HLAProfiler: Column descriptions

Allele1\_Accession: IMGT accession ID for Allele 1 (e.g. HLA00001)

Allele2\_Accession: IMGT accession ID for Allele 2 (e.g. HLA00005)

Allele1: Name for Allele 1 (e.g. A\*01:01:01:01)

Allele2: Name for Allele 2 (e.g. A\*02:01:01:01)

Proportion\_reads: The proportion of the observed reads filtering to kmers from allele in the gene, and also supported by the proposed allele pair (e.g. there might be 10k reads that filter to HLA-A, but only 9k of these are supported by the A\*01:01:01:01 / A\*02:01:01:01 combination)

Proportion\_signal: The proportion of the kmers in the allele pair that have support from the observed reads (e.g. what fraction of the kmers in the A\*01:01:01:01 / A\*02:01:01:01 combination are supported by the sequencing reads)

Correlation: The correlation in kmer counts between gene-filtered observed reads and simulated reads from the proposed allele pair (e.g. in a separate simulation of A\*01:01:01:01 / A\*02:01:01:01, what is the correlation in the resulting kmer counts and the kmer counts from the observed reads?)

Error: (1 - Proportion\_reads) + (1 - Proportion\_signal) + (1 - Correlation)

Pair\_score: The pairwise alignment score for the top 20 proposed allele pairs. The higher this value, the more reads align uniquely to this allele pair than any other allele pair from the top 20.

Final\_score: An amalgam score, incorporating Error and Pair\_score, that indicates the confidence that the proposed pair is correct. The Final\_score is relative to other proposed allele pairs in the top 20. For example, if the final score for the top candidate was 100 and the second candidate was 10, we would have high confidence that the top candidate was correct. However, if the top and second candidates had very similar final scores, either answer may be correct.

Allele1.comments: These comments are often blank, and when they do exist, they describe detection of novel alleles (not present in the HLA-IMGT database) or updated alleles (only partial sequences are present in the database). In these cases, the observed nucleotide sequence is also available, if desired.

Allele2.comments: Same as Allele1.comments, but for the other allele.

If a sample has < 100 reads filtering to the HLA gene, then predictions will be made, and instead the line will read:

HLA HLA GENE\_NAME GENE\_NAME - - - - - - Not enough reads to make call

Where “GENE\_NAME” is replaced with the appropriate gene name (e.g. DMA).