Guideline for physical map assembly

Simone Scalabrin[1], Jan Bartos[2], Melanie Febrer[3], Daniela Schulte[4] and Etienne Paux[5]*

[1] Applied Genomics Institute, Udine, Italy
[2] Institute of Experimental Botany, Olomouc, Czech Republic
[3] John Innes Centre, Norwich, UK
[4] IPK, Gatersleben, Germany
[5] INRA Genetics, Diversity & Ecophysiology of Cereals, Clermont-Ferrand, France
* to whom correspondence should be addressed (etienne.paux@clermont.inra.fr)

1. BAC naming convention:

BAC naming conventions are as described at http://www.wheatgenome.org/pdf/Triticeae_Annotation_Group_Report_2007.pdf.

Example:
The international name is TaaCsp3DLhA for a single library (here, for a clone from the 3DL BAC library).
Note: CNRGV has used accidently Tae as a prefix instead of Taa, as defined for the international convention. Unfortunately, that "typo" was propagated also at IGA, therefore names of barcodes, plates, and fingerprints contain Tae instead of Taa as a prefix for all the BAC libraries from the TriticeaeGenome project. Though, it is not a big problem, as explained below.

Thus, in the example, TaaCsp3DLhA, the nomenclature means:
- Taa: Triticum aestivum subspecies aestivum
- Csp: Chinese Spring
- 3DL: 3D is the chromosome, L is the arm. There are two possible arms, L and S, standing for Long and Short. Note that chromosome 3B was fingerprinted entirely and is not split in the two arms.
- h: h stands for the HindIII enzyme used for the library construction (it can be also e for EcoRI or b for BamHI, however these two enzymes are still not used for library construction)
- A: is the library code. Normally it is A, standing for first library. It is possible to have B, for a second library. The only chromosome with two libraries, at the moment, is 3B.

Each clone is identified by the library name, followed by the symbol "_" (used as separator) and four digits identifying the plate number. E.g. plate number 23 of library TaeCsp3DLhA is TaeCsp3DLhA_0023 (the four digits are padded with zeros if the plate number contains less than 4 digits). If more than a library is needed for a chromosome then plate numbers are progressive in the libraries, therefore, the plate number is already sufficient to uniquely identify a plate inside a library (e.g. There was only one plate 16 for library A and B of chromosome 3B).

Inside each plate a clone is identified by its well position. E.g. clone A01 of plate 23 of library TaeCsp3DLhA_0023 is labelled as TaeCsp3DLhA_0023A01.

Clone identifiers are used when fingerprints are produced (fsa files). E.g. the corresponding fsa file for the above described clone is TaeCsp3DLhA_0023A01.fsa
NB: A fsa file is the data produced from the AB3730 sequencer as representation of the fingerprint.

2. Data flow

Major steps

- Fsa files are converted to text table data with GeneMapper
- Text table data are cleaned from background and are converted into FPC format (.sizes).
- Cleaned data are processed for contaminant removal (format is maintained as .sizes).
- Clones are renamed to fit the FPC limitations
- Sizes are given in input to FPC for physical map assembly.

2.1 Converting .fsa files to text tables using GeneMapper

The first step is to convert data from binary (.fsa) to text (normally .txt). This is done with GeneMapper. Assuming that fingerprinting was done using a size standard LIZ GS500-250, the parameters used to export data from GeneMapper are:

- Peak Detection:
  - Peak Amplitude Thresholds:
    - B: 10
    - G: 10
    - R: 10
    - Y: 10
    - O: 50
  - Min. Peak Half Width: 2 pts
  - Polynomial Degree: 7
  - Peak Window Size: 15 pts
- Allele Number
  - Max expected alleles: 4000

Fsa files can be visualized and converted with different programs, such as GenoProfiler and FPMiner. Nevertheless, we suggest using data exported from GeneMapper which result in more accurate peak calling. Moreover FPMiner is not freely available.

2.2 Cleaning fingerprints using FPB

Text table data are processed using FPB (http://www.appliedgenomics.org/FPB/; http://www.biomedcentral.com/1471-2105/10/127).
FPB performs the following tasks.

1) Removal of all bands being out of the range of 50 to 500 bp.
2) Discard all clones having less than 40 bands or more than 250 bands, as being either badly fingerprinted or potential chimerical clones.
3) Cleaning of fingerprints from background considering peak amplitude: an iterative procedure taking into account high peaks (putative true peaks) and low peaks (putative background peaks) is run to find a threshold below which peaks are rejected. The rationale behind the procedure is that true and background peaks are normally separated by a gap (region with few peaks) and on the distribution of peaks at different amplitudes. Details are presented in the article cited above.
4) Removal of vector bands from fingerprints:for BAC libraries constructed using the pIndigoBAC-5 vector, fingerprint digestion of this vector results in two “red” bands (XhoI restricted fragments) of 161 and 375 base pairs. Analysis of fingerprints revealed a slight deviation from the *in silico* sequence digestion values, leading to fingerprint vector band sizes of 157.11 and 371.57 bp. These bands are removed from fingerprints as they do not originate from the wheat genomic DNA insert and can therefore result in an overestimation of the overlap between clones.
5) Band sizes conversion to integer values, as FPC cannot handle decimals.

For example, a 302.96 bp band should be converted into an integer number. Rounding to 303 bp is possible but results in a loss of information. To avoid this loss, band sizes have to be multiplied. The limit value for a band to be considered in FPC is 65535. Thus, the multiplication factor is 30.

An extra complication arises from the fact that SNaPshot fingerprints are composed of four dyes and FPC only handles numerical data. As a consequence, two bands labelled with different dyes but having the same size are considered identical by FPC. Thus an offset is used for each dye to remap identical bands from different dyes to different numbers: 0, 15000, 30000, and 45000 to blue, green, yellow, and red, respectively.

Combining multiply factor and offsets results in the following band ranges:

- Blue: 1500-15000
- Green: 16500-30000
- Yellow: 31500-45000
- Red: 46500-60000.

6) Exporting processed fingerprints to .sizes files that are compatible with Genoprofiler and FPC.

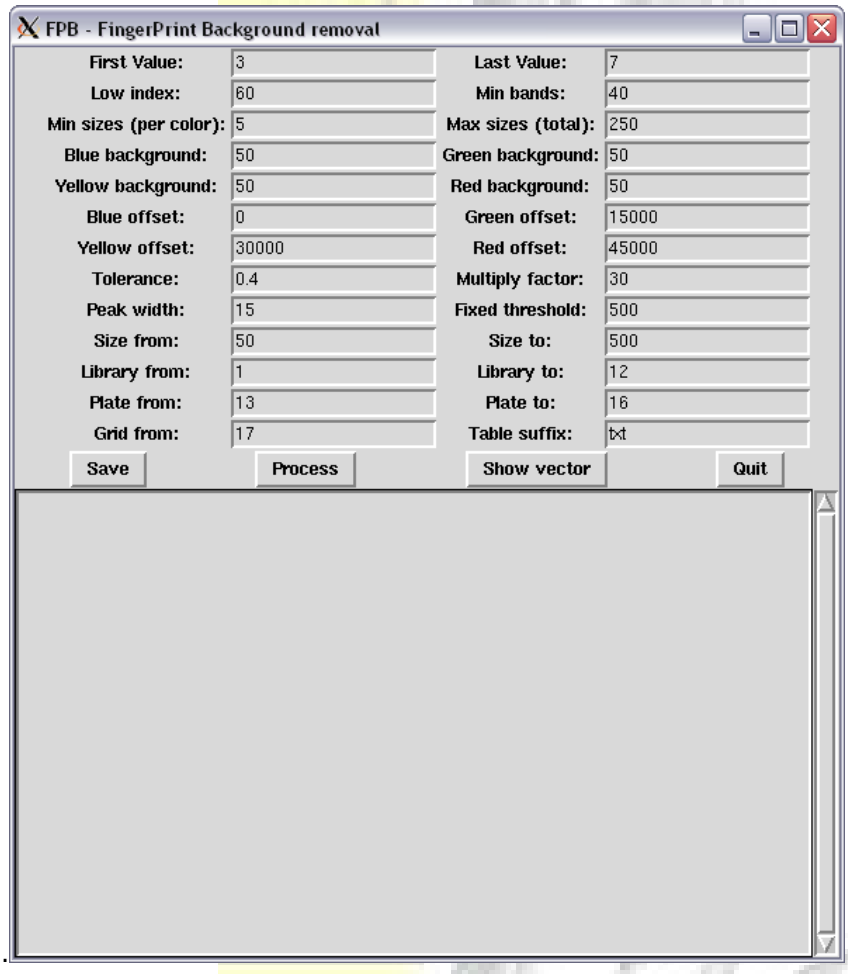


Figure 1: FPB interface with settings

2.3 Editing fingerprints using Genoprofiler: clone renaming and contamination removal

The editing is done using Genoprofiler (http://wheat.pw.usda.gov/PhysicalMapping/tools/genoprofiler/genoprofiler.html).

(Genoprofiler is a bit tricky sometimes, probably due to the Java machine behind it, and it does not analyse data or save them as expected. Therefore, when you run it, be sure it produced/saved what you expected)

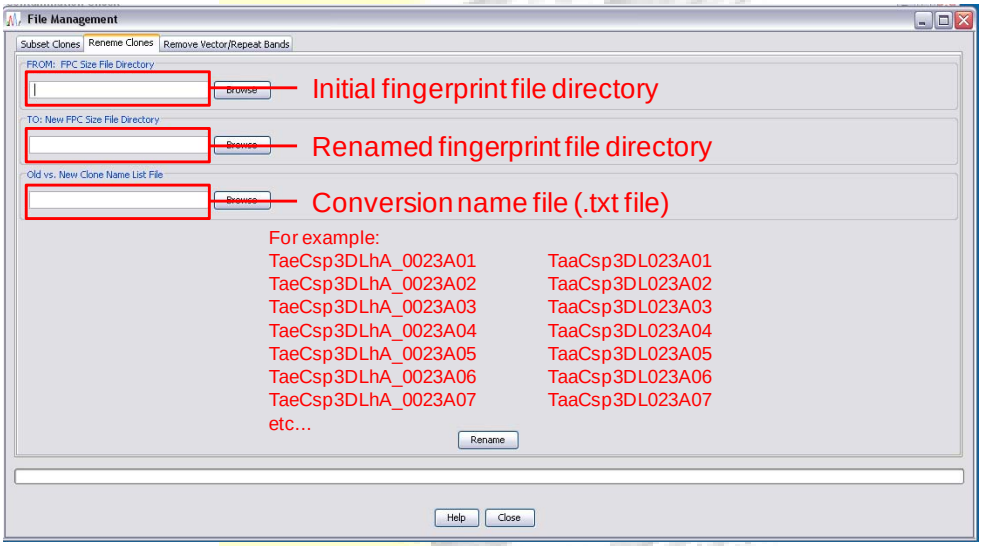
- **Step 1**: clones have to be renamed prior to being used in FPC.

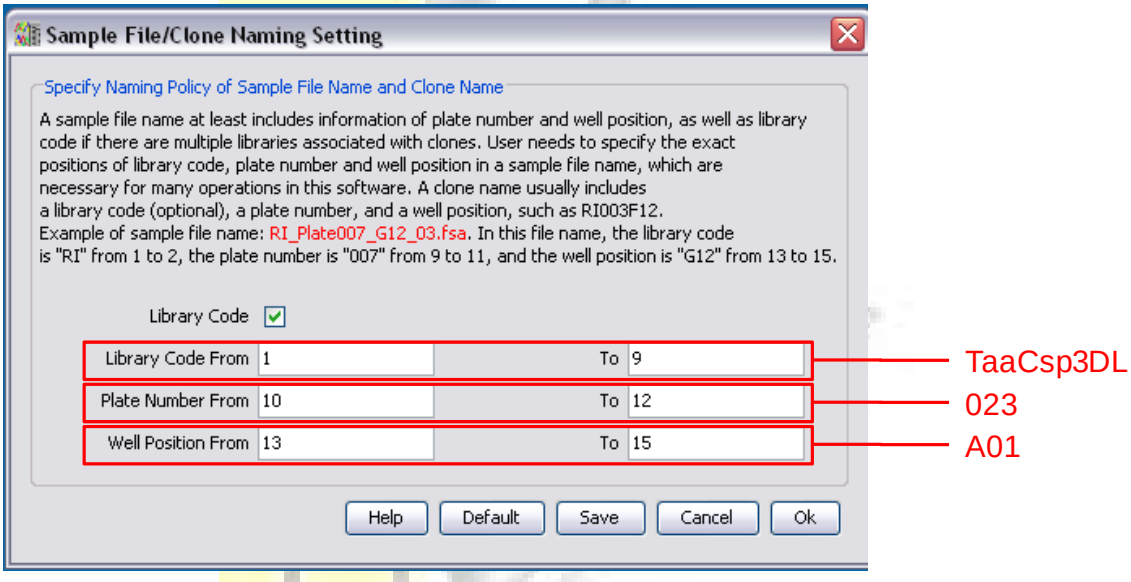
Indeed, BAC clones are named according to the international BAC nomenclature adopted by the IWGSC (http://www.wheatgenome.org/pdf/Triticeae_Annotation_Group_Report_2007.pdf) as reported in the first section of this document.

For example: TaaCsp3DLhA_0023A01.
This nomenclature contains 19 digits. However, FPC cannot deal with clone names longer than 15 characters. Thus, BAC names have to be shortened to 15 digits, with a clear and informative name to avoid any problem.

Below is a proposed simplified coding:
TaaCsp3DLhA_0023A01 → TaaCsp3DL023A01.

This renaming can be done using the 'Rename Clones' function in Genoprofiler.



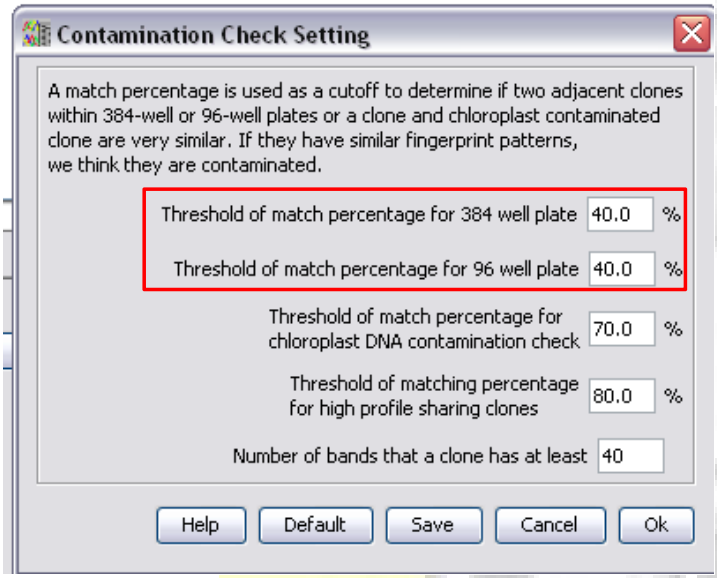


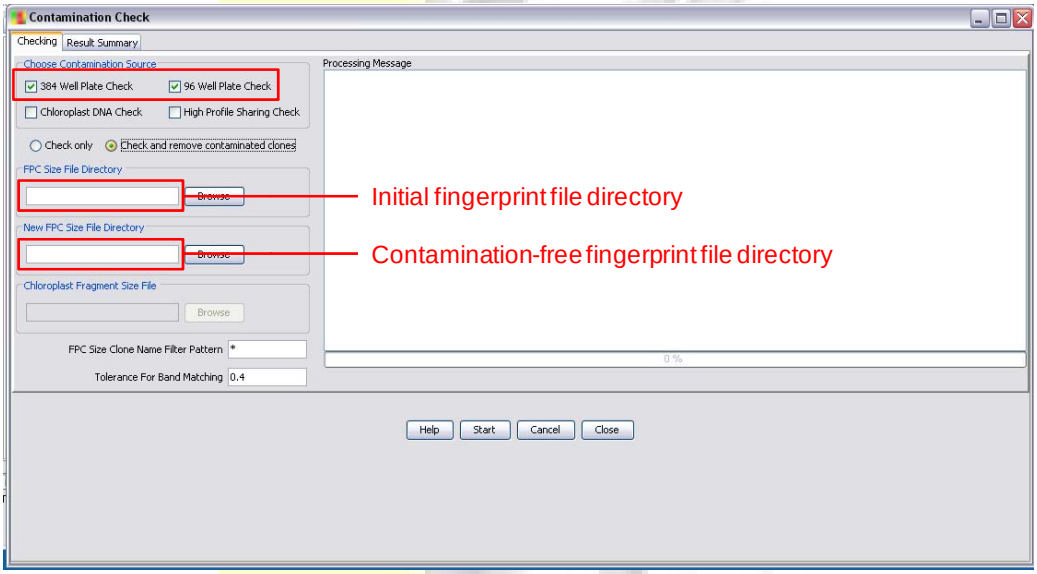
Step 2: Once the setting is done, fingerprints are screened for the presence of 96-well and 384-well contaminations.

(NB: Chloroplast and mitochondrial contamination does not need to be run since the libraries produced in Czech Republic are organellar-DNA free.)

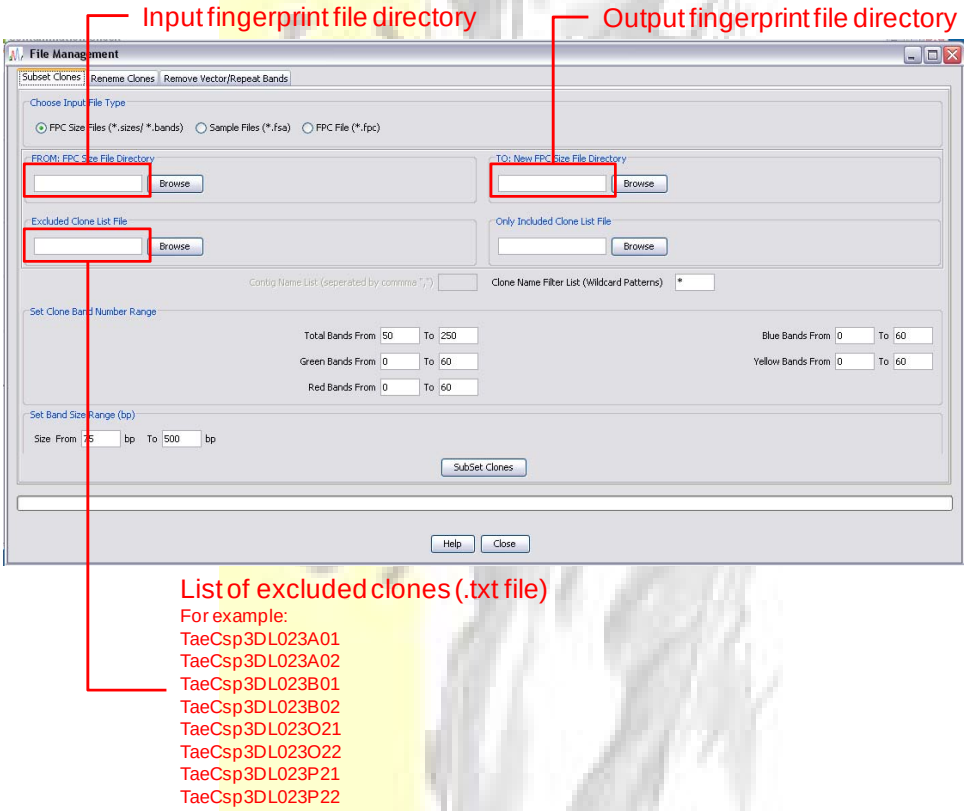
The main parameter here is the percentage of shared bands to consider that two adjacent clones are contaminated. The percentage depends on the number of contamination to be detected. Bidirectional contamination (DNA from one well is added to an adjacent well) and unidirectional contaminations (DNA from one well to an adjacent empty well is added) are the two possibilities; bidirectional contaminations need a lower percentage of shared bands, e.g. if the two adjacent clones A and B contain the same number of bands and clone A contaminated clone B then the number of shared bands to be tested is 50%. However, to cope with the "real life", we suggest using 40% of identity to perform this analysis.

Another important parameter is the tolerance to use to detect overlap of adjacent clones. As already discussed above this parameter should be 0.4 (no need to multiply it, instead the correct multiplier factor, 30, should be provided to Genoprofiler).





If control clones have been added to the plates, they have to be removed.



3. Automated contig assembly

Fingerprinting data (.sizes files) are assembled into contigs with FPC.
The assembly is based on the Sulston formula:

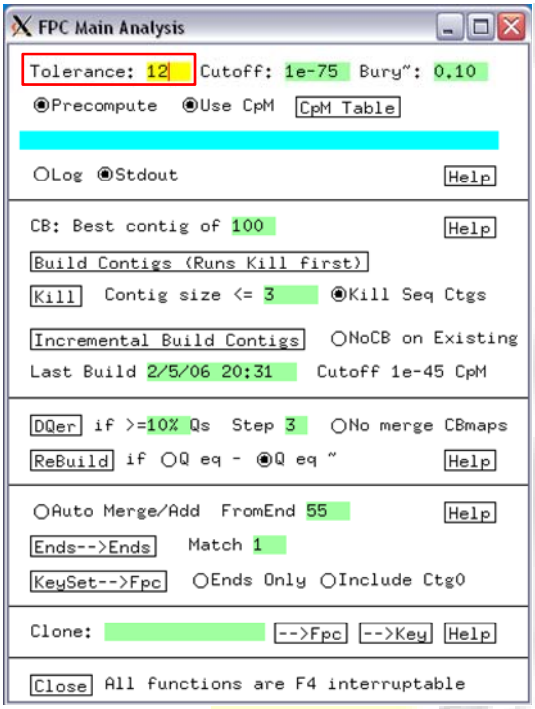
$$\sum_{m=M}^{nL}\left[\binom{nL}{m}\left((1-p)^{m}p^{nL-m}\right)\right]$$

where p = (1 – b)nH, b = 2t/gellen, t is the tolerance, gellen is the number of possible values for bands, nL and nH are the minimum and maximum number of bands for the two clones (nL<nH), and M is the number of shared bands.

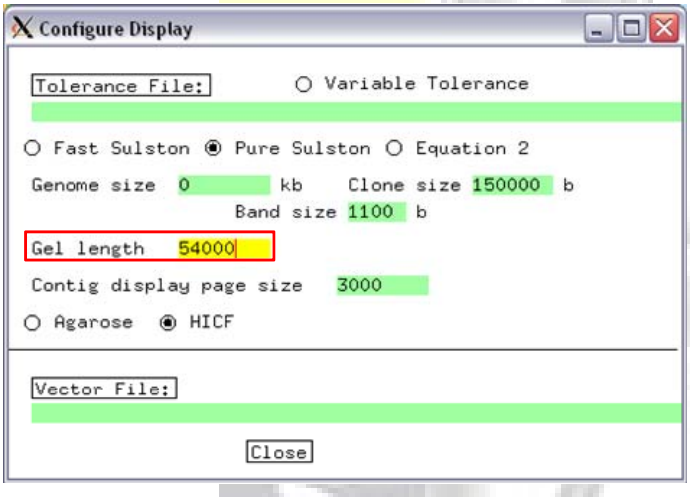
Thus, the main fixed parameters of the Sulston formula are the tolerance (t) and the gel length (gellen). These parameters have to be set up prior to starting any assembly.

- Tolerance determines how close two bands must be to consider them as the same band. Since the tolerance is used in the equation, it is desirable to set it at the beginning of your analysis and never change it; NB: a change requires reassembly of the entire database!!!.

A tolerance of 0.4 bp has already been shown to be well adapted to SNaPshot by Luo et al 2003 and this was confirmed at IGA based on vector bands. Considering the multiply factor used for band sizes, the tolerance has to be set up to 12 (0.4 x 30).



- Gel length (gellen) is the number of possible values for one band. The range for each single dye band is 13500 (1500-15000 for blue bands, 16500-30000 for green bands, 31500-45000 for yellow bands and 46500-60000 for red bands). Thus, the gellen is 54000 (13500 x 4).



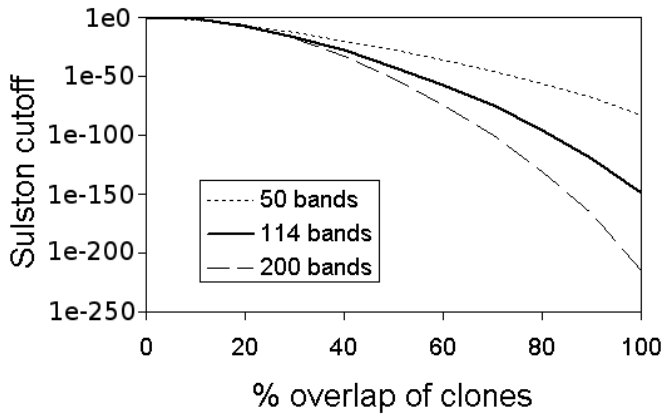
Other parameters that have to be set up are:

- The formula used: Pure Sulston formula with a fixed tolerance.
- Band size: Based on the sequencing of 152 BAC clones, the band length was estimated to be roughly equal to 1.1 kb (R^2 = 0.644), which is consistent with other estimates based on pulse field gel electrophoresis. Remember, this is not the size of a real band but the size of a virtual band produced for FPC. Considering the fact that fingerprinted bands are not real and that bands tend to be slightly bigger than as computed above we suggest using 1150 bp as size for a virtual band.
- Fingerprinting method: HICF.

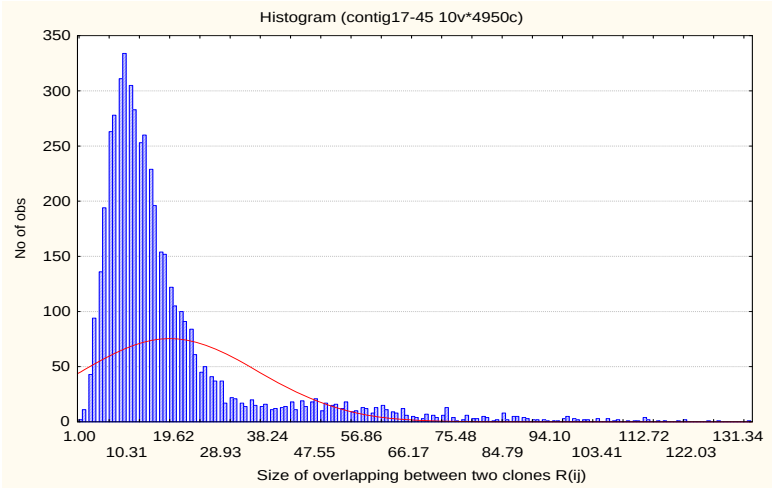
Once the parameters are set up, the initial contig assembly can be run.

The Sulston formula calculates the probability for two clones to be overlapping based on the number of shared bands. The resulting score is compared to a given threshold called cutoff

that discriminates between true overlaps (below the cutoff) and false overlaps resulting from randomly shared bands (above the cutoff).
The cutoff depends on the genome complexity (repetitive nature...), the number of clones in the library. Below is a schematic representation of Sulston scores as a function of clones overlap.

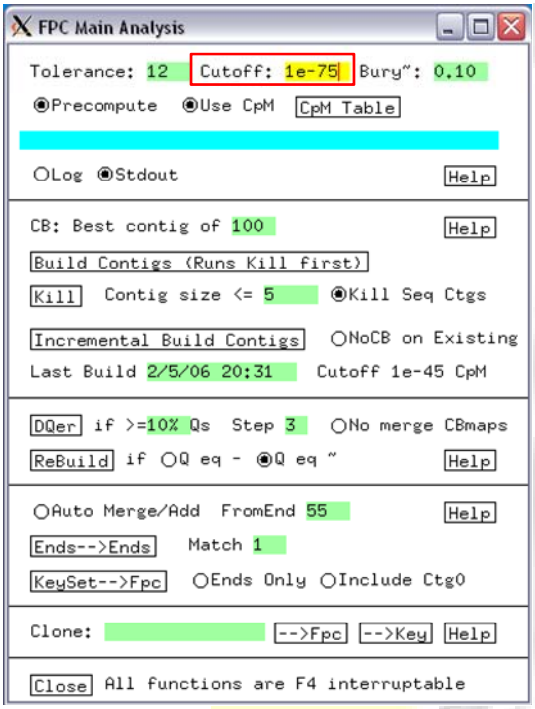


And here is a distribution of randomly shared bands between two non overlapping clones:

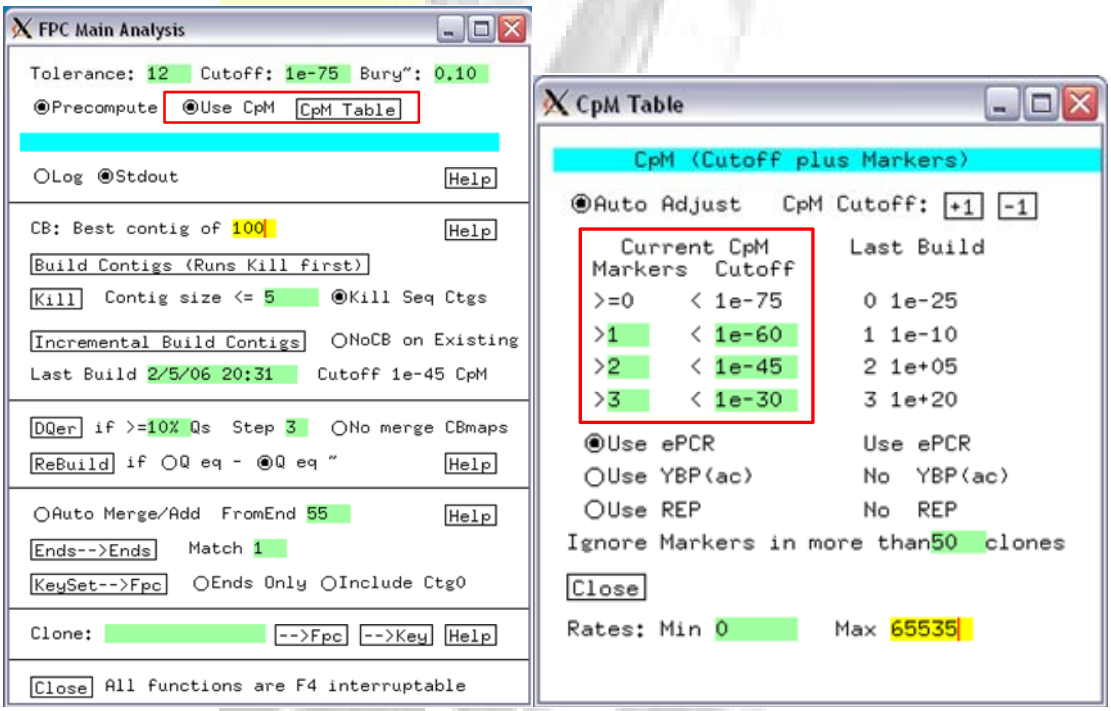


This clearly shows that a significant fraction of clones can randomly share up to 30 bands, corresponding to ~25% of their size and that the probability of false overlap is close to zero with more than 75 bands.

For this reason, **a high stringency should be used to avoid the assembly of chimeric contigs because of random overlaps. The initial cutoff has to be set up to 1e-75.**

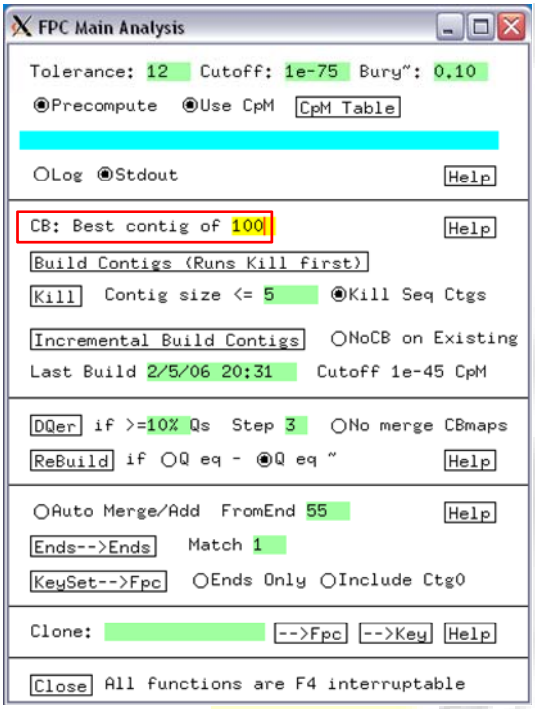


CpM Table should be used to define lower cutoffs based on the number of shared markers between two clones, pending that marker data are highly reliable.

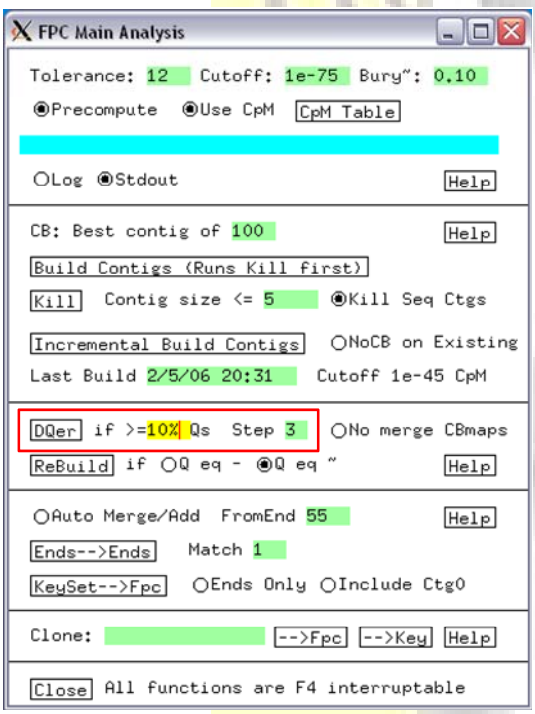


Another parameter to be set up is <u>the number of iterations</u> FPC has to run to find the best solution.

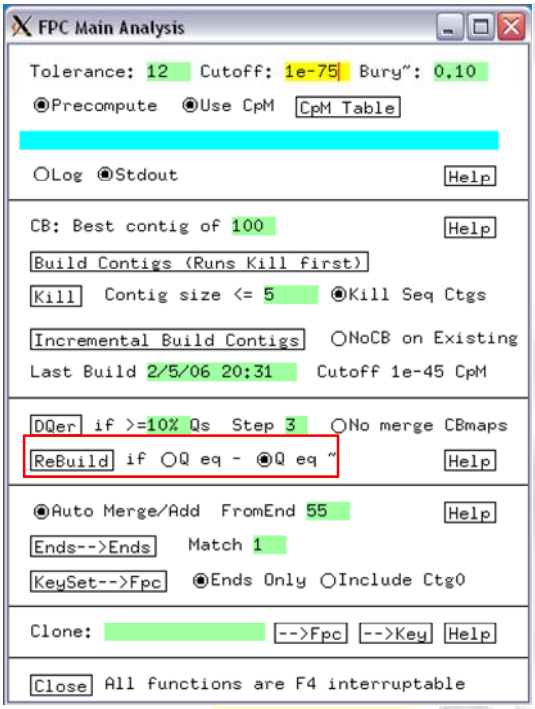
Here, we propose to use 100: FPC will build 100 assemblies of the same contig and will select the best one.



- A quality criterion for the contigs is the number of Q clones. Qs are defined as clones for which the CB algorithm cannot order at least 50% of the bands in the CB map. If there are many Q clones in a contig, it is likely to be chimeric. Thus they have to be removed from contigs. This can be done by rerunning the CB algorithm using more stringent cutoffs.
DQer should be run for contigs having more than 10%Q clones. The step 3 parameter defines the increasing value of the cutoff: -3 at each step, up to 3 times (i.e. 1e-78, 1e-81 and 1e-84).



- After each DQing step, contigs that have been modified by the DQer should be rebuilt to compute the exact number of Qs. This is done using the “ReBuild if Qs eq ~” function with the lowest stringency used so far (at this step, 1e-75).



This "DQer & ReBuild" process should be run until there is no more contigs with more than 10% Qs. If there are still some after the first round, a second round should be done at a higher stringency corresponding to the highest stringency previously used by the DQer.

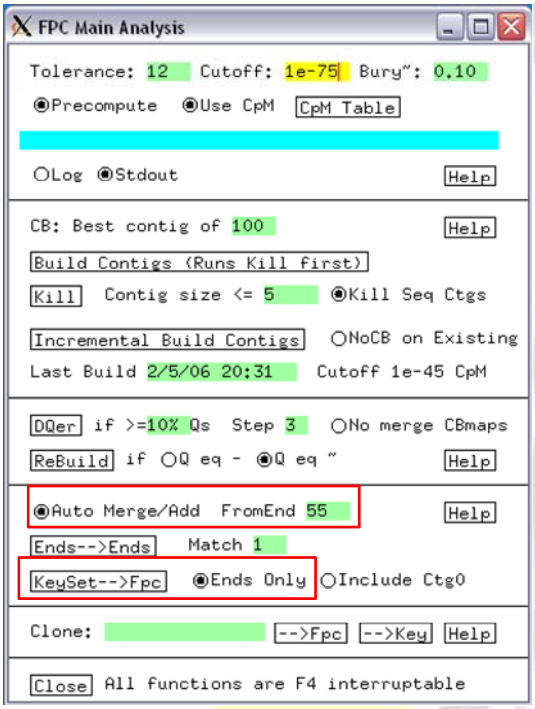
This first process results in a highly reliable assembly with robust contigs that can be confidently used as a backbone for the next steps. These correspond to a series of iterative processes:

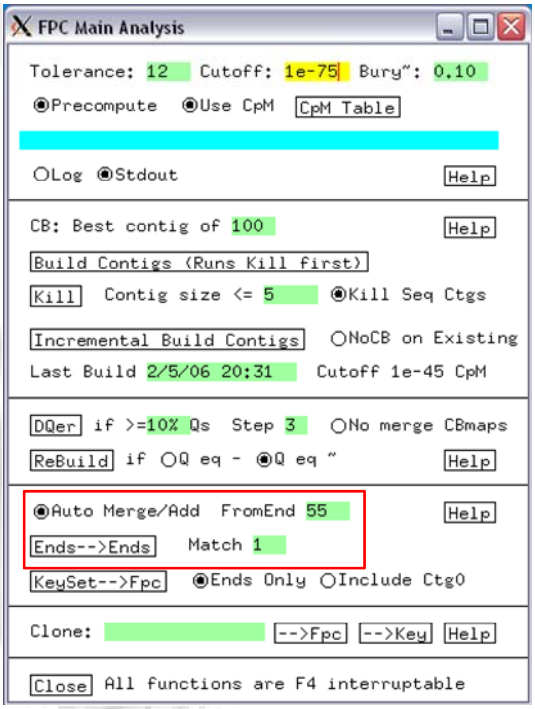
- Single-to-end merging,
- End-to-end merging,
- ReBuild contigs and
- DQer.

At each step, the cutoff has to be increased to relax the stringency from 1e-75 to 1e-45 (*i.e.* 1e-70, 1e-65, 1e-60, 1e-55, 1e-50 and 1e-45).

The <u>single-to-end and end-to-end functions</u> aim at adding singletons to the end of the previously built contigs and at merging contigs, respectively. Two parameters should be set up:

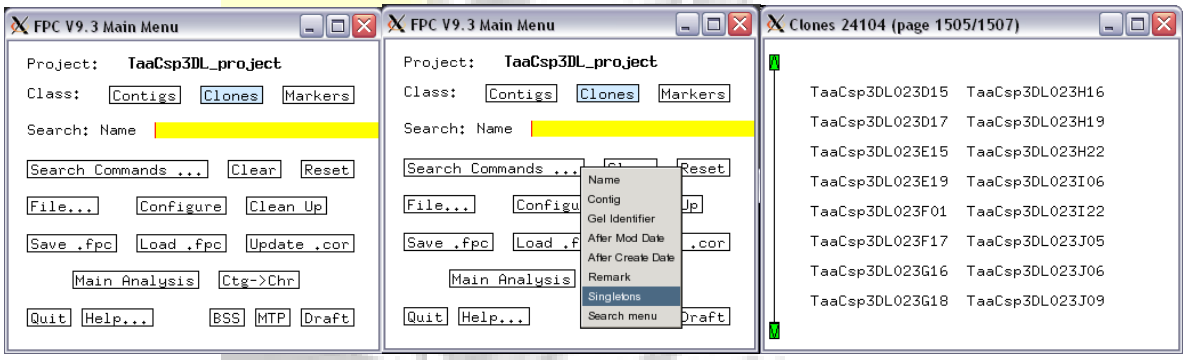
- **Auto Merge: ON** to allow automatic merging
- **FromEnd**: tells how close to the contig end a clone must be in order to count as an end-clone. Its units are CB units and typically it is set to 50% of the number of bands in an average clone. Thus **usually between 50 and 60**.
- **Match**: corresponds to the number of reciprocal matches required to perform merging. **At high stringency (<1e-45), it should be set up to 1**, meaning that only one clone from each contig end is needed.





- There is no direct function to add singletons to contigs in FPC main analysis window. Singletons should be selected in KeySet before they are added to contigs. <u>To add singletons to KeySet:</u>
  - Select search class Clones in FPC Main Menu.
  - Click right Search Commands …
  - Select Singletons

A new window listing all singletons should appear, which corresponds to the KeySet. Now, singletons can be added to contigs using KeySet → Fpc function (see above).



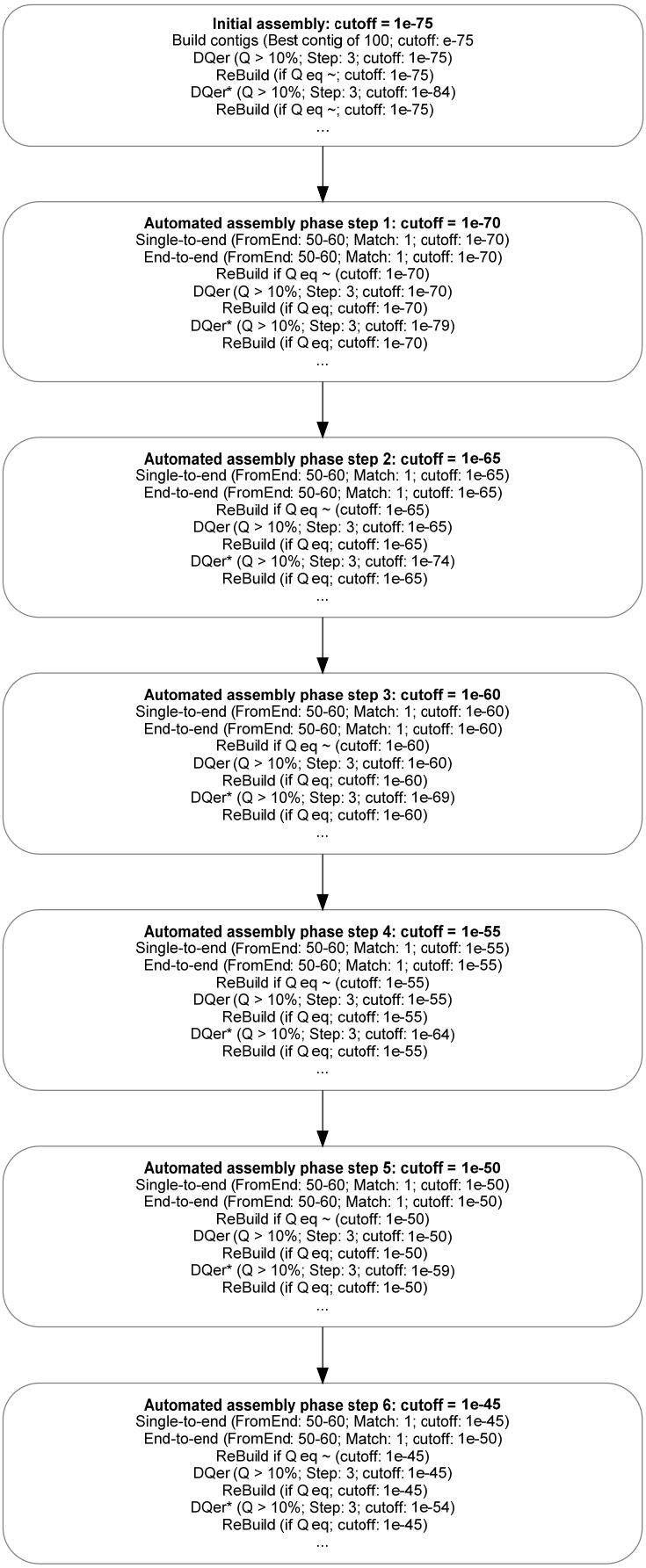
The DQer & ReBuild steps should be run as explained previously.

<u>After the above procedure is completed a draft assembly is obtained.</u> The quality of this assembly should be checked using the following parameters:

- Number of contigs: in general 1000 to 2000 per chromosome arm
- Number of clones in contigs and number of singletons: a ratio of 80/20 is correct but the purity of the library should be taken into consideration
- Average and N50 contig size: approx. 400-600 kb but the higher the better
- Distribution of contig sizes: a large number of contigs with less than 10 clones usually reflect a problem in the assembly

- Total contig length: reflects the coverage; at this stage it could be higher than the actual chromosome (arm) size
- Number of Q clones: should be as low as possible and never higher than 10% of the number of clones in contigs

If this automated assembly meets these general criteria, then it can enter the manual edition (finishing) phase.



** The following steps should be performed if there are still contigs with more than 10% Qs*

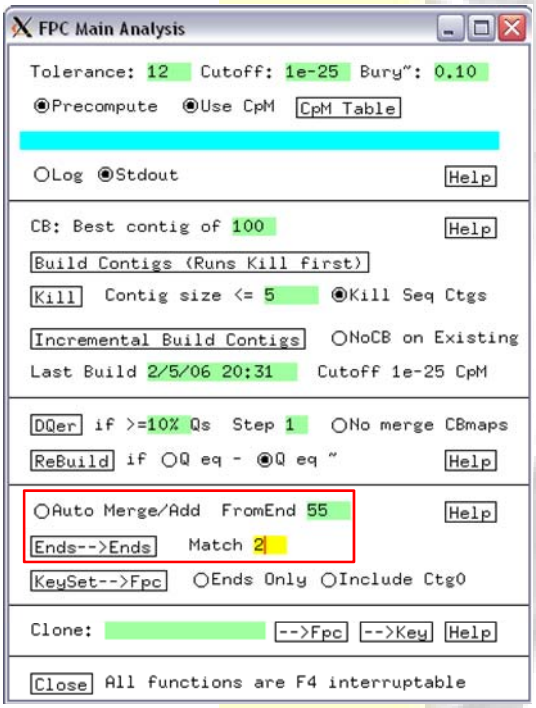
Figure 2: Chart of the automated assembly process

4. Manually edited contig assembly

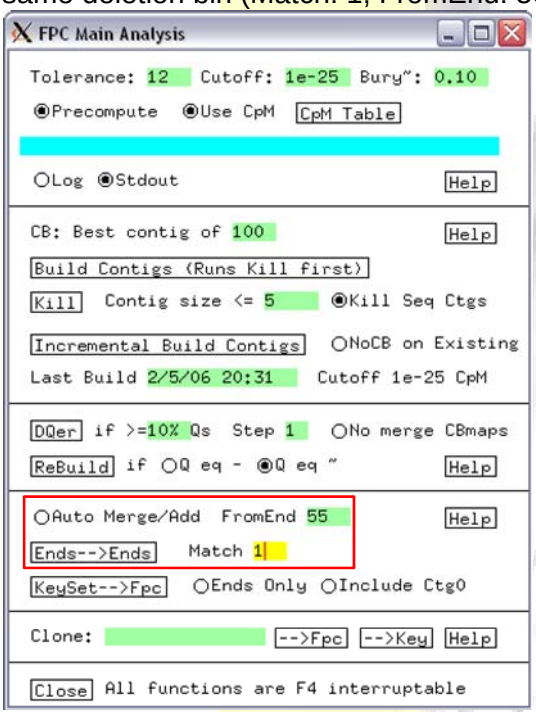
Following the automated assembly, contigs should be merged manually by identifying fingerprint overlaps with a lower stringency (cutoff: 1e-25, corresponding to roughly 25% of overlap) supported by information provided by contig anchoring with molecular markers in deletion bins or genetic maps.

- **This has to be done manually (Auto Merge OFF).**

Two contigs can be merged only if two BAC clones at the end of each contig matched each other in a reciprocal and unique manner at 1e-25 (Match: 2; FromEnd: 50-60).

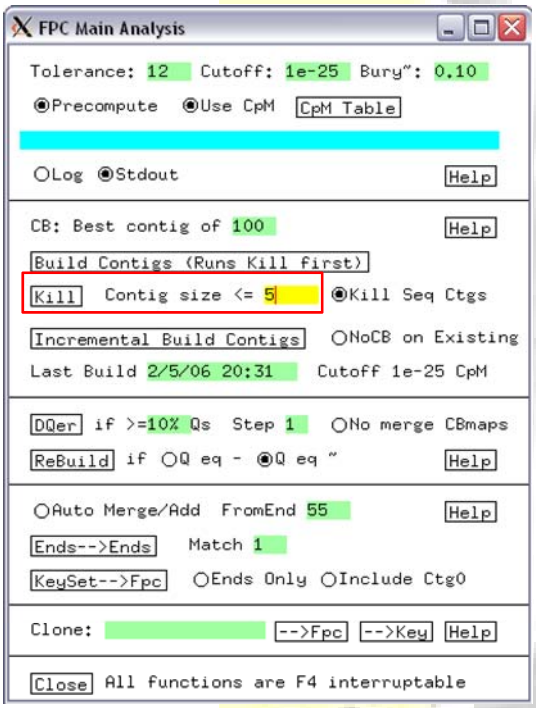


For contigs matching each other with only single BAC clones at their ends, the contigs can be merged only if the match is reciprocal and unique and if both contigs are located in the same deletion bin (Match: 1; FromEnd: 50-60).



Contigs having no fingerprint-based overlap can be manually merged if they share the same marker and are unambiguously assigned to the same deletion bin.

Finally, unanchored small contigs (less than 5 clones and smaller than 300 kb) can be removed (Kill function for the contigs having less than 5 clones; manually for the others) from the final assembly as they did not provide significant information.



5. MTP selection

Selecting an MTP (Minimal Tiling Path) consist in picking a set of minimally overlapping clones that span an entire contig. The MTP can be used efficiently for anchoring the physical map (by screening) and as a template for sequencing.

For anchoring purposes, the MTP has to be defined after the automated assembly and before the manually edited assembly. Indeed, anchoring data based on this MTP will be used to merge contigs (see above). Because of the length of wheat transposable elements (up to 10 kb), one has to select clones with a significant overlap to ensure a correct assembly of sequences.

This overlap is defined by **the Min FPC Overlap parameter** that corresponds to the minimal overlap between two clones based on the FPC coordinates. This parameter **should be set up to 30.** The **Max FPC Overlap should be set up to 250** (corresponding to the maximal length of a BAC clone) to allow for the selection of highly overlapping clones when mandatory to fill a gap.

FromEnd defines the distance from contig end to start picking MTP clones. For example, with a value of 55, all clones being less than 55 bands far from the end of the contig will be ignored and the MTP will only cover the length of the contig – 110 bands (55 at both ends). The main reason for this is that clone ordering at the end of the contigs is quite tricky and not always very robust. One can choose to ignore contig ends to avoid sequencing of clones that are not at the correct location in the contig. On the other hand, selecting these clones is the only way to find small overlaps between contigs through PCR screening (pending they are properly ordered). Thus, **the FromEnd parameter should be set up to 0** in order to define the MTP on the total length of the contig. For the record, on chromosomes 3B and 3DL, a

FromEnd value of 55 resulted in a 90% coverage while a value of 0 resulted in a 99% coverage.

Finally, **the Min Shared Bands parameter** is the number of bands they share by comparing the bands of the two clones. **The value is 12.**