# Designing MHC primers

*Meinolf Ottensmann*

## Preface

This document provides the entire workflow for designing target-specific primer pairs that allow the amplification of Major Histocompatibility Complex II loci in *Arctocephalus gazella*. Computations are based on scripts written in `bash`, `phyton` and `R` respectively. Throughout, the working directory of both may be set to the parent folder of this project `SealMHC`, which contains all relevant data in subfolders.

```r
library(plyr)
library(dplyr)
source("R/primer_functions.R")
```

## Summary

Based on multiple sequence alignments of Major Histocompatibilty II sequences, the seconds exons of `DQA`, `DQB`, `DRA` and `DRB` have been identified on **Contig48** of the draft Antarctic fur seal genome (Humble et al. 2016). Approximate start and end postions of exons are given in table 1. Table 2 shows locus-specific primer sequences.

Table 1: Positions of MHC II exons within the Fur seal genome assembly.

| Locus | Contig | Start | End | bp |
|---|---|---|---|---|
| DQA | Contig48 | 1913465 | 1913711 | 247 |
| DQB | Contig48 | 1937069 | 1937338 | 270 |
| DRA | Contig48 | 1737500 | 1737745 | 246 |
| DRA | Contig48 | 1799016 | 1799261 | 246 |
| DRB | Contig48 | 2002578 | 2002803 | 226 |

Table 2: Locus-specifc primer for seconds exons MHC II loci in Antarctic fur seals.

| Locus | Forward primer | Reverse primer |
|---|---|---|
| DQA | GGCTCTTTTCTCCCTCTGTTTT | TCGTAGGGAGGAAGGGAATG |
| DQB | GCTGTTGGTTGGGCTGAG | CCACCTCAGCAGGAACAGTG |
| DRA | ACTCTCTTCCCTGCCTTTTCA | CATGTCTAGGAGCGCAGCA |
| DRB | GGTGACCGGATCCTCTCTG | GGACGGGAGGAGTCTGTTTC |

## Exploring the fur seal MHC architecture on the genome and transcriptome level

### Set up NCBI BLAST and databases

```
sudo
  apt-get install ncbi-blast+
makeblastdb
```

```
  -in blast/arc_gaz_genome.fasta -dbtype nucl
  -out blast/db/arc_gaz_genome_db
makeblastdb
  -in blast/arc_gaz_transcriptome.fasta -dbtype nucl
  -out blast/db/arc_gaz_transcriptome_db
```

### Blasting sequences against references

Exon II sequences of several carnivores were downloaded from GenBank as single fasta files for each of the targeted loci. These files are used to (i) generate sequence alignments and (ii) map the loci to the *A. gazella* genome and transriptome respectively. *Blastn* results indicate the location of the MHC loci on the genome and the transcriptome and allow to extract the consensus sequences.

```
blastn
  -db blast/db/arc_gaz_genome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dqa.fasta
  -out blast/blast_results/dqa2_arc_gaz_genome.fasta
blastn
  -db blast/db/arc_gaz_genome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dqb.fasta
  -out blast/blast_results/dqb2_arc_gaz_genome.fasta
blastn
  -db blast/db/arc_gaz_genome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/drb.fasta
  -out blast/blast_results/drb2_arc_gaz_genome.fasta
blastn
  -db blast/db/arc_gaz_genome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dra.fasta
  -out blast/blast_results/dra2_arc_gaz_genome.fasta
blastn
  -db blast/db/arc_gaz_transcriptome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dqa.fasta
  -out blast/blast_results/dqa2_arc_gaz_transcriptome.fasta
```

```
blastn
  -db blast/db/arc_gaz_transcriptome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dqb.fasta
  -out blast/blast_results/dqb2_arc_gaz_transcriptome.fasta
blastn
  -db blast/db/arc_gaz_transcriptome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/drb.fasta
  -out blast/blast_results/drb2_arc_gaz_transcriptome.fasta
blastn
  -db blast/db/arc_gaz_transcriptome_db
  -outfmt 6
  -num_threads 8
  -evalue 1e-8
  -word_size 7
  -query blast/dra.fasta
  -out blast/blast_results/dra2_arc_gaz_transcriptome.fasta
```

### Extract regions of interest from blast hits

With consideration of a adequate flanking sequences, 150 bp up/downstream, targets for designing primers are extracted based on the estimated start and end positions of the mhc loci.

```
## list files
files <- list.files("blast/blast_results", include.dirs = FALSE,
    pattern = "\\.fasta$")
## extract targets and write to file
for (i in files) target_extract(file = i, flanking = 150,
    dir = "blast/blast_results")
```

Now, sequences are extracted from the assembled genome and transcriptome respectively, using bedtools.

```
## to avoid compatibility problems between windows and linux
dos2unix blast/blast_results/dqa2_arc_gaz_genome.bed
dos2unix blast/blast_results/dqb2_arc_gaz_genome.bed
dos2unix blast/blast_results/drb2_arc_gaz_genome.bed
dos2unix blast/blast_results/dra2_arc_gaz_genome.bed

dos2unix blast/blast_results/dqa2_arc_gaz_transcriptome.bed
dos2unix blast/blast_results/dqb2_arc_gaz_transcriptome.bed
dos2unix blast/blast_results/drb2_arc_gaz_transcriptome.bed
dos2unix blast/blast_results/dra2_arc_gaz_transcriptome.bed

bedtools getfasta
  -fi blast/arc_gaz_genome.fasta
  -bed blast/blast_results/dqa2_arc_gaz_genome.bed
  -fo blast/seq/dqa_arc_gaz_genome.fasta

bedtools getfasta
  -fi blast/arc_gaz_genome.fasta
  -bed blast/blast_results/dqb2_arc_gaz_genome.bed
  -fo blast/seq/dqb_arc_gaz_genome.fasta

bedtools getfasta
  -fi blast/arc_gaz_genome.fasta
  -bed blast/blast_results/drb2_arc_gaz_genome.bed
  -fo blast/seq/drb_arc_gaz_genome.fasta

bedtools getfasta
  -fi blast/arc_gaz_genome.fasta
  -bed blast/blast_results/dra2_arc_gaz_genome.bed
  -fo blast/seq/dra_arc_gaz_genome.fasta

bedtools getfasta
  -fi blast/arc_gaz_transcriptome.fasta
  -bed blast/blast_results/dqa2_arc_gaz_transcriptome.bed
  -fo blast/seq/dqa_arc_gaz_transcriptome.fasta

bedtools getfasta
  -fi blast/arc_gaz_transcriptome.fasta
  -bed blast/blast_results/dqb2_arc_gaz_transcriptome.bed
  -fo blast/seq/dqb_arc_gaz_transcriptome.fasta
```

```
bedtools getfasta
  -fi blast/arc_gaz_transcriptome.fasta
  -bed blast/blast_results/drb2_arc_gaz_transcriptome.bed
  -fo blast/seq/drb_arc_gaz_transcriptome.fasta

bedtools getfasta
  -fi blast/arc_gaz_transcriptome.fasta
  -bed blast/blast_results/dra2_arc_gaz_transcriptome.bed
  -fo blast/seq/dra_arc_gaz_transcriptome.fasta
```

## Design Primers for DQA, DQB, DRA and DRB

Based on the blasting results obtained by the steps outlined above, all putative regions within the genome and transcriptome of *Arctocephalus gazella* were identified and extracted for a multiple-species alignment using `BioEdit`.
For each of the loci, contig number and the relative position are listed in the first line. The second line denotes the position of the target region on both the extracted sequences as well as the contig. The third rows gives the coordinates used in Primer3Plus to mark the target for designning primers.

### DQA

- Contig48:1913216-1913961
- TARGET Region: 247-496 (1913465-1913711)
- TARGET: 242,256

### DQB

- Contig48:1936819-1937588
- TARGET Region: 251-520 (1937069-1937338)
- TARGET: 251,270 OR 244,251

### DRB

- Contig48:2002402-2003098
- TARGET Region: 169-401 (2002570-2002802)
- TARGET:163 ,239

### DRA

- Contig48:1737350-1737895
- TARGET Region: 150-401 (1737500 - 1737751)
- TARGET:150, 251

The following workflows was conducted to calculate primers with the above mentionen tool.

1. Open `Primer3Plus` with the browser.
2. Paste source sequence from `data/primer_source_seqs.txt` into the designated field.
3. Specify the target coordinates.
4. Upload Primer3Plus settings (`data/Primer3Plus-settings.txt`) to customise parameters under `General Settings`
5. Press the button `activate settings`

6. Press `pick primers`

*The blasting results suggests a duplication of the* **`DRA`** *locus. Primers designed for one position fit to the second region.*

## Check specifity of primers

### Blasting against the genome

Primers sequences are mapped to the genome to ensure specifity for the targeted loci.

```
dos2unix blast/mhc-primer.fasta
blastn
  -db blast/db/arc_gaz_genome_db
  -outfmt 6
  -num_threads 8
  -evalue 10
  -word_size 14
  -query blast/mhc-primer.fasta
  -out blast/blast_results/primer2_arc_gaz_genome.fasta
```

### Analysing hits

In addition to the targeted regions, primers do fit to multiple regions in the genome, but there not a single pair fits elsewhere in such a way that forward and reverse primer are suggested to anneal within the conservative range of 800 bases.

| primer | contig | length | start | end | id | type |
|---|---|---:|---:|---:|---|---|
| DQA_V1_F | Contig48 | 22 | 1913411 | 1913432 | DQA_V1 | F |
| DQA_V1_R | Contig48 | 20 | 1913786 | 1913767 | DQA_V1 | R |
| DQA_V1_F | Contig83 | 19 | 125688 | 125670 | DQA_V1 | F |
| DQA_V1_R | Contig83 | 16 | 6129428 | 6129443 | DQA_V1 | R |
| DQA_V1_R | Contig83 | 16 | 6129568 | 6129583 | DQA_V1 | R |
| DQB_V1_F | Contig48 | 18 | 1937394 | 1937377 | DQB_V1 | F |
| DQB_V1_R | Contig48 | 20 | 1936936 | 1936955 | DQB_V1 | R |
| DRA_V1_F | Contig48 | 21 | 1737478 | 1737498 | DRA_V1 | F |
| DRA_V1_F | Contig48 | 21 | 1798994 | 1799014 | DRA_V1 | F |
| DRA_V1_R | Contig48 | 19 | 1737769 | 1737751 | DRA_V1 | R |
| DRA_V1_R | Contig48 | 19 | 1799285 | 1799267 | DRA_V1 | R |
| DRB_V1_F | Contig48 | 19 | 1842546 | 1842528 | DRB_V1 | F |
| DRB_V1_F | Contig48 | 19 | 2002535 | 2002553 | DRB_V1 | F |
| DRB_V1_R | Contig48 | 20 | 2002919 | 2002900 | DRB_V1 | R |

---

```
sessionInfo()
> R version 3.4.3 (2017-11-30)
> Platform: x86_64-w64-mingw32/x64 (64-bit)
> Running under: Windows 10 x64 (build 16299)
>
> Matrix products: default
>
> locale:
```

```
> [1] LC_COLLATE=English_United Kingdom.1252
> [2] LC_CTYPE=English_United Kingdom.1252
> [3] LC_MONETARY=English_United Kingdom.1252
> [4] LC_NUMERIC=C
> [5] LC_TIME=English_United Kingdom.1252
>
> attached base packages:
> [1] stats     graphics  grDevices utils
> [5] datasets  methods   base
>
> other attached packages:
> [1] dplyr_0.7.4 plyr_1.8.4  knitr_1.17
>
> loaded via a namespace (and not attached):
>  [1] Rcpp_0.12.14     assertthat_0.2.0
>  [3] digest_0.6.13    rprojroot_1.3-1
>  [5] R6_2.2.2         backports_1.1.2
>  [7] formatR_1.5      magrittr_1.5
>  [9] evaluate_0.10.1  highr_0.6
> [11] rlang_0.1.4      stringi_1.1.6
> [13] bindrcpp_0.2     rmarkdown_1.8
> [15] tools_3.4.3      stringr_1.2.0
> [17] glue_1.2.0       yaml_2.1.16
> [19] compiler_3.4.3   pkgconfig_2.0.1
> [21] htmltools_0.3.6  bindr_0.1
> [23] tibble_1.3.4
```

## References

Humble, E., A. Martinez-Barrio, J. Forcada, P. N. Trathan, M. A. S. Thorne, M. Hoffmann, J. B. W. Wolf, and J. I. Hoffman. 2016. "A Draft Fur Seal Genome Provides Insights into Factors Affecting Snp Validation and How to Mitigate Them." *Molecular Ecology Resources* 16 (4): 909–21. doi:10.1111/1755-0998.12502.