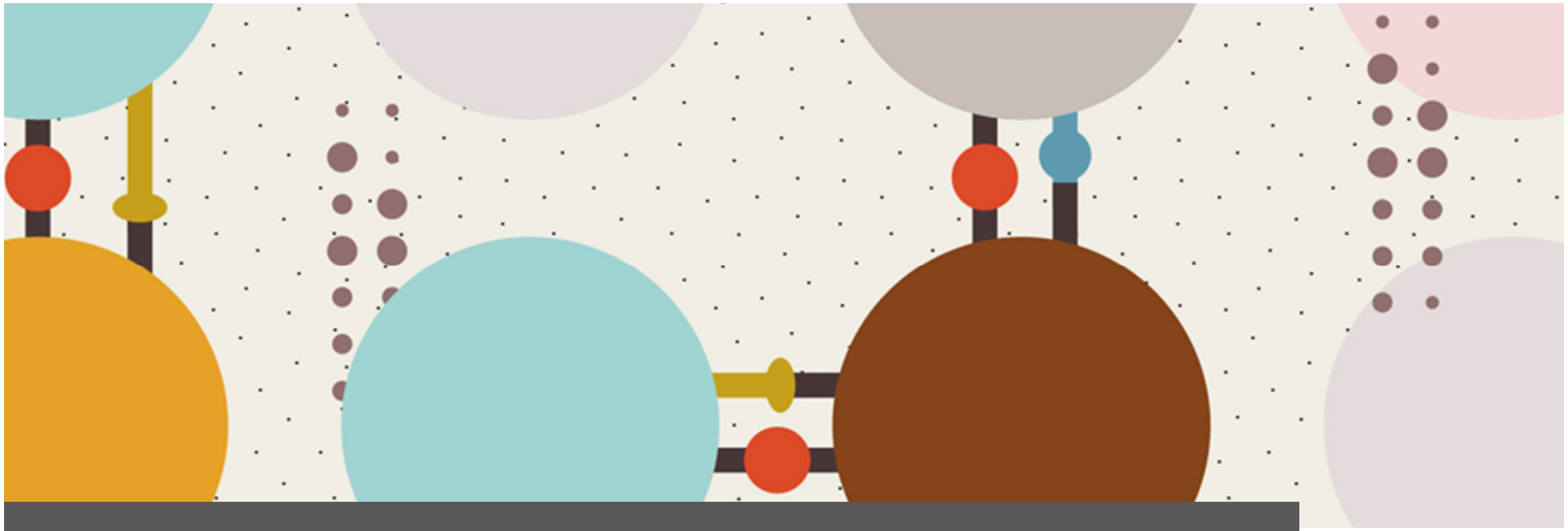


IO17 | Large Scale Bioinformatics for Immuno-Oncology

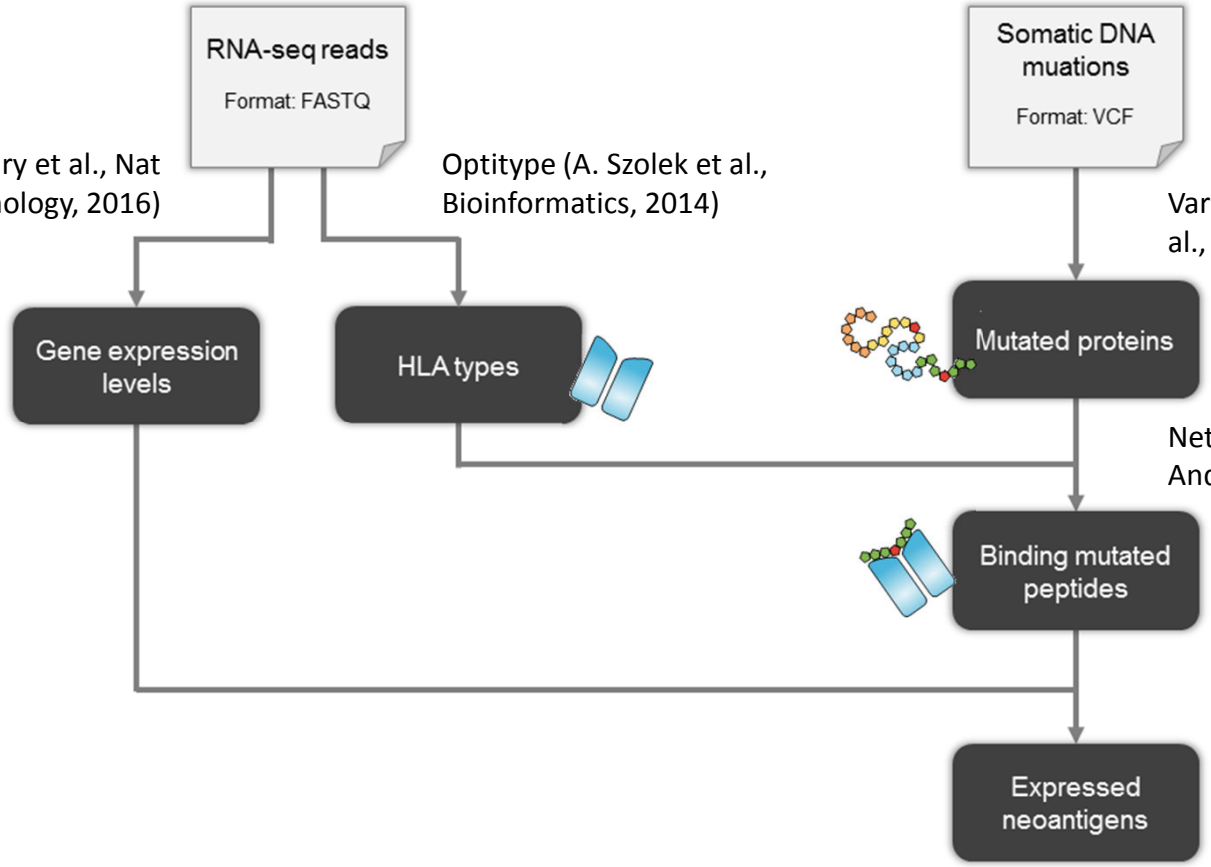
HLA typing

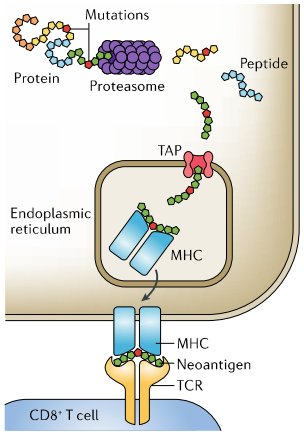
Francesca Finotello, Federica Eduati, and Pedro L. Fernandes

GTPB | The Gulbenkian Training Programme in Bioinformatics
Instituto Gulbenkian de Ciência, Oeiras, Portugal | Sept 19th-22nd, 2017

INSTITUTO GULBENKIAN DE CIÊNCIA

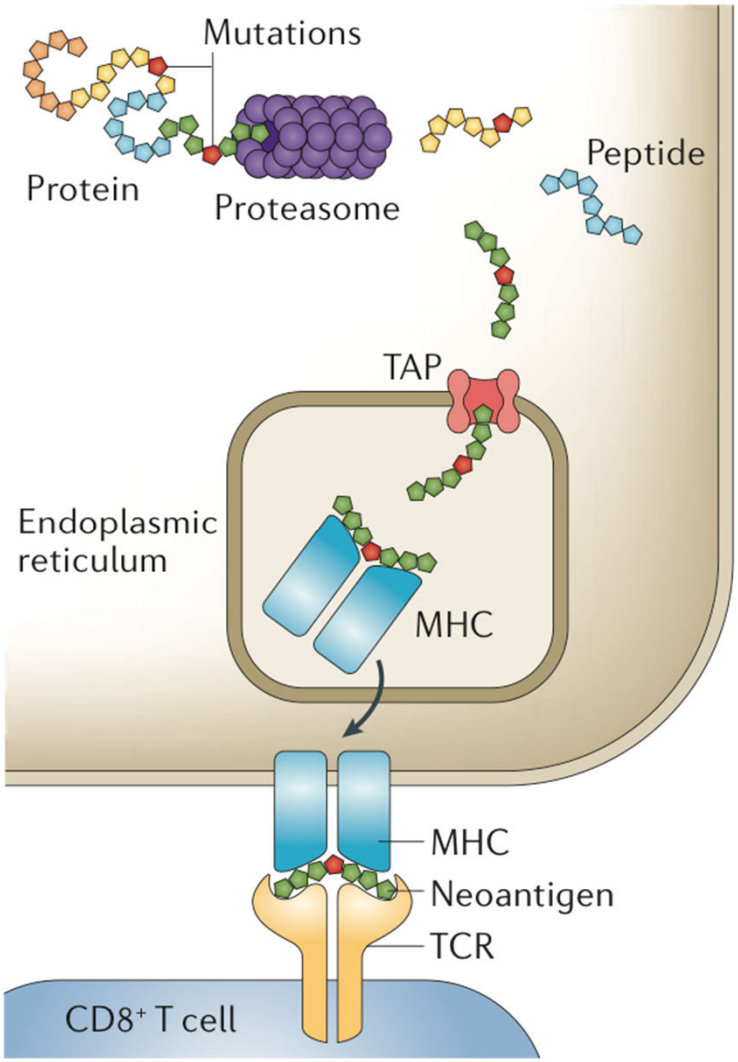
A pipeline for the prediction of class-I neoantigen





E Tappeiner, F Finotello et al., Bioinformatics, 2017

The Human Leukocyte Antigen



MHC: major histocompatibility complex
HLA: human leukocyte antigen (MHC in humans)

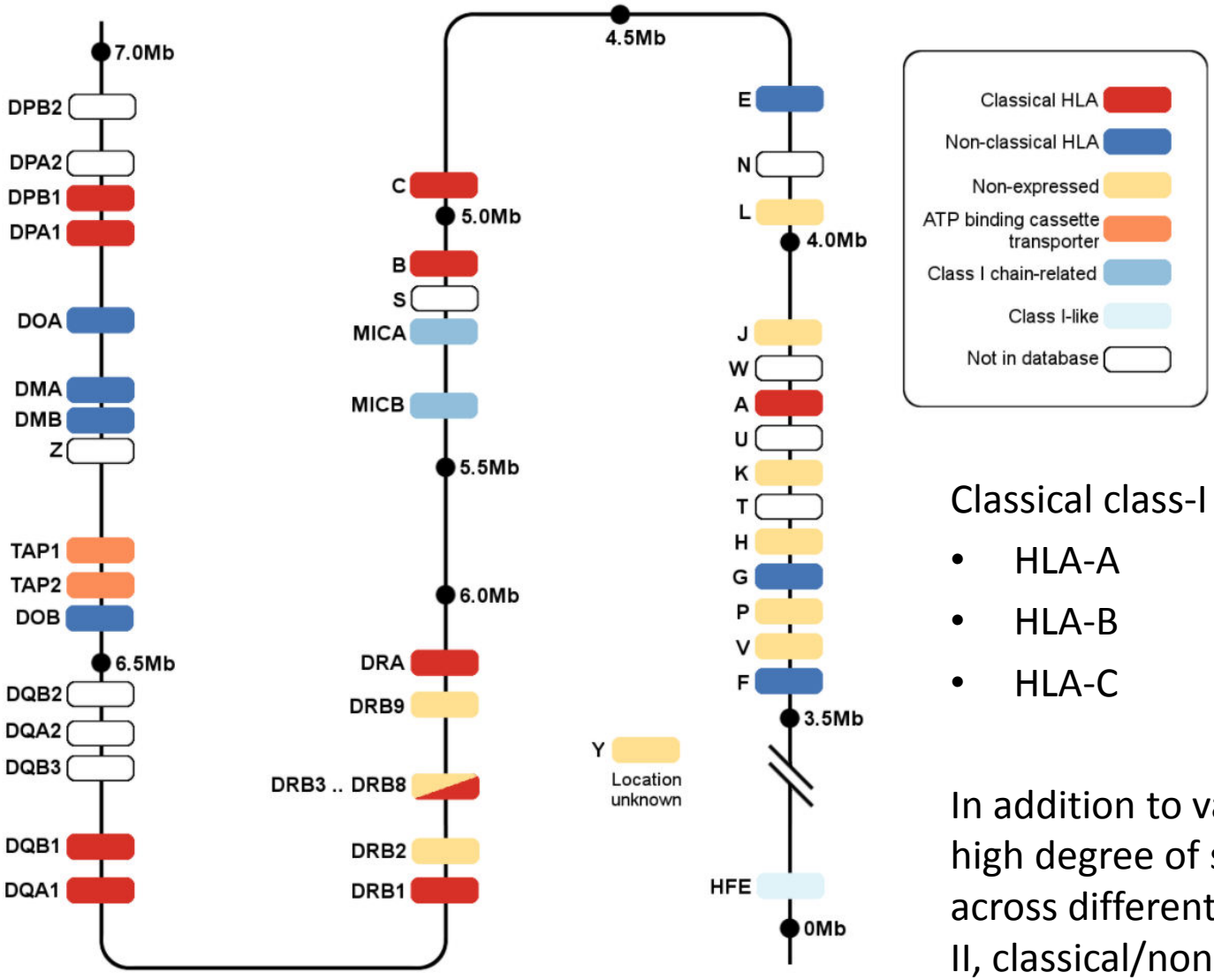
The HLA locus on chromosome 6:

- harbours more than two hundred genes and pseudogenes
- is on of the most polymorphic regions of the human genome

International immunogenetics project HLA (IMGT/HLA) database: collection of more than 13,000 annotated HLA alleles

H Hackl*, P Charoentong*, F Finotello* et al. Nature Reviews Genetics, 2016

The HLA locus



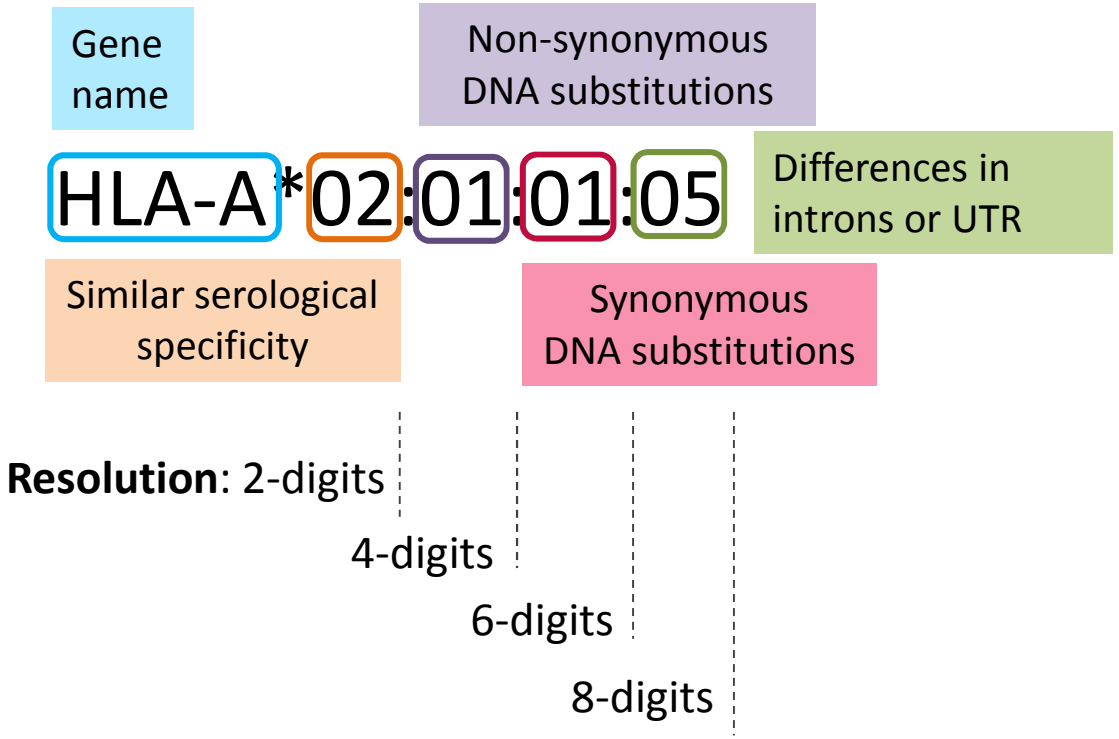
Classical class-I HLA:

- HLA-A
- HLA-B
- HLA-C

In addition to vast allelic variation, high degree of sequence similarity across different loci (class I/class II, classical/non-classical)

Image from: http://hla.alleles.org/alleles/index.html

The HLA nomenclature



Example: HLA-A*02:02 and HLA-A*02:01

- equal at two-digit resolution but not at four-digit resolution
- similar serological specificity for a peptide
- different protein sequence → different T cell recognition of the peptide-MHC complex

http://hla.alleles.org/nomenclature/index.html

Computational tools for HLA typing from NGS dat

ATHLATES	Genotyping of HLA-I and HLA-II alleles, from WGS and WES Illumina data	http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/athlates
HLAforest	Hierarchical reconstruction of HLA-I and HLA-II alleles from RNA-seq data	http://code.google.com/p/hlaforest
HLAminer	Extraction of HLA-I and HLA-II types from non-targeted RNA-seq, WGS and WES data based on read mapping or *de novo* assembly	http://www.bcgsc.ca/platform/bioinfo/software/hlaminer
HLAreporter	WGS- and WES-based genotyping of HLA-I and HLA-II alleles at six-digit resolution	http://paed.hku.hk/genome/software.html
HLA-VBseq	Extraction of eight-digit resolution HLA-I and HLA-II from WGS data	http://nagasakilab.csml.org/hla
Optitype	High-accuracy genotyping of classical HLA-I alleles from RNA-seq, WGS and WES data	http://github.com/FRED-2/OptiType
PHLAT	Genotyping of HLA-I and HLA-II alleles from RNA-seq, WGS, WES and targeted sequencing for different read lengths and coverages	http://sites.google.com/site/phlatfortype
Polysolver	Genotyping of HLA-I alleles from WES data and calling of somatic mutations in the HLA loci	http://www.broadinstitute.org/cancer/cga/polysolver
Seq2HLA	Extraction of HLA-I and HLA-II types from whole-genome RNA-seq, currently optimized also for four-digit resolution	http://bitbucket.org/sebastian_boegel/seq2hla

H Hackl*, P Charoentong*, F Finotello* et al. Nature Reviews Genetics, 2016

Optitype

- Reconstructs the major class-I HLA alleles (HLA-A, HLA-B, and HLA-C) from RNA-seq, WES, or WGS data
- Considers all major and minor HLA-I loci simultaneously (reads can map to alleles of multiple loci equally well)
- Hypothesis: the correct genotype explains the source of more reads than any other genotype

Strategy:

1. Maps the reads against a constructed HLA allele reference (exons 2 and 3, inputed intronic regions for DNA data)
2. Computes a matrix of read-allele correspondences (match with the least number of mismatches)
3. Selects up to two alleles for each locus simultaneously, maximizing the number of mapped reads that can be explained by the predicted genotype

Szolek A, et al. Bioinformatics, 2014

FASTA format

Used to represent medium/long reads and biological sequences as chromosome and proteins

For each sequence/read:

- “>” character followed by the sequence ID (one line)
- Sequence of nucleotides or amino acids in FASTA alphabet (can span multiple lines)

For sequence reads, possible associated file of Phred quality scores $P=-10\log_{10}(e)$, where e is the probability that the base is wrong

```
>seq1
ACGTTCGTAGTAGATAGATATAGAT
AGTAAAGATGATGAGATCATCGATG
GATAGGTAGGGTGGATAGTACGATG
GATA
>seq2
AGTGATGATGACTCTCGAAAAAGCT
GATCTAGATCAGCTGATCGAT
…
```

```
>seq1
40 40 40 37 37 34
…
>seq2
40 40 40 40 29 12
…
```

https://en.wikipedia.org/wiki/FASTA_format

FASTQ format

Main format for short sequencing reads

For each read, 4 lines:

- "@" character followed by the read ID (and possibly by additional information)
- Read sequence in FASTA alphabet
- "+" character (possibly followed by the read ID)
- Quality scores in ASCII encoding (same length of the read sequence)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

For paired-end reads, we (usually) have two files reporting the first and second mate of the pairs, respectively

https://en.wikipedia.org/wiki/FASTQ_format

Run Optitype with TIminer

```
from TIminer import TIminerAPI

TIminerAPI.executeOptitype(…)
```

From TIminer documentation

http://icbi.i-med.ac.at/software/timiner/doc/index.html

`TIminer.TIminerAPI.`**`executeOptitype`**(*inputtype*, *inputFile1*, *inputFile2=None*, *outputFile=None*, *subjectId='unknown'*, *threadCount=2*)

This function takes as input FASTQ files of sequencing reads and predicts class-I HLA types for HLA-A, HLA-B, and HLA-C genes using Optitype.

Parameters:

- **inputtype** (*str*) – 'rna' for RNA sequencing data or 'dna' DNA sequencing data.
- **inputFile1** (*str*) – Path to the first `FASTQ file` containing the NGS reads.
- **inputFile2** (*str*) – For paired-end data, path to the second `FASTQ file` of NGS reads (optional).
- **outputFile** (*str*) – Path to the `output file` containing the subject ID and the identified HLAs in tab-delimited columns (optional, default = *HLA-types/*).
- **subjectId** (*str*) – Subject ID to be stored in the result file (optional, default = *unknown*).
- **threadCount** (*int*) – Number of threads to be used (optional, default = *2*).

Optitype@TIminer output: HLA alleles

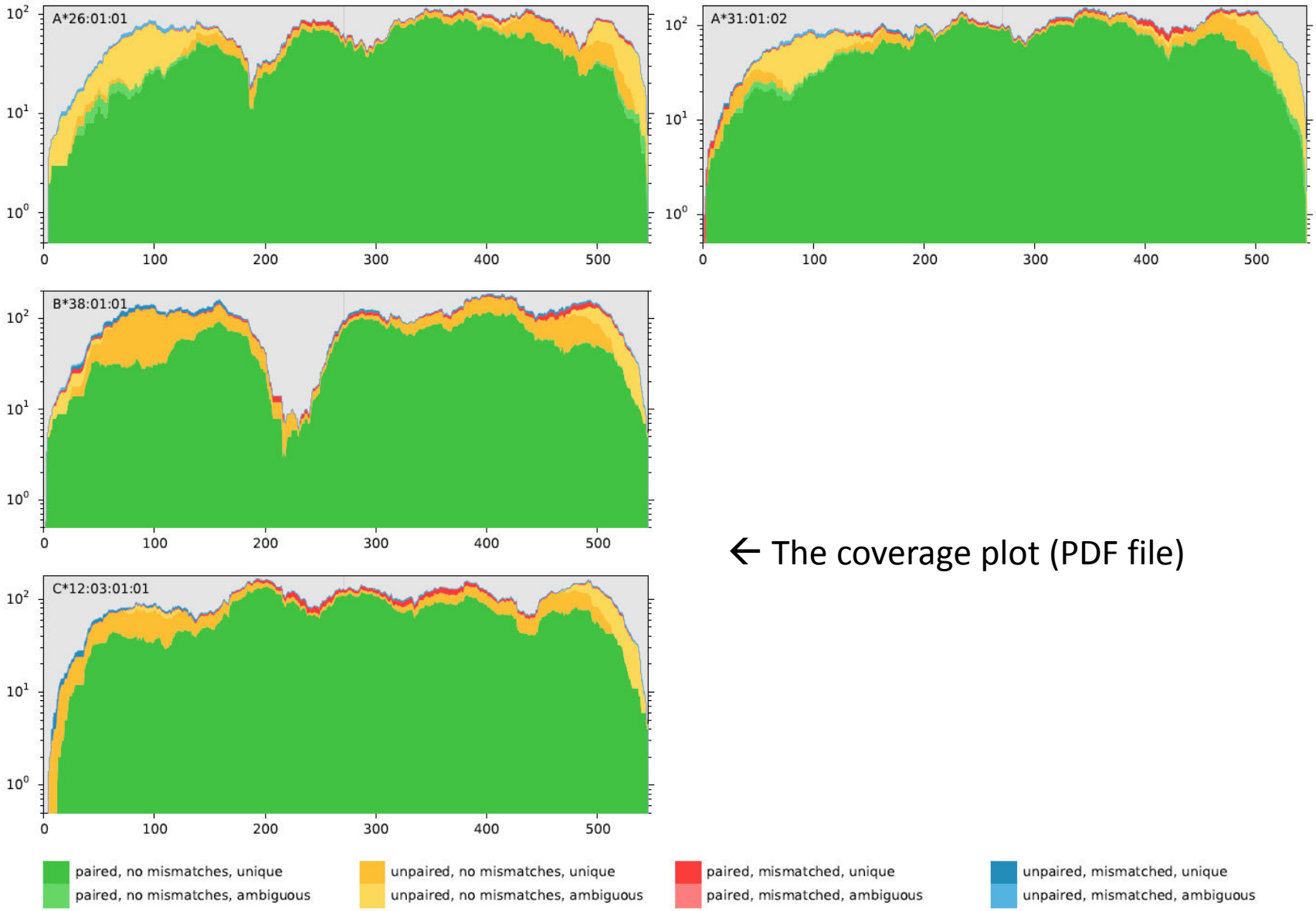
Text file of the HLA alleles at four digits resolution:

heterozygous

HLA-A31:01	HLA-A26:01	HLA-B38:01	HLA-B38:01	HLA-C12:03	HLA-C12:03

homozygous

Optitype@TIminer output: coverage plot



← The coverage plot (PDF file)