



# HAPLOSEARCH SOFTWARE

## User's Manual

<http://www.haplosite.com/haplosearch/>



# INDEX

1. INTRODUCTION .....	3
2. DATA FORMAT .....	4
2.1. SEQUENCES .....	4
2.2. HAPLOTYPES .....	5
a) POINT MUTATIONS .....	5
b) HETEROPLASMY .....	6
c) INDELS .....	7
3. INPUT DATA FILES .....	8
3.1. TRANSFORMING SEQUENCES TO HAPLOTYPES .....	9
3.2. TRANSFORMING HAPLOTYPES INTO SEQUENCES .....	10
4. OUTPUT DATA FILES .....	11
5. IMPORTANT INFORMATION ABOUT DATA FORMAT .....	12
5.1. ALIGNMENT .....	12
5.2. PARTIAL SEQUENCES (POPULATIONS GENETICS) .....	13
5.3. NOMENCLATURE OF DELETIONS (FORENSIC GENETICS) .....	14
6. HAPLOSEARCH INTERFACE .....	15
7. REFERENCES .....	20



## 1. INTRODUCTION

Human mitochondrial DNA (mtDNA) has several characteristics that makes it an invaluable tool for population genetic studies, as high copy number, small size ( $\approx 16,500$  bp) and higher mutation rate than nuclear DNA. Furthermore, mtDNA is maternally inherited without recombination, allowing for the reconstruction of the evolutionary history of populations (Ballard and Whitlock, 2004).

In 1981, the complete sequence of the human mtDNA was published for the first time (Anderson *et al*, 1981). Since that, populations from almost everywhere have been studied from the mtDNA point of view. The comparison of these large sets of mtDNA data have allowed to construct a robust phylogenetic tree (Torrioni *et al*, 2006; van Oven and Kayser, 2009) and to estimate the global distribution and origin of each human mtDNA lineage (Cann *et al*, 1987; Ingman *et al*, 2000; Maca-Meyer *et al*, 2001; Richards *et al*, 2000).

MtDNA analysis has also become an useful tool in forensic genetics, as its mode of inheritance allows testing for a putative exclusion scenario in human identification. On the other hand, when only very limited or severely degraded DNA is present in a sample, mtDNA constitutes the last chance for successful DNA typing (Parson and Bandelt, 2007).

However, published data comparison is frequently complicate as mtDNA results could appear in two different formats: haplotype (detected mutations respect to a reference sequence) and nucleotide sequence data. Manual transformation between formats is time-consuming, complex and likely to introduce mistakes. Moreover, some data analyses, like haplogroup classification or matches between populations, require haplotype data. On the contrary, others, such as genetic diversity calculations, are designed for nucleotide sequences. In all cases, although some data analysis software allow you to use both formats, like Arlequin (Excoffier and Lischer, 2010), the transformation between them is usually needed as published mtDNA results could alternatively appear in both types.

HaploSeach software transforms haplotype and sequence data between them in a quick and easy way, allowing a fast and reliable data comparison. This program admits both partial and complete mtDNA sequences, and recognises substitution mutations (transitions and transversions), heteroplasmies and indels (insertions and deletions).

Although HaploSearch was designed to analyse mtDNA sequences, it is suitable for transforming haplotypes and sequences in any kind of DNA sources. The program only requires a reference sequence from which extract the information, as occurs with the revised Cambridge Reference sequence (CRS) for mtDNA (Andrews *et al*, 1999).

## **2. DATA FORMAT**

### **2.1. Sequences:**

Sequences must be introduced into the commonly used fasta format, following the IUPAC code (Cornish-Bowden, 1985). Using this format in HaploSearch allows you to obtain the complete and partial mtDNA data directly from the main databases (as GeneBank, mtDB...). To be correctly analysed, all sequences have to be equal in length, so they have to be previously aligned with the reference sequence, following the required guidelines (Bandelt and Parson, 2008). Each sequence must be written continuously, without spaces or new paragraphs. For this reason, after performing the alignment, it is encouraged to review the sequences as some aligning programs create new paragraphs into the sequence. Removing spaces or new paragraphs could be easily performed by using the “Replace” tool, which is available for the majority of text processors.

As an example:

```
>CRS
AAAACCCCCTCCCC-ATGCC
>SEC1
AAAACCCCCCCCCCATGCC
>SEC2
AAACCCCCTCCCC-ATGCC
```

### **2.2. Haplotypes:**

Mutations of haplotypes must be arranged from smaller to higher position and separated with spaces. If there are sequences that are exactly the same as the CRS reference, their haplotype would be CRS. This designation for non mutated sequences could be changed when other DNA types are studied.

Using HaploSearch software, mutations could be written using two formats: "Population Genetics Nomenclature" and "Forensic Genetics Nomenclature" (following the DNA Commission of the International Society for Forensic Genetics recommendations as detailed in Carracedo et al. (2000)).

### a) Point mutations

Point mutations are caused when exchanging a single nucleotide for another (Freese, 1959a), in respect to the CRS (or other reference sequence). These changes are classified as transitions or transversions (Freese, 1959b).

- a. Transition: is a mutation changing a purine to another purine nucleotide (A↔G) or a pyrimidine to another pyrimidine nucleotide (C↔T). This is the most common mutation and, for the "Population Genetic Nomenclature", it is only designated by the nucleotide position:

```
0000000001111
1234567890123

CRS   CGACCCCTGTATC
SEC1  CGACCCTTGTGTTC
```

In this example, haplotype would be "SEC1: 7 11", showing that SEC 1 has two transitions, in position 7 and 11, respectively.

However, for the "Forensic Genetic Nomenclature", the haplotype should be designated by the nucleotide position and the mutated base. In this case, it would be "SEC1: 7T 11G".

- b. Transversion refers to the substitution of a purine for a pyrimidine or vice versa.

For both haplotype formats, they are designated by the nucleotide position and the changed base:

```

0000000001111
1234567890123

CRS   CGACCCCTGTATC
SEC1  CGCCCCCTTTATC

```

Thus, haplotype would be “SEC1: 3C 9T”, showing that SEC 1 has one transversion to cytosine in position 3 and one transversion to thymine in position 9.

### b) Heteroplasmy

The presence of more than one mtDNA haplotype in a sample is referred to as heteroplasmy. This phenomenon could be due to differential segregation of pre-existing heteroplasmic variants, to accumulation of new somatic mutations or to a combination of both.

In this situations, it is necessary the use of a single symbol to designate a variety of possible nucleotides at a single position (Table 1).

TABLE 1 - The IUPAC nucleotide code (Cornish-Bowden, 1985)

IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base



For both haplotype formats, they are designated by the nucleotide position and the corresponding IUPAC nucleotide code:

```
0000000001111
1234567890123

CRS   CGACCCCTGTATC
SEC1  CGACCCTGTKTTC
```

Thus, haplotype would be “SEC1: 11K”, showing that SEC1 has heteroplasmy in position 11, where nucleotides G and T are present.

### c) Indels

The term indel includes insertions and deletions, as these two types of genetic mutation are often considered together because of the inability to distinguish between them when comparing two sequences. This problem does not exist when sequences are compared with a reference: insertions add one or more extra nucleotides into the DNA, in respect to the reference; and deletions remove one or more nucleotides from the DNA compared with the reference sequence. Due to indels, the sequences have to be aligned before using HaploSearch, in order to designate a correct haplotype. To perform the alignment it is recommended to use alignment programs as ClustalW (<http://www.ebi.ac.uk/Tools/clustalw2/index.html>).

- a. Insertions: as insertions add one or more extra nucleotides, it is necessary to introduce gaps into the reference sequence to maintain the alignment.

```
0000000--001-111
1234567--890-123

CRS   CGACCCC--TGT-ATC
SEC1  CGACCCCCCTGTCATC
```

To name the insertions in the “Population Genetic Nomenclature”, you must indicate the base position in which the insertion has occurred and the bases that are inserted, preceded by letter “i”. In the above example, haplotype would be “SEC1: 7iCC 10iC”.

For the “Forensic Genetic Nomenclature”, insertions are independently named by first noting the site immediately to the insertion followed by a decimal point and a ‘1’ (for the first insertion), a ‘2’ (if there is a second base inserted), and so on, and then by the nucleotide that is inserted. In the above example, haplotype would be “SEC1: 7.1C 7.2C 10.1C” (Carracedo *et al*, 2000).

- b. Deletions: as deletions remove one or more nucleotides from the DNA, it is necessary to introduce gaps into the studied sequence to maintain the alignment.

```
0000000001111
1234567890123

CRS   CGACCCCTGTATC
SEC1  CGACC--TGT-TC
```

To designate deletions in the “Population Genetic Nomenclature”, you have to write the first base position of the gap and the bases that are deleted, preceded by letter “d”. In the above example, haplotype would be “SEC1: 6dCC 11dA”.

For the “Forensic Genetic Nomenclature”, deletions should be recorded by listing the missing sites followed by a ‘del’. In the example, it would be “SEC1: 6del 7del 11del”.

### **3. INPUT DATA FILES**

Input data files could be written on any text processor, as long as the file is saved as a txt file. However, if a text processor with autocorrection tools (such as Microsoft Word, OpenOffice Writer or Vim) is used, this function has to be disabled in order to avoid modifications that could affect the HaploSearch operation. Indels are prone to be affected by autocorrection tools, as consecutive hyphens are exchanged for only one. This could cause the lost of alignment and, sometimes, the use of characters that are not recognised by HaploSearch. Therefore, it is encouraged to disable the autocorrection tool or to use unformatted text processors.

### 3.1. Transforming sequences into haplotypes:

The input file for transforming sequences into haplotypes has to be a txt file containing the aligned sequences in fasta format. Moreover, we have to indicate what is the reference sequence and the position number for the first nucleotide in the sequence as follows:

- The first line must indicate the nucleotide position for the first base of the reference sequence with the following format: "START:##". This position would be 1 for complete sequences or to begin with the corresponding number for partial sequences. In this line, it is also possible to choose some haplotype nomenclature features (See 5.2 and 5.3 chapters for more information).
- The second line must contain the reference sequence in fasta format. The reference sequence must be named ">reference\_name" and would be the ">CRS" for mtDNA or any consensus sequence for other DNA types.
- In the following lines, the studied sequences have to be introduced in fasta format.

Example:

```
START: 16180
>CRS
AAAACCCCCTCCCCATGCC
>SEC1
AAAACCCCCCCCCCATGCC
>SEC2
AAACCCCCTCCCCATGCC
```

When sequences include indels, they have to be aligned for a correct HaploSearch analysis:

```
START: 16180
>CRS
AAAACCCCCTCCCC-ATGCC
>SEC1
AAAACCCCCCCCC--ATGCC
>SEC2
AAACCCCCTCCCCATGCC
```

### 3.2. Transforming haplotypes into sequences:

The input file for transforming haplotypes into sequences is similar to the previous file, but using haplotype data, with whatever Population Genetic or Forensic Genetic nomenclature. As in the previous file, it has to be indicated what is the reference sequence and the position number for the first nucleotide in the reference sequence:

- The first line must indicate the nucleotide position for the first base of the reference sequence with the following format: "START:##". This position would be 1 for complete sequences or to begin with the corresponding number for partial sequences.
- The second line must contain the reference sequence in fasta format. The reference sequence must be named ">reference\_name" and would be ">CRS" for mtDNA or any consense sequence for other DNA types.
- In the following lines, the haplotypes should be written in a similar way to the fasta format. When sequences do not include mutation, their haplotype would be the reference name. For example, when a mtDNA sequence is identical to the CRS, its haplotype would be "CRS".

Example:

```
START: 16180
>CRS
AAAACCCCCTCCCCATGCC
>SEC1
16189
>SEC2
16183C 16189 16193dC
>SEC3
CRS
```

**IMPORTANT INFORMATION: Reference sequence in haplotype input files would be written without hyphens as HaploSearch automatically adds gaps when indels are present.**

#### 4. OUTPUT DATA FILES

HaploSearch output data files have the same format as the input data file for the opposite transformation. This feature allows you to obtain the original data from the output file, checking if any mistakes were introduced during data manipulation and/or HaploSearch have worked properly.

If the input file is as follows:

```
START: 16180
>CRS
AAAACCCCCTCCCC--ATGCTTACAAGCAAGTACAGCAATCAACCCTCAA
>SEC1
AAACCCCTCCCCCCCCATGCTTACAAGCAAGTACAGCAATCAACCTTCAA
>SEC2
AAAACCCCCTCCCC--ATGCTTACAAGCAAGTACAGCAATCAACCCCAA
>SEC3
AAAACCCCCTCCCC----ATGCTTACAAGCAAGTACAGCAATCAACCCTCAA
```

The output file for "Population Genetic Nomenclature" will be:

```
START: 16180
>CRS
AAAACCCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCCTCAA
>SEC1
16183C 16187 16189 16193iCC 16223
>SEC2
16224
>SEC3
16189 16192dCC
```

Or this one, for "Forensic Genetic Nomenclature":

```
START: 16180
>CRS
AAAACCCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCCTCAA
>SEC1
16183C 16187T 16189C 16193.1C 16193.2C 16223T
>SEC2
16224C
>SEC3
16189C 16192del 16193del
```

Now, if these output files are used as input file, we could obtain the original data source.

## 5. IMPORTANT INFORMATION ABOUT DATA FORMAT:

### 5.1. Alignment:

HaploSearch recognises the indels that are determined by the aligned input sequences. When sequences containing indels are aligned by alignment programs, the gaps are not always placed in the same position as in the commonly used nomenclature.

For instance, SEC1 has four inserted Cs between 301 to 320 mtDNA positions:

```
START: 301
>rCRS
AACCCCCCTCCCCCGC
>SEC1
AACCCCCCCCCCTCCCCCGC
```

As there are several Cs in this position, the alignment could be shown in several ways, all of them being corrected. However, every one would originate different haplotypes:

rCRS: AA--CCCCCCT-CCCCGC	}	302iCC 310iC / 302 .1C 302.2C 310.1C
SEC1: AACCCCCCCCCCTCCCCCGC		
rCRS: AACCCCCC--TCCCC-GC	}	309iCC 315iC / 309.1C 309.2C 315.1C
SEC1: AACCCCCCCCCCTCCCCCGC		
rCRS: AAC--CCCCCCTC-CCCCGC	}	303iCC 311iC / 303.1C 303.2C 311.1C
SEC1: AACCCCCCCCCCTCCCCCGC		

We do not know what mutational event caused these insertions, so all the different alignments are possible. However, certain indels are commonly named in a determined way. In the above example, the correct nomenclature would be “309iCC 315iC” or “309.1C 309.2C 315.1C”. This problem could be overcome by checking the alignment previous to the HaploSearch analysis (for instance, using a sequence editor as BioEdit:

<http://www.mbio.ncsu.edu/BioEdit/BioEdit.html>) and placing the variable indels in the most used place. A later modification of the output file is also possible. For alignment guidelines see Bandelt and Parson (2008).

## 5.2. Partial sequences (Population Genetics nomenclature)

Sometimes, in population genetic studies, when only the hypervariable region I (HVRI) is analysed (positions between 16024 - 16365), the 16### notation could be omitted for clarity reasons. For example, haplotype "SEQ1: 16069 16126" would be "SEQ1: 069 126".

If you want to use this notation in HaploSearch, you should place an asterisk in the START number. For example:

If you use the current notation:

```
START: 16090
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
TATCTCGTACATTACTGCCAGACACCATGA
```

The output would be:

```
START: 16090
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
16093 16111A
```

In other hand, if you eliminate the 16### and add an asterisk to the start number:

```
START: 90*
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
TATCTCGTACATTACTGCCAGACACCATGA
```

The output would be:

```
START: 90
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
093 111A
```

This kind of notation is only possible for Populations Genetic Nomenclature.

### 5.3. Nomenclature of deletions (Forensic Genetics nomenclature)

As recommended by the EMPOP database, deletions are named as “del” in HaploSearch (see “Indels” section in Chapter 2). However, Carracedo et al. (2000) recommends the use of “d” instead of “del”. For this reason, there is the possibility of using “d” in HaploSearch. In this case, it is extremely important to use “D” for the heteroplasmy consisting of a mixture of A, G, and T (following IUPAC code) and “d” for deletions (See the example).

For implementing this feature in HaploSearch, you only have to place a “d” next to the start number. For example:

If you use the current notation:

```
START: 16090
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
TAT-TCGTACATTACTGCCAGACACCCDTGA
```

The output would be:

```
START: 16090
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
16093del 16111A 16116D
```



In other hand, if you add the letter "d" to the start number:

```
START: 16090d
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
TAT-TCGTACATTACTGCCAGACACCCDTGA
```

The output would be:

```
START: 16090
>CRS
TATTTTCGTACATTACTGCCAGCCACCATGA
>SEQ1
16093d 16111A 16116D
```

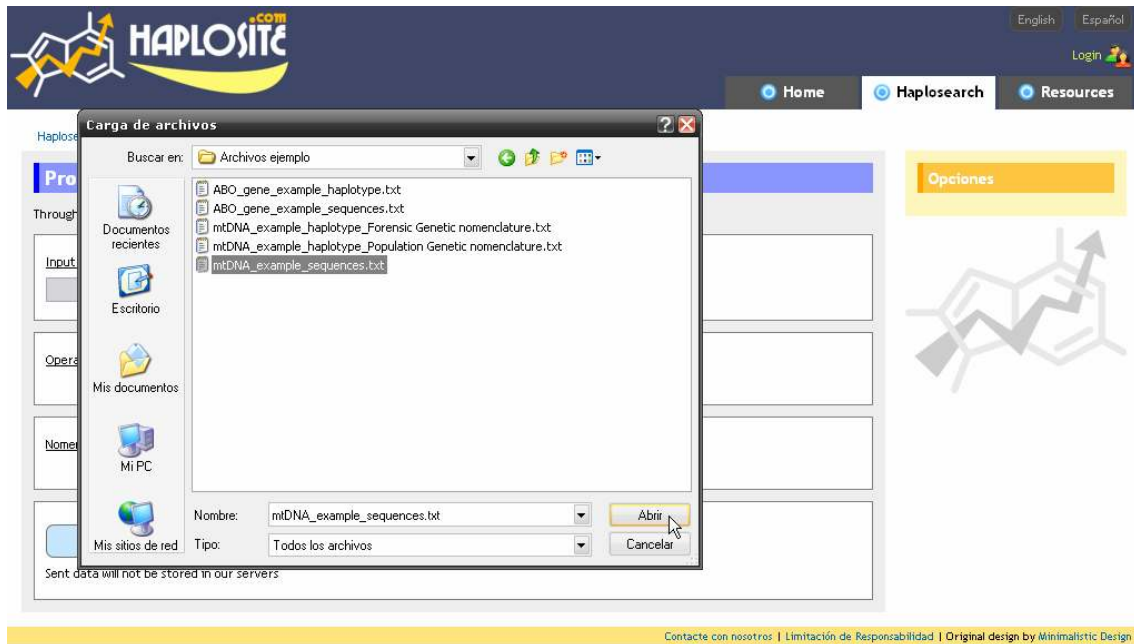
This feature is only possible for Forensic Genetic Nomenclature.

## 6. HAPLOSEARCH INTERFACE

Using HaploSearch interface (<http://www.haplosite.com/haplosearch>) is simple and intuitive. To run HaploSearch, you have to select "Process", in "Topics" box.

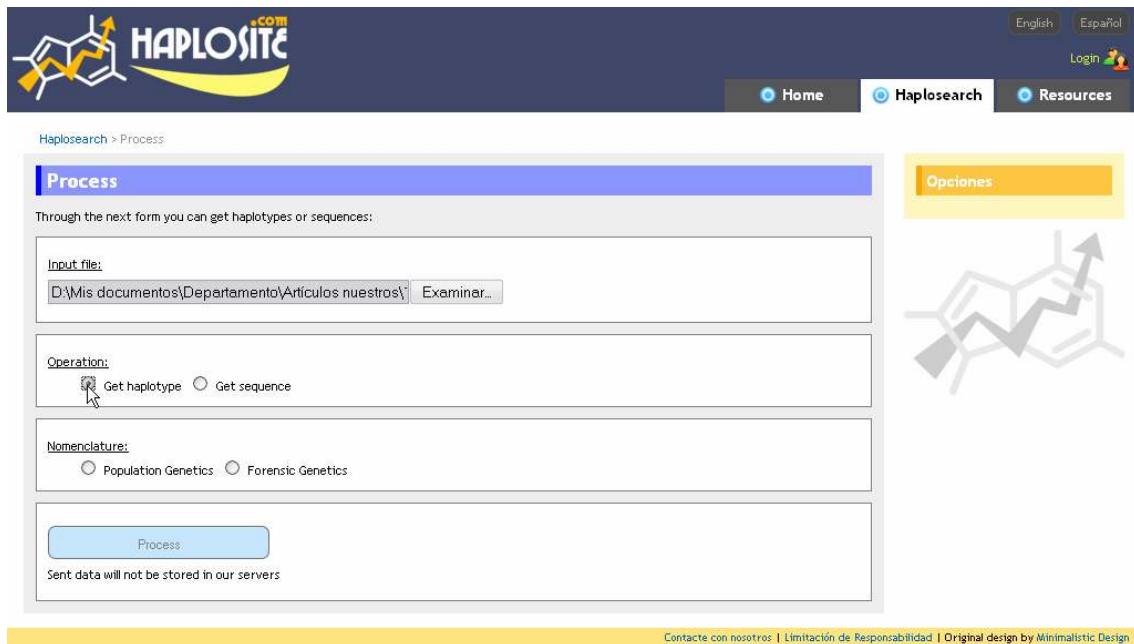
The screenshot displays the HaploSearch web interface. At the top, there is a dark blue header with the HaploSite logo on the left, language selection buttons for 'English' and 'Español' on the right, and a 'Login' button. Below the header is a navigation bar with three buttons: 'Home', 'Haplosearch', and 'Resources'. The main content area is divided into two columns. The left column has a blue 'Home' header and a grey box containing the text: 'This application lets you get haplotypes from sequences and the other way round.' The right column has a yellow 'Opciones' header and a box with 'Process' and 'Help' links. A large, stylized DNA sequence graphic is positioned in the lower right area of the page. At the bottom, a yellow footer bar contains the text: 'Contacte con nosotros | Limitación de Responsabilidad | Original design by Minimalistic Design'.

Then, you have to click "Browse" button and select the input data file:



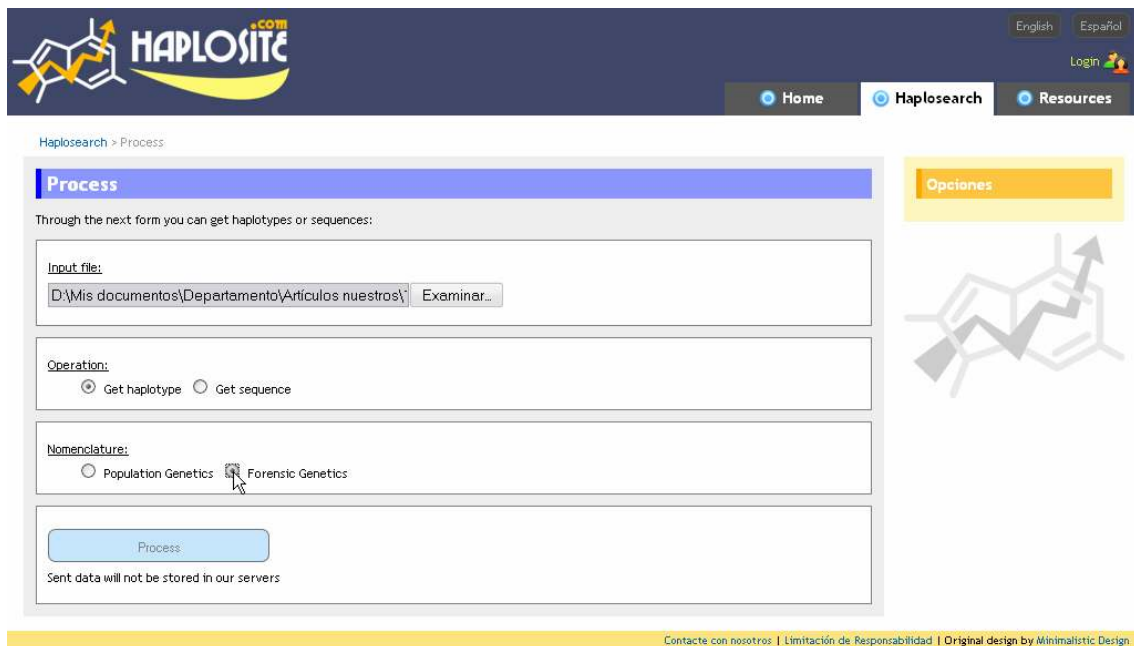
The screenshot shows the HaploSite website interface. At the top, there is a navigation bar with the HaploSite logo, language options (English, Español), a Login button, and menu items for Home, Haplosearch, and Resources. A file selection dialog box titled "Carga de archivos" is open, showing a list of files in the "Archivos ejemplo" folder. The files listed are: ABO\_gene\_example\_haplotype.txt, ABO\_gene\_example\_sequences.txt, mtDNA\_example\_haplotype\_Forensic Genetic nomenclature.txt, mtDNA\_example\_haplotype\_Population Genetic nomenclature.txt, and mtDNA\_example\_sequences.txt. The "mtDNA\_example\_sequences.txt" file is selected. The dialog box also shows the file name "mtDNA\_example\_sequences.txt" and the file type "Todos los archivos". There are "Abrir" and "Cancelar" buttons at the bottom right of the dialog box. A footer at the bottom of the page contains the text "Contacte con nosotros | Limitación de Responsabilidad | Original design by Minimalistic Design".

Once the input data file is selected, you have to select the type of transformation:



The screenshot shows the HaploSite website interface, specifically the "Process" form. The navigation bar at the top is the same as in the previous screenshot. The "Process" form is titled "Process" and contains the following fields: "Input file:" with a text box containing "D:\Mis documentos\Departamento\Articulos nuestros\" and an "Examinar..." button; "Operation:" with two radio buttons, "Get haplotype" (selected) and "Get sequence"; "Nomenclature:" with two radio buttons, "Population Genetics" (selected) and "Forensic Genetics"; and a "Process" button. Below the form, there is a note: "Sent data will not be stored in our servers". A footer at the bottom of the page contains the text "Contacte con nosotros | Limitación de Responsabilidad | Original design by Minimalistic Design".

Then, you have to select the type of nomenclature:

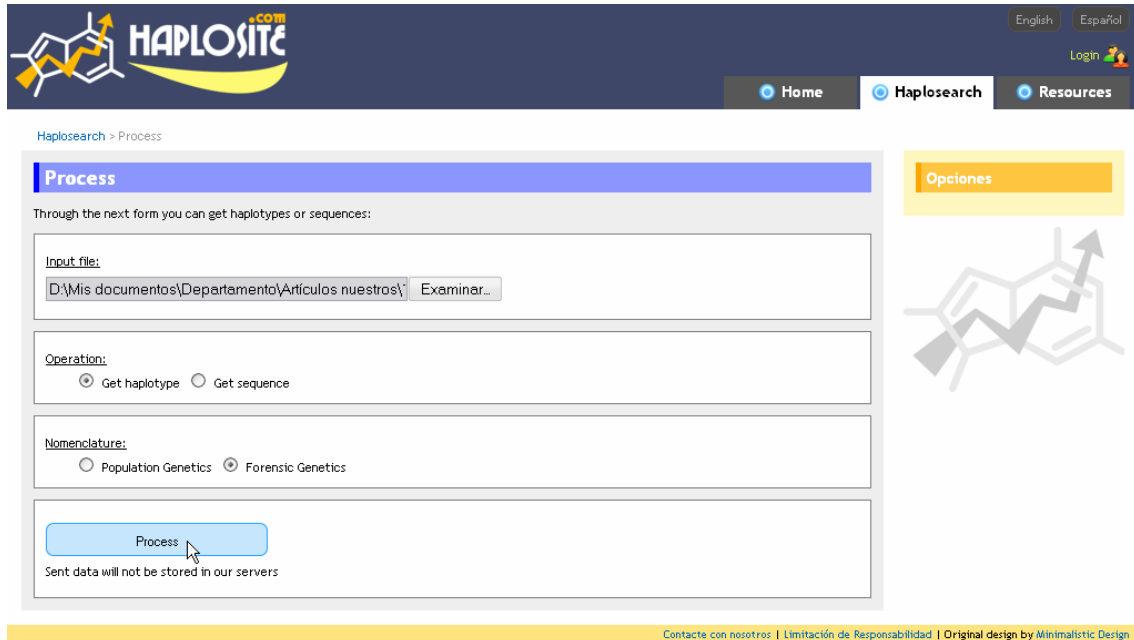


The screenshot shows the HaploSite website interface. At the top, there is a dark blue header with the HaploSite logo on the left, language options for English and Español on the right, and a Login button. Below the header is a navigation bar with Home, Haplosearch, and Resources buttons. The main content area is titled 'Process' and contains a form with the following sections:

- Input file:** A text box containing 'D:\Mis documentos\Departamento\Articulos nuestros\' and an 'Examinar...' button.
- Operation:** Two radio buttons: 'Get haplotype' (selected) and 'Get sequence'.
- Nomenclature:** Two radio buttons: 'Population Genetics' and 'Forensic Genetics' (selected).
- Process:** A blue button labeled 'Process'.

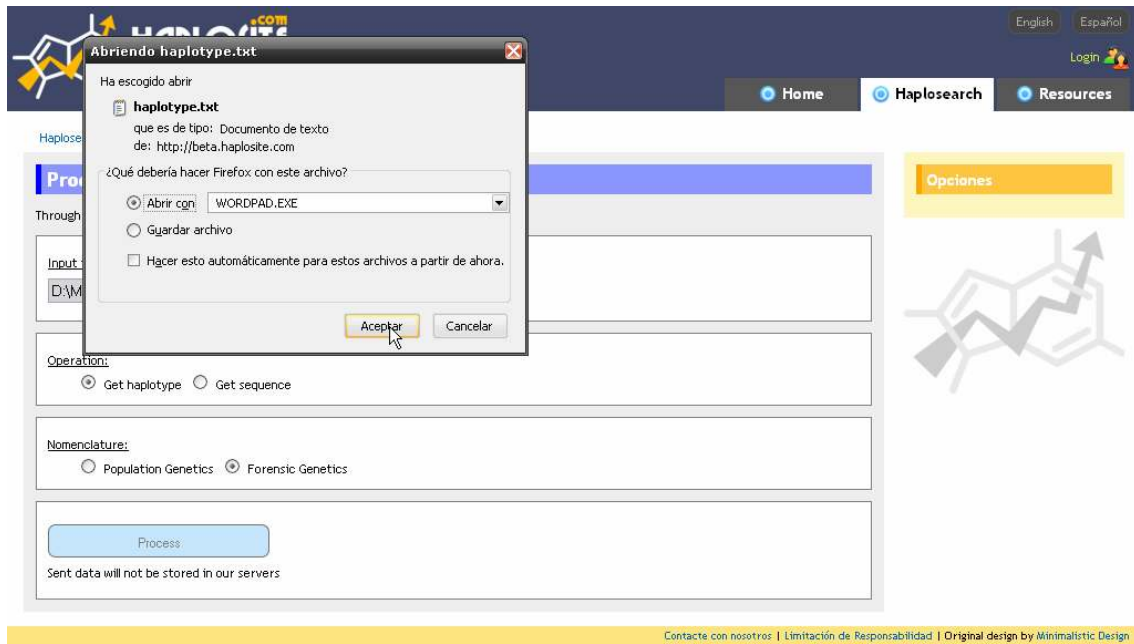
Below the form, a note states: 'Sent data will not be stored in our servers'. To the right of the form is a yellow box labeled 'Opciones' and a large grey graphic of a DNA double helix with an upward-pointing arrow. At the bottom of the page, a yellow footer contains the text: 'Contacte con nosotros | Limitación de Responsabilidad | Original design by Minimalistic Design'.

And finally, press "Process":

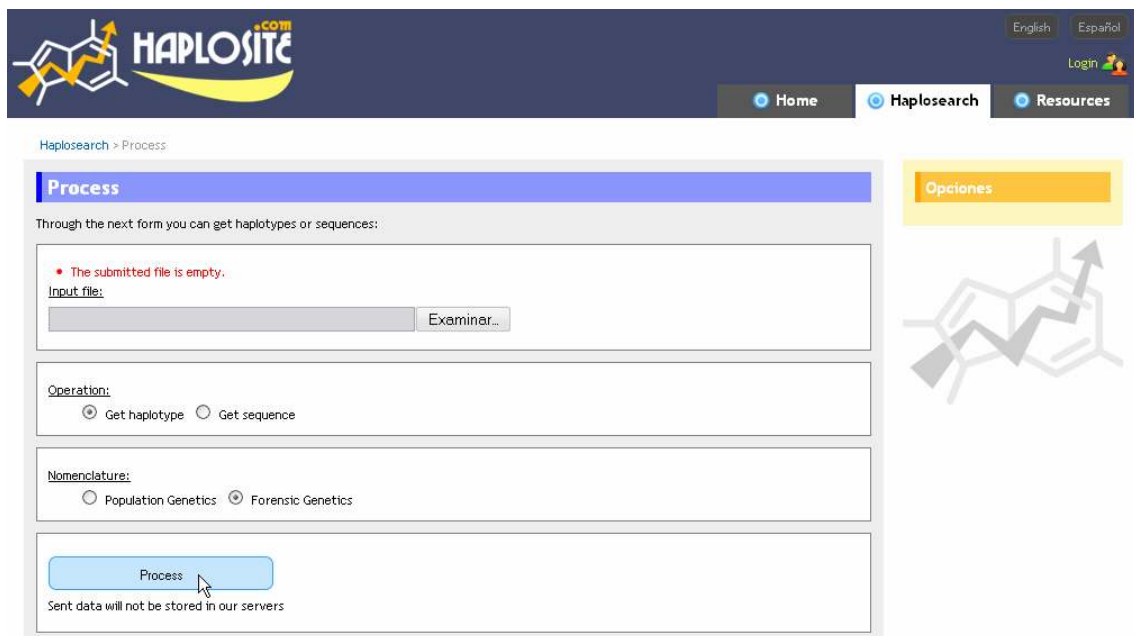


This screenshot is identical to the previous one, showing the HaploSite 'Process' form. The key difference is that a mouse cursor is now clicking on the 'Process' button. The rest of the interface, including the header, navigation, form fields, and footer, remains the same.

If the input data file is correct, the output data file will be ready in a very short time.



When the input data file has some format mistakes, an error message will appear, indicating the origin of the error. If an unexpected error occurs, please contact us. See below some examples.



Haplosearch > Process

### Process

Through the next form you can get haplotypes or sequences:

**This field is required.**  
Input file:  Examinar...

Operation:  
 Get haplotype  Get sequence

Nomenclature:  
 Population Genetics  Forensic Genetics

Process  
Sent data will not be stored in our servers

Opciones

**Syntax error on input file (line: 5).**

[Back](#)

**Base position is not defined on input file.**

[Back](#)

## **7. REFERENCES:**

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J *et al* (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**(5806): 457-465.

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**(2): 147.

Ballard JW, Whitlock MC (2004). The incomplete natural history of mitochondria. *Mol Ecol* **13**(4): 729-744.

Bandelt HJ, Parson W (2008). Consistent treatment of length variants in the human mtDNA control region: a reappraisal. *Int J Legal Med* **122**(1): 11-21.

Cann RL, Stoneking M, Wilson AC (1987). Mitochondrial DNA and human evolution. *Nature* **325**(6099): 31-36.

Carracedo A, Bar W, Lincoln P, Mayr W, Morling N, Olaisen B *et al* (2000). DNA commission of the international society for forensic genetics: guidelines for mitochondrial DNA typing. *Forensic Sci Int* **110**(2): 79-85.

Cornish-Bowden A (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**(9): 3021-3030.

Excoffier L, Lischer HEL (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* **in press**.

Freese E (1959a). The Difference between Spontaneous and Base-Analogue Induced Mutations of Phage T4. *Proc of NAS* **45**(4): 622-633.

Freese E (1959b). The Specific Mutagenic Effect of Base Analogues on Phage T4. *J Mol Biol* **1**(87-105).

Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**(6813): 708-713.

Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C, Cabrera VM (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* **2**: 13.

Parson W, Bandelt HJ (2007). Extended guidelines for mtDNA typing of population data in forensic science. *Forensic Sci Int Genet* **1**(1): 13-19.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V *et al* (2000). Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**(5): 1251-1276.

Torroni A, Achilli A, Macaulay V, Richards M, Bandelt HJ (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet* **22**(6): 339-345.

van Oven M, Kayser M (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* **30**(2): E386-394.