PREDICTION OF PRIMER DIMER FORMATION AND OFF-TARGET AMPLIFICATION APPLIED TO TARGETED SEQUENCING DATA

Rusu EC[1], Cedeño JM[1], Olivares MD[1], Arnau V[2], Diaz-Villanueva W[2], Ivorra C[1]

1. SeqPlexing SL, Catedrático José Beltrán Martínez 2, 46980 Paterna, Valencia, Spain
2. Institute of Integrative Systems Biology (I2Sysbio), University of Valencia and CSIC, Catedrático Agustín Escardino Belloch, 46980 Paterna, Valencia, Spain

INSTITUTE FOR INTEGRATIVE SYSTEMS BIOLOGY

rusu@seqplexing.com

BACKGROUND

Amplicon-based Next-Generation Sequencing (NGS) yields exceptionally high coverage in an affordable way by specifically amplifying the regions of interest in a multiplex PCR. The technique is highly customizable, but new panels require extensive optimization, mainly due to problems in the first multiplex PCR (mPCR) during the library construction (Fig. 1). Primer-dimer species and products arising from off-target amplification (Fig. 2) can take over the library, decreasing the coverage for regions of interest (ROI).

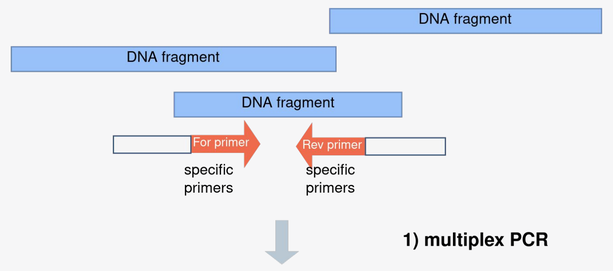

Figure 1. Amplicon library construction. A first multiplex PCR amplifies the ROIs and a second nested PCR adds barcodes and sequencing adapters.

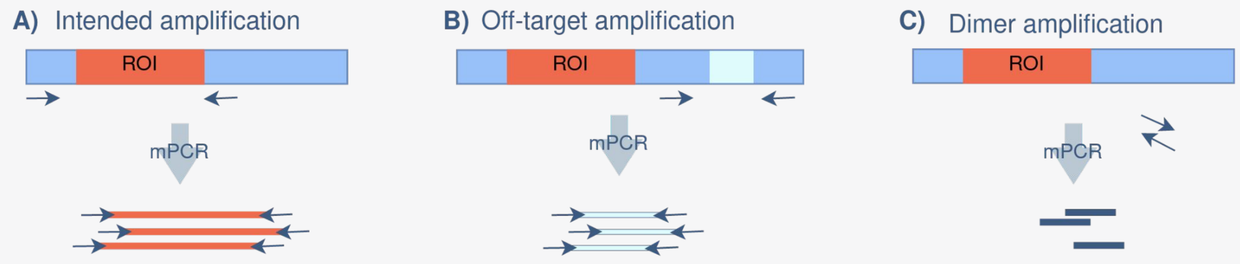

Figure 2. Formation of primer-dimer and off-target species during PCR. A) Intended amplification of the ROI; B) Off-target amplification creates an amplicon outside the ROI; C) primer dimer amplification occurs independently of the genome.

RESULTS

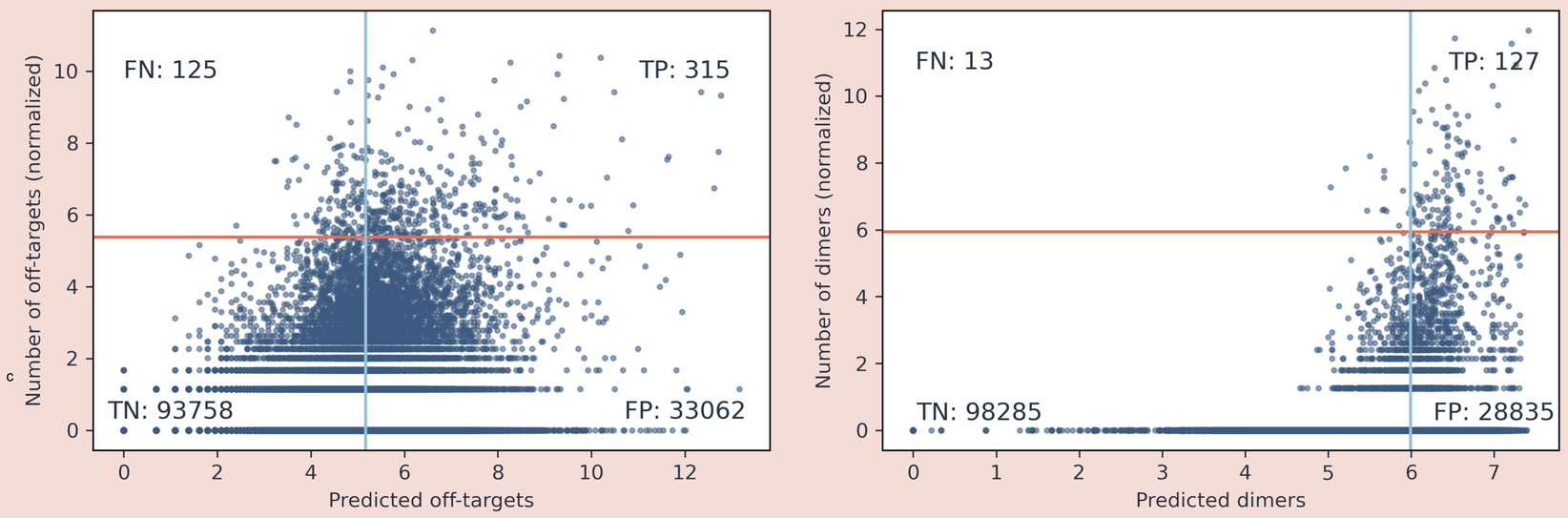

Figure 3. Prediction of dimer species and off-target amplification. *In silico* prediction of dimer species formation (**A**), or off-target product amplification (**B**) between all the interactions of a set of 504 primers (127260 possible combinations). **Off-target species prediction** is based on several alignments of the 3' end of the primers using bowtie (Langmead, 2009), allowing a variable number of mismatches and filtering by a variable number of 3'end matches. **Dimer species prediction** is based on weighted oligonucleotide nearest-neighbor thermodynamics, described by SantaLucia (2004) and widely used thereafter. The actual number of off-targets and dimers were calculated as described in Materials and Methods. **Horizontal orange lines** represent (**A**) the threshold number of off-target reads above which the primer pair was considered problematic (100, normalized), while for (**B**) it shows dimer reads above which a dimer species was considered excessively common (150, normalized). **Blue vertical lines** represent the threshold value for our predicted value, above which an interaction was considered likely problematic. For these thresholds and off-target prediction, we obtained a sensitivity of 0.72 and a specificity of 0.74; while the sensitivity for dimer prediction was 0.91 and the specificity 0.78.

HYPOTHESIS

We believe that *in silico* prediction of dimer species and off-target products can facilitate the design of new amplicon panels, helping reduce time and costs for researchers and clinicians alike.

MATERIALS AND METHODS

Categorization of actual off-target and dimer species

We designed an 252-plex amplicon panel (504 primers) and used for the library preparation of 40 samples, that were later sequenced in paired-end using a MiSeq. For this analysis, fastqs were merged. Adapters were trimmed using cutadapt (Martin, 2011) and reads were split depending on their length. Longer reads were mapped against the amplicon sequences using bowtie2 (Langmead, 2012) and unaligned reads were extracted. Short reads and unaligned long reads were both assessed for possible off-target or dimer origin. In order to do so and as described by Xie (2022), for every paired read we identified the primers at the 5' end. If the length of the read was shorter than the summed length of both primers and if there was an overlap between both primers in the 3' end, **we classified the read as dimer**. Very few cases (n<10) were found without a visible overlap. If the length of the read was longer than the summed length of the primers and the primers came from different amplicons, **we classified the read as off-target**. However, if both primers belonged to the same amplicon, the read was classified as an unaligned on-target.

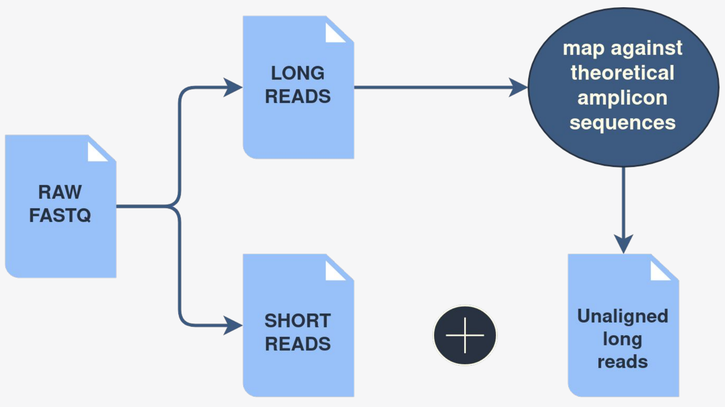

Figure 4. Categorization of off-target and dimer species, step 1.
Short reads and unaligned long reads were extracted from the merged fastq file, in order to be later processed in step 2.

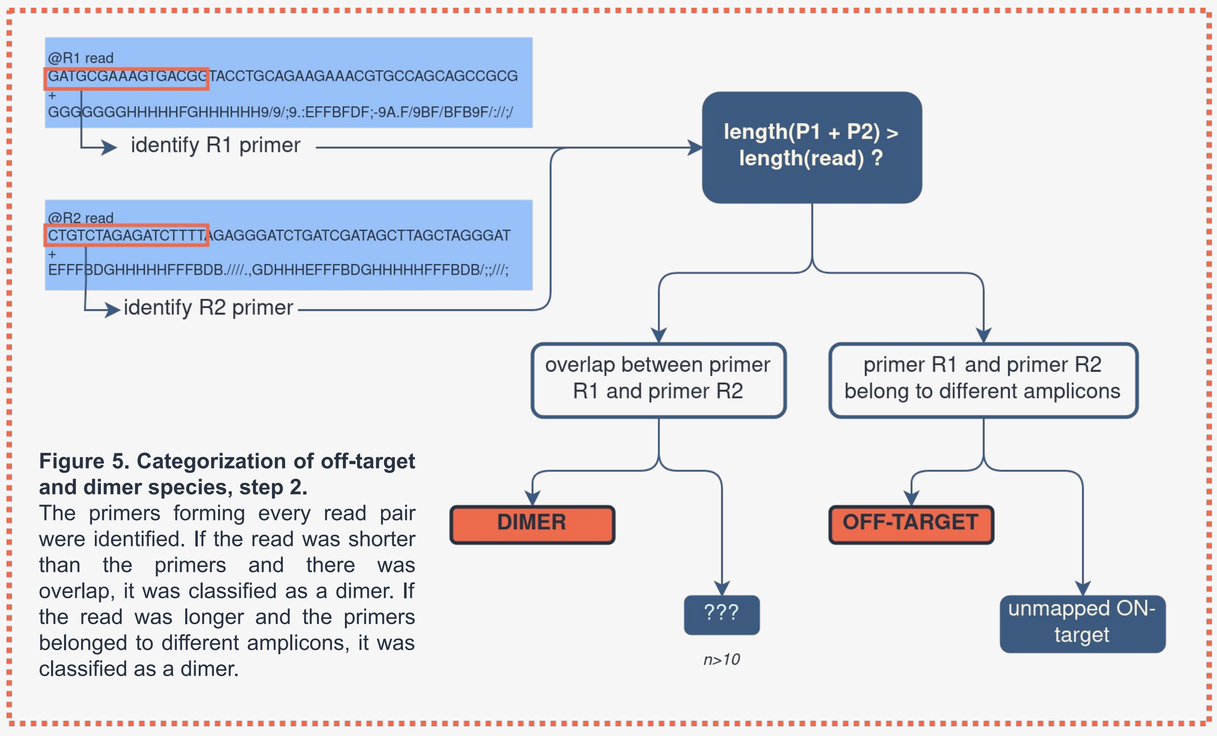

Figure 5. Categorization of off-target and dimer species, step 2.
The primers forming every read pair were identified. If the read was shorter than the primers and there was overlap, it was classified as a dimer. If the read was longer and the primers belonged to different amplicons, it was classified as a dimer.

CONCLUSION

Here we have built a pipeline for dimer and off-target species prediction. This pipeline has the potential to facilitate optimization of new amplicon-based NGS panels, helping reduce time and costs.

REFERENCES

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal, 17*(1), pp. 10-12. doi:https://doi.org/10.14806/ej.17.1.200

SantaLucia, J., Jr, & Hicks, D. (2004). The thermodynamics of DNA structural motifs. *Annual review of biophysics and biomolecular structure*, *33*, 415–440. https://doi.org/10.1146/annurev.biophys.32.110601.141800

Xie, N. G., Wang, M. X., Song, P., Mao, S., Wang, Y., Yang, Y., Luo, J., Ren, S., & Zhang, D. Y. (2022). Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). *Nature communications*, *13*(1), 1881. https://doi.org/10.1038/s41467-022-29500-4