

HLA-DOA data

- [Nucleotide Sequence](#) (Coding sequences)
- [Nucleotide Sequence in PIR format](#)
- [Nucleotide Sequence in FASTA format](#)
- [Peptide Sequence](#)
- [Peptide Sequence in PIR format](#)
- [Peptide Sequence in FASTA format](#)
- [Reference Information](#)

Sequence Formats

FASTA - Sequences in FASTA/Pearson format are represented by two main line types. The first line always begins with a "greater than" (>) sign and contains sequence information. In the files provided, the sequence information contains the name of the HLA allele. The remaining lines contain plain text representing the coding nucleotide sequence. There can be any number of these sequence lines, of any length, to represent the nucleotide sequence.

Example DRB1*0101 in FASTA format:

```
>DRB1*0101
GGGGACACCCGACCACGTTTCTTGTGGCAGCTTAAGTTTGAATGTCATTT
CTTCAATGGGACGGAGCGGGTGC GGTTGCTGGAAAAGATGCATCTATAACC
AAGAGGAGTCCGTGCGCTTCGACAGCGACGTGGGGGAGTACCGGGCGGGT
ACGGAGCTGGGGCGGCCTGATGCCGAGTACTGGAACAGCCAGAAGGACCT
CCTGGAGCAGAGGGCGGGCCGCGGTGGACACCTACTGCAGACACAACCTACG
GGTTTGGTGAGAGCTTACAGTGCAGCGGCGAGTTGAGCCTAAGGTGACT
GTGTATCCTTCAAAGACCCAGCCCCTGCAGCACCAACCTCCTGGTCTG
CTCTGTGAGTGGTTTCTATCCAGGCAGCATTGAAATCAGGTGGTTCCGGA
ACGGCCAGGAAGAGAAGGCTGGGGTGGTGTCCACAGGCCTGATCCAGAAT
GGAGATTGGACCTTCCAGACCCTGGTGTGCTGGAAACAGTTCCTCGGAG
TGGAGAGGTTTACACCTGCCAAGTGGAGCACCAAGTGTGACGAGCCCTC
TCACAGTGAATGGAGAGCACGGTCTGAATCTGCACAGAGCAAGATGCTG
AGTGGAGTCGGGGGCTTCGTGCTGGGCCTGCTCTTCCTTGGGGCCGGGCT
GTTTCATCTACTTCAGGAATCAGAAAGGACACTCTGGACTTCAGCCAACAG
GATTCCTGAGCTGA
```

PIR - The format of sequences in PIR/NBRF format is more complex. The first line of each sequence entry begins with a "greater than" (>) sign. This is immediately followed by a two character sequence type specifier: for the HLA alleles this is "DL", meaning DNA linear. Space four must contain a semi-colon. Beginning in space five is the sequence name or identification code: for HLA alleles this is the official allele name. The second line of each sequence entry contains a brief description, including the sequence length, and an internal checksum for PIR files. The coding nucleic acid sequence begins on the third line. The sequence is free format, but to aid in reading the sequences, the nucleotides have been arranged in blocks of 10 amino acids. The last character is an asterisk (*), and acts as a termination character.

All PIR files have been generated using "ReadSeq", a freely available sequence format conversion program written by D. Gilbert.

Example DRB1*0101 in PIR format.

```
>DL;DRB1*0101
DRB1*0101, 714 bases, A686B796 checksum.
GGGGACACCC GACCACGTTT CTTGTGGCAG CTTAAGTTTG AATGTCATTT
CTTCAATGGG ACGGAGCGGG TGCGGTTGCT GGAAAGATGC ATCTATAACC
AAGAGGAGTC CGTGCGCTTC GACAGCGACG TGGGGGAGTA CCGGGCGGGT
```

ACGGAGCTGG GCGGCCTGA TGCCGAGTAC TGGAACAGCC AGAAGGACCT
CCTGGAGCAG AGGCGGGCCG CCGTGGACAC CTACTIONCAGA CACAACCTACG
GGGTTGGTGA GAGCTTCACA GTGCAGCGGC GAGTTGAGCC TAAGGTGACT
GTGTATCCTT CAAAGACCCA GCCCCTGCAG CACCACAACC TCCTGGTCTG
CTCTGTGAGT GGTTCCTATC CAGGCAGCAT TGAAGTCAGG TGGTTCCGGA
ACGGCCAGGA AGAGAAGGCT GGGGTGGTGT CCACAGGCCT GATCCAGAAT
GGAGATTGGA CCTTCCAGAC CCTGGTGATG CTGGAAACAG TTCCTCGGAG
TGGAGAGGTT TACACCTGCC AAGTGGAGCA CCCAAGTGTG ACGAGCCCTC
TCACAGTGGA ATGGAGAGCA CCGTCTGAAT CTGCACAGAG CAAGATGCTG
AGTGGAGTCG GGGGCTTCGT GCTGGGCCTG CTCTTCCTTG GGGCCGGGCT
GTTCATCTAC TTCAGGAATC AGAAAGGACA CTCTGGACTT CAGCCAACAG
GATTCCTGAG CTGA*

Enquiries to: hladb@ebi.ac.uk