eBURSTv3

http://eburst.mlst.net
Version 3; February 2006

Contents

1. Introduction

BURST was devised and developed by Ed Feil (University of Bath; e.feil@bath.ac.uk), while he was in the laboratory of Brian Spratt (now at Imperial College London; b.spratt@imperial.ac.uk), as a way of displaying the relationships between closely-related isolates of a bacterial species or population*. BURST, unlike cluster diagrams, trees or dendrograms, uses a simple but appropriate model of bacterial evolution in which an ancestral (or founding) genotype increases in frequency in the population, and while doing so, begins to diversify to produce a cluster of closely-related genotypes that are all descended from the founding genotype (Figure 1A). This cluster of related genotypes is often referred to as a "clonal complex".

The BURST algorithm first identifies mutually exclusive groups of related genotypes in the population (typically a multilocus sequence typing [MLST] database), and attempts to identify the founding genotype of each group (see below). The algorithm then predicts the descent from the predicted founding genotype to the other genotypes in the group, displaying the output as a radial diagram, centred on the predicted founding genotype. The procedure was developed for use with the data produced by MLST (sequence types or STs, and their allelic profiles), but using a suitable criterion for the definition of groups of related genotypes, it could be used with some other types of data, particularly data from multilocus enzyme electrophoresis (MLEE) or from molecular typing methods that produce strings of integers by using multiple repeat length polymorphisms.

The approach used in BURST greatly simplifies the problem of depicting the evolutionary relationships among closely related genotypes, which are poorly represented on a tree, as the focus for each genotype is its relationship to, and distance from, its predicted founding genotype, and the great majority of the pair-wise relationships between genotypes in the population are ignored. BURST does not make any inferences about the relationships between the more distantly related genotypes that belong to different groups. In many bacterial populations, the relationships between distantly-related genotypes may be difficult to discern, as the extent of homologous recombination may be sufficiently high that relationships may be poorly represented by a phylogenetic tree, and should be represented by a network.

eBURST does not tell you the truth – it simply produces an hypothesis about the way each clonal complex may have emerged and diversified - and any additional phenotypic, genotypic, or epidemiological data that are available should be used to explore the plausibility of the proposed ancestry and patterns of descent.

The original version of BURST was implemented by Ed Feil and Man-Suen Chan (man-suen.chan@paediatrics.ox.ac.uk). A greatly enhanced version of BURST (eBURST v1) was developed as a Java™ applet by Bao Li (Bao@mac.com), and was integrated into the MLST website by David Aanensen (d.aanensen@imperial.ac.uk), within the laboratory of Brian Spratt, with help and advice from Ed Feil, Jon Evans (jemevans@yahoo.com), Bill Hanage (w.hanage@imperial.ac.uk) and Christophe Fraser (c.fraser@imperial.ac.uk). The original version of BURST is still available at www.mlst.net and, if necessary, may be used in conjunction with eBURST, as the two algorithms differ slightly in the way they display the relationships between STs.

eBURSTv3 has been further developed with funding from the Wellcome Trust by Derek Huntley and David Aanensen at Imperial College London and contains several new features improving on previous versions:

JAVA Web Start Implementation – allowing the most up-to-date version of eBURSTv3 to be installed locally and data streamed over the internet directly to the application.

Comparative eBURST - The ability to compare two datasets and differentially colour those STs unique to one or other of the datasets, or present in both datasets.

Other enhancements –

- Docking of eBURSTv2 floating windows and menu based access to functionality.
- Search and highlight STs within a population.
- Printing directly from eBURST.
- Output formats – images can be saved in a number of bitmap formats and also as Scalable Vector Graphics (SVG) allowing further editing of eBURST diagrams in third party applications.

*Feil, E.J., Li, B., Aanensen, D.M., Hanage, W.P. and Spratt, B.G. 2004. eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bact. **186:** 1518-1530.

2. Defining groups

The MLST data used by eBURST are the STs and their associated allelic profiles. The first step is to divide the input data (e.g. the isolates within a MLST database) into groups of STs that have some user-defined level of similarity in allelic profile, since eBURST focuses on those STs that are similar and which may share descent from the same founding genotype, and provides no information about the more distant relationships between groups.

The definition of the group can be changed, but the default eBURST setting is to identify groups of related STs using the most stringent (conservative) definition, where all members assigned to the same group share identical alleles at ≥ 6 of the 7 loci with at least one other member of the group. A less stringent approach is to define the groups by the sharing of alleles at ≥ 5 of the 7 loci. Whatever group definition is used, this approach results in non-overlapping groups; no ST can be assigned to more than one group.

A 'group' is used here as a neutral term for the collection of STs that are placed together by eBURST, according to the selected group definition, whereas a clonal complex is a set of STs that are all believed to be descended from the same founding genotype. Using the stringent group definition (6/7 shared alleles), isolates in the group defined by eBURST will be considered to belong to a single clonal complex. With a less stringent group definition (e.g. 5/7 shared alleles) all of the STs in an eBURST group cannot be assumed to belong to a single clonal complex.

eBURSTv3 is designed for MLST data, which typically uses seven loci, but there is an option to change the number of loci, and the number of shared alleles used to define a group can also be changed to an appropriate number within the application (see below).

3. Identifying the founder of a group and the patterns of evolutionary descent

The following description of eBURST focuses on groups identified using the stringent default setting, where all STs must be a single locus variant (SLV) of at least one other ST in the group, and eBURST produce a diagram where all STs in the group are linked (a clonal complex). The use of eBURST with the more relaxed group definition of 5/7 shared alleles is discussed later.

The primary founder of a group is defined as the ST that differs from the largest number of other STs at only a single locus (i.e. the ST that has the greatest number of single-locus variants; SLVs). This method of assigning the primary founder takes account of the way in which clones emerge and diversify (Figure 1A); the initial diversification of the founding genotype of a clonal complex will result in variants of the founder that differ at only one of the seven loci (i.e., SLVs of the founder). If two STs have the same number of associated SLVs, the one with the greatest number of double-locus variants (DLVs) is selected as the founding ST. In such cases the confidence in the assignment of the primary founder will be low, as reflected in the bootstrap values (see section 5).

Note that the assignment of the founding ST does not take into account the number of isolates of each ST; this makes the procedure relatively robust to sampling bias. In many cases, however, the predicted primary founder will also be the most prevalent ST within the group. Together with strong bootstrap support, this can provide added confidence in the assignment of the primary founder. The average distance of each ST to all other STs in the eBURST group is also calculated by eBURST v3 and the primary founder is likely to be the ST with the minimum average distance to all other STs.

The eBURST diagrams display the patterns of descent within each group from the predicted founding ST (Figure 1B). The initial version of BURST placed the founder centrally and positioned SLVs and DLVs of the founder within concentric rings, whereas eBURST uses lines to show the radial links from the founder to each of its SLVs. A second difference is that only SLV links are shown in eBURST, and DLVs of the founder will therefore only be linked if the intermediate SLV on the path from the founder to the DLV is present in the input data. All isolates that are directly linked on a eBURST diagram will therefore differ at only a single locus and the cluster of linked STs should represent a clonal complex. A DLV of the predicted primary founder of a group will not be included if the linking SLV has not been sampled and using the stringent (conservative) default group definition some STs that might be part of the clonal complex will therefore not be shown. In the eBURST diagram, the circle representing the predicted primary founder is coloured blue and the areas of each of the circles indicate the prevalence of the ST in the input data (Figure 1B).
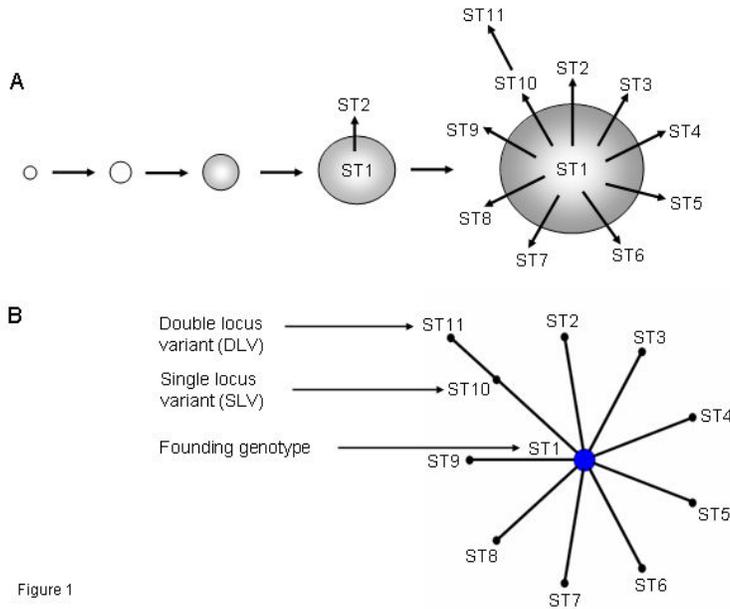
Figure 1

With the less stringent group definition of 5/7 shared alleles, more than one cluster of linked STs (clonal complex) may be displayed in the eBURST diagram, along with a number of unlinked individual STs. The lack of linking between two clusters within a single eBURST group implies that no ST within one cluster is a SLV of any ST in the other cluster(s). Similarly, unlinked individual STs will not be SLVs of any ST in the group. The SLVs and DLVs of any ST can be displayed on the eBURST diagram (see section 9.7.).

6

4. Subgroups and subgroup founders

In larger eBURST groups there may be several STs besides the predicted primary founder that have a number of SLVs of their own. For example, a SLV of the primary founder may have become successful and diversified to produce its own SLVs. A ST that appears to have diversified to produce multiple SLVs is called a subgroup founder. In the default setting for the eBURST diagram, a ST with at least two assigned descendent SLVs (i.e. three SLVs in total, as the link from the ST to its putative progenitor is also a SLV) is defined as a subgroup founder. In larger eBURST groups there may be many STs with at least two assigned descendent SLVs, and the number of descendent SLVs that define whether a ST is shown as a subgroup founder should be increased. The STs that are subgroup founders, according to the default definition of a subgroup, or a user-defined definition, are coloured yellow on the eBURST diagram (see examples below). The primary founder of the group is coloured blue.

5. Estimating the level of confidence in the predicted founding ST

In some cases, particularly in small groups with few SLVs, the predicted primary founder of a group will have little statistical support and a bootstrapping procedure is used in eBURST to provide the level of confidence in the assignment of the primary founder of each group. The bootstrap values are only available using the default group definition as they are not appropriate with less stringent group definitions where an eBURST group may include several unlinked clusters of STs (clonal complexes).

eBURST first divides the input population into groups and, for each group, one example of each ST is extracted, and a user-defined (default = 1000) number of random datasets of the same size as the extracted ST set are produced by re-sampling with replacement. For each re-sampling the predicted primary founder is computed, and a tally is kept of the number or times that each ST in the group is predicted to be the primary founder of the group. The bootstrap values shown for each ST are the percentage of times the ST was predicted to be the primary founder of the group in the bootstrap re-samplings. As a ST cannot be the predicted founder if it is not present in a re-sampled dataset, the calculation of the percentage of times each ST is predicted to be the primary founder omits those re-samplings in which that ST is absent. eBURST therefore produces conditional bootstraps, which scale the bootstrap confidence values to 100%; a ST that is the predicted primary founder in every re-sampling in which it is present will have a bootstrap value of 100%.

Bootstrap values for the primary founders of the groups are shown for re-samplings from the extracted ST set, but are also available for re-samplings taken from the set of all isolates in the group, which often will include multiple isolates of some STs. The bootstrap values for a ST may differ considerably when one of each ST is used, compared to when all isolates are included, since the STs represented in each re-sampling will be influenced by the frequencies of the STs in the database. Considerable caution should therefore be used in evaluating the bootstrap values using all isolates as they are difficult to interpret, and the default setting is to show only the bootstrap values for the resampled datasets obtained from the extracted set containing one of each ST. An option can be set to show the bootstraps calculated using all isolates. We recommend that the bootstrap value obtained using resamplings from a dataset containing one of each ST is used to evaluate the robustness of the assignment of the primary founders.

In some cases two STs may have relatively high bootstrap support and the predicted primary founder may not be the real founder of the group. For example, a SLV of the real founder of a group that becomes antibiotic-resistant, and of much clinical concern, is likely to be massively over-sampled – this can result in a number of its antibiotic-resistant SLVs being sampled. The over-sampling of the latter SLVs can lead to the resistant strain being selected as the founder of the whole group. In this case, mapping antibiotic resistance onto the eBURST diagram would indicate that this may have happened, but the sudden increase in the success of a SLV that is not antibiotic-resistant may cause a similar mis-assignment that would not be so apparent. The presence of two isolates having substantial bootstraps scores provides an alert that the assignment of the primary founder cannot reliably be discerned in the absence of additional information.

6. Building up the eBURST diagram

The original BURST algorithm identifies the primary founder as the ST with the largest number of SLVs.   Those STs that are SLVs of the predicted primary founder are assigned to the primary founder and the new ST that has the greatest number of previously unassigned SLVs is identified. The iterative process of identifying the new ST with the greatest number of previously unassigned SLVs continues until all of the STs that have multiple unassigned SLVs (subgroup founders) have been identified.   DLVs of the primary founder and of the subgroup founders are then identified in a similar iterative manner. The original BURST displayed the primary founder and the subgroup founders, and their SLVs and DLVs, but did not attempt to link all of these clusters of STs together.

In eBURST, the default group definition results in all STs being connected as a single clonal complex and eBURST makes these links. However, if the primary founder and subgroup founders, and their SLVs and DLVs, are assigned as described above for the original BURST algorithm, there can be problems in achieving a fully linked diagram without introducing *ad hoc* linking rules. In order to circumvent this problem, eBURST produces an initial approximation of the above arrangement of STs, which ensures that all STs in the group are linked, and then optimises the arrangement of STs to produce the final eBURST diagram.
The procedure is as follows. The ST that has the greatest number of SLVs is assigned as the primary founder and is positioned centrally with radial links to all of its SLVs. Having assigned all of the SLVs of the primary founder, the SLVs of each of these SLVs are identified (ignoring any STs that have already been assigned to the primary founder) and linked, and this iterative procedure of linking previously unassigned SLVs carries on outwards from the SLVs, to the DLVs and then to the TLVs, until all SLV links have been made.

Optimisation of the initial arrangement of STs is then carried out. The optimisation method looks at each ST (excepting the primary founder and its SLVs, which are unambiguously assigned) and searches for a better positioning of STs that maximises the numbers of SLVs associated with subgroup founders. Optimisation takes account of the simple model of clonal expansion that underpins BURST where some STs within a clonal complex may have increased in frequency and diversified to produce subgroups. It attempts to identify the most likely pattern of subgroups by searching for those subgroup founders that have the greatest numbers of linked SLVs.
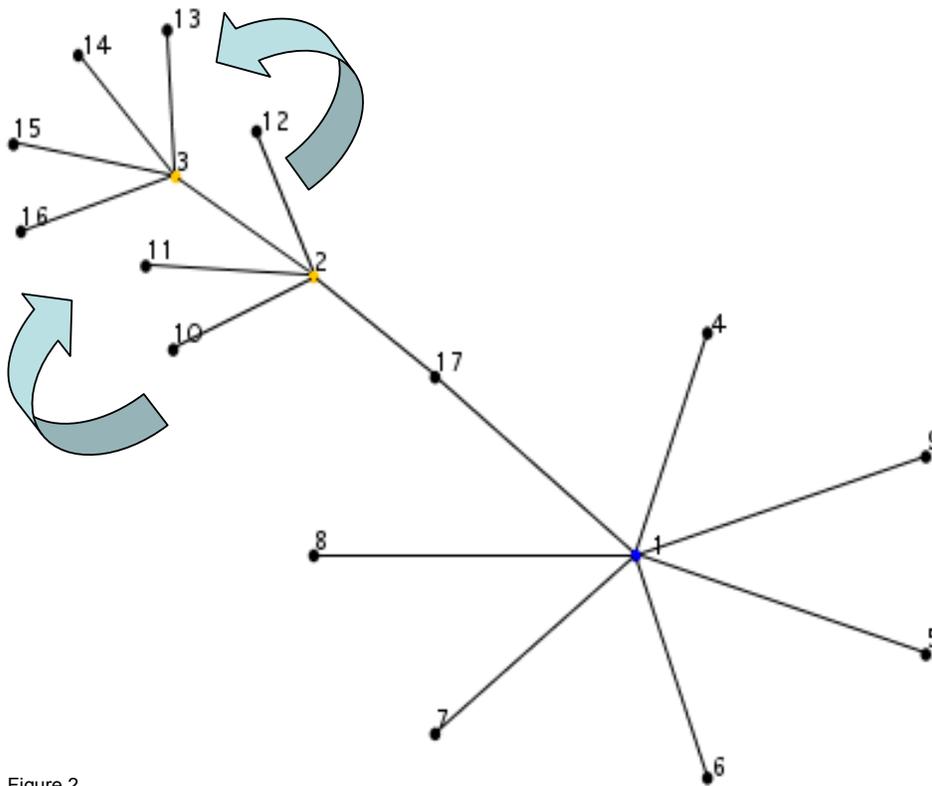
Figure 2

An illustrative example is shown in Figure 2. The initial procedure identifies ST1 as the ST with the greatest number of SLVs (the primary founder) and links ST1 to its seven SLVs. It then assigns the SLVs of each of these seven SLVs, and identifies ST2 as a SLV of ST17 and links it. Progressing further outwards, the four descendent SLVs of ST2 are identified and linked, and the process continues outwards and links the four descendent SLVs of ST3. This initial assignment of SLVs from the primary founder outwards results in STs that are SLVs of more than one ST being preferentially assigned to the more centrally positioned ST (ST2 in Figure 2).

In the example shown in Figure 2, optimisation identifies ST10 and ST12 as SLVs of ST3 as well as of ST2. The optimisation procedure re-assigns STs to maximise the numbers of SLVs associated with ST3 as this subgroup founder has more SLVs than ST2 and thus is a more likely subgroup founder. ST2 and ST3 each start with four SLVs and after optimisation ST3 ends up with six linked SLVs (STs 10, 12, 13, 14, 15 and 16) and ST2 with two linked SLVs (STs 3 and 11). Optimisation re-organises the arrangement of STs to maximise the numbers of SLVs associated with subgroup founders, closely approximating the sub-groups produced by the original BURST algorithm, but providing complete linkage between all of the STs in the group.

7. Exploring alternative patterns of descent

7.1. Re-assigning the primary founder

There may be cases where you believe that the primary founder predicted by eBURST is wrong, or where there are two possible founders with substantial bootstrap support and you would like to explore how the eBURST diagram would look if the alternative ST was assigned as the primary founder. The ability to re-assign the primary founder (see section 9.8 for details of how to do this) can change the diagram substantially as many STs may be SLVs of both the predicted primary founder and an alternative plausible primary founder, but are preferentially assigned to the primary founder (the true number of SLVs of any ST are shown in the Analysis Panel; section 8.2). Re-assigning the primary founder will now preferentially assign all of these SLVs to the user-selected primary founder. Figure 3 shows an example of an eBURST diagram shown with the primary founder assigned by eBURST (blue) and re-drawn with a user-defined ST as the primary founder (red).   The diagrams represent the unedited output of eBURST v3. In this example, the lineage 3 clonal complex of *N. meningitidis*, we proposed in Feil et al. (2004) that ST303 (arrow) may have been the near extinct founding ST of this large clonal complex, although it is not assigned as the primary founder by eBURST. To explore this further, the diagram is re-drawn with ST303 as the user-defined primary founder, which shows that ST303 has many SLVs and looks a plausible founder. This is also supported by the fact that ST303 has the minimum distance to all other STs in the clonal complex.

Figure 3

Original diagram

ST re-assigned as founder

Re-assigned diagram
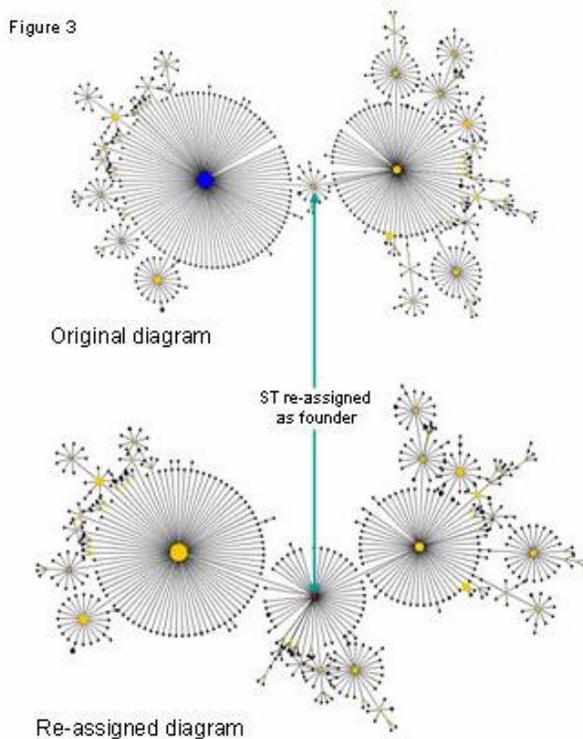
7.2. Using additional data to optimise patterns of descent.

In some cases individual STs could be linked to (descended from) more than one ST and the eBURST diagram provides only one hypothesis for the patterns of descent. eBURST uses a set of rules that attempts to maximise the number of SLVs associated with the primary founder and the subgroup founders and there are many equally plausible minor variants of the arrangement of STs in a complex eBURST diagram. This is somewhat similar to the construction of a phylogenetic tree, where there may be a large number of approximately equally plausible trees, which differ slightly in the branching order, but only one of them is displayed. Additional information can be used in some cases to further explore the patterns of descent implied by an eBURST diagram. For example, one SLV may have gained resistance to antibiotics, and may have diversified to become a subgroup founder, with all of its descendent STs also being antibiotic-resistant. However, there may be another ST that is resistant to the antibiotic but has not been linked by eBURST to the subgroup that includes the antibiotic-resistant STs. The ability to highlight all the SLVs (and DLVs) of this ST (see section 9.7) will show how it relates to the members of the antibiotic-resistant subgroup, and whether it appears to be an independently derived antibiotic-resistant ST, or is an SLV of one of the resistant subgroup and could more plausibly be linked within the subgroup of resistant STs. Similarly, other data (e.g. differences in serotypes within a clonal complex) may suggest equally parsimonious re-arrangements of the linkages between some STs shown in the eBURST diagram that make the patterns of descent more biologically plausible. The ability to show all SLVs (pink) and DLVs (blue) of any selected ST helps to explore alternative equally plausible patterns of descent (Figure 4).

Figure 4

11
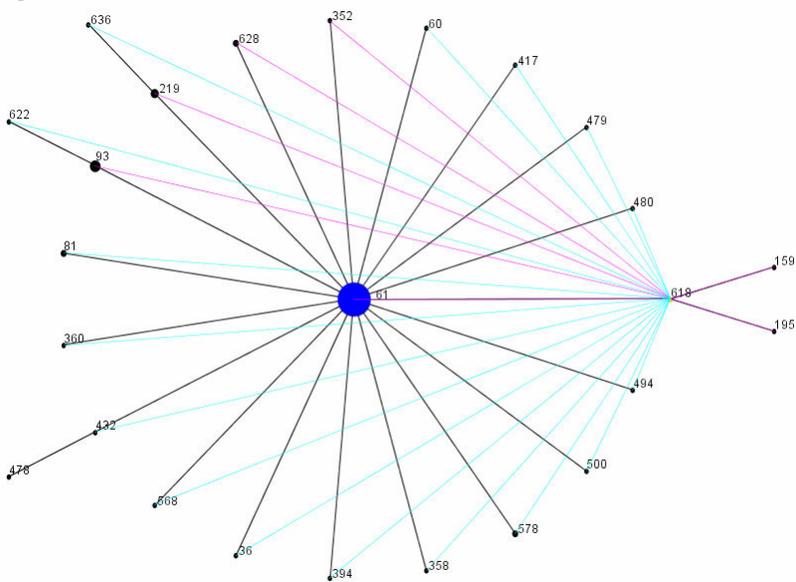
8. Using eBURST

eBURSTv3 can be accessed from the website http://eburst.mlst.net

  Version 3 is implemented as a JAVA Web Start application. On first launching, eBURST will be downloaded and installed on your local machine, and will not need downloading again unless the software has been updated. Web Start allows data to be streamed directly from a website to the locally installed application significantly decreasing the amount of data transferred compared to previous versions.
        This also allows eBURST to be used as a standalone application and details of differing usage methods will be highlighted – details of offline use can be found in section 10.

The interface to the eBURST application is split into three sections –
- The Profiles window.
- The Analysis window.
- The Diagram window

Each of these are accessible via 'TABS' within the interface, and each of the windows contains menu options, a main window and a panel containing context sensitive functions based on which window you are viewing – see Figure 5.


Menus

TABS

Main window

Panel

Formatted: Font: (Default) Arial, 10 pt
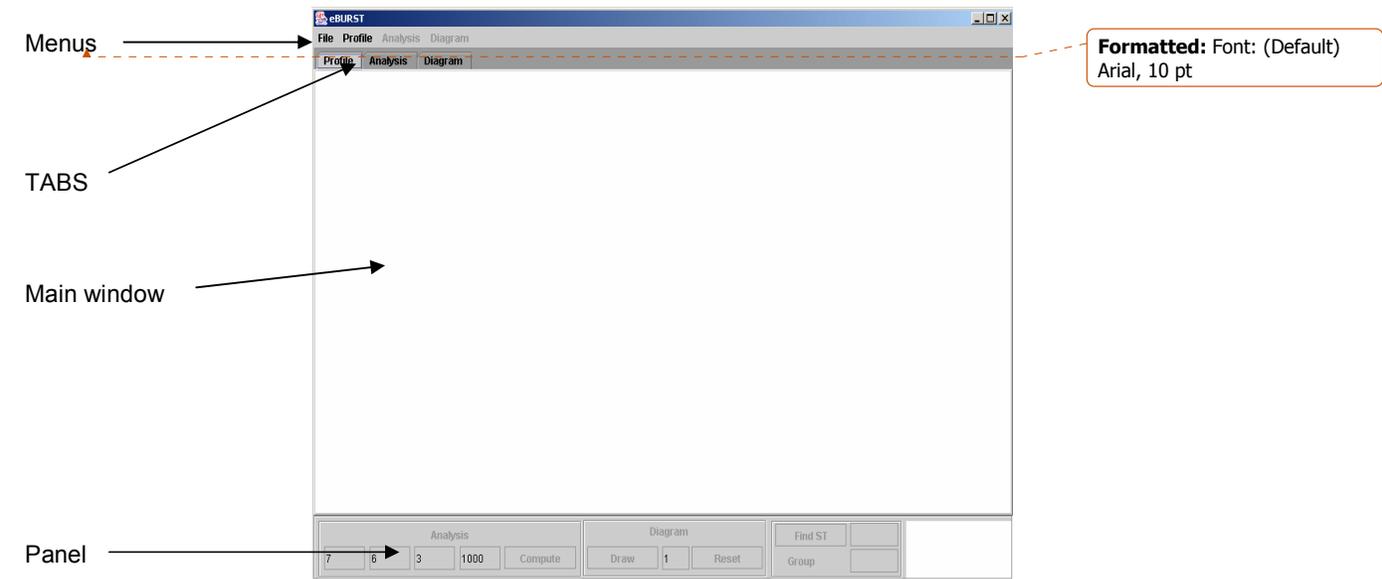
Figure 5 The eBURSTv3 interface

8.1. The Profiles window

The data used by eBURST are a list of STs and their allelic profiles. Data can be entered at the website by 'cut and paste' or can be uploaded and should be tab-delimited (or space-delimited) text, each line having the ST number followed by the allele numbers at the seven MLST loci (in the standard order of loci for the MLST scheme).

An example of correctly formatted input text is shown below:

233	2	4	5	6	7	8	9
88	1	1	1	3	7	23	11
54	12	3	43	32	9	11	3
213	21	22	4	38	5	23	8
54	12	3	43	32	9	11	3

eBURSTv3 is integrated within the mlst.net databases and can be launched after querying the database to examine subsets of a population, or can be launched to view the entire population of a particular species.

The option also exists to combine your own data, added by 'cut and paste' or upload, with all of the data that is in a MLST database. On clicking an eBURST button, the output from the database query, the whole database, or your data is sent to eBURST.

HINT – you may wish to use eBURST to compare your own isolates with all of the isolates of the species in the MLST database, but your MLST data may have alleles and/or STs that are not in the MLST database.   These isolates can still be analysed by eBURST as they can be given temporary allele numbers (we suggest numbers above 1000 to avoid any confusion with the official allele numbers) and temporary ST numbers (we suggest greater than 10,000). Note that these numbers are just to allow you to analyse your data using eBURST and should NEVER be used in any publication (new alleles and STs should be submitted to the curator of the appropriate MLST database for official allele number and ST number assignments).

On submitting the data, eBURST will be launched and the data will appear in the 'Profiles' window. Once the data is loaded it can be sorted by ST number or by allelic profile using the 'Profiles Menu'. Sorting by allelic profile is not optimised and sorts by the allele number at the first locus, then at the second locus, etc.

The 'Open' button in the 'Profiles Menu' allows a file to be uploaded and this option, or 'cut and paste' within the MLST website, allows any MLST data to be analysed by eBURST. The uploaded data will replace any data already in the Profiles window. Uploaded files must be correctly formatted text files, as in the example above.

8.2. The Analysis window

Clicking the 'Analysis' tab produces a blank window and an 'Analysis Panel' with the ability to select the definition used to divide the STs into groups of related STs. The default setting is the sharing of alleles at ≥ 6 of the 7 loci with at least one other ST in the group (see above).

Clicking the 'Compute' button produces a display of the eBURST groups. For each group, the group number is shown, and the total number of isolates and STs in the group, and the predicted primary founder, if one can be assigned. Below this, all the STs in the group are shown, with the number of isolates of each ST in the input data (frequency), the number of STs in the group that differ at a single locus (SLVs), those that differ at two loci (DLVs), and at three loci (TLVs), and those that are more distantly related (satellites).   Some STs may not share alleles at six out of seven loci (or five out of seven loci using the more relaxed group definition) with any other STs in the dataset and are shown at the bottom of the tabular output, below all the groups, as 'singleton STs'.

The average distance of each ST to each other ST in the group is also shown, and is computed from the average number of differences in their allelic profiles. Thus, in a simple eBURST group (e.g. that shown in Table 1), where all STs are SLVs of the primary founder, the average distance of the primary founder will be 1.0 as all other STs differ from the founder at only a single locus.   The ST with the minimum average distance to other STs is likely to be the primary founder and this measure provides another indicator of the likelihood of a ST being the primary founder.

Table 1. Output for an eBURST group.

Group 7: No. Isolates = 20 | No. STs = 9 | Predicted Founder = 354

ST	FREQ	SLV	DLV	TLV	SAT	Average Distance	ST Bootstrap Group	Subgrp
354	12	8	0	0	0	1.0	100%	100%
7	1	2	6	0	0	1.75	0%	0%
339	1	2	6	0	0	1.75	0%	0%
326	1	2	6	0	0	1.75	0%	0%
774	1	2	6	0	0	1.75	0%	0%
310	1	1	7	0	0	1.87	0%	0%
359	1	1	7	0	0	1.87	0%	0%
412	1	1	7	0	0	1.87	0%	0%
528	1	1	7	0	0	1.87	0%	0%

The conditional bootstraps are calculated for each ST being the primary founder of the group and, with the default setting, are displayed for re-samplings from an initial dataset that contains only one example of each ST in the input data (ST Bootstrap in the above table). Bootstrap values for re-samplings that use all isolates in the input data (which will typically include multiple isolates of some STs) can be selected in the 'Analysis Menu'. Bootstrap support is only provided with the default group setting, where all isolates in the group are considered to be members of a single clonal complex.

Bootstrap support is also shown for each ST being a subgroup founder. These conditional bootstraps evaluate the number of times in the resampled datasets each ST is associated with the required number of SLVs to be defined as a subgroup founder. The default value for a subgroup founder is that it has at least two descendent SLVs, but this value can be increased (using the Analysis Panel), with corresponding changes in the level of bootstrap support as a subgroup founder. Thus, bootstrap support for a ST as a subgroup founder may be very high with the default definition but will decrease as the number of STs that are required to define a subgroup increases.

The number of bootstrap re-samplings can be changed from the default value of 1000 in the 'Analysis Panel'. Bootstraps take a few seconds to appear and the progress of the bootstrap calculations can be followed in the 'Analysis Panel'.

Should you wish to know the group that a particular ST falls into, a search box is located in the 'Analysis Panel' allowing you to enter an ST number and click 'Search'. The query ST is shown highlighted in red**.**

8.3. The Diagram window

The 'Diagram' tab takes you to a new blank window and the group number of interest (from the 'Analysis' window) is entered in the 'Diagram Panel' and on clicking the 'Draw' button the initial auto-edited eBURST diagram is produced. Where there is a predicted primary founder this ST is shown as the circle in the centre of the eBURST diagram and is coloured blue. The area of this circle, and of the circles that represent the other STs, is proportional to the abundance of the ST in the input dataset. The numbers are those of the STs.

All of the SLVs of the predicted primary founder are shown arranged in a circle, each linked (connected by a line) to the centrally positioned predicted founder. Unless the group is fairly simple, with a single founder and a set of linked SLVs, the initial auto-edited eBURST diagram will probably need some manual editing before it is saved (see section 9.4).

eBURST is conservative and only shows links between SLVs. It therefore differs from the original BURST in the way it displays the DLVs of the predicted founding ST. BURST showed all of these DLVs in an outer circle, whereas in eBURST, the DLVs of the founding ST are only shown if the linking SLV is present. Thus, if a SLV differs from the predicted founding ST by having allele 14 rather than allele 6 at locus 1, a DLV will only be shown linked to this SLV if it also has allele 14 at locus 1 and a second change at any of the other loci. eBURST therefore suggests pathways by which DLVs may have arisen from a founding ST through known SLVs. Similarly TLVs of the founding ST will be linked only if they connect through an existing DLV.

8.3.1. The default diagram (six out of seven group definition): Using the default definition of a group, all isolates must have the same alleles at six or more of the seven loci with at least one other isolate in the group. Therefore, in the default eBURST diagram, all STs will be linked as a single cluster, which is considered to represent a clonal complex. This stringent default definition of the group means that some STs that may be quite closely related to other members of the group will not be included (e.g. even DLVs of the founding ST will be excluded from the group unless they are SLVs of some other ST in the group). Bootstrapping provides the level of support for the predicted primary founder of the clonal complex and for subgroup founders.

8.3.2. Relaxing the group definition to five out of seven: When the definition of a group is relaxed to five out of seven shared alleles, the output from a large eBURST group will often show several unlinked clusters of STs, and individual unlinked STs (Figure 8). As only links between SLVs are shown, all directly linked STs within a cluster must differ at only one of the seven loci (i.e. they are links between SLVs). If there are two clusters in the eBURST diagram, the fact that the two clusters are not joined into a single cluster implies that no isolate in one cluster is a SLV of any isolate in the other cluster. Some of the isolates in different clusters are likely to be DLVs of each other but eBURST does show links between DLVs (N.B. all of the SLVs and DLVs of any ST can be displayed on the eBURST diagram; see section 9.7). A less conservative approach that allowed links to be drawn between DLVs would almost certainly connect these two clusters, but the validity of the link would be doubtful, as linkages between DLVs are expected to be less robust than those between SLVs. Two clusters within a single group may descend from a common founder, but without additional information the precise evolutionary relationship between them cannot be discerned with confidence. The clusters of linked STs are considered to be clonal complexes. Bootstrapping should be performed using the default group definition to provide the level of support for the predicted primary founder of each of the clonal complexes within a 5/7 eBURST diagram.

Several unlinked clusters can also occur using the original BURST algorithm, but to a lesser extent, as DLVs can be linked without the need for the intermediate SLV, and the unlinked clusters shown using eBURST provide a much more conservative view of the relationships between the STs in the group.

8.3.3. Further relaxation of the group definition: Isolates that share alleles at five of seven loci with at least one other isolate in the group may be quite distantly related and a large group could even include isolates that differ at all loci. We recommend first using the default group definition (6/7) and then, if required, the relaxed definition (5/7). Relaxing the group definition further is not recommended, except to obtain the "population snapshot" (see section 8.5).

8.4. Subgroups and subgroup founders
Each ST that has at least two descendent SLVs (a total of three including the SLV that is its progenitor) will under the default definition be assigned as a subgroup founder and coloured yellow. The number of total SLVs that define a subgroup founder can be changed from this default setting (Analysis Panel) and this is strongly recommended for large groups.

In large MLST databases, large groups will often have a single good candidate for the primary founder, as one ST has many more associated SLVs than the other STs, is the most common ST in the dataset, and has strong bootstrap support. In other cases there may be no predicted primary founder and this is likely to occur with small groups. In some cases there may be two STs that each have a considerable number of SLVs, but one has slightly more SLVs (or the same number of SLVs, but one has more DLVs), and therefore is assigned by eBURST as the primary founder. Due to the preferential assignment rule, STs that are subgroup founders may have considerably more SLVs than are shown in the eBURST diagram. The number of SLVs linked to a subgroup founder in an eBURST diagram will therefore typically be less than the true number of SLVs. The correct number of SLVs of any ST is shown in the tabular output ('Analysis' window). Where one ST has only slightly more SLVs than another possible primary founder, or when the choice of the primary founder is made only on the number of DLVs, the bootstrap support is likely to be unconvincing, with both STs having moderate support as the primary founder. In such cases, either ST may be the primary founder of the group (or the founder of the group may not be in the dataset) and further information may be required to try and understand the patterns of descent within the eBURST group. eBURST v3 allows any group to be re-drawn with any ST selected as the user-defined primary founder (see section 9.8). The user-defined primary founder is coloured red to distinguish it from the founder chosen by the eBURST algorithm (blue).

8.5. Displaying a "Population Snapshot"

The whole input population (e.g., a whole MLST database) can be shown as a single eBURST diagram, to visualise the number and sizes of clusters of linked STs etc. This can be done by setting the group definition in the 'Analysis Panel' from the default value of 6/7 to 0/7. This will put all STs in the whole input data into a single eBURST group, and selecting 'group 1' in the 'Diagram Panel' will display all clusters of STs and unlinked individual STs as a single eBURST diagram – this display is called a "Population Snapshot". The linked clusters within the population snapshot should represent clonal complexes and the primary founders and subgroups founders of these linked clusters are coloured blue and yellow (Figure 9).

8.6. Comparing two datasets

eBURSTv3 provides the capability to compare and differentially highlight two datasets. An initial dataset (REFERENCE) is first loaded and then a second dataset (QUERY) can be loaded and compared to the reference dataset. This is a particularly useful enhancement as it allows comparison of a user dataset (QUERY) with the whole MLST database (REFERENCE) for that species, differentially highlighting those STs that are unique to the user data and those that are also already present in the MLST database. Alternatively, two user datasets can be compared highlighting STs unique to either dataset and those common to both datasets.

Datasets can be uploaded at http://eburst.mlst.net or directly into eBURSTv3 through the File menu.

Two options are offered for comparing datasets at http://eburst.mlst.net.
 A user may -

- ***upload data to compare against an entire MLST database***: allowing a user to upload their own data (QUERY) to compare against a pre-loaded MLST.net database (REFERENCE).
- ***Upload two user datasets*** : allowing a user two upload a REFERENCE dataset and a QUERY dataset

Species-specific mlst.net websites also contain the facility to upload query datasets for comparison against an entire database or a database query allowing you to explore the predicted ancestry of your isolates prior to submission to curators.

The ability to run eBURST on Oxford MLST databases or compare datasets to those in Oxford databases is provided, however, unlike the datasets provided from mlst.net databases these only contain one example of each ST.

After loading both datasets eBURST initially compares the profiles within both to check that there are no identical allelic profiles assigned as different STs or, conversely, no differing allelic profiles assigned as the same ST. Should discrepancies be found, descriptions of the particular profiles are returned in the Profiles Window and you are asked to correct data prior to reloading.

Once consistent data are loaded both datasets are displayed in the profiles window allowing analysis to begin as with a single dataset.   STs in the profiles window are coloured differentially dependant on their membership of the two datasets –

  Black STs found only in the Reference dataset
  Green STs found only in the Query dataset
  Cyan STs found in both the Reference and Query datasets.

When viewing the eBURST diagram of individual groups, or a population snapshot of the two combined datasets, the ST labels also are coloured in this fashion.

To be able to visualise the differences and similarities between the two datasets more clearly, the ST labels can be turned off (From the 'Diagram' menu uncheck 'Show ST labels') and a 'halo' is drawn around the ST circle coloured as follows –

  Green STs found only in the Query dataset

Cyan   STs found in both the Reference and Query datasets..

STs found only in the reference dataset are shown as normal (black).

The border thickness of the halos can be changed through the Diagram menu to produce the optimal visibility of the differential colouring of STs for saving for a publication of powerpoint slide.

For further clarity, the colouring of the primary founders and subgroup founders can be removed (by selecting 'black' in the 'colour options' menu from the 'Diagram' menu.

Figure 6 depicts two datasets, one (REFERENCE) collected at one timepoint against another (QUERY) collected at a later timepoint. It can be seen that there are a number of minor clonal compexes where the SLV's of the predicted founder are only present in the query dataset, collected after the reference dataset.
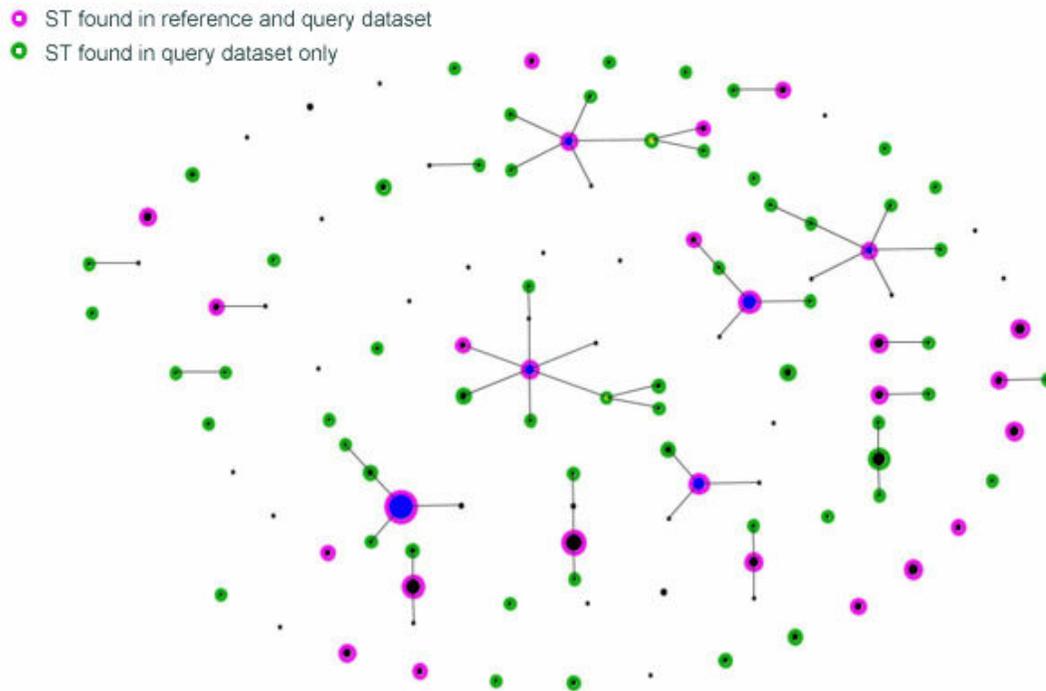

Figure 6. eBURST diagram comparing two datasets.

9. Manipulating and editing the eBURST diagram

9.1. Moving the eBURST diagram: Left clicking on a blank area of the diagram window and moving the mouse allows the diagram to be dragged around the screen. Where there are multiple unlinked clusters, clicking within an individual cluster produces a green rectangle around it and the mouse can be used to drag that individual cluster around the screen. If you have difficulty selecting a clonal complex, you can click on any line between two of the STs in the complex (the line goes green), and then left clicking allows the whole complex to be moved. Clicking on a line also allows pairs or triplets (etc) of STs to be moved together. Clicking in an area outside (or between) the clusters allows you to return to the default where the mouse drags the whole diagram around the screen.

9.2. Horizontal and vertical expansion: Holding down the 'Alt' key, left clicking and moving the mouse allows the diagram to be expanded or contracted in the up-down and left-right dimensions. This can also be used to change the shape of individual selected clonal complexes in diagrams that contain multiple complexes. This is useful as sometimes autoediting produced elliptical diagrams which can be made circular using this editing feature.

9.3. Rotating and zooming: Holding down the shift key, left clicking and moving the mouse allows the diagram to be re-sized and rotated. Individual clonal complexes can be selected and individually resized and rotated. Selecting a subgroup founder by clicking on it (making it red), and then shift-left clicking, allows the whole subgroup to be moved and zoomed. You can always return to the original diagram by clicking on 'draw'. You need to play with this for a while, and being under 30, with experience of PS2 games, or a flight simulator probably helps.

9.4. Manual editing the eBURST diagram: eBURSTv3 automatically optimises the arrangement of STs (auto-editing) and produces eBURST diagrams that need minimal editing. Manual editing features can also be used and where necessary, can convert the initial auto-edited eBURST diagram into an optimal eBURST diagram for use in a publication or a Powerpoint presentation. A new feature of eBURSTv3 is the ability to save a diagram in SVG format. This allows further changes to the diagrams to be undertaken using third party SVG editors (see section 9.5.2)

9.4.1. Moving STs: To move an ST, first click on the circle representing the ST, and a red halo appears around the circle to show it has been selected (you may have to move the mouse around slightly to find exacting the right place to click). The selected ST can then be dragged to the required position by moving the mouse. Where there are linked STs that need to be re-positioned, dragging an external ST will move only that ST, whereas dragging an internal ST will move the whole group of linked STs. This prevents the linkages between STs being inadvertently altered while editing the diagram. All of the STs in a subgroup can now be moved together (see section 9.3).

9.4.2. Highlighting connections between STs: Clicking on a line between STs highlights the connecting line in green.

9.4.4. Text size and font of ST labels: In the 'Diagrams' menu the Colour Options opens the Properties Panel and the text style (font and font size etc) of ST labels can be changed.

9.4.5. Colour Key: A key displaying the default colour coding for elements of the diagram can be viewed to aid interpretation. Should any changes to the default colours be made these will not however be reflected in the key.

9.4.6. Changing default colours and line widths: You can change all colours from their defaults through the 'colour options' section of the diagram menu.

9.4.7. Show ST labels: By default, this option is checked, unchecking removes all text from an eBURST diagram. Should you be comparing two datasets coloured 'halos' will appear surrounding STs to show in which dataset they are represented (see section 8.6)

9.5. Saving and printing output.

9.5.1. Saving text output: The textual output of eBURST groups and bootstraps can be saved to a file through the 'Save' option of the File menu, when the Analysis TAB is selected.

9.5.2. Saving eBURSTv3 diagrams:

eBURST diagrams can be saved in the following formats –
- Jpeg
- PNG
- RAW
- BMP
- SVG

Of particular use is the ability to save in the Scaleable Vector Graphic (SVG) format. SVG is an XML representation of vector images and as such allows the eBURST diagram to be re-edited using 3rd party applications.

To view SVG documents you need an SVG viewer which comes in the form of a plug-in for web browsers. The plug-in is freely available and can be downloaded from http://www.adobe.com/svg/

To further edit eBURST diagrams saved in SVG format we recommend using either the commercial software Adobe Illustrator (www.adobe.com), or freeware editors such as the W3C recommended Amaye – (www.w3.org/Amaya/) or Inkscape (www.inkscape.org).

9.5.3. Printing eBURST diagrams and text: eBURST diagrams can be printed through the File menu 'print' option.

9.6. Obtaining more information about groups or STs: Further information can be obtained about the STs from within an eBURST diagram. The ST of interest should be selected by clicking on it until it becomes red. Then selecting the 'database' option in the 'Diagram menu' will open a new panel with the selected ST number already entered. Isolates of the selected ST within the MLST database are returned as the default. Ticking the '+ SLVs' box returns details of the selected ST and all of its SLVs. Obviously linking from the eBURST diagram to a MLST database only works for data submitted to eBURST from an MLST database

9.7. Highlighting all SLVs and DLVs of a ST: You can view all SLV links or all DLV links of a particular ST or entire clonal complex by using the checkboxes located in the Diagram Panel. After checking either SLV, DLV or both, select an ST by clicking it (a red halo appears around the ST) The diagram now displays lines from the selected ST to all of its SLVs (pink lines) and/or DLVs (blue lines). You can remove the lines by unchecking the boxes. Several STs can be selected in turn to show the SLVs and DLVs of multiple STs. Shift and clicking within a clonal complex shows all of the SLVs and DLVs of all of the STs within the clonal complexes. Shift and clicking on an empty area of the diagram shows all of the SLV and DLV links of all of the STs.
Hint – you can always remove all the lines if you press 'draw' to refresh the diagram.

9.8. Re-assigning the primary founder: From the Diagram TAB, click/highlight the ST you wish to assign as the primary founder and then click 'Founder' in the 'Diagram menu'. The new auto-edited eBURST diagram is drawn, with the user-defined primary founder' in red, to show that you, rather than eBURST, have made this assignment of the primary founder. Note that the number of SLVs linked to the selected primary founder may well increase considerably (compared to the original eBURST diagram) as all its SLVs will now be preferentially assigned to the new 'founder'.

10. Using eBURSTv3 standalone

10.1. Launching eBURSTv3 standalone: When eBURSTv3 is initially downloaded and installed you will be asked whether you would like to install a shortcut icon on your desktop (Windows) - clicking this will launch eBURST as a standalone application. On Mac OS X after using eBURST more than three times you will be asked whether you would like to treat eBURST as an application and relevant shortcuts will be installed.

10.2. Loading data: On startup a blank Profiles window will be displayed allowing you to load your data through the Profiles menu. To load a single dataset select 'Open Ref' and browse to your correctly formatted text file.   To load a query dataset to compare against your reference dataset select 'Open Query' after your reference data is loaded. Once data are loaded you may progress with analysis as above (See section 8). Prior to loading subsequent datasets it is advisable to clear all data from within eBURST by selecting the 'Clear Data' option from the Profile menu.

10.3. Changing the number of loci: You may wish to analyse data that contains a differing number of loci to the mlst default of 7. While this can be achieved by uploading data to the eBURST website you can also change the number of loci by selecting 'Edit Loci Number' in the Analysis menu

10.4. Differences when using standalone eBURSTv3: Database connectivity is disabled by default for standalone eBURST useage as data of different origins can be analysed without a corresponding database being available. Should you with to utilise the database functionality within eBURST we recommend that you launce BURST from within the organism-specific mlst website or from the 'MLST datasets' section at http://eburst.mlst.net

11. Software requirements

11.1. eBURSTv3 application eBURSTv3 is a JAVA Webstart application and requires that you have a recent version of the JAVA runtime environment (versioin 1.4 or greater) which includes webstart installed.

Note that the Java permissions file required in eBURSTv2 is no longer required in v3.
JAVA can be downloaded and installed from http://www.java.com

11.2. http://eburst.mlst.net To access the full functionality of the website , please make sure you have a Javascript enabled Web Browser

eBURSTv3 has been tested on Internet Explorer 6.0, Mozilla Firefoz 1.1, Safari (Mac OS X) and Opera 8.1.

12. **Further examples of eBURST.**

Figure 7A. A simple auto-edited eBURST group (default group definition) from *S. pneumoniae* with a central founder (ST199; bootstrap confidence = 100%), and 12 linked SLVs of the founder. Two of the SLVs of the founder have diversified to produce DLVs. The size of the circles shows that the founder is also the most prevalent ST in the group.
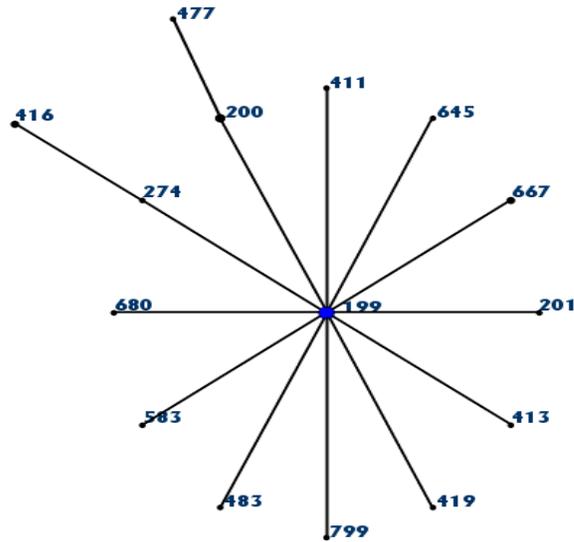

Figure 7B. Using the more relaxed group definition of 5 out of 7 shared alleles some additional unlinked STs are included in the group, some of which are DLVs of the predicted founder.
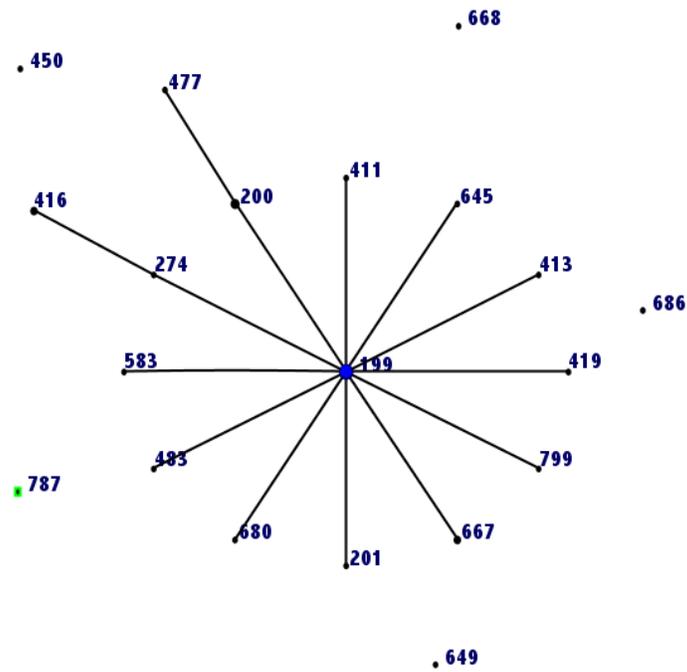
Figure 8. The auto-edited eBURST diagram produced using the 5/7 group definition shows two unlinked clonal complexes and two unlinked STs.   The option to display all SLV and DLV links has been used to show that these two clonal complexes are in the same group using this definition as ST543 has DLV links to both of the clonal complexes (shown in blue).
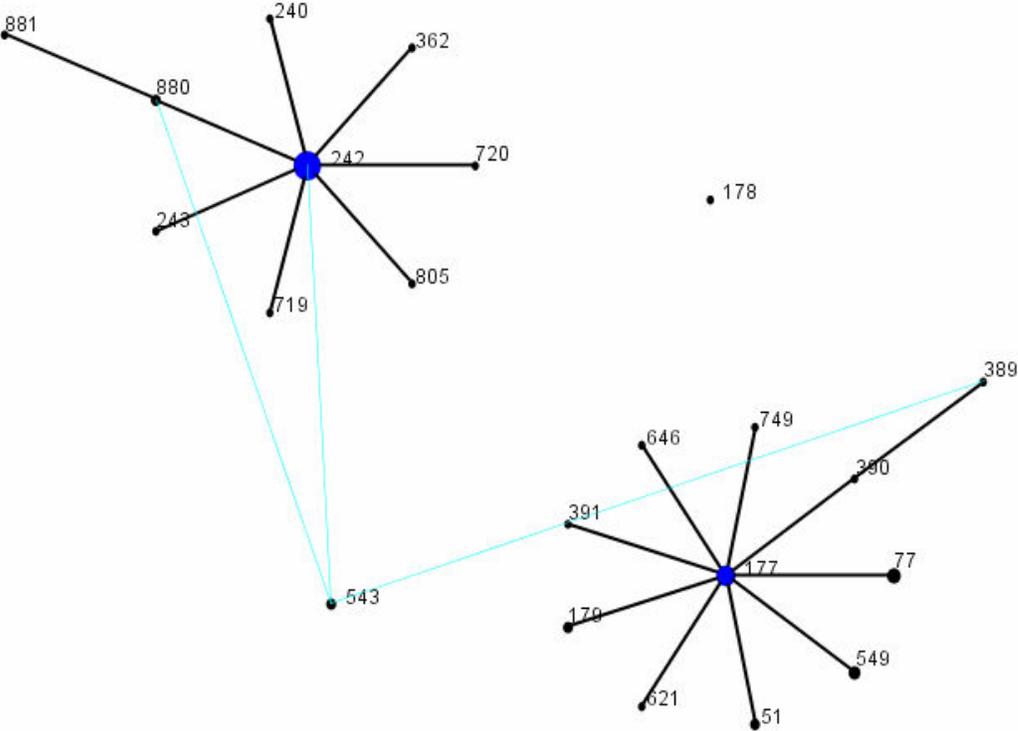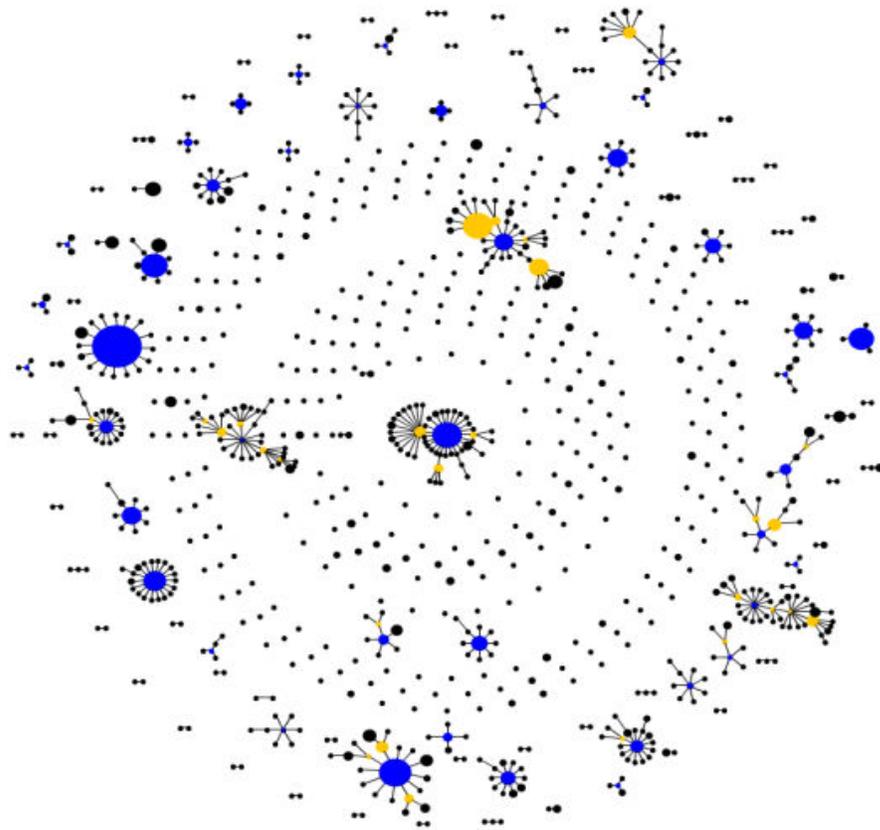
Figure 9. A 'Population Snapshot' showing the clusters of linked STs and unlinked STs in the whole *S. pneumoniae* MLST database (1508 isolates). ST labels have been removed and the autoedited diagram has been optimised by manual edited.