
BE A
MATCH
SAVE
A LIFE
HLA & NGS
EFI 2013
ANTHONY
NOLAN
BE A MATCH, SAVE A LIFE

James Robinson
Senior Bioinformatics Scientist
Anthony Nolan Research Institute

IMGT/HLA XML

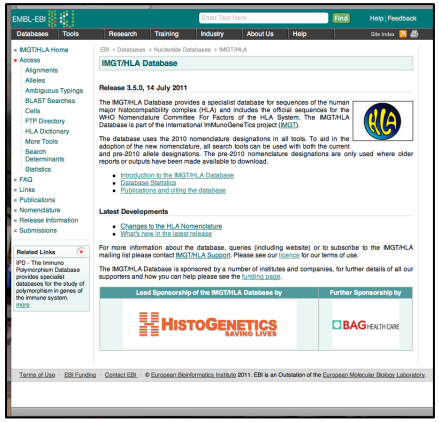
- XML - Extensible Markup Language
  - rules for encoding documents
  - machine-readable form.
  - emphasize simplicity, generality, and usability over the Internet.
  - textual data format with strong support via Unicode
  - widely used for the representation of arbitrary data structures, for example in web services.

- HIG have worked with the NMDP Bioinformatics and some Biotech companies to develop a standardised machine readable export of the HLA alignments with supporting data
  - NMDP looking to move their infrastructure to be based upon these exports
  - Biotech companies looking to use the XML to aid in the compilation and updating of the libraries used for primer design and software
  - EBI moving search functions to only use XML platforms

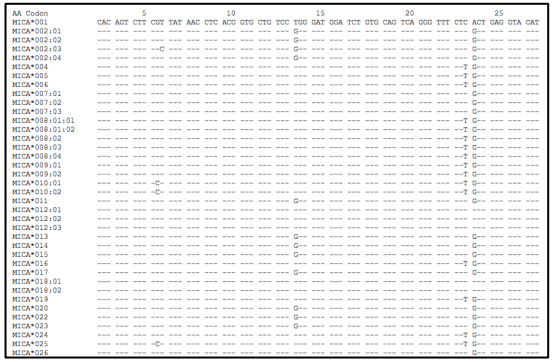
Objective

- To translate these, into single export file, which all parties can use:

Nomenclature Reports



IMGT/HLA Database



Sequence Alignments

ANTHONY NOLAN
BE A MATCH, SAVE A LIFE

Progress

- Final stages of beta testing

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<alleles xmlns:xs="http://www.w3.org/2001/XMLSchema"  xs:noNamespaceSchemaLocation="http://hla.alleles.org/xml/hla.xsd
  <allele id="HLA00001" name="A*01:01:01:01" dateassigned="1989-08-01">
    <releaseversions firstreleased="1.0.0" lastupdated="1.0.0" currentrelease="3.5.0" />
    <locus genesystem="HLA" locusname="A" class="I" />
    <citations>
      <citation pubmed="3375250" authors="Parham P, Lomen CE, Lawlor DA, Ways JP, Holmes N, Coppin HL, Salter RD, Wan
      <citation pubmed="2251137" authors="Girdlestone J" title="Nucleotide sequence of an HLA-A1 gene" location="Nucle
      <citation pubmed="9349617" authors="Laforet M, Froelich N, Parissiadis A, Pfeiffer B, Schell A, Faller B, Woehl-
      <citation pubmed="15140828" authors="Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunha
      <citation pubmed="19735485" authors="Zhu F, He Y, Zhang W, He J, He J, Xu X, Yan L" title="Analysis of the compl
    </citations>
    <sourcexrefs>
      <xref acc="AJ278305" pid="CAB93537.1" />
      <xref acc="AL645935" />
      <xref acc="CR759913" />
      <xref acc="EU445470" pid="ACA34990.1" />
      <xref acc="GU812295" pid="ADE80886.1" />
      <xref acc="M24043" pid="AAA59652.1" />
      <xref acc="X55710" pid="CAA39243.1" />
      <xref acc="Z93949" pid="CAB07989.1" />
    </sourcexrefs>
    <sourcematerial>
      <species latinname="Homo sapiens" commonname="Human" ncbitaxon="9606" />
      <ethnicity>Oriental, Caucasoid</ethnicity>
      <samples>
        <sample name="7550800303" />
        <sample name="APD" />
        <sample name="B4702" />
        <sample name="COX" />
        <sample name="LCL721" />
        <sample name="MOLT-4" />
        <sample name="PP" />
      </samples>
    </sourcematerial>
    <sequence>
      <alignmentreference allelename="A*01:01:01:01" alleleid="HLA00001" />
      <nucsequence>ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCCCACTCCATGAGGTATTTCTTC
      <feature id="1.1" order="1" type="Exon" name="Exon 1" status="complete">
          <SequenceCoordinates start="1" end="73" />  <cDNACoordinates start="1" end="73" readingframe="1" />
      </feature>
```

Progress – sample XSD

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified" xmlns:xs="htt
  <xs:import namespace="http://www.w3.org/2001/XMLSchema" schemaLocation="xs.xsd"/>
  <xs:element name="alleles">
    <xs:complexType>
      <xs:sequence>
        <xs:element maxOccurs="unbounded" ref="allele"/>
      </xs:sequence>
      <xs:attribute ref="xs:noNamespaceSchemaLocation" use="required"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="allele">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="releaseversions"/>
        <xs:element ref="locus"/>
        <xs:element minOccurs="0" ref="citations"/>
        <xs:element minOccurs="0" ref="sourcexrefs"/>
        <xs:element ref="sourcematerial"/>
        <xs:element ref="sequence"/>
      </xs:sequence>
      <xs:attribute name="dateassigned" use="required" type="xs:NMTOKEN"/>
      <xs:attribute name="id" use="required" type="xs:NCName"/>
      <xs:attribute name="name" use="required"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="releaseversions">
    <xs:complexType>
      <xs:attribute name="currentrelease" use="required" type="xs:NMTOKEN"/>
      <xs:attribute name="firstreleased" use="required" type="xs:NMTOKEN"/>
      <xs:attribute name="lastupdated" use="required" type="xs:NMTOKEN"/>
    </xs:complexType>
  </xs:element>
  <xs:element name="locus">
    <xs:complexType>
      <xs:attribute name="class" use="required" type="xs:NCName"/>
      <xs:attribute name="genesystem" use="required" type="xs:NCName"/>
      <xs:attribute name="locusname" use="required" type="xs:NCName"/>
    </xs:complexType>
  </xs:element>
```


ANTHONY NOLAN
BE A MATCH, SAVE A LIFE

Progress – sample entry

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<alleles xmlns:xs="http://www.w3.org/2001/XMLSchema"  xs:noNamespaceSchemaLocation="http://hla.alleles.org/xml/hla.xsd">
  <allele id="HLA00001" name="A*01:01:01:01" dateassigned="1989-08-01">
    <releaseversions firstreleased="1.0.0" lastupdated="1.0.0" currentrelease="3.5.0" />
    <locus genesystem="HLA" locusname="A" class="I" />
    <citations>
      <citation pubmed="3375250" authors="Parham P, Lomen CE, Lawlor DA, Ways JP, Holmes N, Coppin HL, Salter RD, Wan AM, 
      <citation pubmed="2251137" authors="Girdlestone J" title="Nucleotide sequence of an HLA-A1 gene" location="Nucleic A
      <citation pubmed="9349617" authors="Laforet M, Froelich N, Parissiadis A, Pfeiffer B, Schell A, Faller B, Woehl-Jaeg
      <citation pubmed="15140828" authors="Stewart CA, Horton R, Allcock RJ, Ashurst JL, Atrazhev AM, Coggill P, Dunham I, 
      <citation pubmed="19735485" authors="Zhu F, He Y, Zhang W, He J, He J, Xu X, Yan L" title="Analysis of the complete g
    </citations>
    <sourcexrefs>
      <xref acc="AJ278305" pid="CAB93537.1" />
      <xref acc="AL645935" />
      <xref acc="CR759913" />
      <xref acc="EU445470" pid="ACA34990.1" />
      <xref acc="GU812295" pid="ADE80886.1" />
      <xref acc="M24043" pid="AAA59652.1" />
      <xref acc="X55710" pid="CAA39243.1" />
      <xref acc="Z93949" pid="CAB07989.1" />
    </sourcexrefs>
    <sourcematerial>
      <species latinname="Homo sapiens" commonname="Human" ncbitaxon="9606" />
      <ethnicity>Oriental, Caucasoid</ethnicity>
      <samples>
        <sample name="7550800303" />
        <sample name="APD" />
        <sample name="B4702" />
        <sample name="COX" />
        <sample name="LCL721" />
        <sample name="MOLT-4" />
        <sample name="PP" />
      </samples>
    </sourcematerial>
```

Progress – sample entry

```
<sequence>
  <alignmentreference allelename="A*01:01:01:01" alleleid="HLA00001" />
  <nucsequence>ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCCCACTCCATGAGGTATTTCTTCACATC
  <feature id="1.1" order="1" featuretype="Exon" name="Exon 1" status="complete">
      <SequenceCoordinates start="1" end="73" />
      <cDNACoordinates start="1" end="73" readingframe="1" />
  </feature>
  <feature id="1.2" order="2" featuretype="Exon" name="Exon 2" status="complete">
      <SequenceCoordinates start="74" end="343" />
      <cDNACoordinates start="74" end="343" readingframe="3" />
  </feature>
  <feature id="1.3" order="3" featuretype="Exon" name="Exon 3" status="complete">
      <SequenceCoordinates start="344" end="619" />
      <cDNACoordinates start="344" end="619" readingframe="3" />
  </feature>
  <feature id="1.4" order="4" featuretype="Exon" name="Exon 4" status="complete">
      <SequenceCoordinates start="620" end="895" />
      <cDNACoordinates start="620" end="895" readingframe="3" />
  </feature>
  <feature id="1.5" order="5" featuretype="Exon" name="Exon 5" status="complete">
      <SequenceCoordinates start="896" end="1012" />
      <cDNACoordinates start="896" end="1012" readingframe="3" />
  </feature>
  <feature id="1.6" order="6" featuretype="Exon" name="Exon 6" status="complete">
      <SequenceCoordinates start="1013" end="1045" />
      <cDNACoordinates start="1013" end="1045" readingframe="3" />
  </feature>
  <feature id="1.7" order="7" featuretype="Exon" name="Exon 7" status="complete">
      <SequenceCoordinates start="1046" end="1093" />
      <cDNACoordinates start="1046" end="1093" readingframe="3" />
  </feature>
  <feature id="1.8" order="8" featuretype="Exon" name="Exon 8" status="complete">
      <SequenceCoordinates start="1094" end="1098" />
      <cDNACoordinates start="1094" end="1098" readingframe="3" />
  </feature>
  <feature id="1.9" name="Translation" featuretype="Protein">
    <translation>MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEGPEYWDQETRNMKAHSQTDRANL
  </feature>
</sequence>
```


ANTHONY NOLAN
BE A MATCH, SAVE A LIFE

Progress – sample entry

```
<sequence>
  <alignmentreference allelename="A*01:01:01:01" alleleid="HLA00001" />
  <nucsequence>ATGGCCGTCATGGCGCCCCGAACCCTCCTCCTGCTACTCTCGGGGGCCCTGGCCCTGACCCAGACCTGGGCGGGCTCCCACTCCATGAGGTATTTCTTCACAT
  <feature id="4.1" order="1" featuretype="Exon" name="Exon 1" status="complete">
      <SequenceCoordinates start="1" end="73" />
      <cDNACoordinates start="1" end="73" readingframe="1" />
  </feature>
  <feature id="4.2" order="2" featuretype="Exon" name="Exon 2" status="complete">
      <SequenceCoordinates start="74" end="343" />
      <cDNACoordinates start="74" end="343" readingframe="3" />
  </feature>
  <feature id="4.3" order="3" featuretype="Exon" name="Exon 3" status="complete">
      <SequenceCoordinates start="344" end="619" />
      <cDNACoordinates start="344" end="619" readingframe="3" />
  </feature>
  <feature id="4.4" order="4" featuretype="Exon" name="Exon 4" status="complete">
      <SequenceCoordinates start="620" end="896" />
      <cDNACoordinates start="620" end="895" readingframe="3"
      <cDNAindel start="627" end="628" size="1"  type="insertion"/>
  </feature>
  <feature id="4.5" order="5" featuretype="Exon" name="Exon 5" status="complete">
      <SequenceCoordinates start="897" end="1013" />
      <cDNACoordinates start="896" end="1012" readingframe="3" />
  </feature>
  <feature id="4.6" order="6" featuretype="Exon" name="Exon 6" status="complete">
      <SequenceCoordinates start="1014" end="1046" />
      <cDNACoordinates start="1013" end="1045" readingframe="3" />
  </feature>
  <feature id="4.7" order="7" featuretype="Exon" name="Exon 7" status="complete">
      <SequenceCoordinates start="1047" end="1094" />
      <cDNACoordinates start="1046" end="1093" readingframe="3" />
  </feature>
  <feature id="4.8" order="8" featuretype="Exon" name="Exon 8" status="complete">
      <SequenceCoordinates start="1095" end="1099" />
      <cDNACoordinates start="1094" end="1098" readingframe="3" />
  </feature>
  <feature id="4.9" name="Translation" featuretype="Protein">
    <translation>MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQKMEPRAPWIEQEGPEYWDQETRNMKAHSQTDRAN
  </feature>
</sequence>
```


ANTHONY NOLAN
BE A MATCH, SAVE A LIFE

Additional Modules

- Added additional XML and XSD to encode SBT ambiguities
  - Provides machine readable format of PDF and XLS documents
  - Provides diploid sequence of ambiguity
- Looking into other datasets provided by IMGT/HLA that could also be provided as XML.

Example XML

```
<?xml version="1.0" encoding="UTF-8"?>  <tns:ambiguityData xmlns:tns="http://www.example.org/ambig-aw"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.example.org/ambig-aw ambig-aw.xsd ">
    <tns:releaseVersion currentRelease="3.12.0" date="2013-04-17" />
    <tns:geneList>
        <tns:gene geneSystem="HLA" name="A">
            <tns:gGroupsList>
                <tns:gGroup name="A*01:01:01G" gid="HGG00001">
                    <tns:gGroupAllele name="A*01:01:01:01" alleleid="HLA00001" />
                    <tns:gGroupAllele name="A*01:01:01:02N" alleleid="HLA02169" />
                    <tns:gGroupAllele name="A*01:01:38L" alleleid="HLA03587" />
                    <tns:gGroupAllele name="A*01:01:51" alleleid="HLA08781" />
                    <tns:gGroupAllele name="A*01:04N" alleleid="HLA00004" />
                    <tns:gGroupAllele name="A*01:22N" alleleid="HLA02878" />
                    <tns:gGroupAllele name="A*01:32" alleleid="HLA03522" />
                    <tns:gGroupAllele name="A*01:37" alleleid="HLA03831" />
                    <tns:gGroupAllele name="A*01:45" alleleid="HLA04137" />
                    <tns:gGroupAllele name="A*01:56N" alleleid="HLA05224" />
                    <tns:gGroupAllele name="A*01:81" alleleid="HLA05908" />
                    <tns:gGroupAllele name="A*01:87N" alleleid="HLA06531" />
                    <tns:gGroupAllele name="A*01:103" alleleid="HLA07356" />
                    <tns:gGroupAllele name="A*01:107" alleleid="HLA07776" />
                    <tns:gGroupAllele name="A*01:109" alleleid="HLA07807" />
                </tns:gGroup>
```

Example XML

```
<tns:ambiguousComboGroup>
    <tns:ambiguousComboElement>
        <tns:ambigAllele1 name="A*01:01:01G" alleleid="HGG00001" />
        <tns:ambigAllele2 name="A*02:06:02" alleleid="HLA01956" />
    </tns:ambiguousComboElement>
    <tns:ambiguousComboElement>
        <tns:ambigAllele1 name="A*01:01:04" alleleid="HLA02540" />
        <tns:ambigAllele2 name="A*02:06:01G" alleleid="HGG00005" />
    </tns:ambiguousComboElement>
    <tns:ambiguousComboElement>
        <tns:ambigAllele1 name="A*01:43" alleleid="HLA04122" />
        <tns:ambigAllele2 name="A*02:01:49" alleleid="HLA05243" />
    </tns:ambiguousComboElement>
    <tns:glElement glstring="A*01:01:01G+A*02:06:02|A*01:01:04+A*02:06:01G|A*01:43+A*02:01:49" />
    <tns:diploidsequence ambigsequence="GCTCYCACTCCATGAGGTATTTCTWCACMTCCGTGTCCCGGCCCGGCCGCGGGGAGCCCCGCTTC
    ATCGCMGTGGGCTACGTGGACGACACGCAGTTCGTGCGGTTCGACAGCGACGCCGCGAGCCAGARGATGGAGCCGCGGGCGCCGTGGATAGAGCAGGAGGG
    KCCGGAGTATTGGGACSRGGAGACACGGAAWRTGAAGGCCCACTCACAGACTSACCGAGYGRAYCTGGGGACCCTGCGCGGCTACTACAACCAGAGCGAGG
    MCG|GTTCTCACACCRTCCAGAKRATGTATGGCTGCGACGTGGGGYCGGACKGGCGCTTCCTCCGCGGGTACCRSCAGKACGCCTACGACGGCAAGGATTA
    CATCGCCCTGAAMGAGGACCTGCGCTCTTGGACCGCGGCGGACATGGCAGCTCAGAYCACCAAGCRCAAGTGGGAGGCGGYCCATGYGGCGGAGCAGYKGA
    GAGYCTACCTGGAGGGCMSGTGCGTGGASKGGCTCCGCAGATACCTGGAGAACGGGAAGGAGACGCTGCAGCGCACGG" />
</tns:ambiguousComboGroup>
```

http://hla.alleles.org/xml/