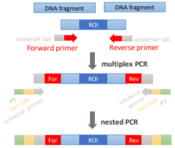
Primer design parameters effect on target sequencing and novel pipelines to ensure amplicon coverage

Rusu EC[1,2], Arnau V[1], Díaz-Villanueva W[1], Fuentes-Trillo A[3], Olivares D[2], Chaves F[2,3], Ivorra C[2]

1 Institute of Integrative Systems Biology (I2Sysbio), University of Valencia and CSIC, 46980 Valencia, Spain
2 SeqPlexing SL, 46980 Valencia, Spain
3 Unit of Genomic and Diabetes, INCLIVA Health Research Institute, University of Valencia, 46010 Valencia, Spain

1 BACKGROUND



Amplicon-based libraries (AbL) are widely-used technique for library preparation for targeted next-generation sequencing, both in research and clinical practice. AbL preparation includes a multiplex PCR and a nested PCR that, in a sequential manner, enrich the samples with the regions of interest (ROIs) and add barcodes and other helper sequences at both ends of every DNA fragment, called *amplicons* (Hess *et al*, 2020; Chang *et al*, 2013). The technique is highly customizable and the panel of ROIs can be unique to a specific experiment; but we need to design the primers for the multiplex PCR accordingly.

Considering that a typical panel includes hundreds of amplicons and that there are no software tools that consider all the necessary requirements, **primer design** turns into a time-consuming and ongoing bottleneck. In addition, it has the potential to impact the final coverage obtained from the sequencing reaction.

2 AIMS OF THIS WORK

This work develops around the problem of primer design for targeted sequencing with amplicon-based libraries. The aims are to:

1. Gain insight into the **link** between different **primer parameters** and the **final coverage** of targeted sequencing with AbL

2. Present **DocPrimer**, a novel pipeline that automates the process of primer design for library preparation of AbL

3 MATERIALS & METHODS

We examined a total of 544 amplicons coming from 4 different panels (targeting different single nucleotide polymorphisms -SNPs-) and assessed the effects of different primer parameters have on the final coverage of the amplicons, classified as high or low (threshold of 100x). The parameters were primer length, GC content, melting temperature (Tm), GC clamp and different mappings to the sample genome and local alignments between primers and helper sequences. Given that some of these are single primer qualities, the least optimal value was kept to in order to represent the primer pair. Statistical analyses were performed in R (version 4.0.0) using the test of equal or given proportions of the *stats* v3.6.2. package.

REFERENCES

- Chang F, Li MM. Clinical application of amplicon-based next-generation sequencing in cancer. Cancer Genet. 2013;206(12):413-419.
- Hess JF, Kohl TA, Kotrová M, et al. Library preparation for next generation sequencing: A review of automation strategies. Biotechnol Adv. 2020;41:107537.
- Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40(15):e115.

4 RESULTS

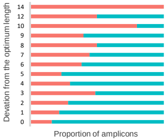


Figure 1. Primer length. Difference between the actual length of the primers and the optimum value (20 nucleotides).

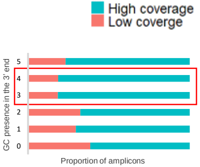


Figure 2. GC clamp. Number of Gs and Cs in the last 5 bases of the 3' end sequence (*y* axis) and different proportions of high and low coverage amplicons (*x* axis).

*Primers' **length** deviation (Fig 1) from their optimum value negatively impacts the final coverage of the amplicons. As expected, similar results were found regarding **GC content** and **Tm**.*

*The presence of a stronger **GC clamp** (Fig 2; 3 or 4 Gs or Cs in the last 5 nucleotides) enhances the final coverage.*

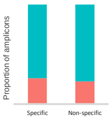


Figure 3. Individual inespecific mappings of the full primer sequence. We used *bwa aln* (single alignment) against the human genoma, and primers were classified as "non-specific" if secondary mappings were found.

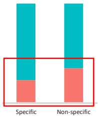


Figure 4. Individual inespecific 3' end mappings. The number of appearances in the genome of the 15 bases at the 3' end of the primers were calculated, and primers with more tan 25 appearances in the genome were classified as "non-specific".

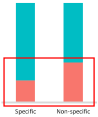
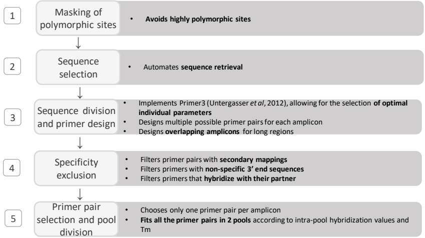


Figure 5. Primer pair inespecific mappings. We used *bwa aln* (paired alignment) against the human genome and i primers were classified as "non-specific" if secondary mappings were found.

*Mapping the primers' full sequence in an individual manner (Fig 3) against the sample genome does not inform on the final coverage obtained. **Specificity of the 3' end of the primers** (Fig 4) and **pair specificity** (Fig 5) seem to play a crucial role in obtaining a good coverage.*

5 DocPrimer

The challenges of primer design for multiplex PCR mainly come from the need to design **hundreds of primer pairs,** many times in an **overlapping** manner, and join them in a small number of reaction pools. DocPrimer is a novel pipeline that faces these challenges computationally, following the steps described below.



1 Masking of polymorphic sites
- **Avoids highly polymorphic sites**

2 Sequence selection
- Automates **sequence retrieval**

3 Sequence division and primer design
- Implements Primer3 (Untergasser *et al*, 2012), allowing for the selection **of optimal individual parameters**
- Designs multiple possible primer pairs for each amplicon
- Designs **overlapping amplicons** for long regions

4 Specificity exclusion
- Filters primer pairs with **secondary mappings**
- Filters primers with **non-specific 3' end sequences**
- Filters primers that **hybridize with their partner**

5 Primer pair selection and pool division
- Chooses only one primer pair per amplicon
- **Fits all the primer pairs in 2 pools** according to intra-pool hybridization values and Tm

Steps 1 and 2 involved the creation of a small **database** containing the human genome, masked in its polymorphic sites. Step 3 involves the use of a **greedy algorithm,** that, when overlapping amplicons are needed, dictates the position of amplicon *n* considering amplicons *n-1* and *n-2*. Step 4 excludes non-specific primer pairs according to the findings presented in this work. Step 5 uses a novel **branch and bound algorithm** to reach the best combination of primers and reaction pools.

Preliminar tests have returned promising results; *in silico* analyses have proven correct overlapping amplicon design, while early *in vitro* experiments have returned a coverage above the threshold for more than 90% of amplicons, when tested with a panel of 254 amplicons for SNP detection.

6 CONCLUSION

Multiplex primer design for targeted sequencing directly affects the final coverage of our samples. Some frequently overlooked parameters, such as the presence of a GC clamp, or the method used for the specificity check might play a bigger role that previously thought. DocPrimer has offered promising results and will continue to be improved, in order to achieve a full automation and control over primer design for amplicon-based libraries.