

# Hapl-o-Mat – Data Preparation

Please refer to [detailedGettingStartedLinux](#), [gettingStarted](#), or [detailedGettingStartedWindows](#) for information on how to use Hapl-o-Mat

## *Data Preparation*

Hapl-o-Mat relies on information on the HLA nomenclature. This information is provided by data files, which we are going to create. As the HLA nomenclature evolves over time, e.g. by finding new alleles or adding new multiple allele codes, it is important to update data from time to time. Hapl-o-Mat relies on the following files, which must be placed in the folder “Hapl-o-Mat/data” for Hapl-o-Mat to work:

File name	Description
AllAllelesExpanded.txt	A list of relevant existing HLA alleles with their enclosed more-digit typing resolutions
AlleleList.txt	If your input data in GLS format includes a missing single-locus genotype, it can be replaced by combining all alleles of the same locus from this file. You only must create it in this case.
Ambiguity.txt	Data basis for the ambiguity filter
LargeG.txt	A list of G-groups with their enclosed alleles in 8-digit resolution
MultipleAlleleCodes.txt	A list of multiple allele codes and their translation to alleles in 4-digit resolution
P.txt	A list of P-groups with their enclosed alleles in 8-digit resolution
Smallg.txt	A list of g-groups with their enclosed alleles in 8-digit resolution

In the following we are going to create these data files. Enter the folder “prepareData”. Everything is going to happen from here.

As the data-processing is a little bit tedious, we provide you with an automated script, see “Automatic Way”. If you insist on doing it all on your own head to “Manual Way”.

## *Automatic Way*

If you are the lazy guy, you can just run the python script “BuildData.py”, which does the whole job for you.

## *Manual Way*

Here, we perform the data preparation step by step.

### **Download Data**

First, we need some input data from the web. Download the following files:

- 1) Go to the website [http://hla.alleles.org/wmda/hla\\_nom\\_p.txt](http://hla.alleles.org/wmda/hla_nom_p.txt) and save the file hla\_nom\_p.txt by right-clicking and choosing “Save as...”. Move the file “hla\_nom\_p.txt” to the folder “Hapl-o-Mat/prepareData”.

```
# file: hla_nom_p.txt
# date: 2016-04-15
# version: IPD-IMGT/HLA 3.24.0
# origin: http://hla.alleles.org/wmda/hla_nom_p.txt
# author: WHO, Steven G. E. Marsh (steven.marsh@ucl.ac.uk)
A*01:01:01:01:01:01:03:01:02:01:01:03:01:04:01:05:01:06:01:01:
8/01:01:19/01:01:20/01:01:21/01:01:22/01:01:23/01:01:24/01:01:25/01:01:26/01:
:38/01:01:39/01:01:40/01:01:41/01:01:42/01:01:43/01:01:44/01:01:45/01:01:46/
01:58/01:01:59/01:01:60/01:01:61/01:01:62/01:01:63/01:01:64/01:01:65/01:01:6
01:81/01:01:83/01:01:87/01:01:89/01:01:92/01:01:94/01:01:95/01:01:97/01:01:
A*01:02;
A*01:03;
A*01:06;
A*01:07;
A*01:08;
A*01:09;
A*01:10;
A*01:12;
A*01:13;
A*01:14;
A*01:17;
A*01:19;
A*01:20;
A*01:21;
A*01:23;
A*01:24;
A*01:25;
A*01:26;
A*01:28;
A*01:29;
A*01:30;
A*01:33;
A*01:35;
A*01:36;
A*01:38;
A*01:39;
A*01:40;
A*01:41;
A*01:42;
A*01:43;
```

- Go to the website [http://hla.alleles.org/wmda/hla\\_nom\\_g.txt](http://hla.alleles.org/wmda/hla_nom_g.txt) and save the file hla\_nom\_g.txt (same as in 1)). Move the file "hla\_nom\_g.txt" to the folder "Hapl-o-Mat/prepareData".
- Go to the website <https://bioinformatics.bethematchclinical.org/HLA-Resources/Allele-Codes/Allele-Code-Lists/Allele-Code-List-in-Alphabetical-Order/>. Click on "Alphabetical Allele Code List (ZIP) (new nomenclature)" and save alpha.v3.zip.

Allele Code List in Alphabetical Order

Example:  
AA 01:02/03/05  
AB 01:02  
AC 01:03

The allele code list in alphabetical order is provided in a variety of formats:

**HTML Format (.html)**

This format is recommended if you simply want to view the allele code list online.

- [Alphabetical Allele Code List \(HTML\) \(new nomenclature\)](#)  
**Note:** New window will open.
- [Alphabetical Allele Code List \(HTML\) \(old nomenclature\)](#)  
**Note:** New window will open.

**Text Format (.txt)**

Download a zip compressed file. Once extracted, the text file will be called "alpha.txt."

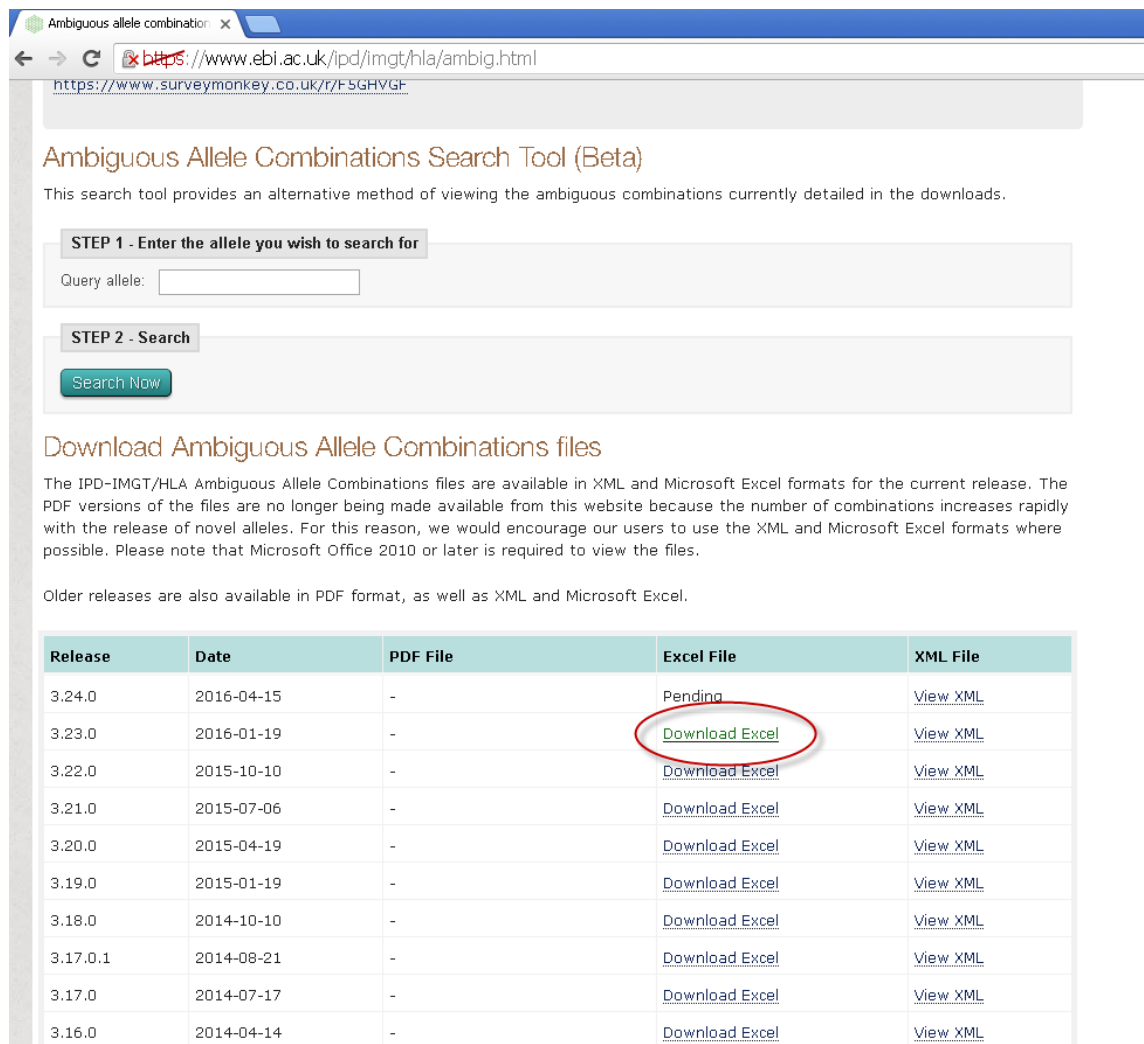
- [Alphabetical Allele Code List \(ZIP\) \(new nomenclature\)](#)  
**Note:** Extraction requires a data compression program such as WinZip.
- [Alphabetical Allele Code List \(ZIP\) \(old nomenclature\)](#)  
**Note:** Extraction requires a data compression program such as WinZip.

The self-extracting executable file has been removed as of 10/21/03. The allele code lists will no longer be available for download in this format. Please mail [new-allelecodes@nmdp.org](mailto:new-allelecodes@nmdp.org) with any questions or concerns regarding this change.

Extract the archive alpha.v3.txt. This should be straightforward in Windows. Using a Terminal under Linux you can use the command “unzip alpha.v3.txt”. You can remove the archive “alpha.v3.zip” afterwards. We only need the file “alpha.v3”. Move it to the folder “Hapl-o-Mat/prepareData”.

4) You have two options to download the next file. The first approach is faster.

- a) Download the file [https://github.com/jrob119/IMGTHLA/raw/Latest/xml/hla\\_ambigs.xml.zip](https://github.com/jrob119/IMGTHLA/raw/Latest/xml/hla_ambigs.xml.zip). Extract it as in 3) including removing the archive. Move the file hla\_ambigs.xml to folder “Hapl-o-Mat/prepareData”.
- b) Go to the website <https://www.ebi.ac.uk/ipd/imgt/hla/ambig.html>. Click on "Download Excel" for the wanted release (usually the latest) and save ambiguity\_v<>.xls (replace <> by version).



**Ambiguous Allele Combinations Search Tool (Beta)**

This search tool provides an alternative method of viewing the ambiguous combinations currently detailed in the downloads.

**STEP 1 - Enter the allele you wish to search for**

Query allele:

**STEP 2 - Search**

[Search Now](#)

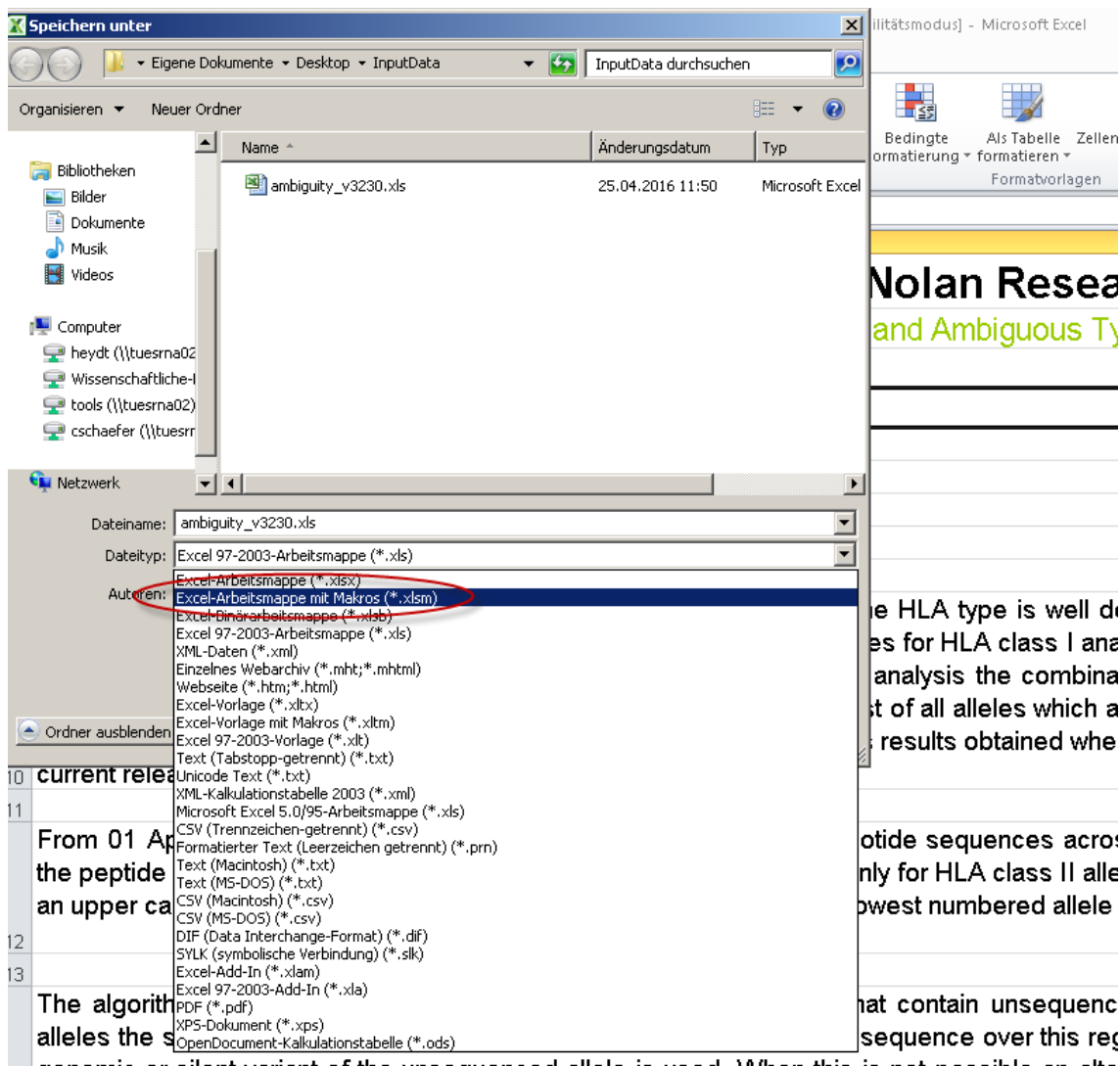
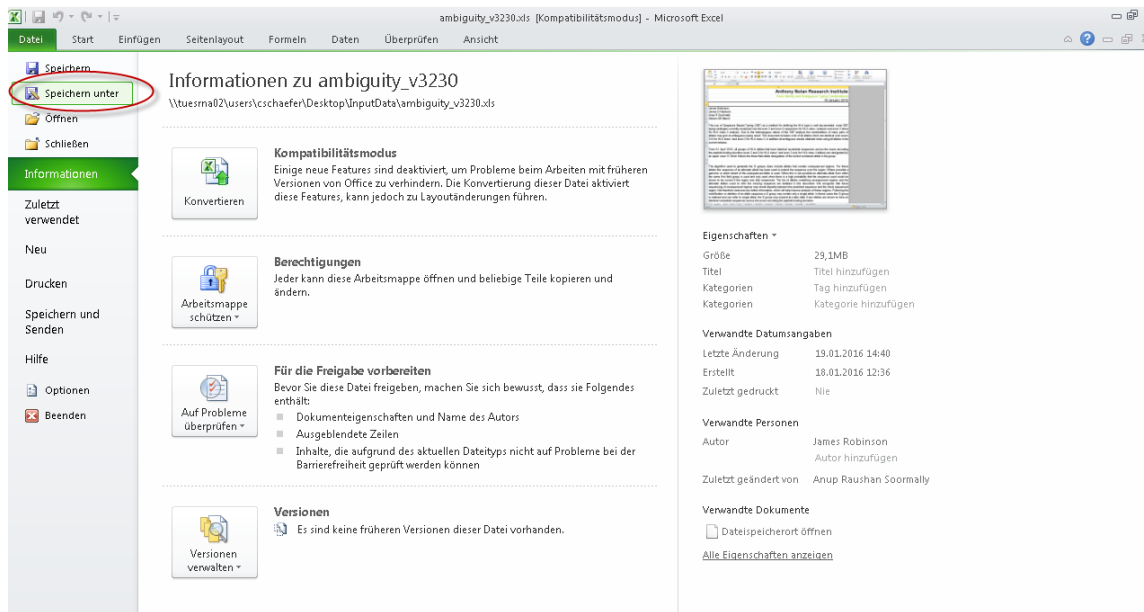
**Download Ambiguous Allele Combinations files**

The IPD-IMGT/HLA Ambiguous Allele Combinations files are available in XML and Microsoft Excel formats for the current release. The PDF versions of the files are no longer being made available from this website because the number of combinations increases rapidly with the release of novel alleles. For this reason, we would encourage our users to use the XML and Microsoft Excel formats where possible. Please note that Microsoft Office 2010 or later is required to view the files.

Older releases are also available in PDF format, as well as XML and Microsoft Excel.

Release	Date	PDF File	Excel File	XML File
3.24.0	2016-04-15	-	Pending	<a href="#">View XML</a>
3.23.0	2016-01-19	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.22.0	2015-10-10	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.21.0	2015-07-06	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.20.0	2015-04-19	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.19.0	2015-01-19	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.18.0	2014-10-10	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.17.0.1	2014-08-21	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.17.0	2014-07-17	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>
3.16.0	2014-04-14	-	<a href="#">Download Excel</a>	<a href="#">View XML</a>

Next we extract the information from the Excel sheet. Open ambiguity\_v<>.xls in Excel and save as ambiguity\_v<>.xlsm to run macros.



Now insert the following macro, which saves relevant information from the Excelsheets as text files:

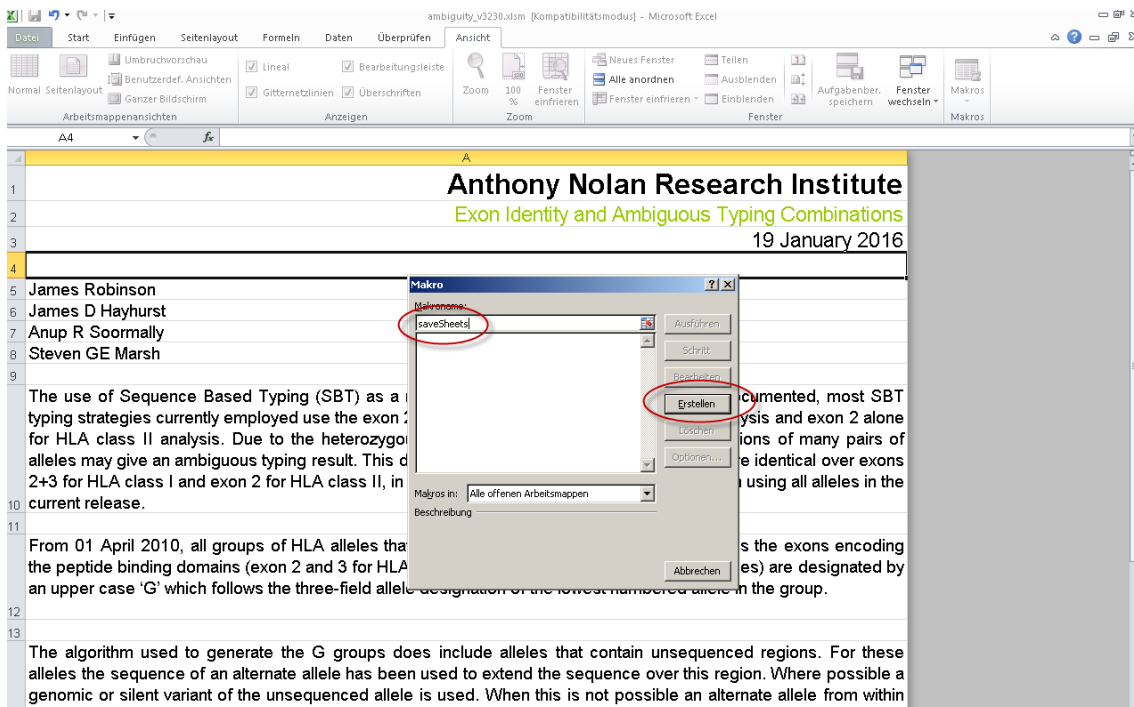
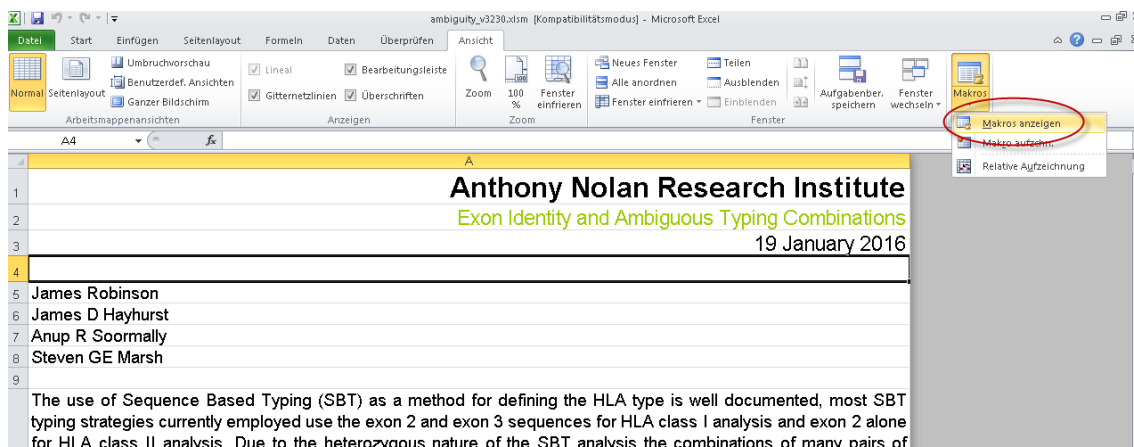
```
Sub saveSheets()
```

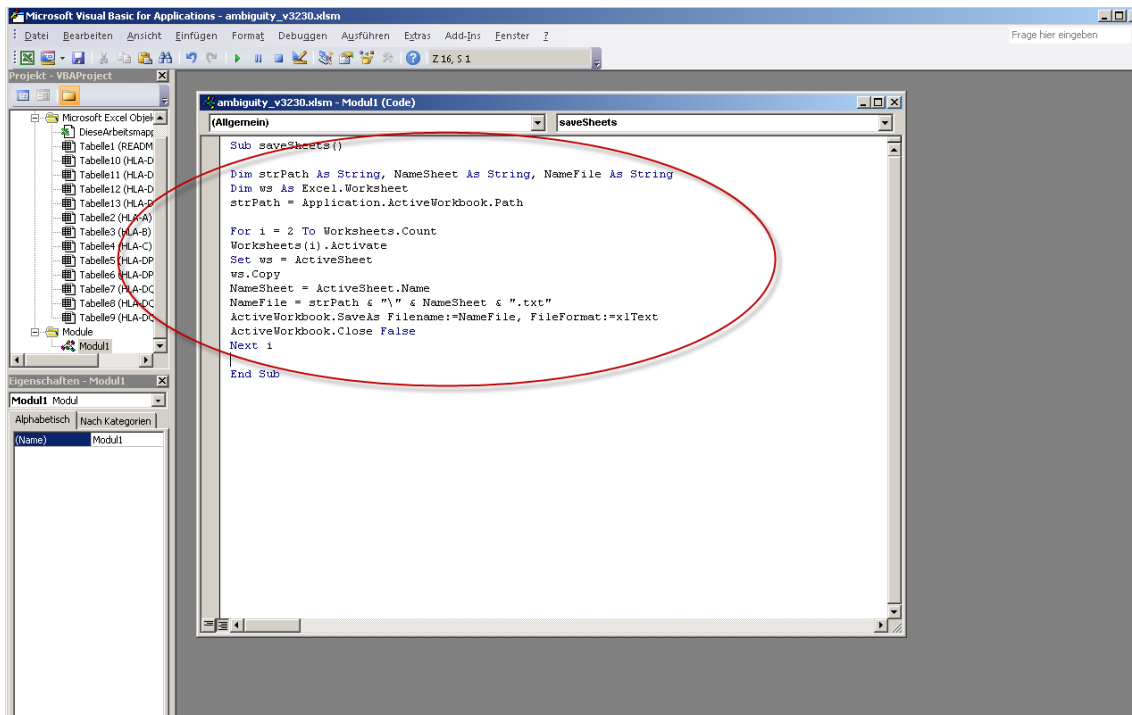
```

Dim strPath As String, NameSheet As String, NameFile As String
Dim ws As Excel.Worksheet
strPath = Application.ActiveWorkbook.Path

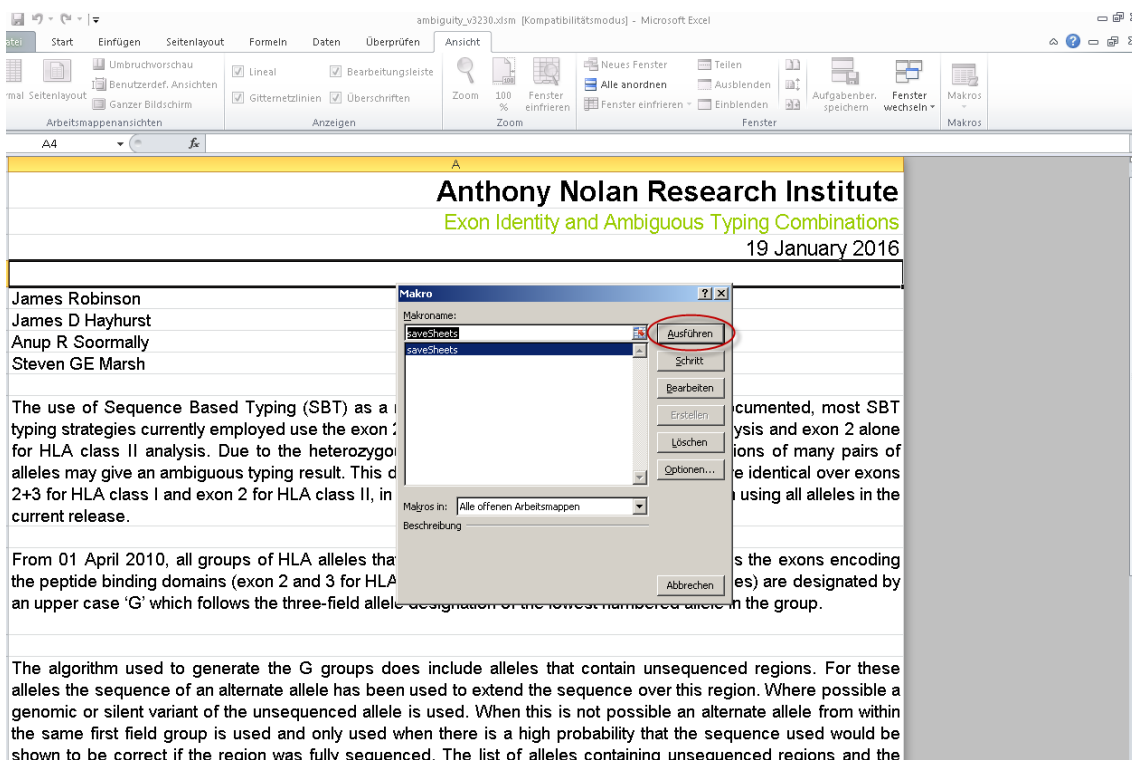
For i = 2 To Worksheets.Count
Worksheets(i).Activate
Set ws = ActiveSheet
ws.Copy
NameSheet = ActiveSheet.name
NameFile = strPath & "\" & NameSheet & ".txt"
ActiveWorkbook.SaveAs Filename:=NameFile, FileFormat:=xlText
ActiveWorkbook.Close False
Next i
End Sub

```





Close the window and execute the macro:



A bunch of text files as “HLA-A.txt” and so on should have appeared. Move these to the folder “Hapl-o-Mat/prepareData”. Afterwards you can remove the Excel file.

## Build Data for Hapl-o-Mat

Now you are ready to build the data. Enter the folder “Hapl-o-Mat/prepareData” and run the following python scripts:

- 1) `python BuildAllAllelesFrom_hla_nom_g.py`
- 2) `python BuildAllAllelesExpanded.py`
- 3) `python BuildP.py`
- 4) `python BuildLargeG.py`
- 5) `python BuildSmallg.py`
- 6) `python AddAllelesMissingIngCode.py`
- 7) `python TransferAlphaToMultipleAlleleCodes`
- 8) If you went in the last section with the xml-file run `python BuildAmbiguityFromXML.py` , if you went with the Excel sheet run `python BuildAmbiguityFromTextFiles.py`.
- 9) `python AddGToAmbiguity.py`

Next, create the folder “Hapl-o-Mat/data” and move the freshly created files LargeG.txt, P.txt, Smallg.txt, Ambiguity.txt, MultipleAlleleCodes.txt, and AllAllelesExpanded.txt there. You can remove the files `alpha.v3.txt`, `hla_ambigs.xml`, `hla_nom_g.txt`, `hla_nom_p.txt`, `allAlleles.txt`, and `OneElementG.txt`.

If you want to analyse data in GL-format with unresolved genotypes (GL-id=0), you can prepare the file `AlleleList.txt` from the GL-id input file by running `BuildAlleleList.py`. Then move `AlleleList.txt` to data.