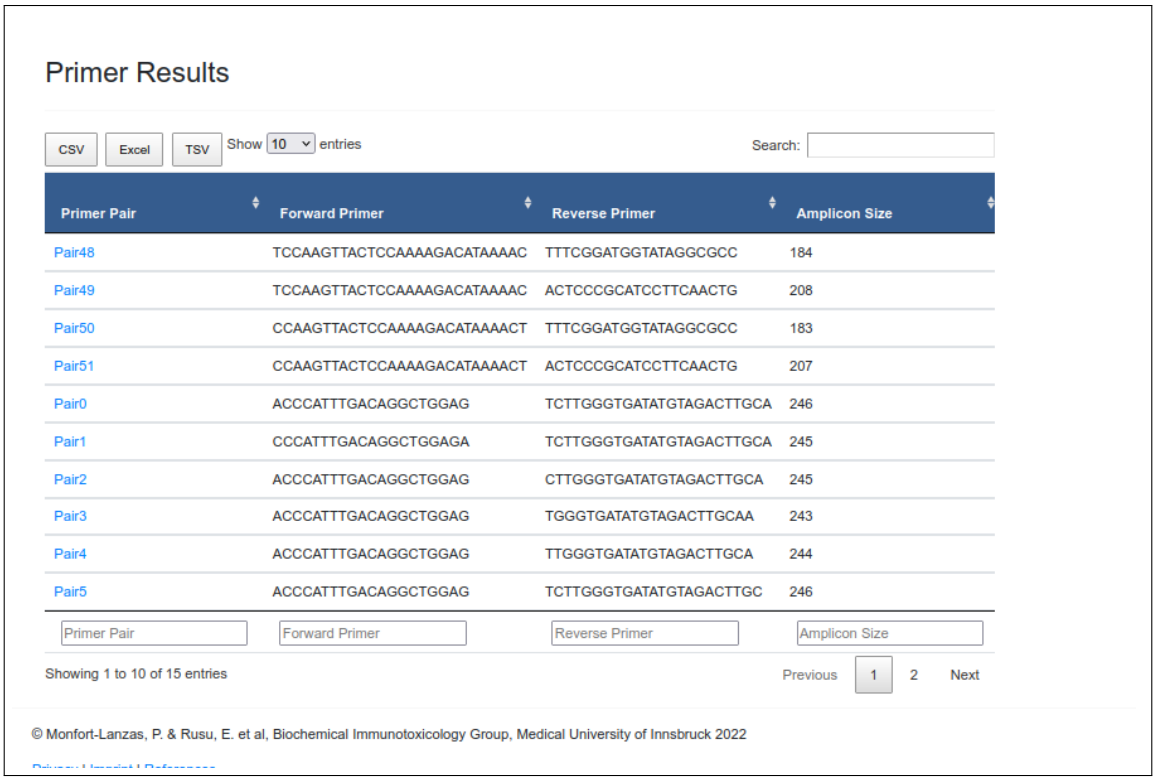

Figure 4: Results display. Complete primer pairs.

The complete primer pairs table contains the following columns:

- **Primer Pair:** Pair identifier.
- **Forward Primer:** Forward primer sequence.
- **Reverse Primer:** Reverse primer sequence.
- **Amplicon Size:** Product size (with primer length included).
- **Amplicon Tm:** Amplicon melting temperature, as calculated by primer3-py.
- **Forward Tm:** Forward primer melting temperature, as calculated by primer3-py (using the provided PCR conditions).
- **Reverse Tm:** Revcerse primer melting temperature, as calculated by primer3-py (using the provided PCR conditions).
- **Forward GC:** Forward primer GC percentage.
- **Reverse GC:** Reverse primer GC percentage.
- **Individual Alignment Score:** Number of times the primers align to transcripts from other genes (sum of forward and reverse alignments). The closer to 0, the better.
- **Detected Transcripts:** Transcripts of the target gene that the primer pair amplifies.
- **Not Detected Transcripts:** Transcripts of the target gene that the primer pair does not manage to amplify.
- **Pair Score:** Primer pair scoring according to optimal primer length, GC and Tm values, as well as dimer and hairpin interactions. The higher, the better the primer pair is (less dimers and hairpins, closer to optimal values).

To view the amplicon sequence and position of a specific primer pair, click on ”Click here for cDNA and amplicon sequence” (marked in green in Figure 3). The view is showed in Figure 5.

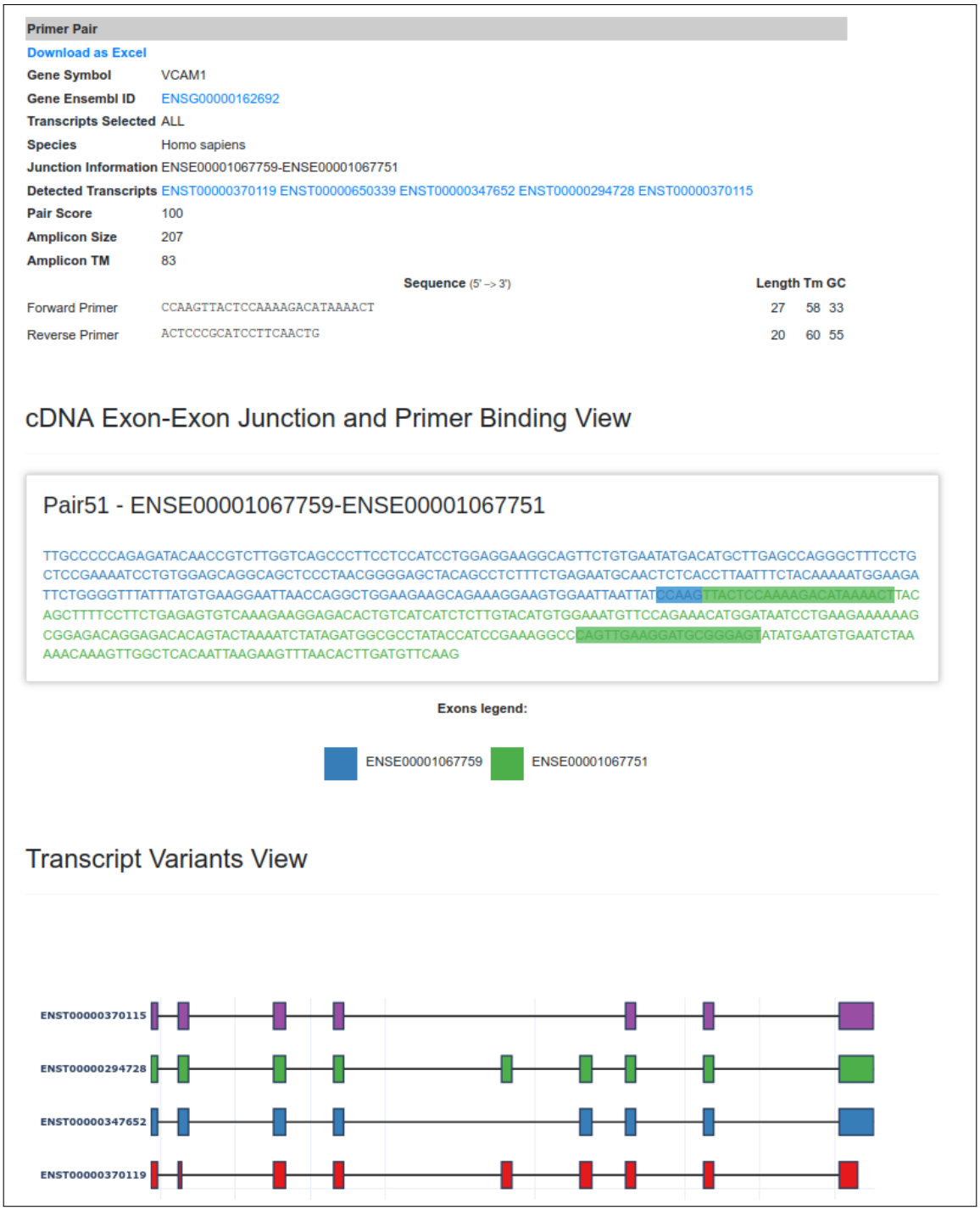

Figure 5: cDNA and amplicon sequence display. Amplicon sequence of a specific primer pair.

If a primer pair has any off-target amplification a warning is issued and we can check the off-target sequences by clicking on it.

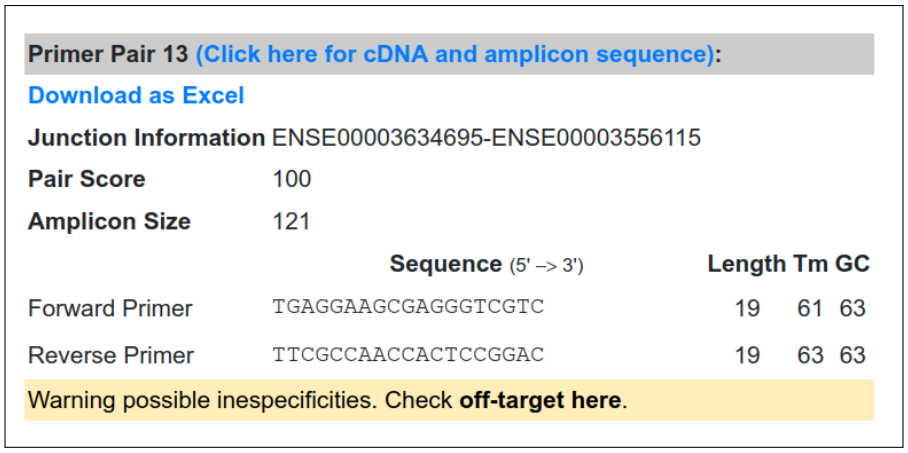

Figure 6: Off-target amplification warning.

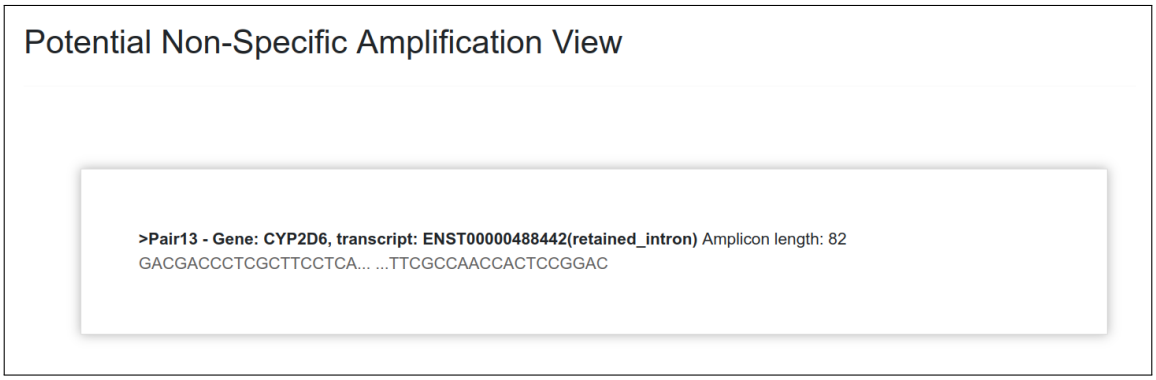

Figure 7: Off-target amplification view.

3 ExonSurfer's workflow

3.1 Workflow overview

ExonSurfer's internal workflow consists of 4 main modules, as shown in Figure 7:

1. **Target selection:** ExonSurfer chooses the best exon junctions depending on the target transcript(s) the user inputs. For the human genome, the sequence has polymorphisms with a minor allele frequence bigger than 1% masked (optionally).
2. **Primer design:** We use primer3-py[1] to design many possible primer pairs, in order to filter them in the following steps and report the best ones.
3. **Specificity checks:** Using BLAST[2], all the primers are aligned against the mRNA of the selected species and possible off-target amplification is annotated.
4. **Primer filtering and reporting:** Only the most specific primers are returned, sorted by the pair score.

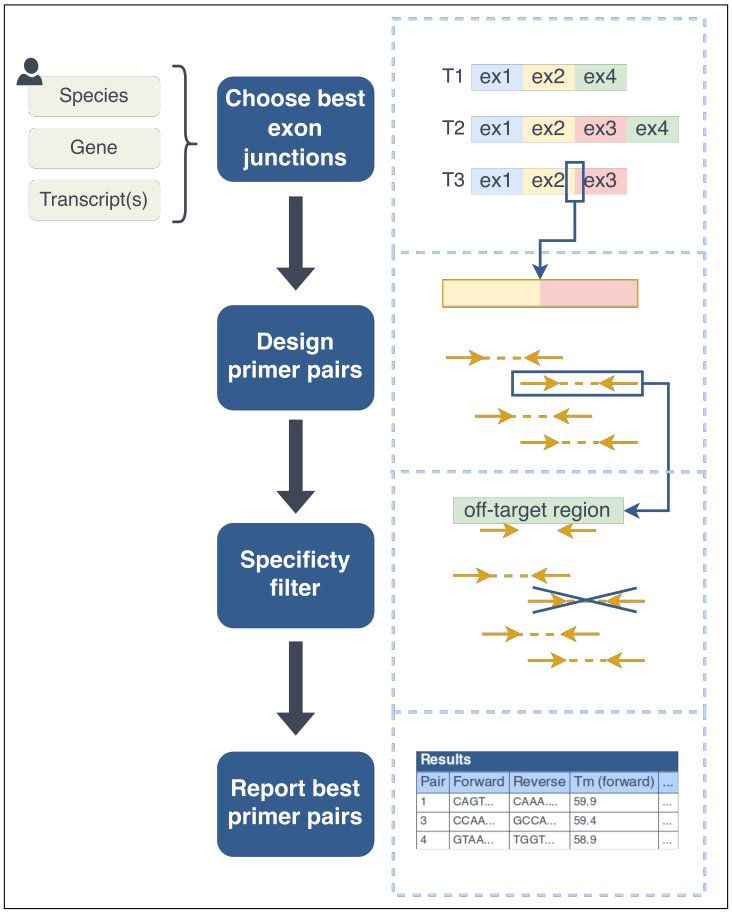

Figure 7: ExonSurfer workflow. The user inputs the desired species, gene symbol, and target transcript(s). ExonSurfer identifies the optimal junctions; generates multiple primer pairs that either overlap or flank the junctions, avoiding common polymorphisms (for *Homo sapiens*). ExonSurfer evaluates each primer pair for potential off-target amplification or nonspecific interactions, and filters out any problematic pairs. The user obtains a list of primers that are most suitable for qPCR, sorted based on the optimal parameters.

[1]Untergasser A, Cutcutache I, Koressaar T, et al. Primer3–new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115. doi:10.1093/nar/gks596

[2]Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421. Published 2009 Dec 15. doi:10.1186/1471-2105-10-421

3.2 Target selection

ExonSurfer uses a unique algorithm to determine the best exon junctions on which to design primers. The first step is to build, for each transcript, a list of consecutive exon junctions, where each junction is depicted by a list of ensembl exon identifiers. The number of exons used for each junction is dependant on exon length and the optimum amplicon length specified by the user (200bp by default). These junctions will be further filtered, depending on whether we want the most conserved or the most unique ones. Note that this selection process is done taking into account the exon identifiers.

3.2.1 All possible transcripts

When we target all transcripts (”ALL” option in Figure 2), ExonSurfer will search for the most conserved exon junction among the protein coding transcripts. In many cases, there is no exon junction ubiquitous among all transcripts, and in these case the junctions present in the canonical transcript, that cover the most other transcripts, are prioritized.

3.2.2 Specific transcripts

When we target one or some specific transcripts, ExonSurfer will search for a junction present among all our targets and not present in other transcripts from the same gene. This is often impossible, so ExonSurfer settles for the exon junctions present in all or most of the targets, as a first priority. Secondly, and if possible, it tries to exclude exon junctions that target a large number of not selected transcripts.

3.3 Masking common polymorphisms

For *homo sapiens*, ExonSurfer avoids placing primers on polymorphisms with a MAF (minor allele frequency) higher than 1% in any population group. For this, we extracted the SNPs from dbSNP (NCBI, reference GRCh38.p12, dbSNP build 154) and masked their locations in the chromosome sequences.

3.4 Primer design

Primers are designed using primer3-py (v 0.6.1) using the configuration (primer, amplicon and PCR parameters) provided by the user. ExonSurfer places the primers on the junctions in two distinct manners (Fig. 8). The preferred option is for one of the primers to overlap with the exon junction, amplifying in this way only the cDNA. However, given the need for further specificity checks, ExonSurfer also designs primers that flank the exon junction, ensuring a wider variety of oligos.

Figure 8: Design option. Primers are placed on the exon junction and flanking it. Primer sequences are marked as red arrows.

3.5 Self-complementarity check

ExonSurfer uses the default values from primer3-py for discarding primer pairs that form hairpin or dimer-like structures.

3.6 Specificity check

ExonSurfer uses BLAST (2.12.0+) in order to map the designed primers to the specified mRNA. The alignments are filtered according to the e-value and identity values. In order to boost performance, the primers with the most blast alignments are discarded from the design; this is repeated until the blast results decrease to a certain value.

3.6.1 Primer sequestration

For every primer pair, ExonSurfer counts the number of alignments each primers has to genes different from the target. This value is returned in the results table (Fig. 4) as ”Individual Alignment Score”.

3.6.2 Off-target amplification

If primers map to the same target, in opposite strand, and close to one another, they lead to off-target amplification. ExonSurfer assesses this event for every possible combination within a primer pair (forward with reverse, reverse with reverse or forward with forward) and annotates the possible off-targets. ExonSurfer classifies as off-target:

- Any possible amplification in a different gene.
- Any amplification in transcripts (of the target gene) not select by the user. This category is not taken into account with the option ”ALL”. If the number of any of the above categories is higher than 0, we also annotate the number of protein coding off-targets.

3.7 Primer filtering

ExonSurfer takes a flexible approach when filtering the final primers pairs. For some genetic regions it is feasible to expect primers with 0 alignments to other genes, but for other regions this is hardly possible (e.g. CYP2D6). This flexible approach ensures that the algorithm is as strict as possible for each designed region.

3.7.1 Filtering parameters

The filtering parameters (FP), sorted according to their importance, are:

- **FP1:** Number of possible off-target amplicons present in different genes.
- **FP2:** Number of protein coding off-target amplicons present in different genes.
- **FP3:** Number of possible off-target amplicons present in different transcripts (only when specific transcripts are selected).
- **FP4:** Number of protein coding off-target amplicons present in different transcripts (only when specific transcripts are selected).
- **FP5:** Number of alignments each primer pair has to other genes (independent of off-targets).
- **FP6:** Type of design (Fig. 8), where option 1 is preferred.

3.7.2 Filter overview

The filtering is done according to the parameter importance. Ideally, for specific transcripts we would have FP[1-5] = 0 and FP6 = 1, while for ”ALL” transcript design type we would disregard FP3 and FP4. However, this level of strictness does not adapt well to more difficult and challenging genomic regions. Therefore, primer pairs are filtered in a ”sorted” fashion, where ExonSurfer tries to keep the FP1 = 0 pairs first (independently of the values of the subsequent filters); if this is not possible if disregards FP1 and tries to keep the FP2 = 0 primer pairs; if this is possible, it further tries to keep FP3 = 0 primer pairs, and so on.

3.7.3 Primer pair score

The score reported is a normalization of primer3's pair penalty. For more information on its calculation, please visit this manual.

4 Customizing parameters

4.1 Primer parameters

Users can modify the following primer parameters:

Primer length Values between 18-25 are advised. If the target gene has extremely GC rich regions, lower the minimum length to 17. On the contrary, if the region is very poor in GC, primers can be as long as 36 bases, though more than 30 are not advised.

GC content (%) We advise keeping this parameter with the loose default values (20-80), since primer length and melting temperature control it. In addition, it is taken into account to calculate the "primer pair score".

Melting temperature (°C) Primer melting temperature, calculated with the tables of thermodynamic values and the method published in "A unified view of polymer, dumbbell and oligonucleotide DNA nearest-neighbor thermodynamics".

4.2 Amplicon parameters

Product size (bp) Length of the amplicon (primer sequences included) to generate. Typical values for qPCR are between 120 - 250bp.

Melting temperature (°C) Melting temperature of the product. Calculated using equation (iii) from the paper "Optimization of the annealing temperature for DNA amplification in vitro". This parameter is usually at fault when primers cannot be designed for a specific region (specially in rich GC genes). However, the PCR setup should be adaped if product Tm is higher than 90°C.

4.3 PCR parameters

These parameters are used as specified in the primer3 manual.

Salt Divalent (mM) The millimolar concentration of divalent salt cations in the PCR.

Salt Monovalent (mM) The millimolar (mM) concentration of monovalent salt cations (usually KCl) in the PCR.

dNTP Concentration (mM)] The millimolar concentration of the sum of all deoxyribonucleotide triphosphates.

4.4 BLAST parameters

Maximun E-value The Expect value is the number of hits that are expected by chance when search a database of a particular size. The smaller the E-value, the smaller is the possibility of finding the sequence in the database by chance (i.e., the smaller the E-value, the more meaningfull the hit is). For more information, please refer to BLAST's frequently asked questions. Hits with a larger E-value than the specified will not be reported nor used in the off-target assessment.

Minimun percentage of identity (%) The minimum identity to report a hit is the minimum percentage of base pairs that are the same between the query sequence (primer) and the subject sequence (specific mRNA).

Maximun target amplicon size The maximum distance between two hits in order to consider them as an off-target.

5 Troubleshooting

5.1 Why are my primers not being designed?

The design parameters are usually at fault for unsuccessful designs. Here we report a list of the most common issues and how to solve them.

- **Excessive amplicon length** Take into account that if your target gene has a very short coding sequence, the amplicon length parameters should be adjusted accordingly.
- **Amplicon Tm outside the expected length** For genes with very extreme GC contents, there might be difficult to design amplicons with a Tm lower than 90℃. However, take into account that PCR setup should adapt to amplicons with Tm higher than 90℃.
- **Primer length outside the expected length** Also for genes with very extreme GC contents, you might need to accept primers as long as 32-34 nucleotides, or as short as 17.

5.2 Contact support

If you are having issues with your specific design and these have not been solved by the previous specifications, of you are finding any other issue whithin the application, please contact pablo.monfort@i-med.ac.at or ecrisru@alumni.uv.es.

6 Acronyms

cDNA Complementary DNA

MAF Minor allele frequency

qPCR Quantitative polymerase chain reaction

SNP Single nucleotide polymorphism