## Guideline for physical map assembly

Simone Scalabrin<sup>1</sup>, Jan Bartos<sup>2</sup>, Melanie Febrer<sup>3</sup>, Daniela Schulte<sup>4</sup> and Etienne Paux<sup>5</sup>\*

<sup>1</sup> Applied Genomics Institute, Udine, Italy

<sup>2</sup> Institute of Experimental Botany, Olomouc, Czech Republic

<sup>3</sup> John Innes Centre, Norwich, UK

<sup>4</sup> IPK, Gatersleben, Germany

<sup>5</sup> 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).

<u>Note:</u> 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 (<u>http://www.appliedgenomics.org/FPB/;</u> <u>http://www.biomedcentral.com/1471-2105/10/127</u>). FPB performs the following tasks.

- 1) <u>Removal of all bands being out of the range of 50 to 500 bp.</u>
- 2) <u>Discard all clones having less than 40 bands or more than 250 bands</u>, as being either badly fingerprinted or potential chimerical clones.
- 3) <u>Cleaning of fingerprints from background</u> 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) <u>Removal of vector bands from fingerprints</u>:for BAC libraries constructed using the plndigoBAC-5 vector, fingerprint digestion of this vector results in two "red" bands (Xhol 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) <u>Band sizes conversion to integer values</u>, 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:

- Bl<mark>ue: 1500-15000</mark>
- Green: 16500-30000
- Yellow: 31500-45000
- Red: 46500-60000.
- 6) <u>Exporting processed fingerprints to .sizes files</u> that are compatible with Genoprofiler and FPC.

🔆 FPB - FingerPrint Bac	ckground removal				
First Value:	3	Last Value:	7		
Low index:	60	Min bands:	40		
Min sizes (per color):	5	Max sizes (total):	250		
Blue background:	50	Green background:	50		
Yellow background:	50	Red background:	50		
Blue offset:	0	Green offset:	15000		
Yellow offset:	30000	Red offset:	45000		
Tolerance:	0.4	Multiply factor:	30		
Peak width:	15	Fixed threshold:	500		
Size from:	50	Size to:	500		Figure 1: FPB interface
Library from:	1	Library to:	12		with settings
Plate from:	13	Plate to:	16		
Grid from:	17	Table suffix:	t×t		
Save	Process	Show vector	Quit		
			2	r	
		1001	5 12 1	1.	

# 2.3 Editing fingerprints using Genoprofiler: clone renaming and contamination removal

(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

(<u>http://www.wheatgenome.org/pdf/Triticeae\_Annotation\_Group\_Report\_2007.pdf</u>) 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 propo<mark>sed simplified co</mark>ding: TaaCsp3DLhA\_0023A01 → TaaCsp3DL023A01.

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

		AND A REAL PROPERTY OF A REAL PR			3
File Management Subset Clones Remove Vector/Repe	a Dec de l				
FROM: FPC Size File Directory	st banus				
Erows	- Initial fing	erprint file director	v		
TO: New FPC Size File Directory			<b>y</b>		
TO: New FPC Size Hile Directory		e			
	Renamed	fingerprint file dire	ectory		
Old vs. New Clone Name List File					
- Brown	Conversion	on name file (.txt fil	e)		
			•/		
	For example:	24.04 TO 2DI	0004.04		
	TaeCsp3DLhA_002				
	TaeCsp3DLhA_002				
	TaeCsp3DLhA_002				
	TaeCsp3DLhA_002 TaeCsp3DLhA 002				
	TaeCsp3DLhA_002				
	TaeCsp3DLhA_002				
	etc	Idausp3DL	020001		
	610	Rename			
		Help Close			
			Million -	_	
🕼 Sample File/Clone Na	ming Setting		<b>•</b>	<	
Specify Naming Policy of Sa	ample File Name and Clo	ne Name			
A sample file name at least	includes information of	plate number and well positi	ion, as well as library		
code if there are multiple lib					
positions of library code, pla	ate number and well pos	sition in a sample file name,	which are		
necessary for many operation	ons in this software. A	clone name usually includes			
a library code (optional), a					
Example of sample file name					
is "RI" from 1 to 2, the plate	number is "007" from t	9 to 11, and the well positio	n is "G12" from 13 to 15.		
Library Code					
· · · · · · · · · · · · · · · · · · ·					
Library Code From	1	To 9			TaaCsp3E
Disks Musshaw D	10	<b>T</b> 42			
Plate Number From	10	To 12	<u> </u>		023
Well Position From	13	To 15	5		A01
		.0 1			
	Help	Default Save	Cancel Ok		
	<u> </u>				

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

(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 <u>tolerance to use to detect overlap of adjacent</u> <u>clones</u>. 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).

	1 Jack 10 1 1 1 1 1 1 1
🏭 Contam	ination Check Setting 🛛 🛛 🔀
within 384- clone are v	ercentage is used as a cutoff to determine if two adjacent clones well or 96-well plates or a clone and chloroplast contaminated ery similar. If they have similar fingerprint patterns, ey are contaminated.
	Threshold of match percentage for 384 well plate $40.0$ %
	Threshold of match percentage for 96 well plate $[40.0]$ %
	Threshold of match percentage for chloroplast DNA contamination check $70.0$ %
	Threshold of matching percentage for high profile sharing clones $\%$
	Number of bands that a clone has at least 40
	Help Default Save Cancel Ok
	1.000

Contamination Check	
Checking Result Summary	
Choose Contamination Source Processing Message	4
☑ 384 Well Plate Check ☑ 96 Well Plate Check	
Chloroplast DNA Check High Profile Sharing Check	
Check only 🥝 Check and remove contaminated clones	
FPC Size File Directory	
Initial fingerprint file directory	
New FPC Size File Directory	
Contamination-free fingerprint file d	rectory
Chloroplast Fragment Size File	
Browse	
FPC Size Clone Name Filter Pattern	
Tolerance For Band Matching 0.4	
Help Start Cancel Close	

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



e Management		
et Clones Reneme Clones Remove Vector/Rep	eat Bands	
ose Input File Type		
) FPC Size Files (*.sizes/ *.bands) 🛛 Sample F	iles (*.fsa) O FPC File (*.fpc)	
DM: FPC Sze File Directory		TO: New FPC Size File Directory
Browse		Browse
luded Clone List File		Only Included Clone List File
Browse		Browse
		Clone Name Filter List (Wildcard Patterns)
Clone Band Number Range		
Carrier Carrier Street	Total Bands From 50 To 250	Blue Bands From 0 To 60
	Total bands From 50 To 250	
	Green Bands From 0 To 60	Yellow Bands From D To 60
Band Size Range (bp)	Green Bands From 0 To 60	
	Green Bands From 0 To 60	
and the second	Green Bands From 0 To 60	Vellow Bands From 0 To 60
and the second	Green Bands From 0 To 60 Red Bands From 0 To 60	Vellow Bands From 0 To 60
	Green Bands From 0 To 60 Red Bands From 0 To 60	Vellow Bands From 0 To 60
and the second	Green Bands From 0 To 60 Red Bands From 0 To 60	Vellow Bands From 0 To 60
9 From 25 bp To 500 bp	Green Bands From 0 To 60 Red Bands From 0 To 60 SubSet 0	Vellow Bands From 0 To 60
e From 25 bp To 500 bp	Green Bands From 0 To 60 Red Bands From 0 To 60 Subset Help ( excluded clones (.tx	Vellow Bands From 0 To 60
e From 75 bp To 500 bp	Green Bands From 0 To 60 Red Bands From 0 To 60 Subset Help ( excluded clones (.tx	Vellow Bands From 0 To 60
e From 75 bp To 500 bp	Green Bands From 0 To 60 Red Bands From 0 To 60 SubSet 0 Help ( excluded clones (.tx mple:	Vellow Bands From 0 To 60
e From 75 bp To 500 bp	Green Bands From 0 To 60 Red Bands From 0 To 60 SubSet ( Help ( excluded clones (.tx mple: 3DL023A01	Vellow Bands From 0 To 60
e From 25 bp To 500 bp List of For exar TaeCsp TaeCsp TaeCsp TaeCsp	Green Bands From 0 To 60 Red Bands From 0 To 60 SubSet ( Free Cluded clones (.tx mple: 3DL023A01 3DL023A02 3DL023B01 3DL023B02	Vellow Bands From 0 To 60
e From 25 bp To 500 bp List of For exar TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi	Green Bands From 0 To 60 Red Bands From 0 To 60 Subset Help C C C C C C C C C C C C C C C C C C C	Vellow Bands From 0 To 60
e From 75 bp To 500 bp List of For exar TaeCspt TaeCspt TaeCspt TaeCspt TaeCspt TaeCspt TaeCspt TaeCspt	Green Bands From 0 To 60 Red Bands From 0 To 60 Subset Help ( BL023A01 3DL023A02 3DL023B02 3DL023B02 3DL023021 3DL023O21	Vellow Bands From 0 To 60
e From 75 bp To 500 bp List of For exar TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi TaeCspi	Green Bands From 0 To 60 Red Bands From 0 To 60 Subset Help C C C C C C C C C C C C C C C C C C C	Vellow Bands From 0 To 60

### 3. Automated contig assembly

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

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

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 ( $0.4 \times 30$ ).



🗙 FPC Main Analysis	
Tolerance: 12 Cutoff: 1e-75 B @Precompute @Use CpM CpM Tak	the second s
OLog @Stdout	Help
CB: Best contig of 100	Help
Build Contigs (Runs Kill first)         Kill       Contig size <= 3	ll Seq Ctgs CB on Existing
DQer if >=10% Qs Step 3 ○No ReBuild if ○Q eq - @Q eq ~	
OAuto Merge/Add FromEnd 55 Ends>Ends Match 1 KeySet>Fpc OEnds Only OInc	Help lude Ctg0
Clone:>Fpc -	>Key Help
Close All functions are F4 inte	rruptable

- 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).

🕻 Configure Display		
Tolerance File:	🔿 Variable Tolerance	
Genome size <mark>0</mark>	Pure Sulston () Equation 2 kb (Clone size 150000 Band size 1100 b	ь
Gel length 54000 Contig display pag O Agarose () HICF	e size 3000	
Vector File:		
	Close	

Other parameters that have to be set up are:

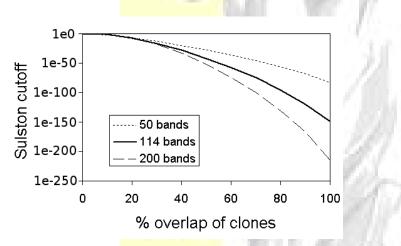
- The formula used: Pure Sulston formula with a fixed tolerance.
- <u>Band size</u>: Based on the sequencing of 152 BAC clones, the band length was estimated to be roughly equal to 1.1 kb (R<sup>2</sup> = 0.644), which is consistent with other estimates based on pulse field gel electrophoresis. Remember, 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.

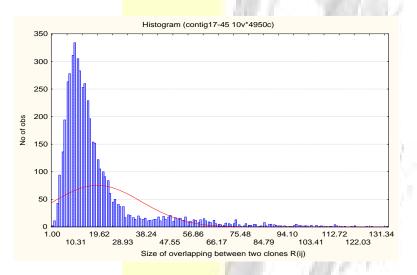
The Sulston formula calculates the probability for two clones to be overlapping based on the number of shared bands. The resulting score is compared to a given threshold called cutoff

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  $\sim$ 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.



🗙 FPC Main Analysis	
Tolerance: 12 Cutoff: 1e-75 Bur @Precompute @Use CpM <u>CpM Table</u>	
OLog @Stdout	Help
CB: Best contig of <mark>100</mark> Build Contigs (Runs Kill first) Kill Contig size <= <b>5</b> @Kill	Help Seg Ctas
Incremental Build Contigs ONoCB Last Build 2/5/06 20:31 Cutoff	on Existing
DQer]if>=10% Qs Step 3No m ReBuild if OQ eq - ⊛Q eq ~	erge CBmaps Help
○Auto Merge/Add FromEnd 55 Ends>Ends Match 1 KeySet>Fpc ○Ends Only ○Inclu	Help de Ctg0
Clone:>Fpc]>	Key Help
Close All functions are F4 interr	uptable

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.

🗙 FPC Main Analysis 📃 🗆 🔀	ACT
Tolerance: 12 Cutoff: 1e-75 Bury": 0.10	
Precompute	🗙 CpM Table 📃 🗆 🔀
	CpM (Cutoff plus Markers)
OLog @Stdout Help	Auto Adjust CpM Cutoff: +1 -1
CB: Best contig of <mark>100</mark> Build Contigs (Runs Kill first)	Current CpM Last Build Markers Cutoff
<u>Kill</u> Contig size <= <mark>5</mark> ●Kill Seq Ctgs	>=0 < 1e-75 0 1e-25
Incremental Build Contigs ONoCB on Existing	>1 < 1e-60 1 1e-10
Last Build 2/5/06 20:31 Cutoff 1e-45 CpM	>2 < 1e-45 2 1e+05
	>3 < 1e-30 3 1e+20
DQer if >=10% Qs Step 3 ONo merge CBmaps	OUSE ePCR USE ePCR
ReBuild if OQ eq - @Q eq "Help	OUse YBP(ac) No YBP(ac)
OAuto Merge/Add FromEnd 55 Help	OUse REP No REP
Ends>Ends Match 1	Ignore Markers in more than <mark>50</mark> clones
KeySet>Fpc OEnds Only OInclude Ctg0	Close
Clone:>Fpc>Key Help	Rates: Min <mark>0</mark> Max <mark>65535</mark>
Close All functions are F4 interruptable	

Another parameter to be set up is the number of iterations 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.

🗙 FPC Main Analysis	_ 🗆 🛛
Tolerance: 12 Cutoff: 1e-75 Bury": @Precompute @Use CpM <u>CpM Table</u>	0.10
OLog ®Stdout	Help
CB: Best contig of 100 Build Contigs (Runs Kill first) Kill Contig size <= 5	Existing
DQer if >= <mark>10% Q</mark> s Step <mark>3</mark> ○No merg ReBuild if ○Q eq - @Q eq ~	e CBmaps Help
○Auto Merge/Add FromEnd 55 Ends>Ends Match 1 KeySet>Fpc ○Ends Only ○Include	Help Ctg0
Clone:>Fpc]>Key	Help
Close All functions are F4 interrupt	able

- 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).

🗙 FPC Main Analysis	
Tolerance: 12 Cutoff: 1e-75 Precompute  Use CpM CpM Ta	Contraction of the second s
OLog @Stdout	Help
CB: Best contig of 100	Help
Build Contigs (Runs Kill first) Kill Contig size <= 5 @K	
	NoCB on Existing
DQer if >= <mark>10%</mark> Qs Step 3 ON <u>ReBuild</u> if OQ eq - @Q eq ~	lo merge CBmaps Help
OAuto Merge/Add FromEnd 55 Ends>Ends Match 1 KeySet>Fpc OEnds Only OIn	Help Helude Ctg0
Clone:>Fpc	>Key Help
Close All functions are F4 int	erruptable

- 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).

🗙 FPC Main Analysis	
Tolerance: 12 Cutoff: <mark>1e-75</mark> B @Precompute @Use CpM <u>CpM Ta</u>	Contraction of the second s
OLog ®Stdout	Help
CB: Best contig of 100 Build Contigs (Runs Kill first)	Help
Kill     Contig size <= 5	Lll Seq Ctgs DCB on Existing
	°f 1e-45 CpM
DQer if >=10% Qs Step 3 ONo ReBuild if OQ eq - @Q eq "	o merge CBmaps Help
⊛Auto Merge/Add FromEnd <mark>55</mark> Ends>Ends Match 1	Help
KeySet>Fpc @Ends Only OInc	
Clone:>Fpc ·	>Key Help
LIOSE HII FUNCTIONS are F4 Inte	BILITADIE

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 endclone. 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.



🗙 FPC Main Analysis 📃 🗆 🔀	🗙 FPC Main Analysis 📃 🗆 🔀
Tolerance: 12 Cutoff: 1e-75 Bury": 0.10	Tolerance: 12 Cutoff: 1e-75 Bury": 0,10
@Precompute @Use CpM <u>CpM Table</u>	@Precompute @Use CpM <u>CpM Table</u>
OLog @Stdout Help	OLog @Stdout Help
CB: Best contig of 100 Help	CB: Best contig of 100 Help
Build Contigs (Runs Kill first)	Build Contigs (Runs Kill first)
Kill Contig size <= 5	Kill Contig size <= 5
DQer if >= <mark>10% Q</mark> s Step <b>3</b> ONo merge CBmaps	DQer if >=10% Qs Step 3 ONo merge CBmaps
ReBuild if OQ eq - @Q eq " Help	ReBuild if OQ eq - @Q eq " Help
Auto Merge/Add FromEnd 55     Help     Ends>Ends Match 1     KeySet>Fpc	Auto Merge/Add FromEnd 55     Help     Ends>Ends Match 1     KeySet>Fpc    Ends Only OInclude Ctg0
Clone:>Fpc>Key Help	Clone:>Fpc>Key Help
Close All functions are F4 interruptable	Close All functions are F4 interruptable

- 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  $\rightarrow$  Fpc function (see above).

🔀 FPC V9.3 Main Menu 📃 🗆 🔀	🗙 FPC V9.3 Main Menu 📃 🗖 🔀	🗙 Clones 24104 (page 1505/1507) 📃 🗔 🔀
Project: TaaCsp3DL_project	Project: TaaCsp3DL_project	φ.
Class: Contigs Clones Markers	Class: <u>Contigs</u> <u>Clones</u> <u>Markers</u>	TaaCsp3DL023D15 TaaCsp3DL023H16
Search: Name	Search: Name	TaaCsp3DL023D17 TaaCsp3DL023H19
		TaaCsp3DL023E15 TaaCsp3DL023H22
Search Commands Clear Reset	Search Lommands Name Keset	TaaCsp3DL023E19 TaaCsp3DL023I06
File Configure Clean Up	File Configu Gel Identifier	TaaCsp3DL023F01 TaaCsp3DL023I22
Save .fpc Load .fpc Update .cor	Save .fpc Load .f After Mod Date .cor	TaaCsp3DL023F17 TaaCsp3DL023J05
Main Analysis Ctg->Chr	Main Analysis Remark	TaaCsp3DL023G16 TaaCsp3DL023J06
	Singlebons Quit Help Search menu Draft	TaaCsp3DL023G18 TaaCsp3DL023J09
Quit Help BSS MTP Draft	Quit Help Search menu Draft	l 🗖

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

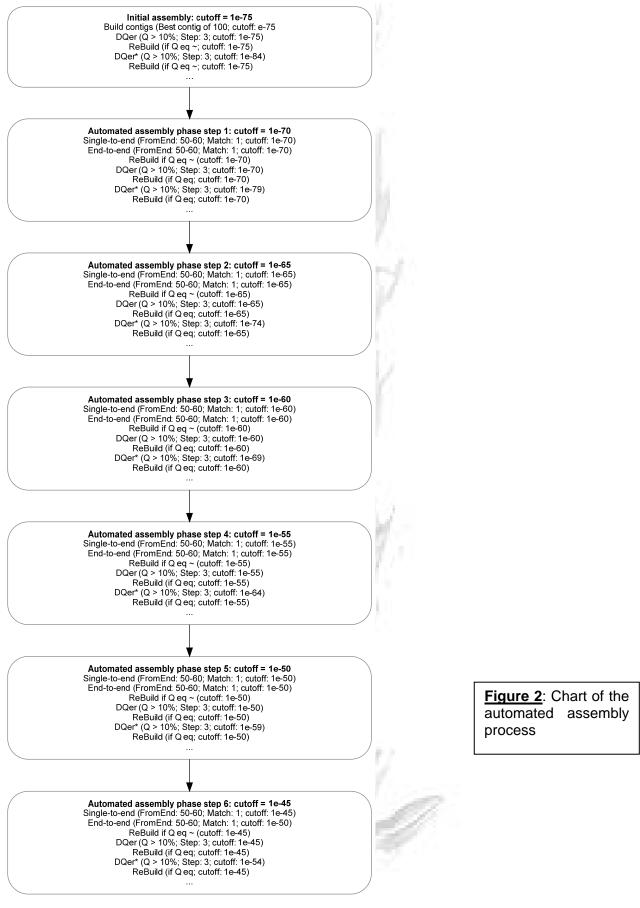
<u>After the above procedure is completed a draft assembly is obtained.</u> The quality of this assembly should be checked using the following parameters:

- 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



### 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).

🗙 FPC Main Analysis		
Tolerance: 12 Cutoff: 1e-25 Precompute  Use CpM  CpM T		
OLog @Stdout	Help	
CB: Best contig of 100	Help	
[Inter entering of the second get]	allowers and	
DQer if >=10% Qs Step 1 _ Of ReBuild if OQ eq - ⊛Q eq "	No merge CBmaps Help	
OAuto Merge/Add FromEnd 55 Ends>Ends Match 2	Help	
KeySet>Fpc OEnds Only OInclude Ctg0		
Clone:>Fpc	>Key Help	
Close All functions are F4 interruptable		

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).

🗙 FPC Main Analysis 📃 🗖
Tolerance: 12 Cutoff: 1e-25 Bury": 0.10 Precompute  Use CpM <u>CpM Table</u>
OLog @Stdout Help
CB: Best contig of 100 Help Build Contigs (Runs Kill first) Kill Contig size <= 5
DQer if >=10% Qs Step 1 ONo merge CBmaps ReBuild if OQ eq - @Q eq " Help
OAuto Merge/Add FromEnd 55 Help Ends>Ends Match 1 KeySet>Fpc OEnds Only OInclude Ctg0
Clone:>Fpc]>Key] Help
Close All functions are F4 interruptable

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 (<u>Kill function</u> for the contigs having less than 5 clones; manually for the others) from the final assembly as they did not provide significant information.

🗙 FPC Main Analysis	- 🗆 🛛	
Tolerance: 12 Cutoff: 1e-25 Bury": 0.10 Precompute  Use CpM CpM Table		
OLog @Stdout	Help	
CB: Best contig of 100	Help	
Build Contigs (Runs Kill first) Kill Contig size <= 5	eq Ctgs	
Incremental Build Contigs ONoCB o Last Build 2/5/06 20:31 Cutoff 1e	n Existing -25 CpM	
DQer if >=10% Qs Step 1 ONo mer ReBuild if OQ eq - ⊛Q eq ~	ge CBmaps Help	
OAuto Merge/Add FromEnd 55	Help	
Ends>Ends Match 1 KeySet>Fpc OEnds Only OInclude Ctg0		
Clone:>Fpc>Ke	y Help	
Close All functions are F4 interrup	table	

### 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.

1

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.

X Select MTP	
STEP 1: Find overlapping pairs	111 30
Min FPC Overlap: 30 Max FPC Overlap: 250	18
FromEnd: 0	11. 10
Use Fingerprints:	
Min Shared Bands: 12	1
O Use BSS results	B
BSS File:	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Filter BLAST Score: 400 Identity: 97	363 A
Advanced settings	W. Wasser
	1 B.S.S.
Or use existing Pairs File:	1 1 1 2 2 3 3 1 1
Load	Los Ballo Jr.
STEP 2: Pick MTP clones	The second of the
All contigs O Contig 1	1 E. M. B. B. T. S.
Give preference to large clones	and the second sec
Mandatory clones 5 options selected	- 10 C
	10071007
STEP 3: View results (if desired)	1. All all and
Mini Window Contig 1	1.
Step through pairs: Next Previous Pick start	10.00
Step through MTP: Next Previous All	100007
Show fingerprints (Fp only)	10000
Set MTP clone status to TILE for contig	1 11.5
STEP 4: Save results	and the second s
Set MTP clone status to TILE Clear TILE	P #
File of Pairs: Save	1 1 / ABR
File of MTP clones: Save	
	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1
Eval Clone Seq->BES Results	119 11 1
Close Help	
and the second se	111
	MR
	178 C 17 C 1
the second se	1 K.C
100 million (100 m	
the second se	
	1.1 1977 107 1174
	1 1 1 A
	and the second sec
	10000
	- F. S. C.
	10.1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1.2
Comments of the local sector of the local sect	
	Sec. 1