

# **Minimum SNPs**

**Version 2043**

## **User Manual**

For assistance or additional information, contact Phil Giffard at [p.giffard@qut.edu.au](mailto:p.giffard@qut.edu.au)

Largely written by John Bamber, November 2006.  
Last updated by Phil Giffard, September 2007

# Contents

1. <u>Definitions.</u> Definitions of terminology, background information, references.	P3
2. <u>Getting Started.</u> System and software requirements, how to start the program	P6
3. <u>Creating a mega-alignment.</u> How to down load and concatenate MLST data.	P7
4. <u>Obtaining and loading a pre-concatenated MLST database.</u>	P10
5. <u>Deriving highly informative sets of SNPs.</u>	P11
6. <u>Working backwards method.</u> How to go from SNP profile to sequences.	P16
7. <u>Tools and options.</u> These functions facilitate efficient and effect searches.	P29

# 1 Definitions:

## 1.1 Minimum SNPs

**Minimum SNPs is proprietary software for deriving discriminatory sets of SNPs from DNA sequence alignments, as described in:**

Price E.P., Inman-Bamber, J., Thiruvenkataswamy, V., Huygens, F and Giffard, P.M. 2007. Computer-aided identification of polymorphism sets diagnostic for groups of bacterial and viral genetic variants. BMC Bioinformatics 8:278.

**Earlier versions of the software are described in:**

Robertson, G.A., Thiruvenkatswamy, V, Shilling, H., Price E.P., Huygens, F., Henkens, F.A. and Giffard, P.M. 2004. Identification and interrogation of highly informative single nucleotide polymorphism sets defined by bacterial multilocus sequence typing databases. J. Med. Microbiol. **53**:35-45.

Price, E.P., Thiruvenkatswamy, V., Micken, L., Unicomb, L., Rios, R.E., Huygens, F., and Giffard, P.M. 2006. Genotyping of *Campylobacter jejuni* using seven single-nucleotide polymorphisms in combination with *flaA* short variable region sequencing. J. med. Microbiol. **55**:1061-1070.

**Additional applications of the software are described in:**

Stephens, A.J., Huygens, F., Inman-Bamber, J., Price, E.P., Nimmo, G.R., Schooneveldt, J, Minckhof, W and Giffard, P.M. 2006. Methicillin-resistant *Staphylococcus aureus* genotyping using a small set of polymorphisms. J. Med. Microbiol. **55**:43-51.

Price, E.P., Huygens, F and Giffard, P.M. 2006. Fingerprinting of *Campylobacter jejuni* using resolution-optimized binary gene targets derived from comparative genome hybridization studies. Appl. Env. Microbiol. **72**: 7793-7803.

Huygens, F., Inman-Bamber, J., Nimmo, G.R., Minckhof, W., Schooneveldt, J., Harrison, B., McMahon, J.A and Giffard, P.M. 2006. *Staphylococcus aureus* genotyping using novel real-time PCR formats. J. Clin. Microbiol. **44**:3712-3719

Stephens A.J., Huygens F., and Giffard, P.M. 2007. Systematic derivation of marker sets for Staphylococcal Cassette Chromosome *mec* typing. Antimicrob. Agents Chemother. **51**:2954-2964.

The software was developed by researchers at the Queensland University of Technology (Brisbane, Queensland, Australia) node of the Cooperative Research Centre for Diagnostics, with input from researchers at the University of Newcastle, Newcastle, New South Wales, Australia.

Minimum SNPs can only be obtained from [www.ihbi.qut.edu.au/research/cells\\_tissue/phil\\_giffard/](http://www.ihbi.qut.edu.au/research/cells_tissue/phil_giffard/)

## 1.2 SNP

A single nucleotide polymorphism is a variation in the genetic code at a specific point on the DNA. In principle, SNPs could be bi-, tri-, or tetra-allelic polymorphisms. However, tri-allelic and tetra-allelic SNPs are rare and so SNPs are sometimes simply referred to as bi-allelic markers. However, we have identified two tri-allelic SNP that have proved very useful in genotyping *Staphylococcus aureus*.

## 1.3 MLST

Multi Locus Sequence Typing:

*“The original MLST web software was developed by Man-Suen Chan (Oxford University) and this version has been developed by David Aanensen (Imperial College) who is funded by The Wellcome Trust*

*Multilocus sequence typing (MLST) is an unambiguous procedure for characterising isolates of bacterial species using the sequences of internal fragments of seven house-keeping genes. Approx. 450-500 bp internal fragments of each gene are used, as these can be accurately sequenced on both strands using an automated DNA sequencer. For each house-keeping gene, the different sequences present within a bacterial species are assigned as distinct **alleles** and, for each isolate, the alleles at each of the seven loci define the **allelic profile** or **sequence type (ST)**.*

*Each isolate of a species is therefore unambiguously characterised by a series of seven integers which correspond to the alleles at the seven house-keeping loci.*

*In MLST the number of nucleotide differences between alleles is ignored and sequences are given different allele numbers whether they differ at a single nucleotide site or at many sites. The rationale is that a single genetic event resulting in a new allele can occur by a point mutation (altering only a single nucleotide site), or by a recombinational replacement (that will often change multiple sites) - weighting according to the number of nucleotide differences between alleles would imply that the latter allele was more distantly-related to the original allele than the former, which would be true if all nucleotide changes occurred by mutation, but not if the changes occurred by a recombinational replacement.*

*Most bacterial species have sufficient variation within house-keeping genes to provide many alleles per locus, allowing billions of distinct allelic profiles to be distinguished using seven house-keeping loci. For example, an average of 30 alleles per locus allows about 20 billion genotypes to be resolved.*

*MLST is based on the well established principles of multilocus enzyme electrophoresis, but differs in that it assigns alleles at multiple house-keeping loci directly by DNA sequencing, rather than indirectly via the electrophoretic mobility of their gene products.”* (Extracted from <http://www.mlst.net/misc/further.asp> on November 1, 2006)

## 1.4 Locus/loci:

The loci are internal fragments of seven house-keeping genes (450-500bp) common to all isolates within a species. More broadly speaking, a locus is position on a chromosome of a gene or other chromosome marker; also, the DNA at that position.

## **1.5 Allele**

For each housekeeping gene, alleles are the different sequences present within a bacterial species. More broadly speaking, an allele is any one of a number of alternative forms of the same gene occupying a given locus (position) on a chromosome.

## **1.6 Allelic profile:**

The alleles at each of the seven loci (housekeeping genes) in a particular isolate. More broadly speaking, the allelic profile is the combination of alleles that any one individual possesses.

## **1.7 ST:**

Sequence Type:

Sequence type is a number given to represent a unique allelic profile → analogous to allelic profile.

## **1.8 Method**

There are four basic applications, or methods, that we have developed for use with Minimum SNPs. Below is a brief description of each.

### **1.8.1 Percent Method:**

What set of SNPs should be tested to differentiate a single known sequence type from all other sequence types available on the MLST database?

### **1.8.2 “D” method:**

What set of SNPs should be tested to differentiate an unknown sequence type from any other sequence type in the MLST database?

### **1.8.3 Not-N method:**

What set of SNPs should be tested to identify a defined group or complex of sequence types such that a false negative result cannot occur?

### **1.8.4 Working backwards method:**

Which sequence types available on the MLST database share a defined set of SNP alleles?

## 2 Getting Started

Minimum SNPs uses the Java Runtime Environment so you will need this software installed on your computer before you begin. You can download the latest version of Java from [http://www.java.com/en/download/windows\\_ie.jsp](http://www.java.com/en/download/windows_ie.jsp). If you experience problems using the latest version of Java you may wish to try using an older version. Java 2 Runtime Environment, SE v1.4.2\_01 was used for the creation of this manual.

There are many versions of Minimum SNPs that were released at various stages in its evolution. Be aware that some older versions have serious bugs that will make some of the applications described in this manual erroneous. The version of Minimum SNPs used at the time of writing this manual was MLST2043\_\_\_Not\_N.

Once you have unzipped the parent file (MLST2043\_\_\_Not\_N) you will notice that a new folder has been created called "classes". There are many files within the "classes" folder but there is only one executable Jar file called "MLST". When you open the "MLST" file make sure that you open with Java ("javaw"). The following icon should appear if the file is ready to be opened with Java.



MLst.jar

### Important note

At the time the early versions of Minimum SNPs were written, it was not possible to download concatenated databases from MLST web sites. As a consequence the software contains its own concatenation facility. (In this manual, a concatenated MLST database is termed a "mega-alignment"). Although the concatenation function in Minimum SNPs is completely functional, it is somewhat inconvenient and counter-intuitive to use, with the major issue being that the concatenated data cannot be stored and must be re-assembled from the allele sequence and ST profile data every time the software is used. (Although this only takes a few minutes, it is quite annoying). In consequence, we recommend that analyses be carried out on pre-concatenated databases where possible. This allows analyses to be carried out using much fewer key-strokes and in an inherently simpler fashion. The only disadvantage is that while the "on board" concatenation function keeps track of where each locus starts and stops, and defines SNPs in terms of location within MLST loci, this obviously cannot be done with pre-concatenated data. In this case, the output SNPs are defined only in terms of their location within the concatenated data.

This manual is currently structured around using the on-board concatenation function. However, sections that detail how to obtain and use pre-concatenated MLST data have been added throughout.

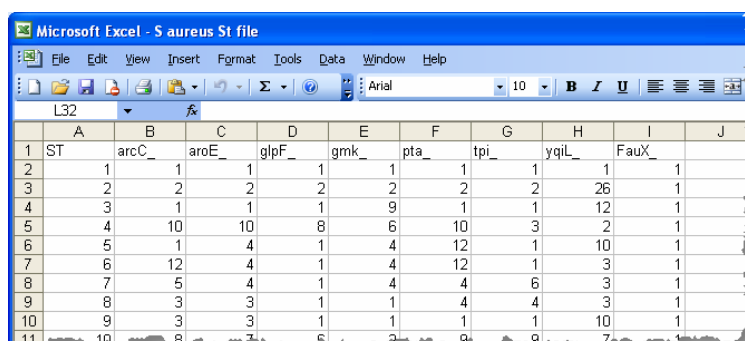
### 3 Creating a Mega-Alignment:

**Note:** if you intend to use pre-concatenated MLST data, you do not need to create a mega-alignment. Go to Section 4.

Minimum SNPs concatenates MLST data by processing the individual allele sequence alignments, with reference to a file of MLST allele profiles. This section describes how to download these data, ensure that they are in the correct format, and assemble a mega-alignment i.e. a concatenated MLST database.

#### 3.1 Obtaining and preparing a file of MLST allele profiles :

Sequence types and allelic profiles can be downloaded from the Multi Locus Sequence Typing website (<http://www.mlst.net/>). Follow the link to “Databases” and select the database for the organism of interest. Depending on the host website, allelic profiles may be downloaded in tab-delimited form, comma-delimited form or as a csv file. The sequence type and allelic profile list must then be converted to a csv file with column headers “ST” for column **A** and the loci names for columns **B** through **H**. The software has a bug which results in the last locus (column **H**) not being included in the mega-alignment. To circumvent this problem, simply insert a fake locus into column **I**. Column **I** then becomes the last column and is not included in the analysis.



	A	B	C	D	E	F	G	H	I	J
1	ST	arcC	aroE	glpF	gmk	pta	tpi	yqiL	FauX	
2		1	1	1	1	1	1	1	1	1
3		2	2	2	2	2	2	2	26	1
4		3	1	1	1	9	1	1	12	1
5		4	10	10	8	6	10	3	2	1
6		5	1	4	1	4	12	1	10	1
7		6	12	4	1	4	12	1	3	1
8		7	5	4	1	4	4	6	3	1
9		8	3	3	1	1	4	4	3	1
10		9	3	3	1	1	1	1	10	1
11		10	8	7	6	2	9	9	7	1

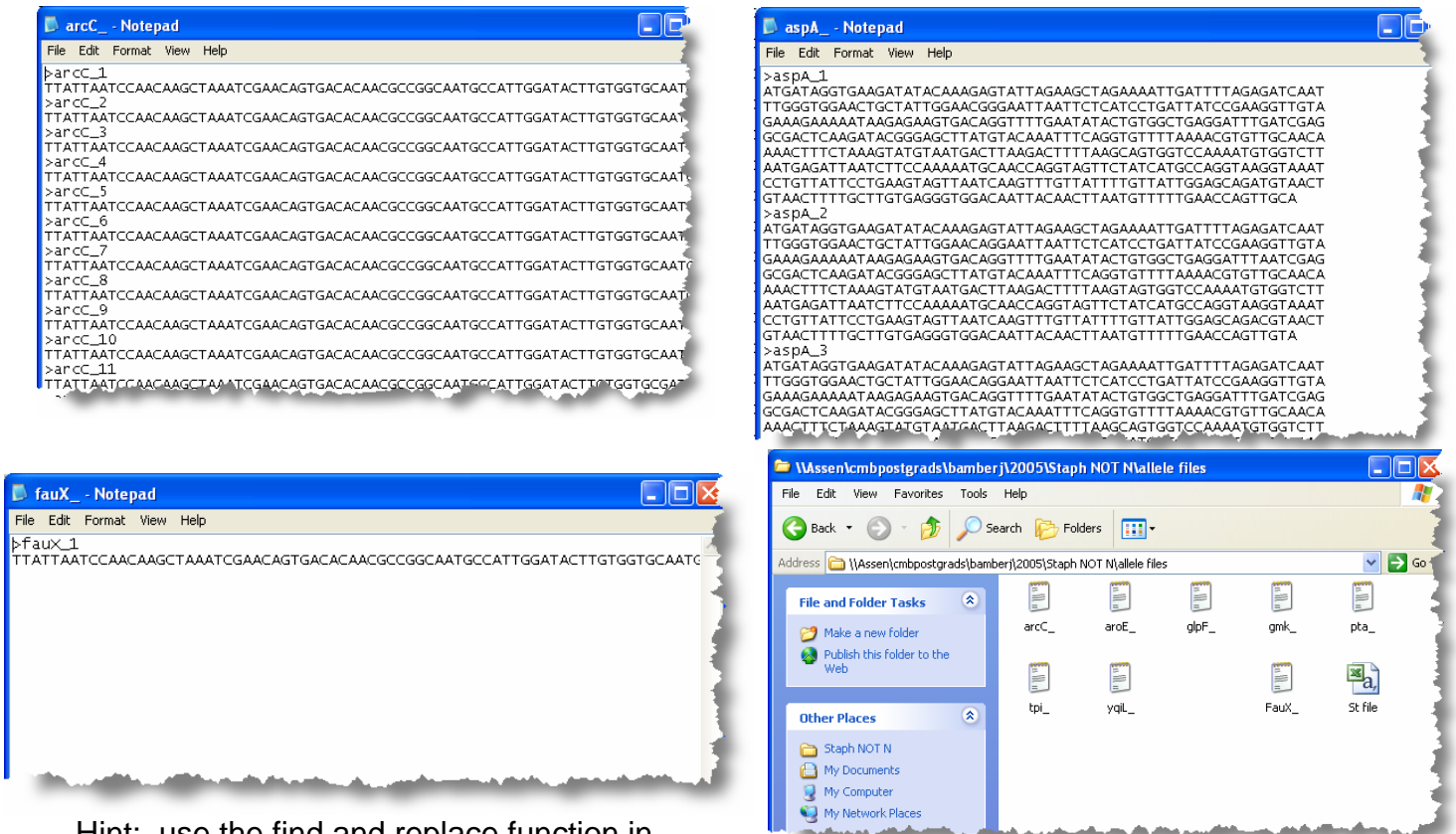
The software has a bug which results in the last locus (column **H**) not being included in the mega-alignment. To circumvent this problem, simply insert a fake locus into column **I**. Column **I** then becomes the last column and is not included in the analysis.

**NB:** The names used to identify the loci must be exactly the same as those used in the allele files.

#### 3.2 Obtaining and preparing allele sequence alignment files:

Allele sequences are also downloaded from the MLST web site (<http://www.mlst.net/>). Once again, follow the link to “Databases” and select the database for the organism of interest. Allele sequences can be downloaded in FASTA format from all host websites. Some downloaded allele sequences will have paragraph endings (¶) at regular intervals thought each sequence while others will have no paragraph endings within the sequence. Either version of the download can be used with Minimum SNPs.

**NB:** The names used for the loci in FASTA format allele sequences must be exactly the same as those used in the ST file above.



Hint: use the find and replace function in Microsoft Word to change the allele sequences into the correct format.

### 3.3 Loading Allele Files:

Loading the Allele files into Minimum SNPs follows the same procedure for all four methods (applications). As described in Section 3.1, a false locus should be the last allele file to be loaded into Minimum SNPs. The last allele will not be included in the Mega-Alignment, but will act as a bookend or a full stop, in a manner of speaking.

- Open Minimum SNPs.
- Click “File” on the menu bar.
- Select the option “Load Allele File” from the dropdown menu.
  - A new window opens which allows you to brows for the location of you allele files.
- Open the first allele file.
  - The sequences with which the allele files are loaded does not matter at this point. However, I usually load the false allele last for consistency, as you will soon see.



- Repeat the steps above until all allele files have been loaded.
  - You can keep track of where you are up to in loading allele files by the title given to the Minimum SNPs window.

Once you have loaded all the allele files, individual alleles can be viewed in task pane. A locus is selected by

- Click “File” on the menu bar.
- Select the option “Alleles” from the dropdown menu.
- From the second dropdown menu select the locus that you would like to use.
  - Notice that the incorrect terminology has been used here. The option “Alleles” on the “File” dropdown menu should really read “Loci”.
- You will notice that each allele of the selected locus can be viewed by selecting from the drop down menu adjacent to the word “Allele” in the task bar.

### **3.4 Assembling Alleles and Loading ST File:**

Once all the alleles have been loaded into Minimum SNPs, they need to be assembled into a Mega-Alignment.

This is done by sequentially ‘joining’ the loci together followed by loading the sequence type file which of which alleles from each locus to include in the sequence of a particular sequence type. It is important to note that the loci do not need to be joined in any particular order except for the final locus, which is a false locus, acting as a bookend.

- Change the mode from “%” to “D” by clicking the “D” button on the task bar.
- Click “File” on the menu bar.
- Select the option “Alleles” from the dropdown menu.
- From the second dropdown menu select the locus that you would to add first.
  - I usually work from top to bottom as they are listed in the drop down menu – the order in which they were loaded..
- On the task bar click “Add”
- On the task bar click “Start”
- On the task bar click “Accept”
- Click “File” on the menu bar.
- Select the option “Alleles” from the dropdown menu.
- From the second dropdown menu select the locus that you would to add next.
- Click “Add”
- Click “Accept”

- Repeat the above 5 steps until all loci have been joined (added) including the false locus.
  - Once again, you can keep track of where you are up to in joining loci together (adding loci) by the title given to the Minimum SNPs window.
- On the task bar click “Finish”
- At this point a new dialog box opens enabling you to locate and open the sequence time file (called a strain file).
- Once you have opened the sequence type file the word “Mega-alignment” will appear in the title if the process has been successfully completed. If you do not see the word “Mega-Alignment”, then there has been an error in either the data within your allele or ST files or in the sequence of loading and adding alleles in Minimum SNPs.

This mega-alignment is now ready for analysis using the %, D or Not-N modes.

## 4 Obtaining and loading a pre-concatenated MLST database

Most MLST sites now provide the option of downloading concatenated data. The early versions of Minimum SNPs were designed to analyse data from individual MLST loci. As a consequence of this, Minimum SNPs can deal with a pre-concatenated MLST database as if it were, in effect, a single giant MLST locus, and each sequence type is an allele. This approach is very easy and straightforward.

### 4.1 Downloading a pre-concatenated MLST database

- Download and save the concatenated MLST data in FASTA format. In order for the program to run, all the sequences must be same length and composed only of G's, A's, T's and C's. We have found that sequences of the incorrect length are indeed present in some concatenated databases. These must be removed or "doctored" so they are the correct length.

### 4.2 Loading pre-concatenated data into Minimum SNPs

- Open Minimum SNPs.
- Click "File" on the menu bar.
- Select the option "Load Allele File" from the dropdown menu.
  - A new window opens which allows you to browse for the location of your concatenated data file.
- Select the file that is the concatenated data in FASTA format.

The file will then load. If the file is large (e.g. thousands of STs), the loading process may take 20 seconds or more.

The concatenated MLST data is now ready for analysis using the %, D or Not-N modes.

## 5. Deriving highly informative sets of SNPs.

Minimum SNPs assembles highly informative sets of SNPs by carrying out an empirical search i.e. by assessing the informative powers of thousands of different SNP combinations. Three different algorithms – the %, D and Not-N modes – are available for measuring the informative powers of candidate SNP sets. For complete details see Robertson et al (2004), Price et al (2006), Price et al (2007).

### 5. 1. % Mode

This mode is used to identify SNP set(s) diagnostic for a single user-specified ST. It is called the % mode because the informative power is stated as the % of STs that are discriminated from the user-specified ST by the SNP(s).

To obtain and load data files, see Section 3.

#### **5.1.1. Using a mega-alignment (An MLST database that has been concatenated using the Minimum SNPs on-board concatenation facility)**

- Change the mode from “D” to “%” by clicking the “D” button on the task bar.
- Select the ST that you are interested from the dropdown menu adjacent to the word “Allele” in the top right corner of the task bar.
- Select percent.
- Click “Identify Allele” in the task bar.

The results pane will display the following information:

- The loci included in the analysis
- The ST that has been queried
- The concatenated sequence of that sequence type (according to the order in which the loci were joined together (added).
- The identification constraints – see [Tools and Options](#)
- The SNP’s returned by the Minimum SNP’s analysis and the corresponding confidence of identifying the sequence type queried from all other sequence types in the database using those SNP’s.
- Click “Clear Report” on the results bar to start over OR;
- [Save and Print your Results](#)

#### **5.1.2. Using a pre-concatenated MLST database downloaded from an MLST web site.**

- Select the sequence type of interest in the “allele” drop-down menu.

- Ensure that the “%” mode has been selected (top right hand corner). The symbol appears ghosted when it is selected.
- Click “identify allele”.

The results pane will display the following information:

- The ST that has been queried (labelled as an allele).
- The sequence of that ST.
- The identification constraints – see [Tools and Options](#)
- The SNP’s returned by the Minimum SNP’s analysis and the corresponding confidence of identifying the sequence type queried from all other sequence types in the database using those SNP’s.
- Click “Clear Report” on the results bar to start over OR;
- Save and Print your Results

## **5. 2 “D” Method:**

This is used to identify SNP sets that are optimised towards discriminating all STs from all STs. This is accomplished by assessing the informative power of candidate SNP(s) by calculating the Simpsons Index of Diversity ( $D$ ) with respect to the input sequence alignment. In this context,  $D$  is the probability that any two sequences selected at random from the alignment (without replacement), will be discriminated by the SNP set under test.

To obtain and load data files, see Section 3.

### **5.2.1. Using a mega-alignment (An MLST database that has been concatenated using the Minimum SNPs on-board concatenation facility)**

- Make sure that the mode selected in the task bar is “D” (The symbol appears ghosted when it is selected).
- Click “Identify Allele” in the task bar. You may need to wait some time at this point.

**Note: although it is possible to select an ST in the drop down menu, this has no effect on the results, because the D algorithm does not identify SNPs with reference to any particular ST.**

The results pane will display the following information:

- The loci included in the analysis
- The identification constraints – see [Tools and Options](#)
- The SNP’s returned by the Minimum SNP’s analysis and the corresponding Simpson’s index of diversity associated when the MLST database is assessed with the resulting SNPs.
- Click “Clear Report” on the results bar to start over OR;

- Save and Print your Results

### **5.2.2. Using a pre-concatenated MLST database downloaded from an MLST web site.**

- Ensure that the “D” mode has been selected (top right hand corner). The symbol appears ghosted when it is selected.
- Click “identify allele”.

**Note: although it is possible to select an “allele” (actually an ST) in the drop down menu, this has no effect on the results, because the D algorithm does not identify SNPs with reference to any particular ST.**

The results pane will display the following information:

- The identification constraints – see [Tools and Options](#)
- The SNP’s returned by the Minimum SNP’s analysis and the corresponding Simpson’s index of diversity associated when the MLST database is assessed with the resulting SNPs.
- Click “Clear Report” on the results bar to start over OR;
- Save and Print your Results

### **5.3 Not-N Method: (Note – the key strokes for this method are very different from the % and D modes)**

#### **5.3.1. Using a mega-alignment (An MLST database that has been concatenated using the Minimum SNPs on-board concatenation facility)**

- Change the mode form “D” to “%” by clicking the “D” button on the task bar.
- Click “Consensus” in the task bar.
- In the data box adjacent to the “Consensus” button, type (or paste) the list of sequence types (ST’s) that you want to identify as a group.
  - The list of ST’s must have the format:
    - ST  $x_1$ , ST  $x_2$ , ...,ST  $x_n$
    - Where x is the MLST sequence type and n the n<sup>th</sup> sequence type in the series.
    - Make sure there is a space after every “ST” and a space after every “,” and no other spaces.
- Click “Not-N” in the task bar.

The results pane will display the following information:

- The loci included in the analysis

- The identification constraints – see [Tools and Options](#)
  - The results pane is as for the % and D modes, with the exception that the selected group of sequences is listed, and the SNPs are defined in “NOT” mode, i.e. “NOT – AGT” means “C”, and “NOT – GC” means “A or T”.
- Click “Clear Report” on the results bar to start over OR;
  - Save and Print your Results

### **5.3.2. Using a pre-concatenated MLST database downloaded from an MLST web site.**

- (A quirk of the software is that it does not matter which of the D or % buttons is selected.)
- Click on the “consensus” button
- In the data box adjacent to the “Consensus” button, type (or paste) the list of sequence types that you wish to identify as a group.
  - **Important: these must be typed EXACTLY as they are listed in the “allele” drop down menu. This is case sensitive, and spaces must be included if they are within the name. In addition, the “>” symbol that denotes sequence names in the FASTA format must also be included. The sequence names must be separated by commas, and there MUST be a space after each comma. If you are doing many Not-N analyses, it is helpful to use an input file in which the sequence names (in the Fasta format) are very short e.g. just “>x” where x is the ST number. You can then type the sequence as e.g. “>5, >28, >79, >766” in the consensus box.**
- Click the Not-N button.
- The results pane is as for the % and D modes, with the exception that the selected group of sequences is listed, and the SNPs are defined in “NOT” mode, i.e. “NOT – AGT” means “C”, and “NOT – GC” means “A or T”.

**Helpful hint.** It is common for Not-N analyses to fail to yield highly resolving sets of SNPs – such sets may simply not exist. Therefore, it is important to not set the “confidence” level (under the “tools” drop down menu, “allele options” menu item) too high. If it is too high, the program will simply provide no results. It is prudent to start with the confidence at about 50% and work up from there.

## **6. Working Backwards Method:**

“Working backwards” is calculating the STs defined by a particular user-defined SNP profile. The methods for doing this in Minimum SNPs are a little counter-intuitive, as is explained below:

1. The software cannot work backwards using a mega-alignment that has been constructed using the on-board concatenation function. However, it can work backwards using the separate locus-specific alignments that are used to construct mega-alignments. Although this is fully functional, it is not very convenient. In particular, it requires that the software be re-started, and the sequence alignments reloaded.
2. Working backwards using a pre-concatenated MLST database (i.e. a single alignment) is extremely quick and easy. The ease of working backwards is a major argument in favour of using pre-concatenated data.
3. This mode does not operate by the user actually inputting the SNP profile. Rather, it operates as follows: if the program is given one or more SNPs and a particular ST, it will determine the SNP profile defined by that ST, and also determine which STs have the same SNP profile.

### **6.1 Working backwards if you are using separate locus specific sequence alignments and an ST allele profiles file i.e. you usually work using the on-board “mega-alignment” concatenation function**

If you have already created a Mega-Alignment you will need to close Minimum SNPs and start over as working backwards does not make use of a Mega-Alignment.

- Load allele files
- Change the mode from “D” to “%” by clicking the “D” button on the task bar.
- Click “File” on the menu bar.
- Select the option “Alleles” from the dropdown menu.
- From the second dropdown menu select a locus that you would like to work with.
- In the data box adjacent to “Identity Check” type the position of the SNP within the selected locus that you would like add to the SNP profile you are characterising.
  - If there is more than one position within that particular locus you must query both at once. Separate the SNP positions with a coma. Do not use any spaces. Example:
    - 36,241,243
- Click “Add” in the task bar.
- Click “Start” in the task bar.
- Click “Insert” in the task bar.



- Check that the polymorphs presented in the data box adjacent to the “Start” button represent the profile that you are characterising. If they do not, modify the data by replacing the letter for the nucleotide at the appropriate position with the desired polymorph. Do not change the format of the data in the box at all.
- Click “Accept” in the task bar.
- Click “File” on the menu bar.
- Select the option “Alleles” from the dropdown menu.
- From the second dropdown menu select the next locus that you would like to work with.
- In the data box adjacent to “Identity Check” type the position of the SNP within the second locus that you would like add to the SNP profile you are characterising.
- Click “Add” in the task bar.
- Click “Insert” in the task bar.
- Check that the polymorphs presented in the data box adjacent to the “Start” button represent the profile that you are characterising
- Click “Accept” in the task bar.
- Repeat the above process for all loci included in the profile.
- Click “Finish” on the menu bar.
- At this point a new dialog box opens enabling you to locate and open the sequence time file (called a strain file).
- Once you have opened the sequence type file the process is complete and the results will be presented in the results pane.

The results pane will display the following information:

- A list of alleles that share the same profile at each selected locus
- The indistinguishable STs based on the SNPs that have been included in the profile.
- Click “Clear Report” on the results bar to start over OR;
- Save and Print your Results

### ***6.2 Working backwards using a pre-concatenated MLST database downloaded from an MLST web site.***

- Type the SNP positions in the “identity check” box. These must be separated by commas (in this instance, do NOT put a space next to the comma).

- Select the ST that defines the SNP profile you are interested in, using the “allele” drop down menu
- Click “add”
- The results pane contains the SNP profile of the selected ST, and the other STs that share the same profile.
- Click “Clear Report” on the results bar to start over OR;
- Save and Print your Results

## 7. Tools and Options

The “allele options” page under the “tools” menu is very useful. The functions of these options are listed below

### 7.1 Number of results

Minimum SNPs assembles SNP sets by first finding the most informative SNP and calling that SNP1, and then finding the SNP that is most informative in combination with the SNP1, and calling that SNP2, and so forth. If there is a “tie” (i.e. the same score – a draw) between one or more SNPs, it uses that to define alternative SNP sets. In effect, each of the tied alternatives seeds new pathways of SNP set assembly. The “number of results” option simply sets an upper limit on the number of SNP sets that are listed in the results. Nearly all ties occur late in the process of SNP set assembly, so that if the the SNP set is kept small using e.g. the Confidence option, then often only one result will be returned.

### 7.2 Paragraph width

Affects the format of the results.

### 7.3 Inclusions and Exclusions

This option is extremely useful. It allows the user to tell the program to ignore SNPs (exclusions), or to include SNPs in the output set, irrespective of resolving power (inclusions). This allows SNP sets to be easily adapted and optimised. For example, a SNP that is technically difficult to interrogate can easily be replaced by the next most informative SNP, without the necessity for redesigning the whole SNP set. It is important to note that when a SNP is forced to be included in the set using the “include” function, its informative power is used in the derivation of other SNPs to be added to the set. “Included” SNPs can therefore be used to “seed” SNP sets. The “include” function is also a very easy way of determining the resolving power of SNP sets that have been identified by “Minimum SNPs”-independent methods.

### 7.4 Time out (seconds)

This also can be very important, although primarily for the trouble it causes, rather than the benefits. In essence, if it is set too low, then the program will terminate before finishing its calculation, and will return a result of “no results found”. It usually makes sense to set it to a high value – in particular if you are doing D mode searches on large databases. These searches can take up to several hours, especially on older computers (although more usually it is a few minutes at most). % and not-N searches should only take a few seconds.

### 7.5 Confidence (1-100)

This option is very important. It is the maximum resolving power that is aimed for in % and not-N searches. It is important to have this set at an appropriate value. This is because Minimum SNPs does not have good error-handling functions. If this value is set higher than can ever be reached – and this can happen if the not-N mode is used, or there are identical sequences in the input file, then the program will never finish its calculation, and

will simply proceed until it times out, and returns no result. An efficient and low stress approach to doing searches is to start with the confidence set low, and work up. E.g. start with 95 for % searches, and 50 for not-N searches.

## **7.6 Simpsons Index (0-1)**

This is the D value analog of the “confidence” option above. This option is even more important, because it can take a large number of SNPs to reach a D value of 1.0 with most MLST databases. As the calculation time increases exponentially with the number of SNPs, it can be impractical in terms of processing time to reach  $D = 1.0$ . Once again, it is prudent (and efficient) to begin with a low D (e.g. 0.9), and work up from there to see how high it is practical to achieve. A search that is limited by  $D = 0.9$  will generally only take a few seconds, even on very large data sets.

## **7.7 Search Depth**

We do not use this option very often, but it can be useful. This limits the size of the SNP set that is returned. It apparently over-rides the limits set by the Confidence or Simpsons Index options above, so e.g. if the Confidence is set to 99%, and the Search Depth to 2, then it may return a SNP set that does not reach 99%. Also (somewhat confusingly), the search depth is incorrect by 1, so that a search depth of 1 will actually allow the return of a two member SNP set, and so forth. It is most useful as a time out or infinite loop error avoidance strategy, although similar aims can be achieved with judicious use of the Confidence and Simpsons Index options.

## **7.8 Number of loci**

This sets an upper limit on the number of loci that can be used in assembling mega-alignments. There is no problem with having it set higher than the number of loci used – in particular having it set to “7” (the number of loci in MLST schemes) causes no problem in using pre-concatenated MLST datasets in the form of single alignments. Therefore, it is better to simply set it to 7 and then leave it alone.