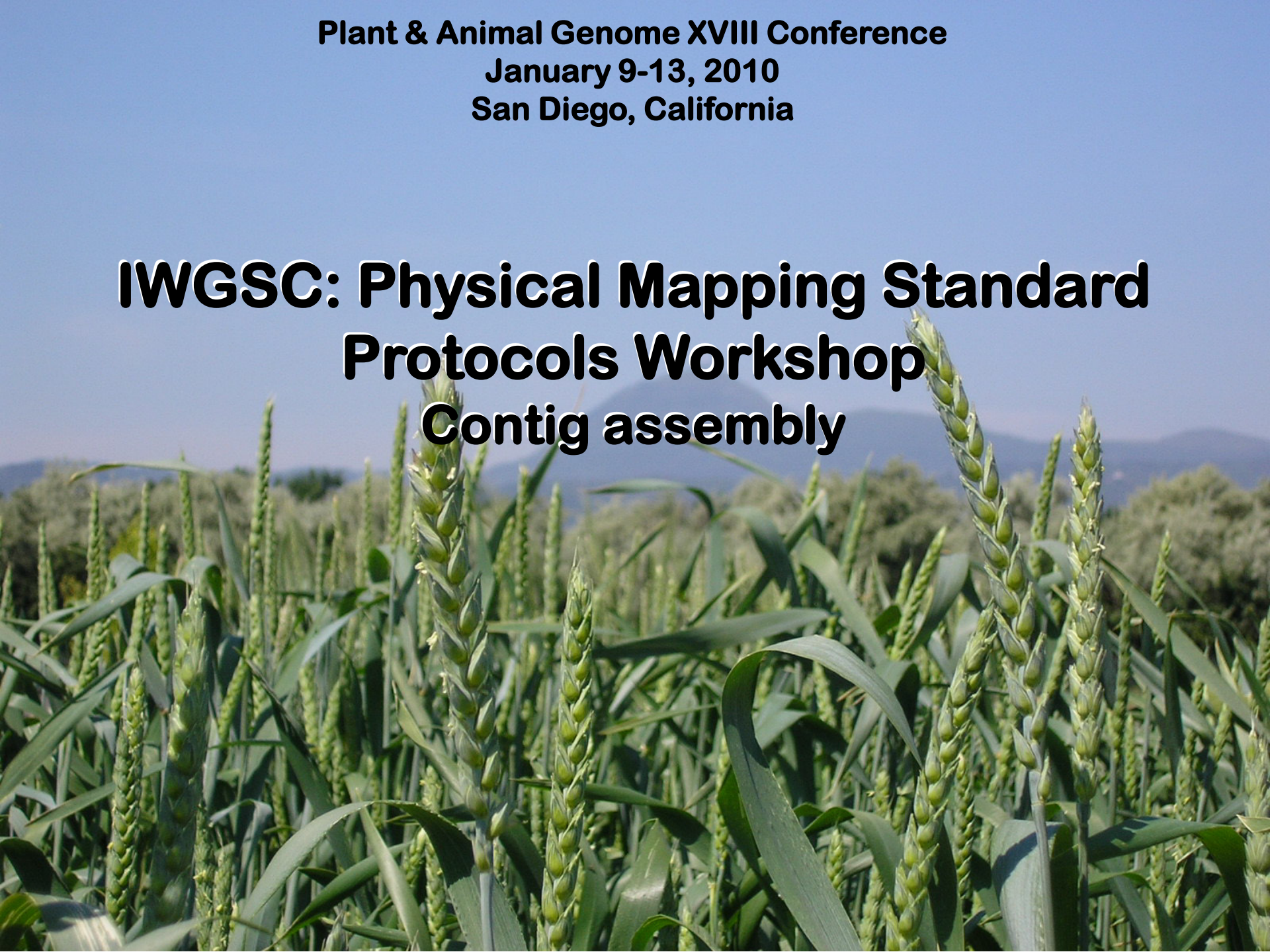


Plant & Animal Genome XVIII Conference
January 9-13, 2010
San Diego, California

**IWGSC: Physical Mapping Standard
Protocols Workshop**
Contig assembly



The background of the slide is a photograph of a lush green wheat field. The wheat stalks are in the foreground, slightly out of focus, with their green heads and long leaves visible. In the distance, a range of blue mountains is visible under a clear, light blue sky. The overall scene is bright and natural.

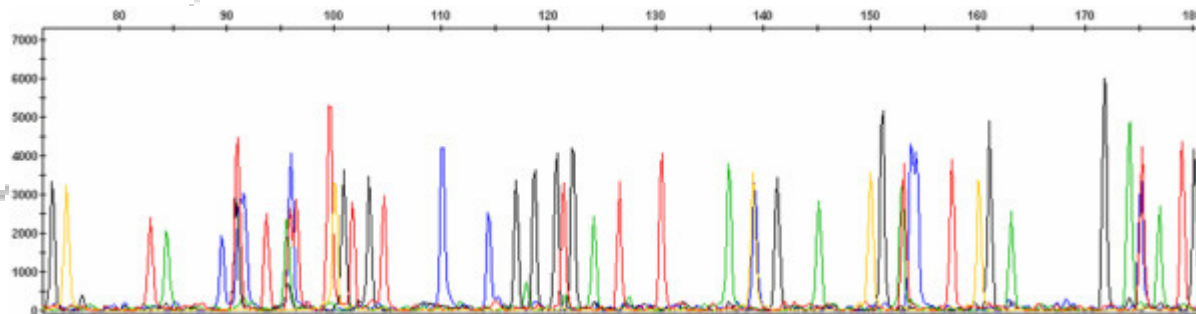
Editing fingerprints

1- FPB

Different sources of peaks

Each peak represents a fragment with a certain size and intensity and it can derive from different sources:

- ✓ "true peak" derived from a DNA insert digested band;
- ✓ low signal peak produced by the machine;
- ✓ partial digestion related peak;
- ✓ star activity by-product;
- ✓ *E. coli* genomic DNA band;
- ✓ vector band;
- ✓ out of size standard range band (with unreliable sizing);
- ✓ wide area peak (unreliable, resulting from co-migrating fragments).



Cleaning fingerprints using FPB

Automated FingerPrint Background removal: FPB

Scalabrin *et al.* (2009) BMC Bioinformatics, 10:127

Background removal

✓ "true peak" derived from a DNA insert digested band; **BAC fingerprint**

✓ low signal peak produced by the machine;

✓ partial digestion related peak;

✓ star activity by-product;

✓ *E. coli* genomic DNA band;

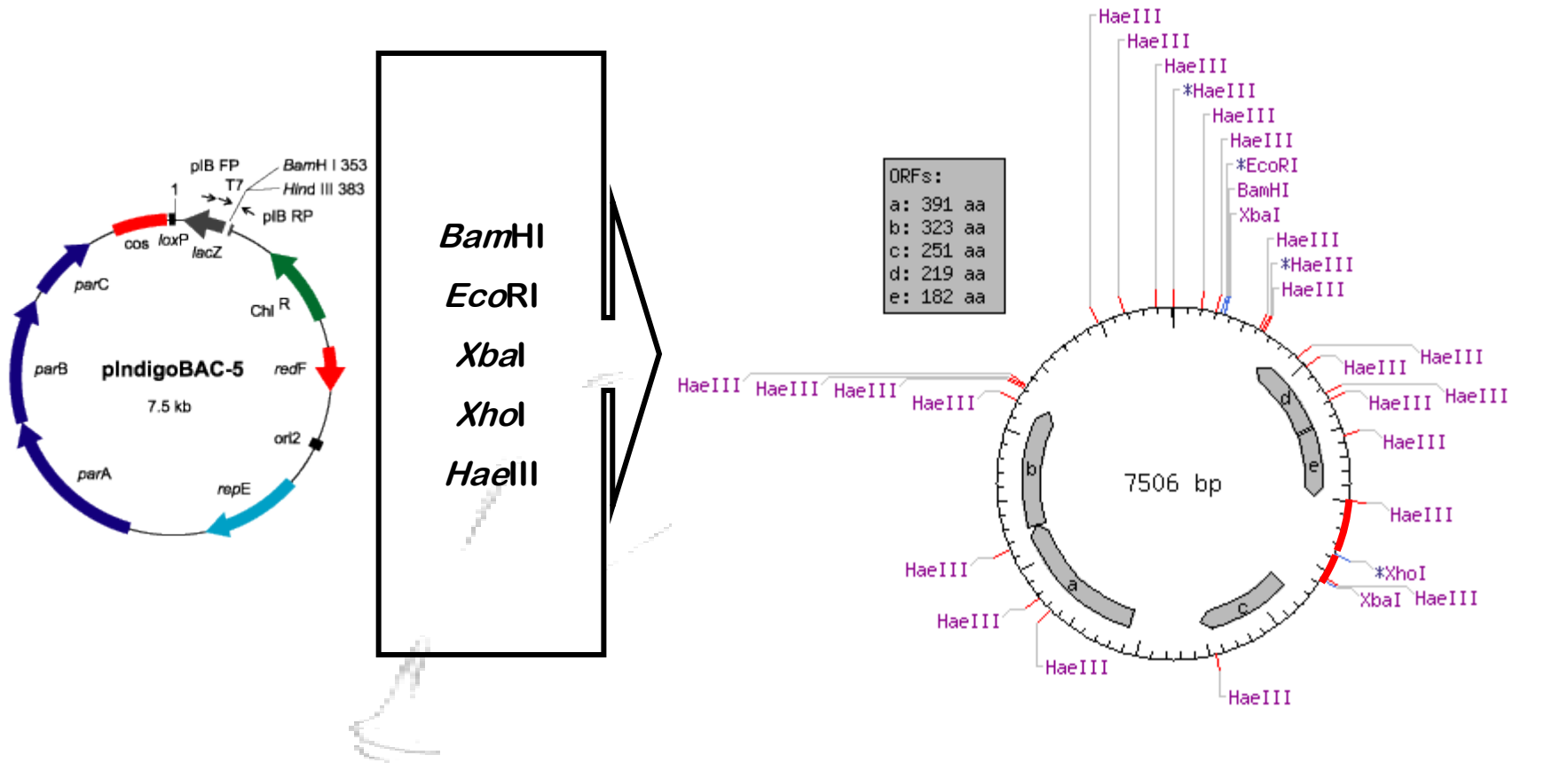
✓ vector band;

✓ out of size standard range band (with unreliable sizing);

✓ wide area peak (unreliable, resulting from co-migrating fragments).

Pre-processing

Vector bands



Two red fragments (*Xho*I): 161 & 375 bp

common to all fingerprints

(all the other labelled fragments are too short to be selected)

Removing vector bands

FPB - FingerPrint Background 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
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	txt

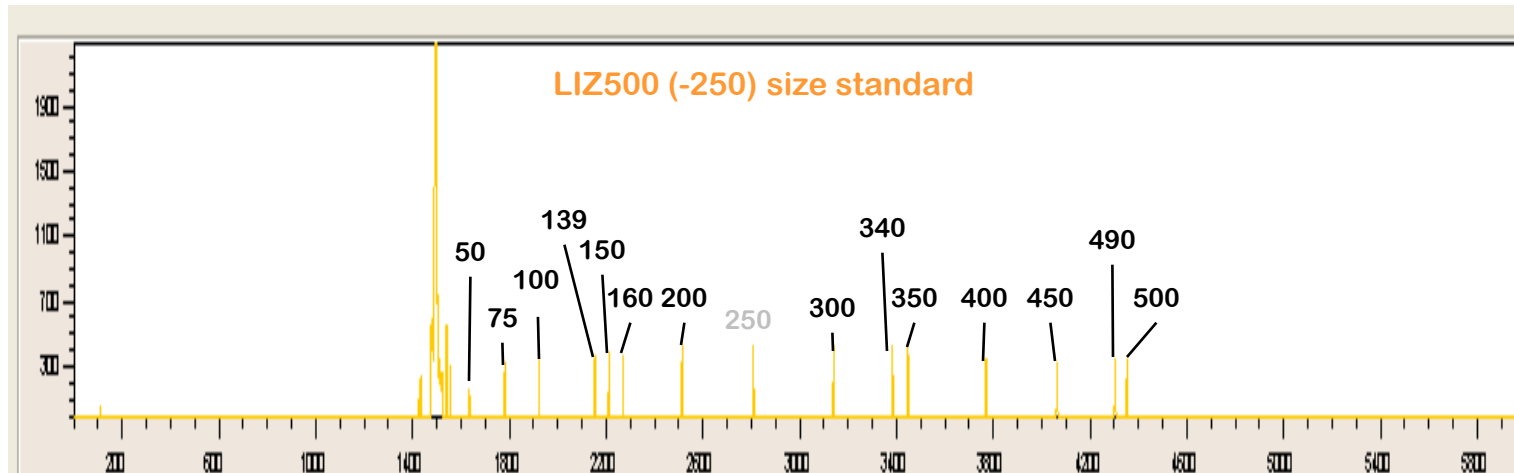
Save Process **Show vector** Quit

```
#Vector File
Red
157.11
371.57
-1
```

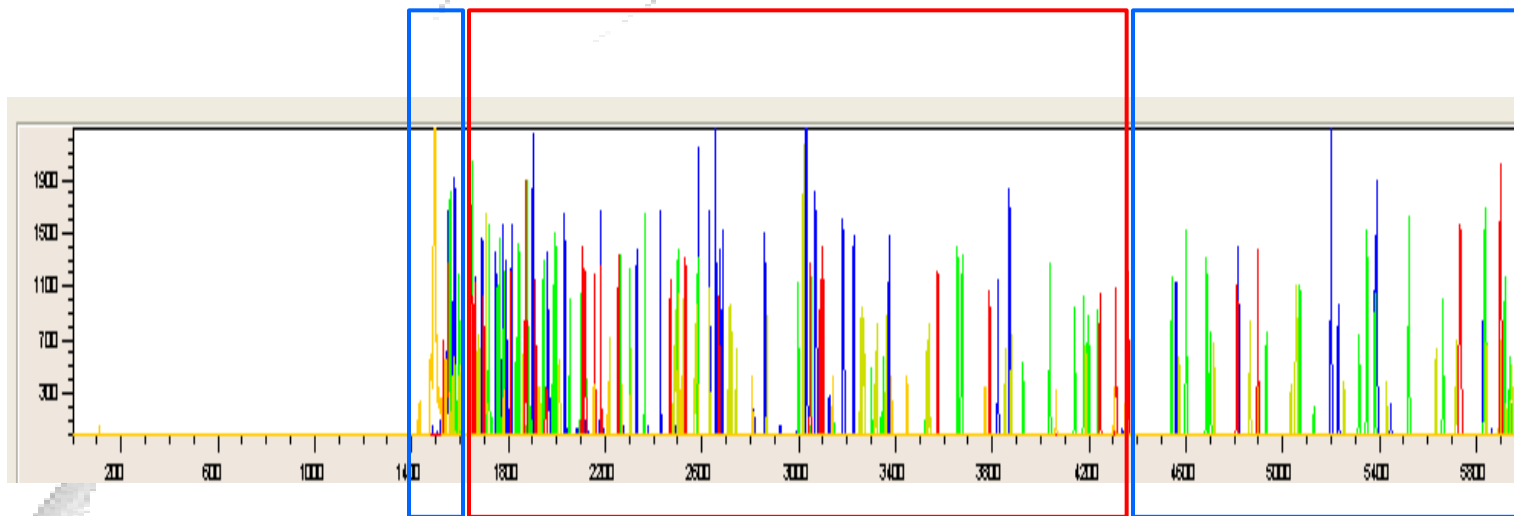
Observed values vs. Expected values

vector.cfg

“Out of range” bands



Out of range 50-500 bp range Out of range

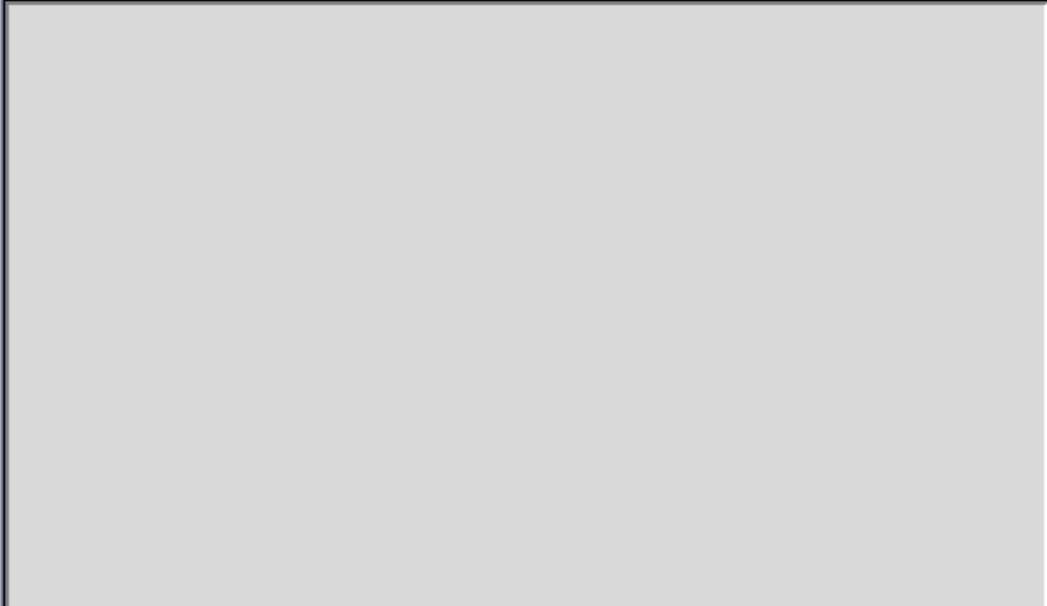


Removing “out of range” bands

FPB - FingerPrint Background 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
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	txt

Save Process Show vector Quit




Removing wide peaks

FPB - FingerPrint Background removal

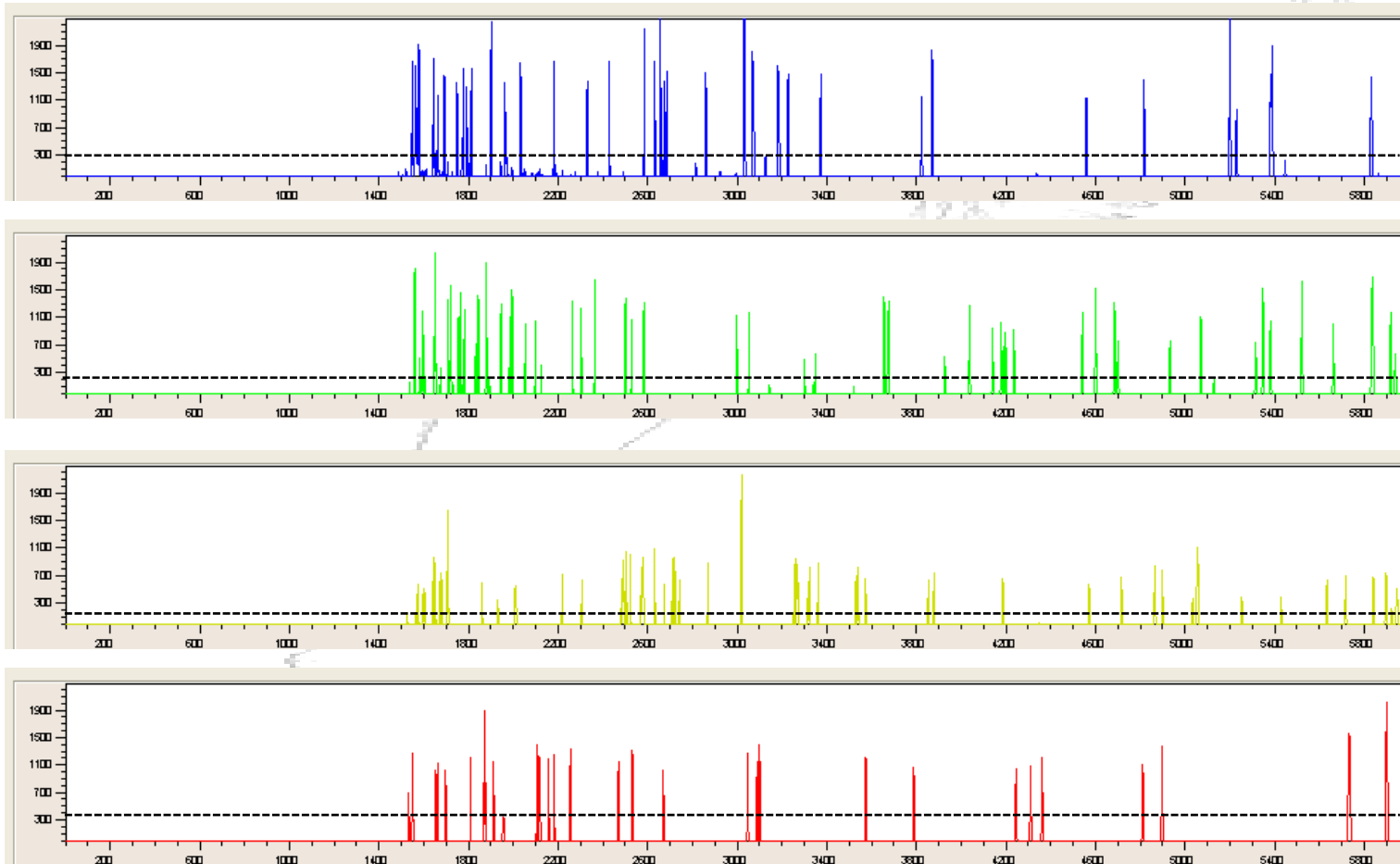
First Value:	3	Last Value:	7
Low index:	60	Min bands:	40
Min sizes (per color):	5	Max sizes (total):	250
Blue background:	50	Green background:	50
Yellow background:	50	Red background:	50
Blue offset:	0	Green offset:	15000
Yellow offset:	30000	Red offset:	45000
Tolerance:	0.4	Multiply factor:	30
Peak width:	15	Fixed threshold:	500
Size from:	50	Size to:	500
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	txt

Save Process Show vector Quit



True signal vs. background

Calculation of the background threshold for each dye



Removal of all peaks below the threshold

(adapted from Scalabrin *et al.*, *BMC Bioinformatics*, 2009)

Multiplication factor & color shift

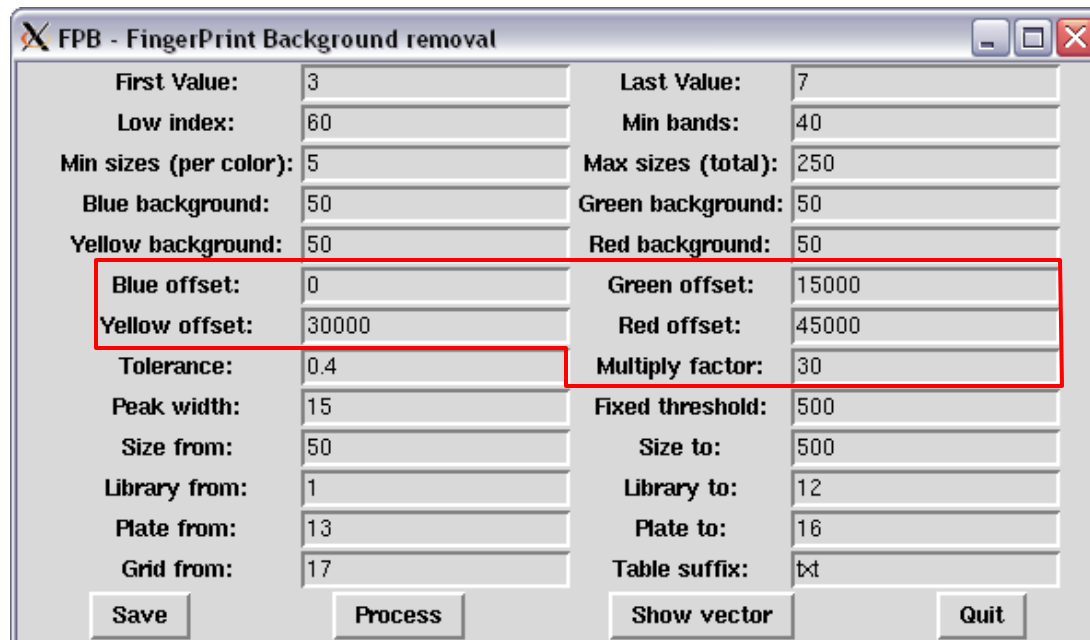
FPC does not accept **color labels** or **fractional sizes**, so the fragments must be manipulated before being loaded into FPC.

First, every size is **multiplied by 30**, after which the decimal part can be dropped without losing significant information.

This results in a set of fragments ranging from 1500 to 15000 instead of the 50-500 bp.

Then the color labels are converted to non-overlapping numeric ranges by adding a **different offset value** for each color: 0 to blue; 15,000 to green; 30,000 to yellow and 45,000 to red.

This puts each color into its own range, not overlapping with fragments of other colors. The total range is then 0-60,000, with 4 gaps of length 1500 (0-1500; 15,000-16,500; 30,000-31,500 and 45,000-46,500).

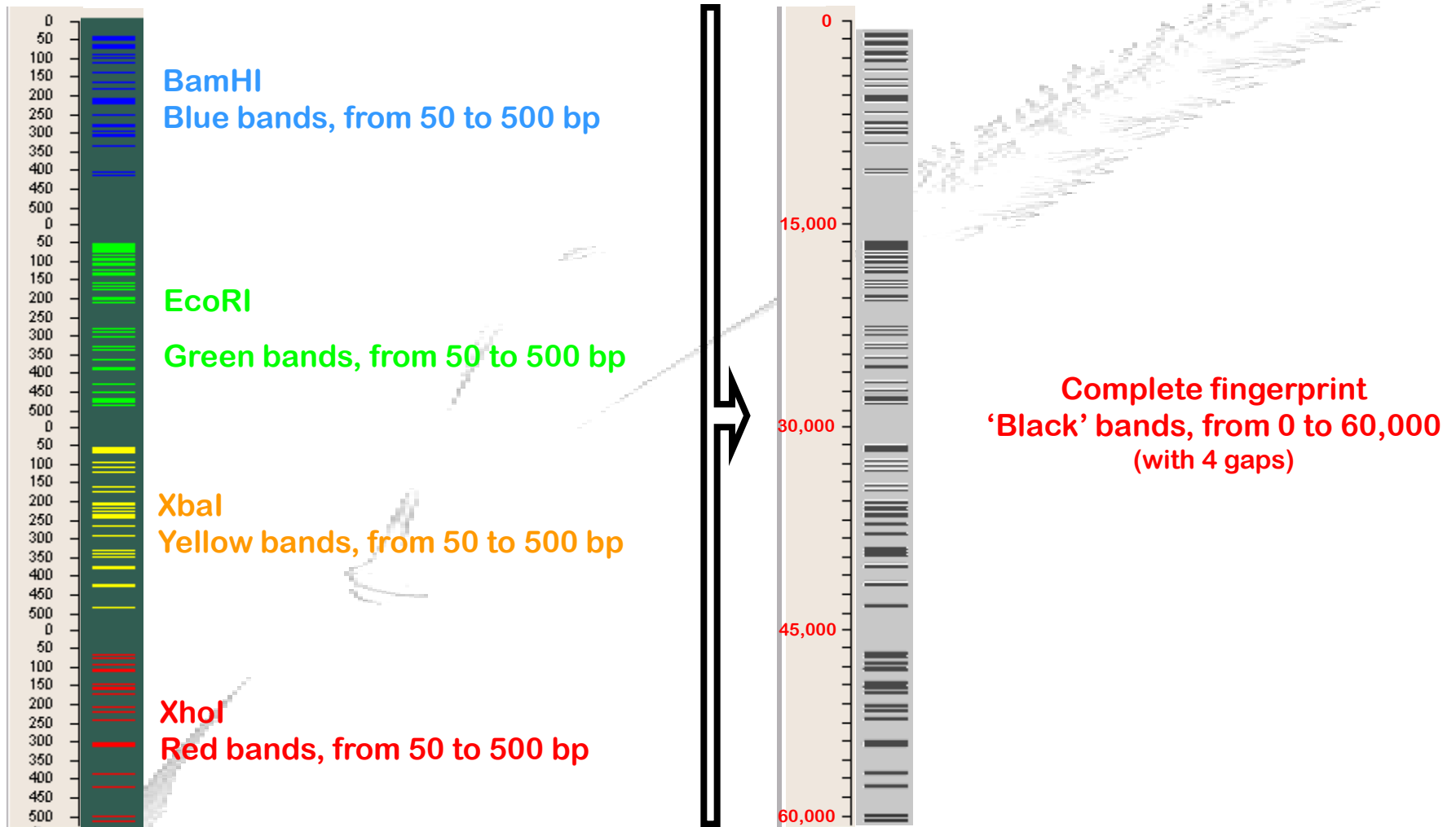


FPB - FingerPrint Background removal

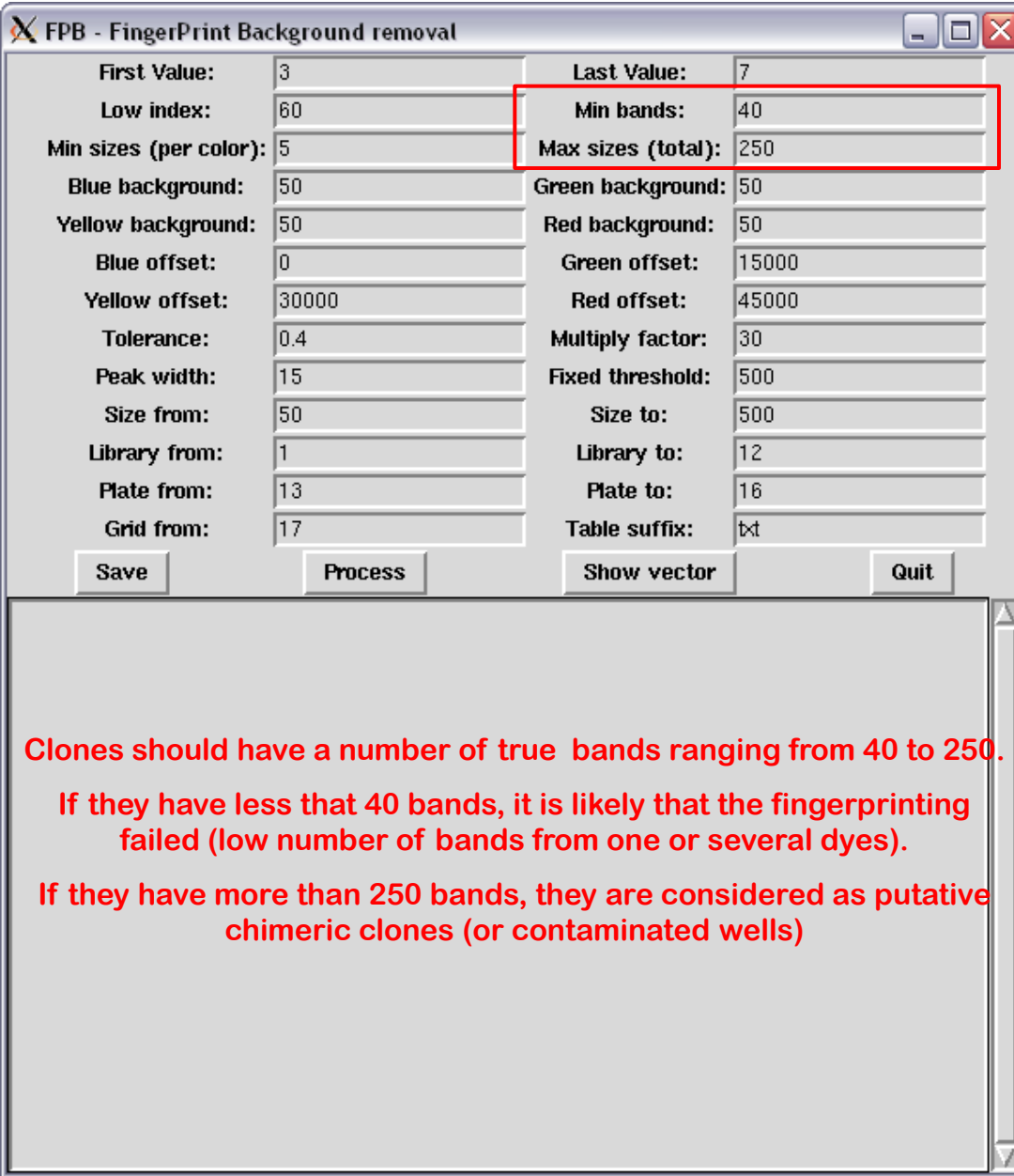
First Value:	3	Last Value:	7
Low index:	60	Min bands:	40
Min sizes (per color):	5	Max sizes (total):	250
Blue background:	50	Green background:	50
Yellow background:	50	Red background:	50
Blue offset:	0	Green offset:	15000
Yellow offset:	30000	Red offset:	45000
Tolerance:	0.4	Multiply factor:	30
Peak width:	15	Fixed threshold:	500
Size from:	50	Size to:	500
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	xt

Save Process Show vector Quit

Multiplication factor & color shift



Removing low quality fingerprints



FPB - FingerPrint Background 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
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	txt

Save Process Show vector Quit

Clones should have a number of true bands ranging from 40 to 250.

If they have less than 40 bands, it is likely that the fingerprinting failed (low number of bands from one or several dyes).

If they have more than 250 bands, they are considered as putative chimeric clones (or contaminated wells)

International naming convention (IWGSC)

TaaCsp3BFhA_0001A23 is a specific BAC with the following specifications:

✓ **Digits 1-3 define the genus/species (Taa).**

Three characters are used since there was concern two would not be enough to clearly define all possible cases (*e.g.* Taa = *Triticum aestivum* ssp. *aestivum*).

✓ **Digits 4-6 define the cultivar (Csp).**

Three characters since we're concerned two won't be enough in future, and to handle cultivars that already have a standard 3 letter designation (*e.g.* Csp = Chinese Spring).

✓ **Digits 7-9 define the chromosomal source of DNA (3BF).**

F for full chromosome, L for long arm, S for short arm, ALL for whole genome and 146 for 1D-4D-6D (*e.g.* 3BF = whole chromosome 3B).

✓ **Digits 10-11 define the restriction enzyme used to make the library and the number of the library (hA).**
(*e.g.* hA s the first library made with *HindIII*, hB the second one).

✓ **Digit 12 separates the library name from the specific clone identification within that library (_).**

Its main function is to improve readability, instead of the continuous long stream of characters which the eye will tend to blur.

✓ **Digits 13-19 identify plate number and well position within the plate (0001A23).**

Four digits are used for the plate number (*e.g.* 0001A23 = clone A23 from the plate 1).

Setting up clone name in FPB

FPB - FingerPrint Background removal

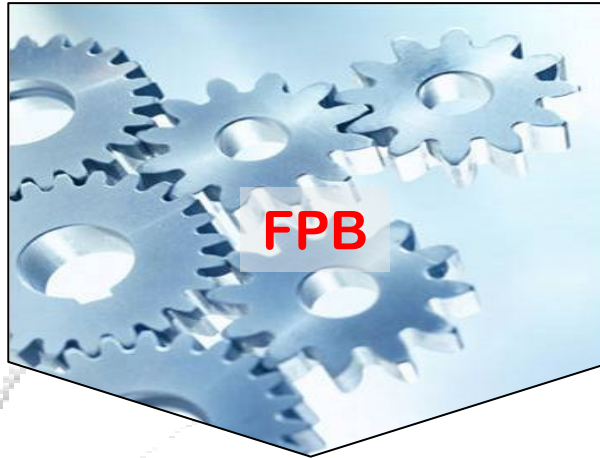
First Value:	3	Last Value:	7
Low index:	60	Min bands:	40
Min sizes (per color):	5	Max sizes (total):	250
Blue background:	50	Green background:	50
Yellow background:	50	Red background:	50
Blue offset:	0	Green offset:	15000
Yellow offset:	30000	Red offset:	45000
Tolerance:	0.4	Multiply factor:	30
Peak width:	15	Fixed threshold:	500
Size from:	50	Size to:	500
Library from:	1	Library to:	12
Plate from:	13	Plate to:	16
Grid from:	17	Table suffix:	txt

Save Process Show vector Quit

TaaCsp3BFhA_0001A23

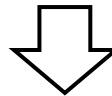
FPB output

GeneMapper .txt files



FPB .sizes files

- ✓FPC-compatible
- ✓Background-free
- ✓Vector-free
- ✓Ranging from 50 to 500 bp...



Genoprofiler

A background image of a lush green wheat field. The wheat stalks are in the foreground, slightly out of focus, with their green heads and long leaves visible. In the background, there are rolling hills or mountains under a bright, clear sky. The overall tone is bright and natural.

Editing fingerprints

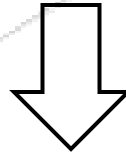
2- Genoprofiler

Clone renaming

FPC cannot handle BAC names longer than **15 digits**.

Thus BAC names have to be **shortened** to be used in FPC.

TaaCsp3BFhA_0001A23

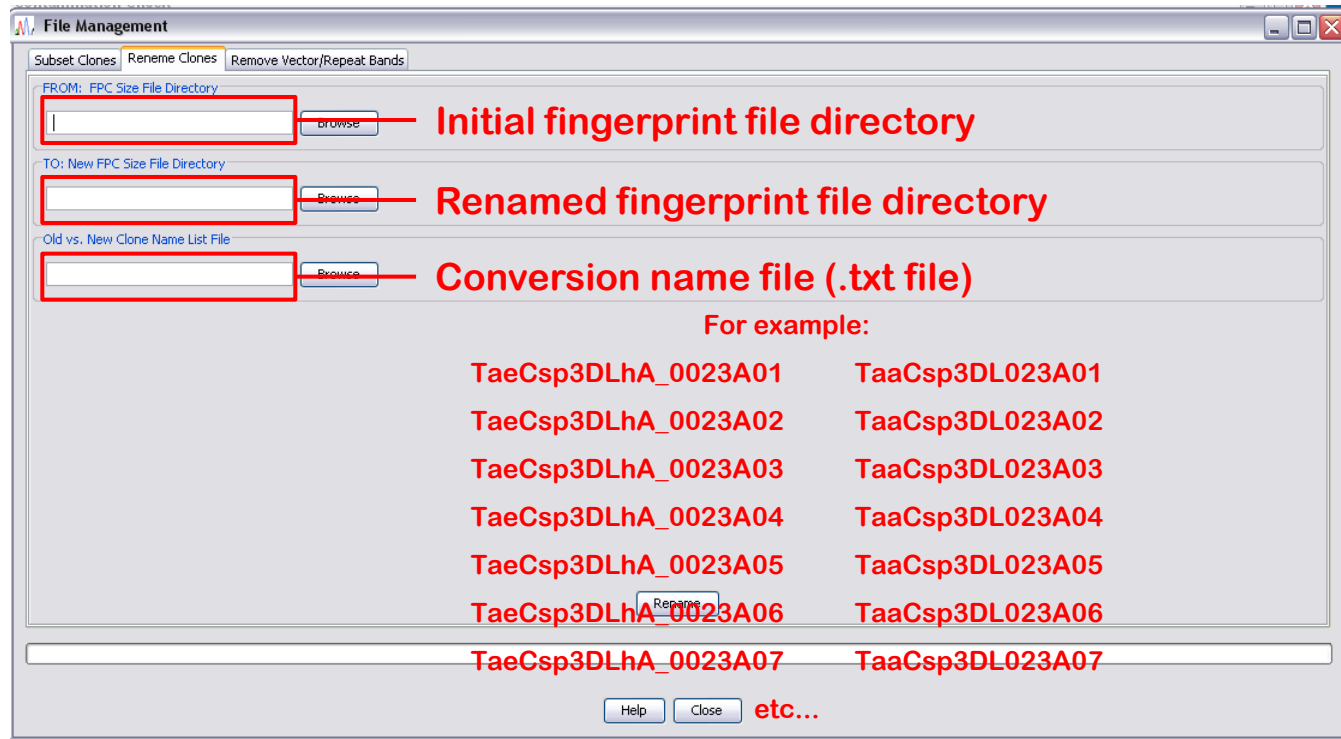


TaaCsp3BF001A23

Short names are **informative** enough for FPC analysis.

However, clones have to be **renamed according the international nomenclature** prior to being released in the public domain.

Clone renaming using Genoprofiler



But the 'rename clone' function of Genoprofiler does not work with names longer than 10 digits!!

Clone renaming using *perl*

TaaCsp3BFhA_0001A01

TaaCsp3B001A01

TaaCsp3BFhA_0001A02

TaaCsp3B001A02

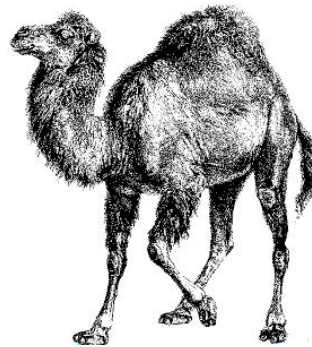
TaaCsp3BFhA_0001A03

TaaCsp3B001A03

...

Command line:

```
> perl -pe "s/TaaCsp3BFhA_0/TaaCsp3B/g" File_to_be_renamed.sizes > Renamed_file.sizes
```



Configuring Genoprofiler

Sample File/Clone Naming Setting

Specify Naming Policy of Sample File Name and Clone Name

A sample file name at least includes information of plate number and well position, as well as library code if there are multiple libraries associated with clones. User needs to specify the exact positions of library code, plate number and well position in a sample file name, which are necessary for many operations in this software. A clone name usually includes a library code (optional), a plate number, and a well position, such as RI003F12. Example of sample file name: **RI_Plate007_G12_03.fsa**. In this file name, the library code is "RI" from 1 to 2, the plate number is "007" from 9 to 11, and the well position is "G12" from 13 to 15.

Library Code ☒

Library Code From	1	To	9
Plate Number From	10	To	12
Well Position From	13	To	15

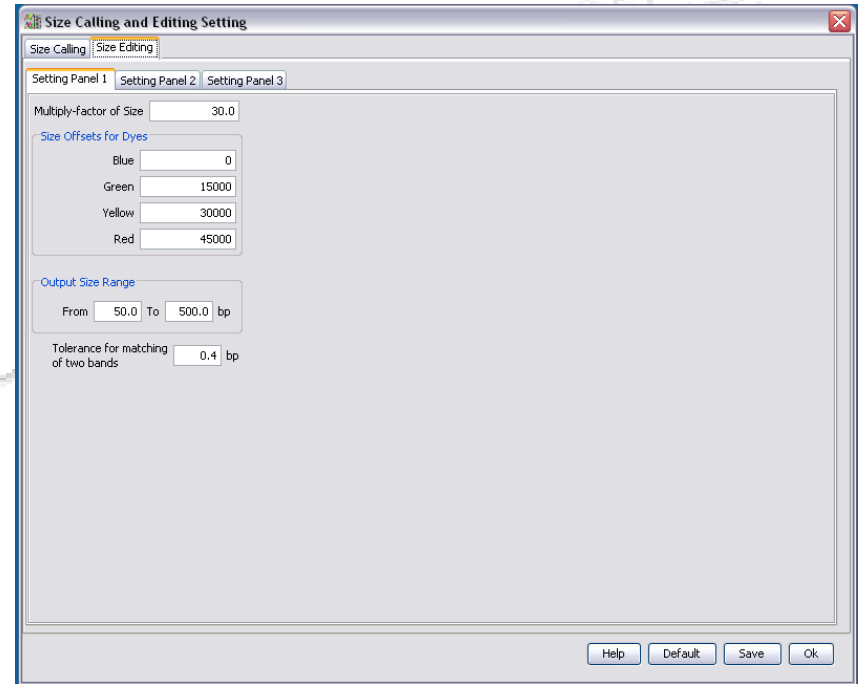
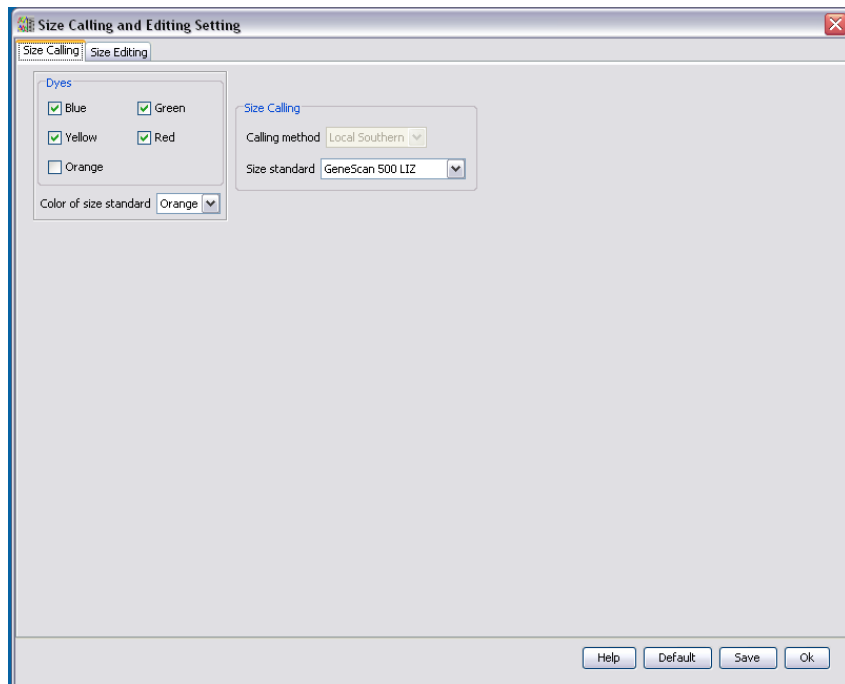
Help Default Save Cancel Ok

TaaCsp3DL

023

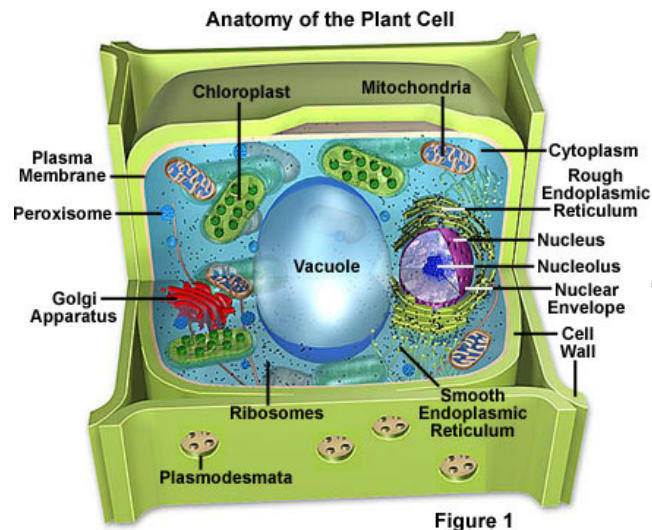
A01

Configuring Genoprofiler

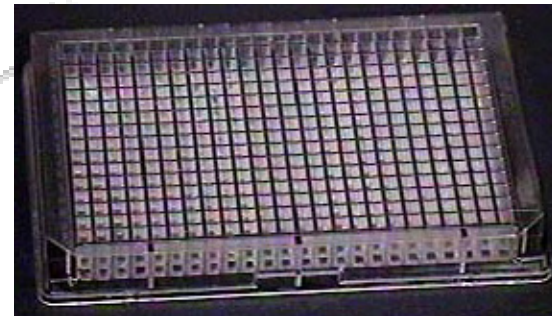


Sources of DNA contamination

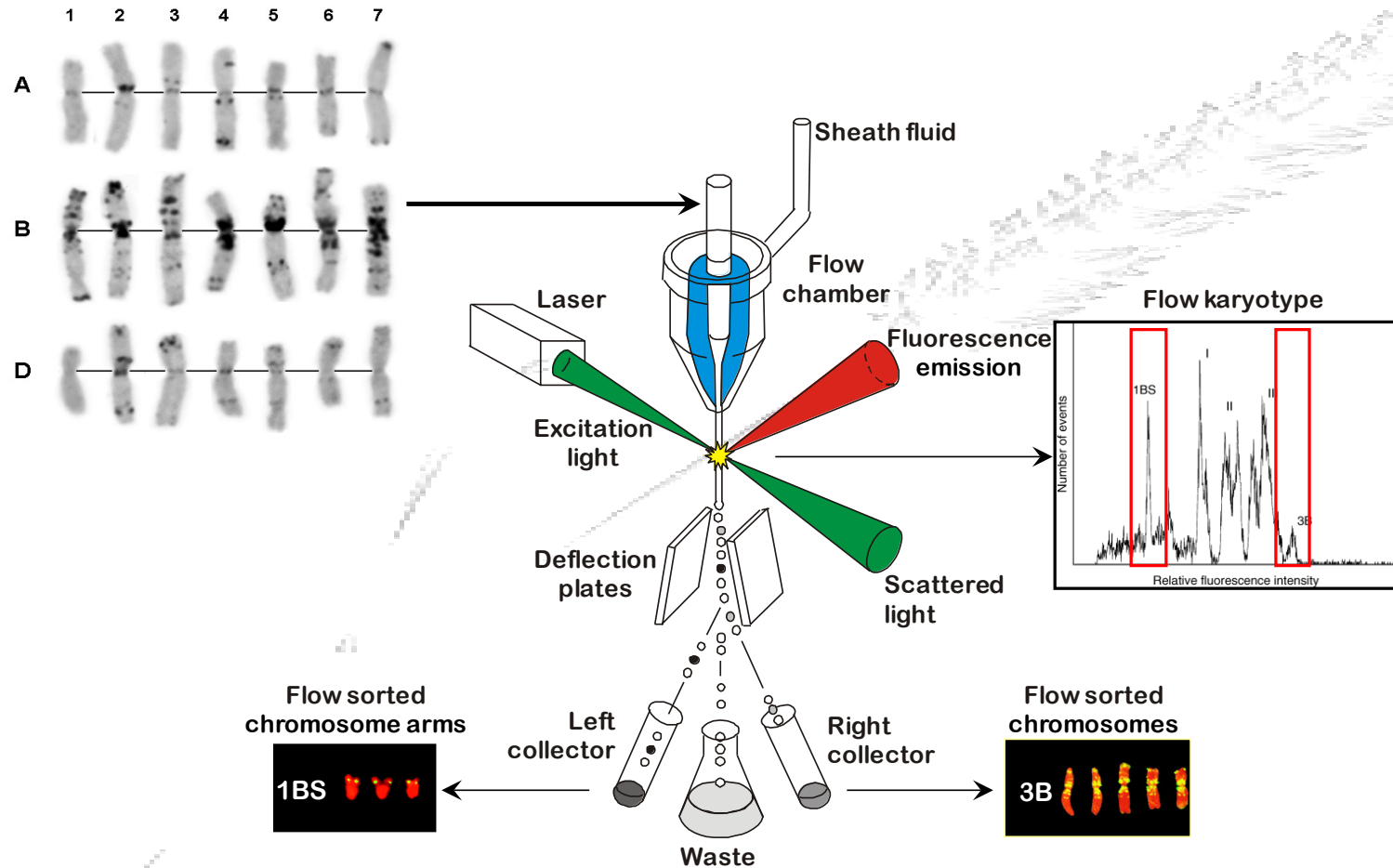
✓ Chloroplasmic DNA contamination



✓ Well-to-well contamination



Chloroplast DNA contamination

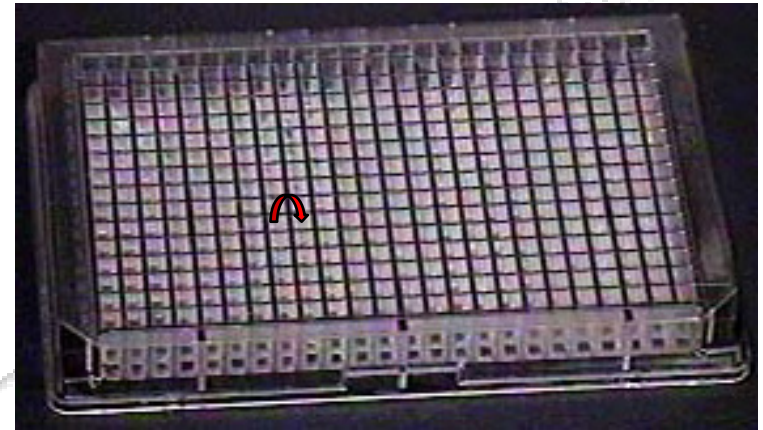
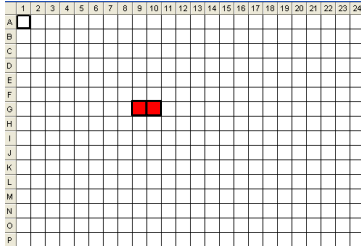


No chloroplast DNA contamination since chromosomes are flow-sorted and not simply extracted

(kindly of J. Dolezel)

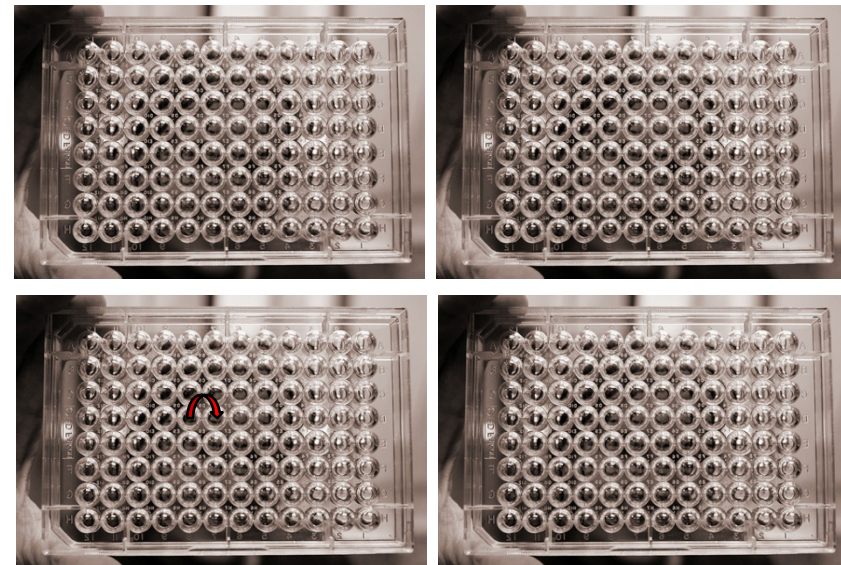
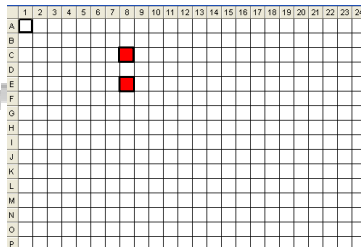
Well-to-well contamination

- ✓ Well-to-well contamination in **384-well plate** format
 - Adjacent wells showing similar profiles



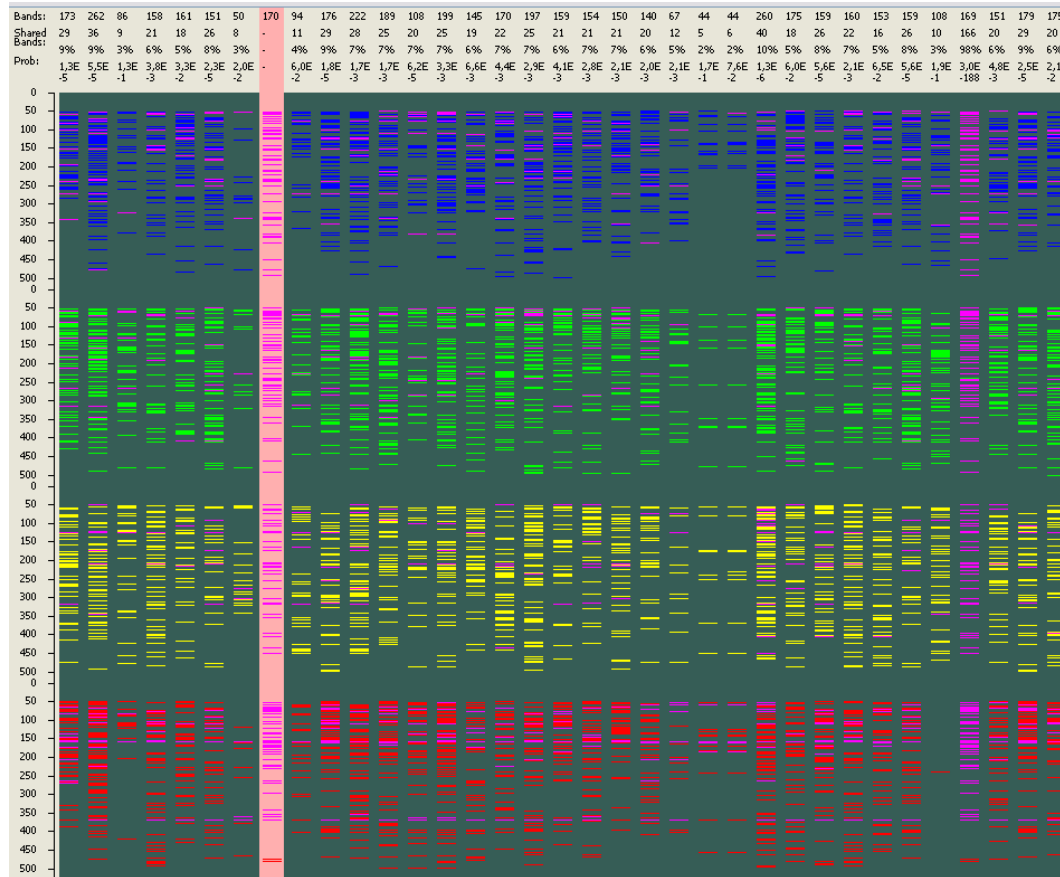
Splitting of 384-well plate into four 96-well plate during DNA extraction process.

- ✓ Well-to-well contamination in **96-well plate** format
 - Non-adjacent wells showing similar profiles



'One-to-one' contamination

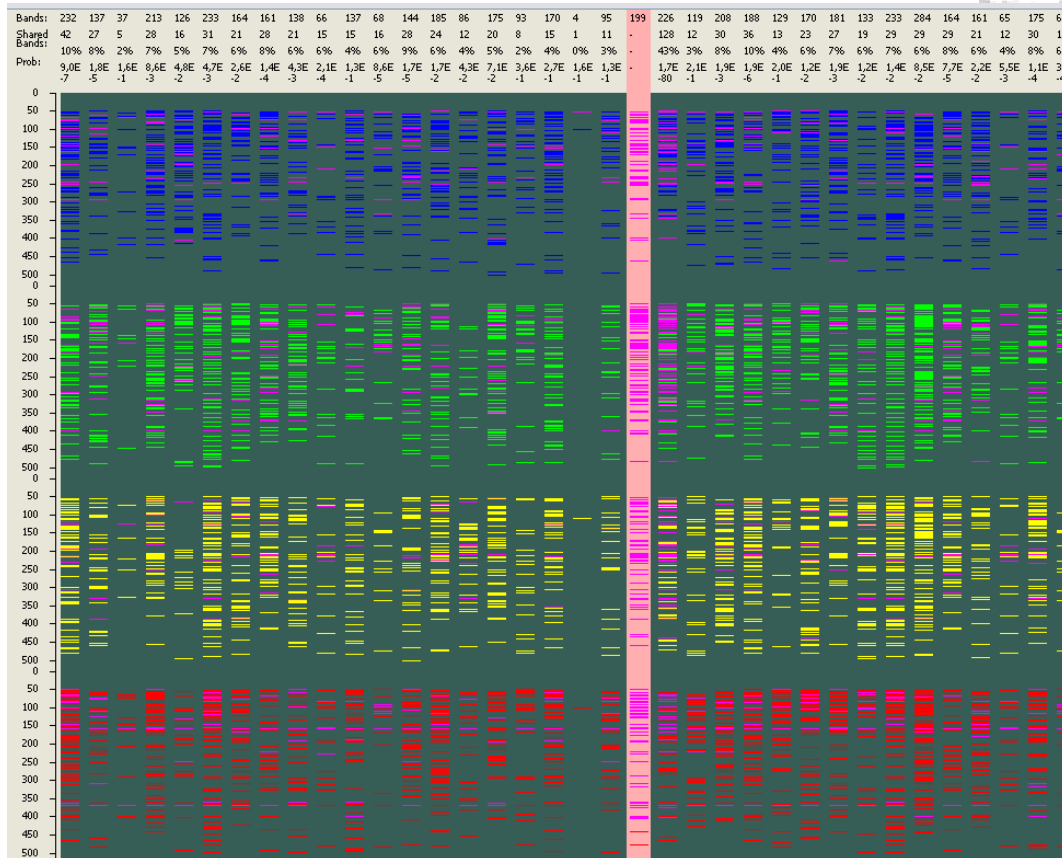
Two adjacent wells contain the same clone B1



80-100% identity of fingerprints

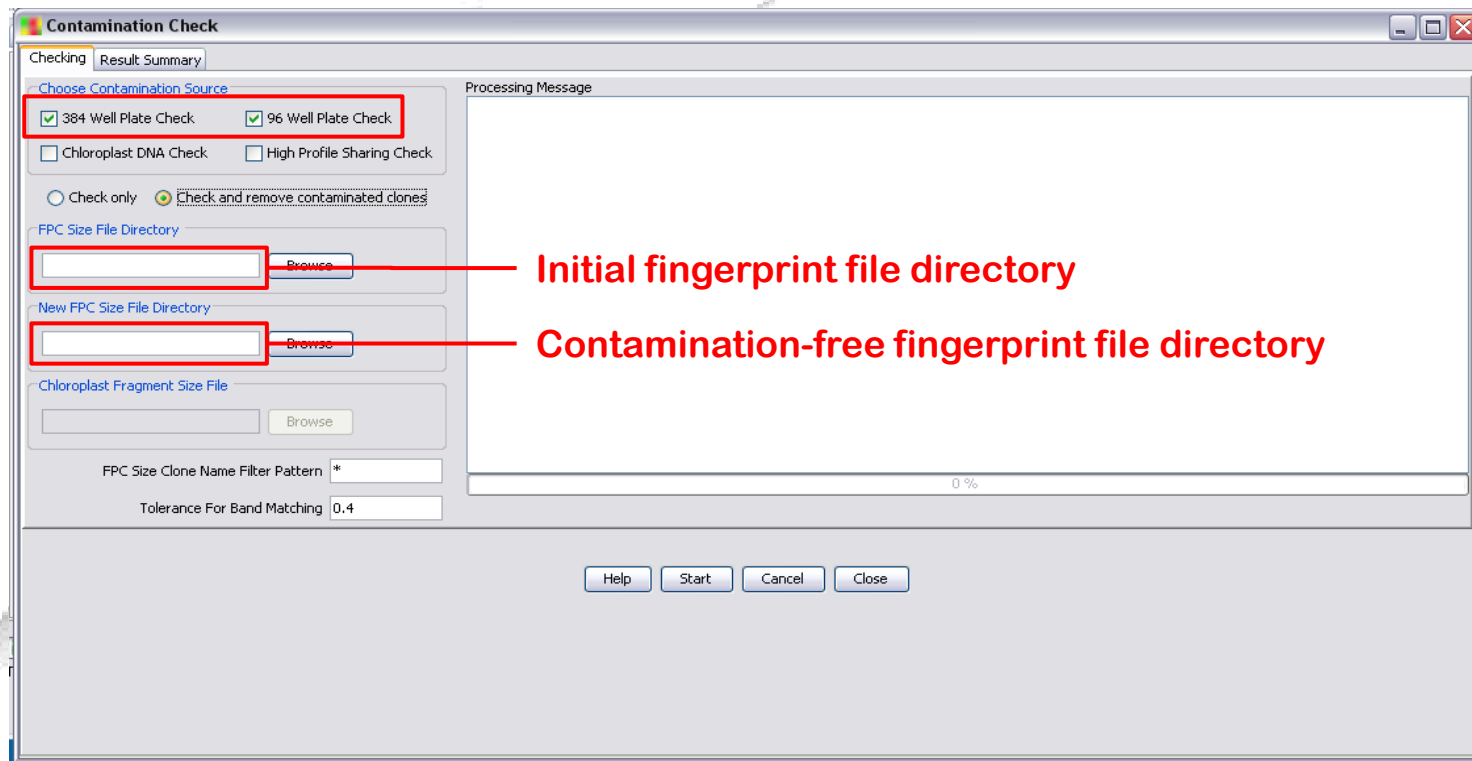
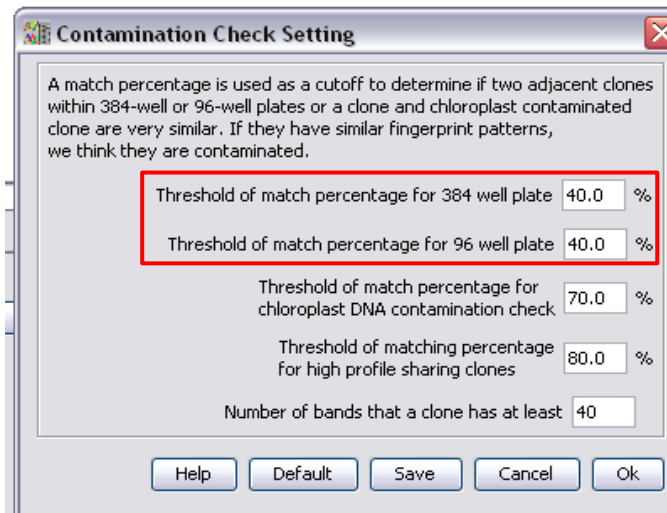
'One-to-two' contamination

One well contains one clone B1 and the adjacent one contains the same clone B1 and another one B2



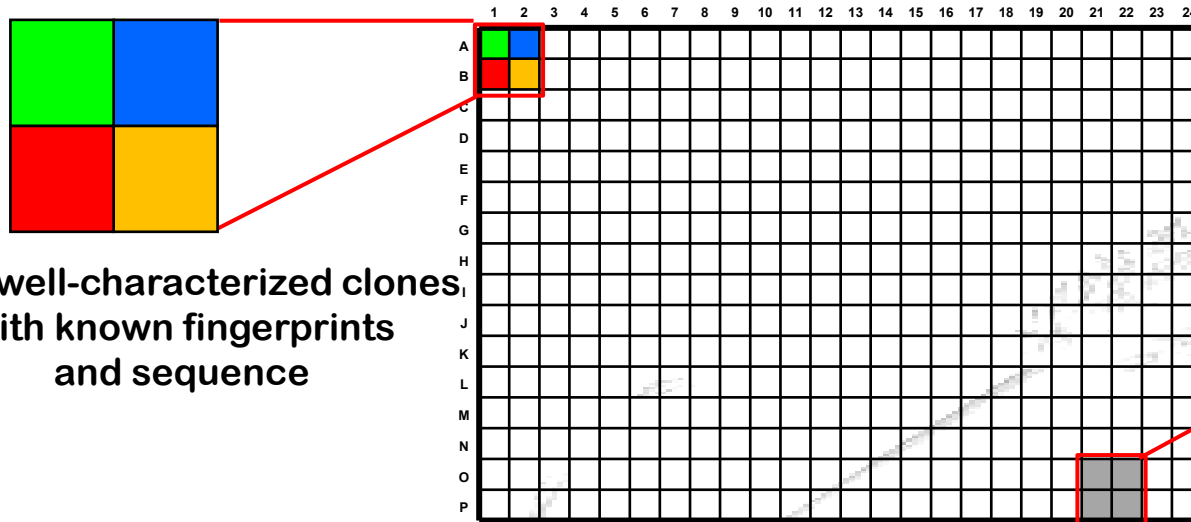
35-50% identity of fingerprints:
one of the well displays two merged fingerprints

Contamination removal using Genoprofiler





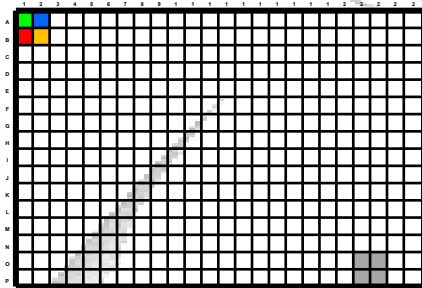
Control clones for quality check



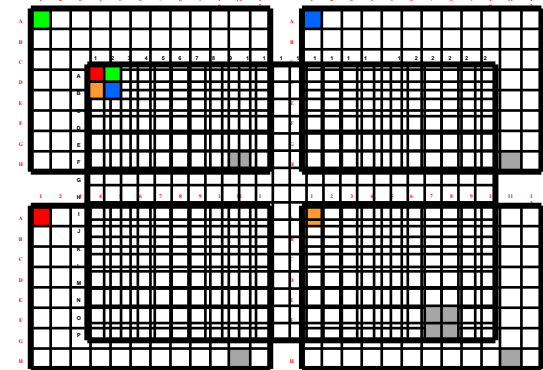
Four well-characterized clones
with known fingerprints
and sequence

Four empty wells

✓ Control of plate **rotation** or **inversion**



✓ Calculation of **contamination** rate



Removing control clones using Genoprofiler

The screenshot shows the 'File Management' window of Genoprofiler. Red lines and text annotations highlight specific fields:

- Input fingerprint file directory:** A red line points to the 'FROM: FPC Size File Directory' text and the text box below it, which is also enclosed in a red box.
- Output fingerprint file directory:** A red line points to the 'TO: New FPC Size File Directory' text and the text box below it, which is also enclosed in a red box.
- List of excluded clones (.txt file):** A red line points to the 'Excluded Clone List File' text box, which is also enclosed in a red box.

The window contains the following sections and controls:

- Subset Clones:** Includes tabs for 'Subset Clones', 'Rename Clones', and 'Remove Vector/Repeat Bands'.
- Choose Input File Type:** Radio buttons for 'FPC Size Files (*.sizes/ *.bands)' (selected), 'Sample Files (*.fsa)', and 'FPC File (*.fpc)'.
- FROM: FPC Size File Directory:** A text box and a 'Browse' button.
- TO: New FPC Size File Directory:** A text box and a 'Browse' button.
- Excluded Clone List File:** A text box and a 'Browse' button.
- Only Included Clone List File:** A text box and a 'Browse' button.
- Contig Name List (seperated by comma ","):** A text box.
- Clone Name Filter List (Wildcard Patterns):** A text box with an asterisk '*'.
- Set Clone Band Number Range:** Includes input fields for 'Total Bands From' (50) 'To' (250), 'Green Bands From' (0) 'To' (60), 'Red Bands From' (0) 'To' (60), 'Blue Bands From' (0) 'To' (60), and 'Yellow Bands From' (0) 'To' (60).
- Set Band Size Range (bp):** Includes input fields for 'Size From' (75) 'bp' 'To' (500) 'bp'.
- SubSet Clones:** A button.
- Help** and **Close** buttons at the bottom.

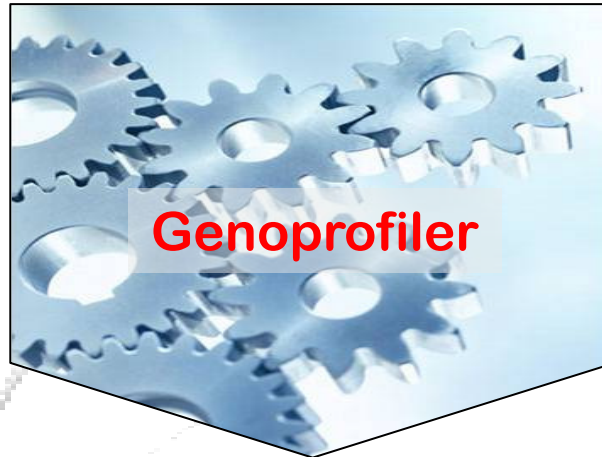
List of excluded clones (.txt file)

For example:

TaeCsp3DL023A01
TaeCsp3DL023A02
TaeCsp3DL023B01
TaeCsp3DL023B02
TaeCsp3DL023O21
TaeCsp3DL023O22
TaeCsp3DL023P21
TaeCsp3DL023P22

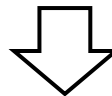
Genoprofiler output

FPB .sizes files



Genoprofiler.sizes files

- ✓ Contamination-free
- ✓ Control clone-free...



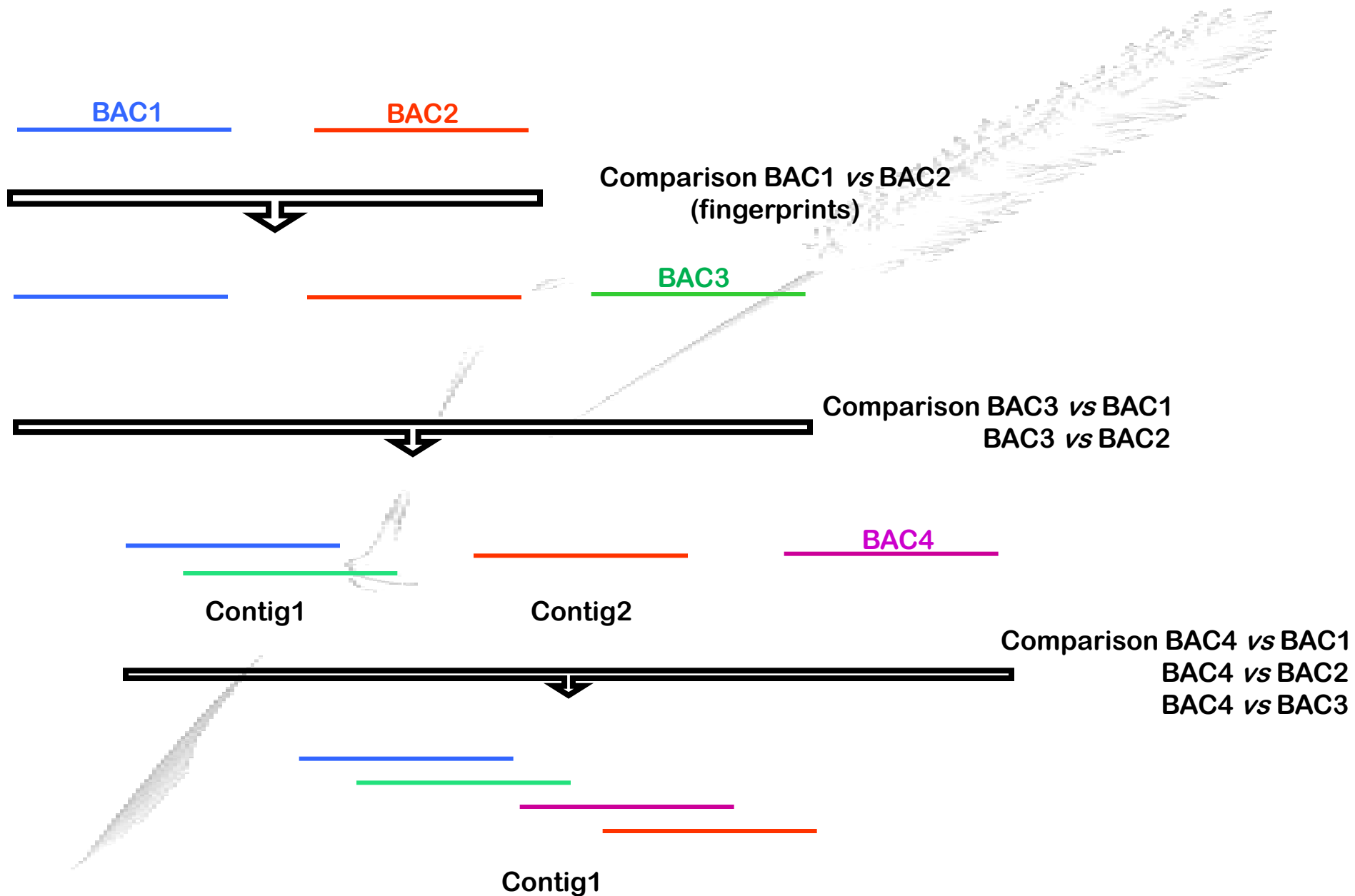
FPC

The background of the slide is a photograph of a lush green wheat field. The wheat stalks are in the foreground, slightly out of focus, with their green heads and long, narrow leaves visible. In the background, more rows of wheat stretch towards a distant, hazy horizon under a bright, clear sky. The overall tone is bright and natural.

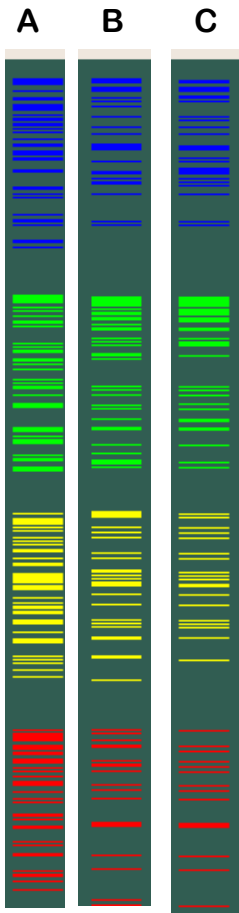
Contig assembly

1- Overview

Pairwise comparison and contig assembly



Overlap calculation: the Sulston score



A —————

B —————

C —————

FingerPrinted Contigs (FPC)

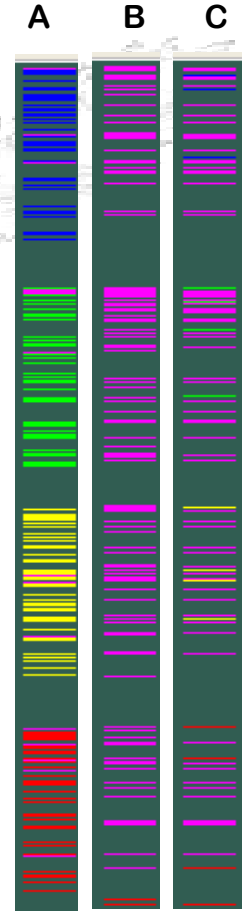
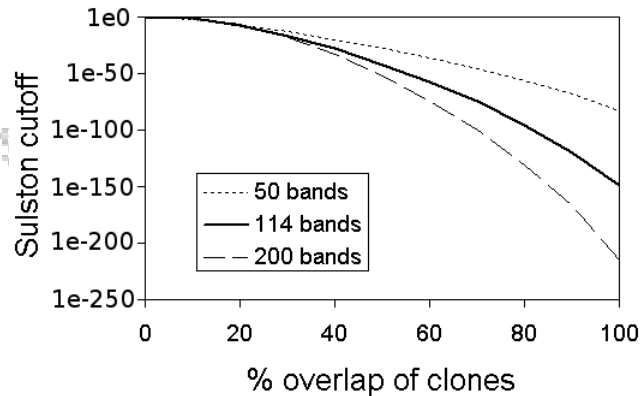
$$\sum_{m=M}^{nL} \left[\binom{nL}{m} ((1-p)^m p^{nL-m}) \right]$$

Tolerance for two bands to be identical

Number of possible values for bands

Number of bands for two clones

Number of shared bands

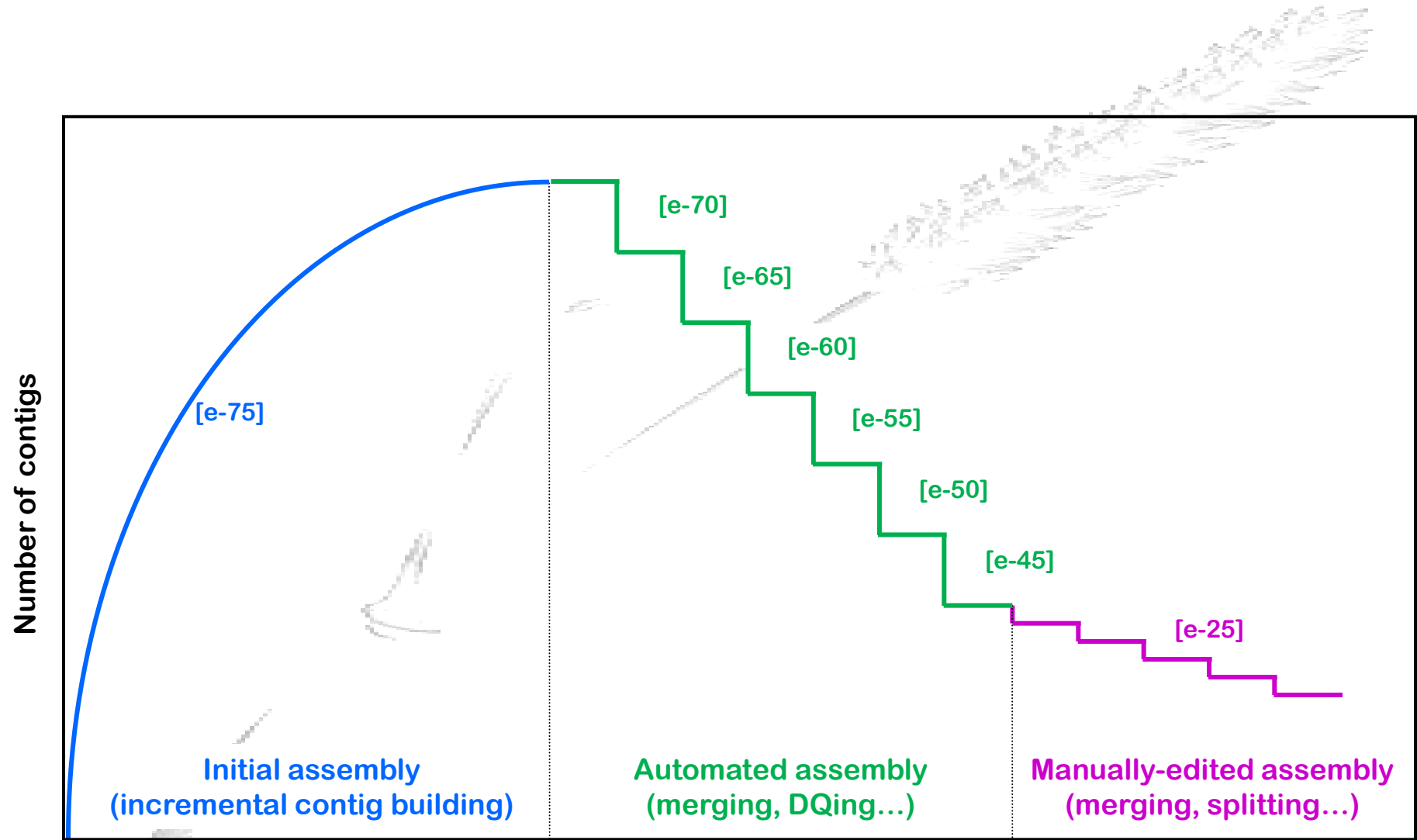


A —————

B —————

C —————

Assembly of the physical map



The background of the slide is a photograph of a lush green wheat field. The wheat stalks are in the foreground, slightly out of focus, with their green heads and long, narrow leaves visible. In the background, more rows of wheat stretch towards a distant, hazy horizon under a bright, clear sky. The overall tone is bright and natural.

Contig assembly 2- FPC overview

The background of the slide is a photograph of a lush green wheat field. The wheat stalks are in the foreground, slightly out of focus, with their green heads and long, narrow leaves visible. In the background, more rows of wheat stretch towards a distant, hazy horizon under a bright, clear sky. The overall tone is bright and natural.

Contig assembly

3- Initial assembly

Configuring FPC: configure window

For clones larger than 100 bands

Average band size
(based on fingerprints and sequences)

The screenshot shows the 'Configure Display' window with the following settings and annotations:

- Tolerance File:** [Empty text box]
- Variable Tolerance:** ☐
- Fast Sulston:** ☐ **Pure Sulston:** ☒ **Equation 2:** ☐ (An arrow points from the text 'For clones larger than 100 bands' to this section.)
- Genome size:** 0 kb
- Clone size:** 150000 b
- Band size:** 1100 b (An arrow points from the text 'Average band size (based on fingerprints and sequences)' to this field.)
- Gel length:** 54000 (An arrow points from the text 'Number of possible values for one band: (15,000 - 1500) x 4 = 54,000' to this field.)
- Contig display page size:** 3000
- Agarose:** ☐ **HICF:** ☒ (An arrow points from the text 'SNaPshot labelling & capillary sequencer' to this section.)
- Vector File:** [Empty text box]
- Close:** [Button]

Number of possible values for one band:
 $(15,000 - 1500) \times 4 = 54,000$

SNaPshot labelling & capillary sequencer

Building contigs

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-75 Bury~: 0.10

☒Precompute ☒Use CpM CpM Table

☐Log ☒Stdout Help

CB: Best contig of 100 Help

Build Contigs (Runs Kill first)

Kill Contig size <= 5 ☒Kill Seq Ctgs

Incremental Build Contigs ☐NoCB on Existing

Last Build 2/5/06 20:31 Cutoff 1e-45 CpM

DQer if >=10% Qs Step 3 ☐No merge CBmaps

ReBuild if ☐Q eq - ☒Q eq ~ Help

☐Auto Merge/Add FromEnd 55 Help

Ends-->Ends Match 1

KeySet-->Fpc ☐Ends Only ☐Include Ctg0

Clone: -->Fpc -->Key Help

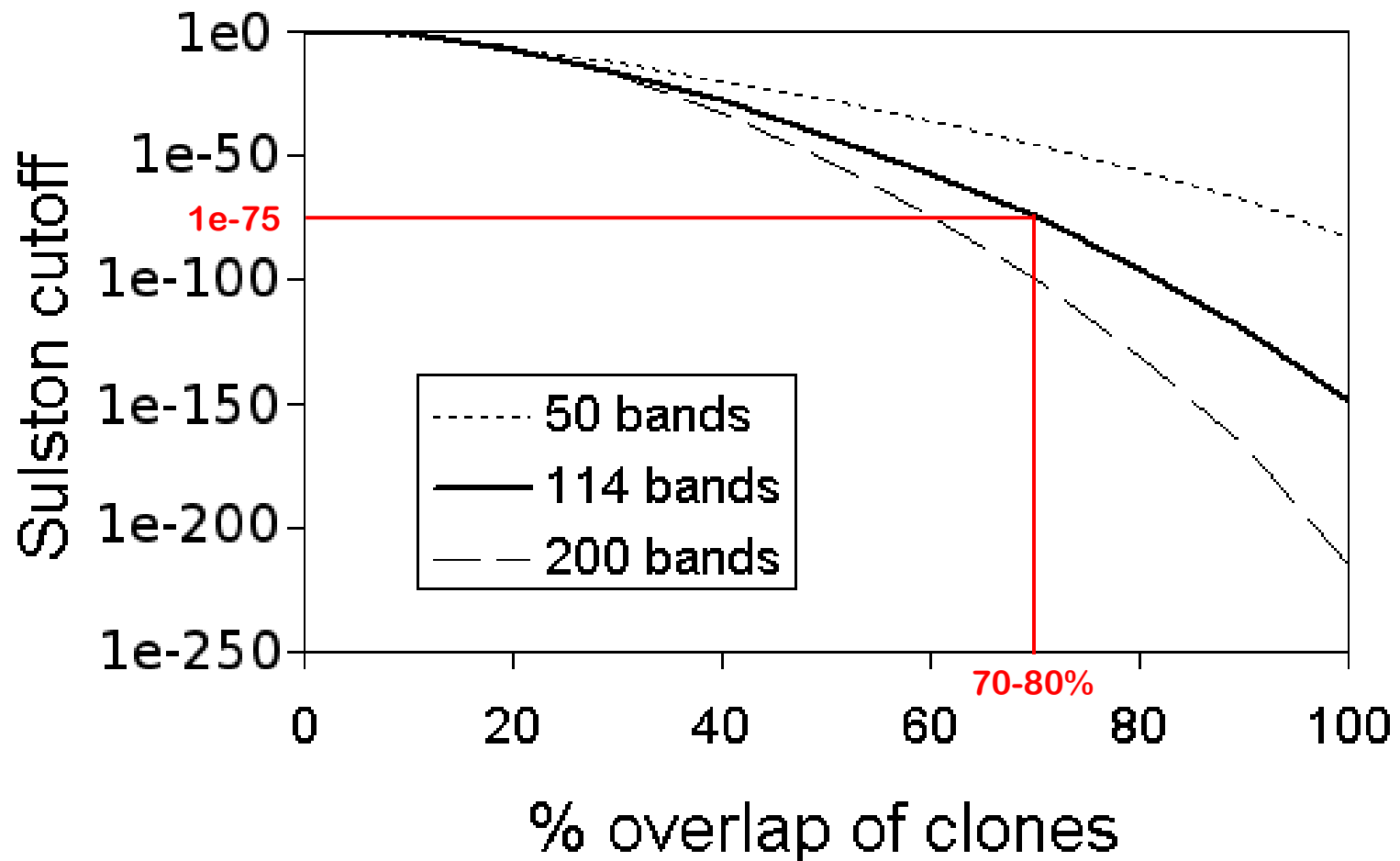
Close All functions are F4 interruptable

Start at very high stringency (1e-75)

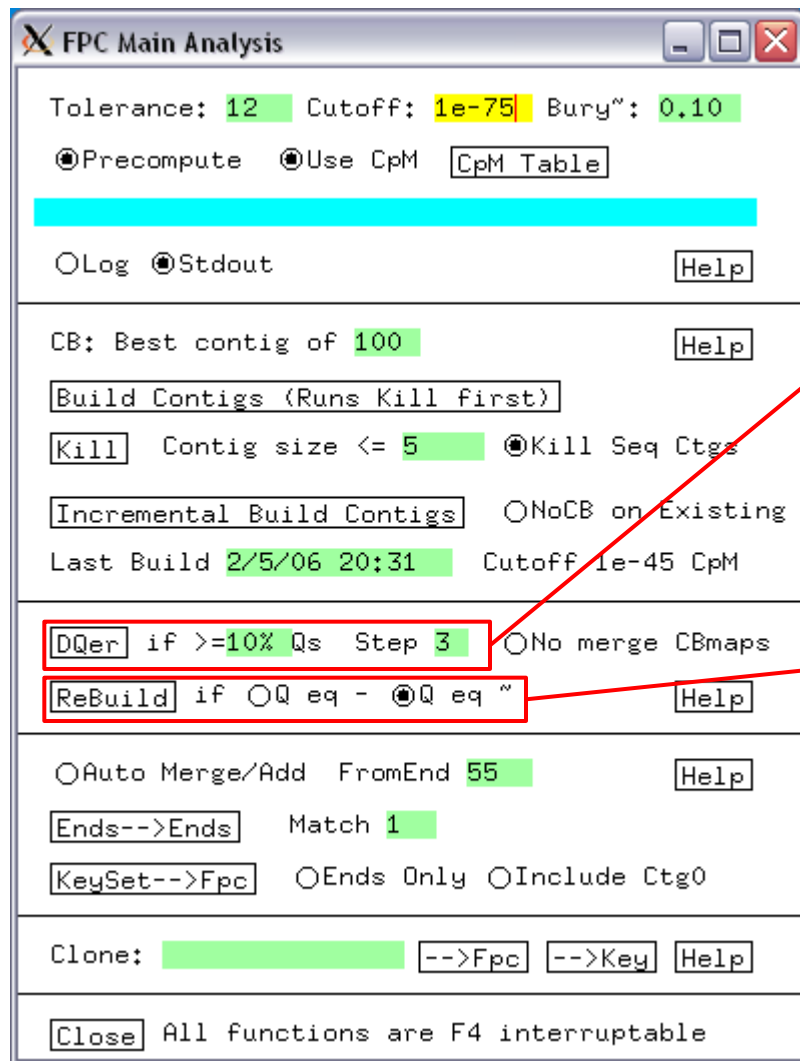
Start a new assembly

Compute newly added fingerprints

Sulston score overlap



DQing contigs



The screenshot shows the 'FPC Main Analysis' window. A red line points from the 'DQer' button to the text '1- DQer'. Another red line points from the 'ReBuild' button to the text '2- Rebuild modified contigs as the number of Qs is no longer reliable'. A third red line points from the 'DQer' button to the text '3- If necessary, perform a new DQer step, starting at 1e-84, followed by Rebuild...'. The window contains the following fields and buttons:

- Tolerance: 12 Cutoff: 1e-75 Bury~: 0.10
- ☒ Precompute ☒ Use CpM CpM Table
- ☐ Log ☒ Stdout Help
- CB: Best contig of 100 Help
- Build Contigs (Runs Kill first)
- Kill Contig size <= 5 ☒ Kill Seq Ctgs
- Incremental Build Contigs ☐ NoCB on Existing
- Last Build 2/5/06 20:31 Cutoff 1e-45 CpM
- DQer if >=10% Qs Step 3 ☐ No merge CBmaps
- ReBuild if ☐ Q eq - ☒ Q eq ~ Help
- ☐ Auto Merge/Add FromEnd 55 Help
- Ends-->Ends Match 1
- KeySet-->Fpc ☐ Ends Only ☐ Include Ctg0
- Clone: -->Fpc -->Key Help
- Close All functions are F4 interruptable

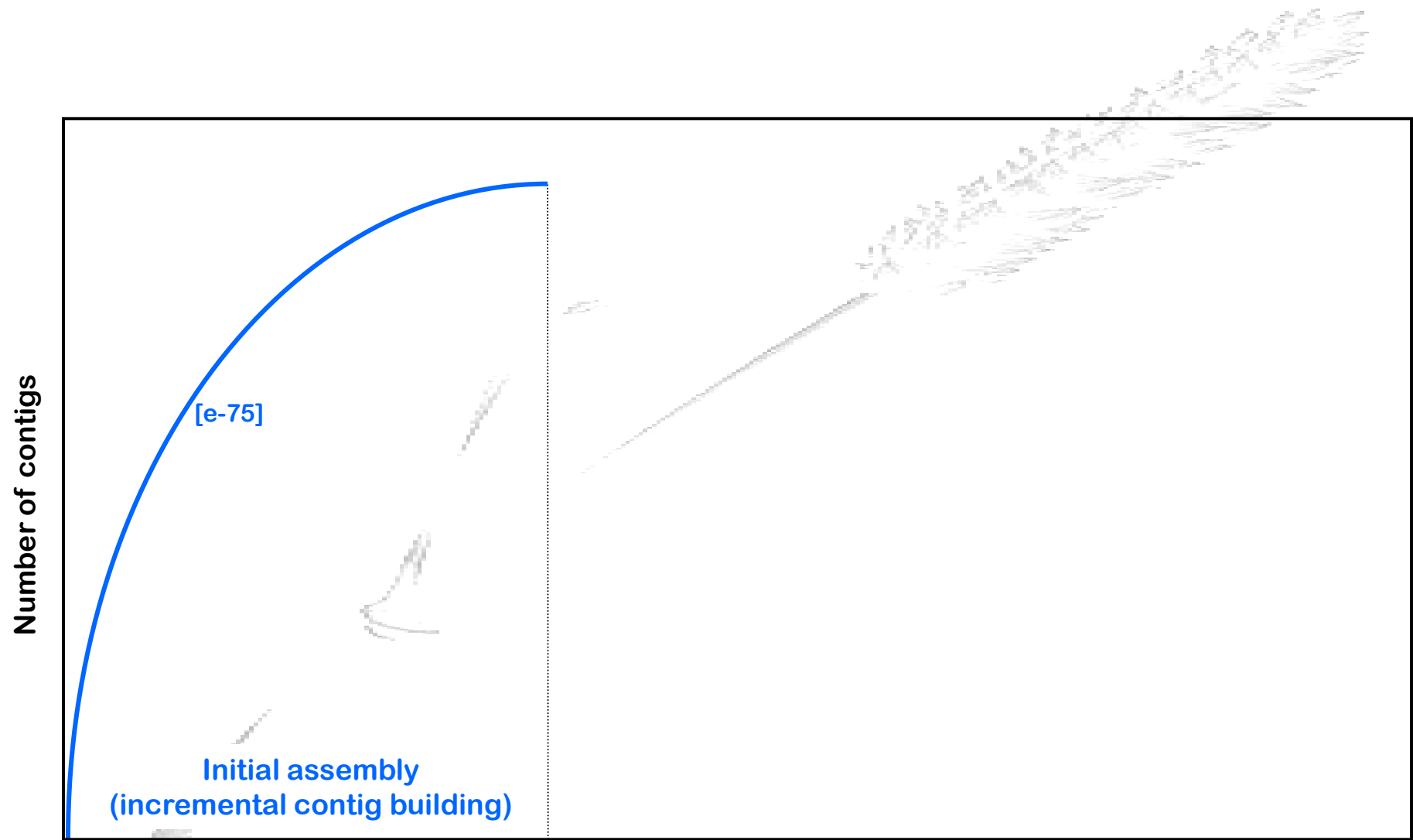
1- DQer

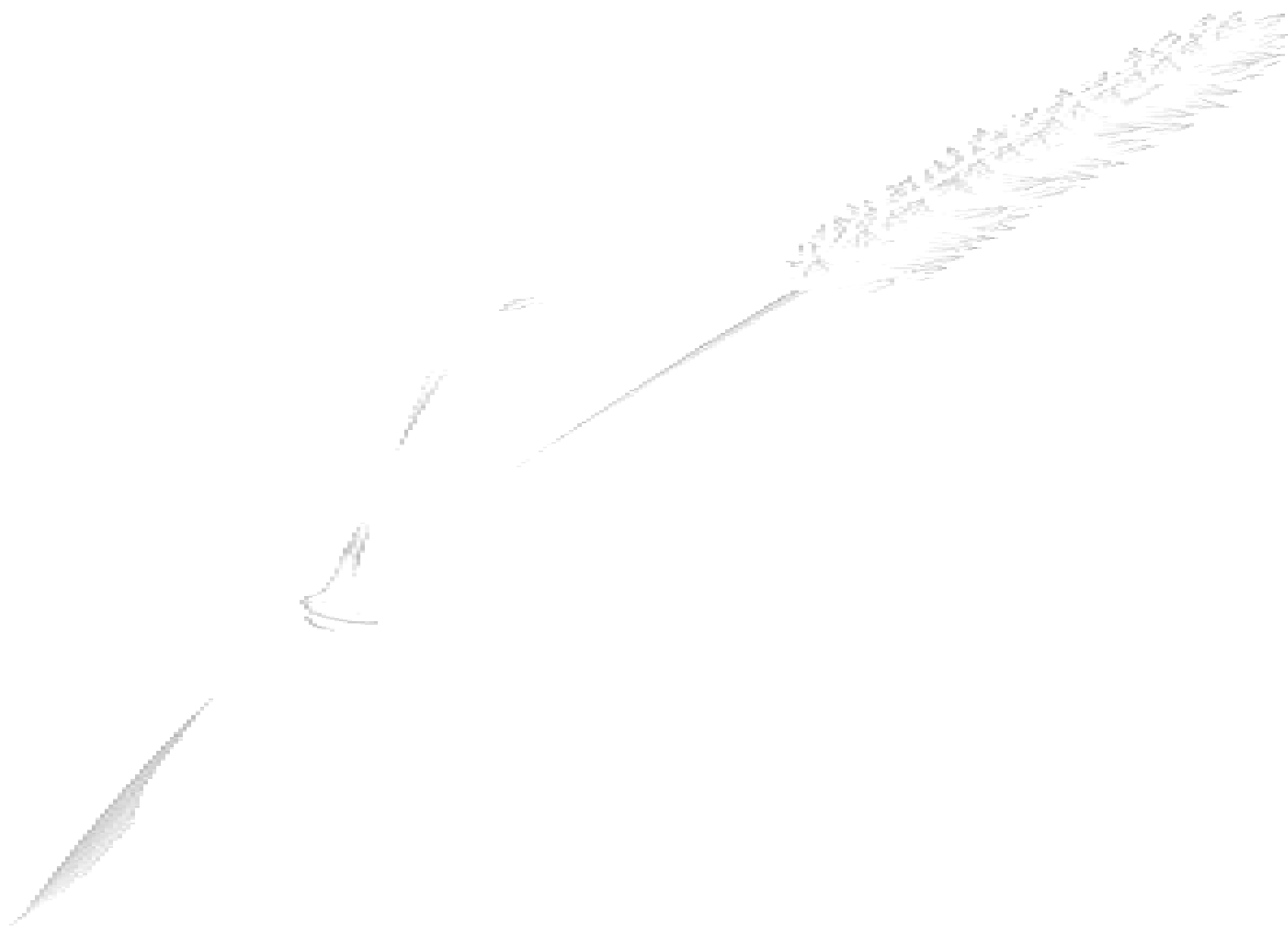
- ✓ decreasing the cut-off to remove Qs
- ✓ only for contigs having more than 10% Qs
- ✓ Three times (1e-78, 1e-81, 1e-84)

2- Rebuild modified contigs as the number of Qs is no longer reliable

3- If necessary, perform a new DQer step, starting at 1e-84, followed by Rebuild...

Assembly of the physical map





A background image of a lush green wheat field with several stalks in sharp focus in the foreground. The sky is a pale, clear blue.

Contig assembly

4- Automated assembly

Single-to-end merging

The screenshot shows the 'FPC Main Analysis' window. Several fields and buttons are highlighted with red boxes and red arrows pointing to explanatory text on the right. The fields include 'Tolerance: 12', 'Cutoff: 1e-70', 'Bury~: 0.10', 'FromEnd 55', 'Match 1', and 'KeySet-->Fpc'. Buttons include 'Precompute', 'Use CpM', 'CpM Table', 'Log', 'Stdout', 'Help', 'Build Contigs (Runs Kill first)', 'Kill', 'Incremental Build Contigs', 'ReBuild', 'Auto Merge/Add', 'Ends-->Ends', 'Clone: -->Fpc', and 'Close'.

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-70 Bury~: 0.10

☒Precompute ☒Use CpM CpM Table

☒Log ☒Stdout Help

CB: Best contig of 100 Help

Build Contigs (Runs Kill first)

Kill Contig size <= max ☒Kill Seq Ctgs

Incremental Build Contigs ☐NoCB on Existing

Last Build 13/8/09 13:54 Cutoff 1e-45 CpM

DQer if >=10% Qs Step 3 ☐No merge CBmaps

ReBuild if ☐Q eq - ☒Q eq ~ Help

☒Auto Merge/Add FromEnd 55 Help

Ends-->Ends Match 1

KeySet-->Fpc ☒Ends Only ☐Include Ctg0

Clone: -->Fpc -->Key Help

Close All functions are F4 interruptable

Decrease the stringency stepwise
(1e-70, 1e-65, 1e-55, 1e-50, e-45)

Select Automerge for automatic merging

FromEnd tells how close to the contig
end a clone must be in order to
count as an end-clone (1/2 the
number of bands in an average
clone)

Match tells the number of clones from
one contig that have to match with
another contig for merging

Start single-to-end merging (singletons
are added to contig end only)

End-to-end merging

The screenshot shows the 'FPC Main Analysis' window. Red boxes and arrows highlight specific settings: 'Tolerance: 12', 'Cutoff: 1e-70', 'Bury~: 0.10', 'Auto Merge/Add', 'FromEnd 55', 'Match 1', and the 'Ends-->Ends' button. A blue bar is also present above the 'Log' and 'Stdout' options.

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-70 Bury~: 0.10

☒ Precompute ☒ Use CpM CpM Table

☒ Log ☒ Stdout Help

CB: Best contig of 100 Help

Build Contigs (Runs Kill first)

Kill Contig size <= max ☒ Kill Seq Ctgs

Incremental Build Contigs ☐ NoCB on Existing

Last Build 13/8/09 13:54 Cutoff 1e-45 CpM

DQer if >=10% Qs Step 3 ☐ No merge CBmaps

ReBuild if ☐ Q eq - ☒ Q eq ~ Help

☒ Auto Merge/Add FromEnd 55 Help

Ends-->Ends Match 1

KeySet-->Fpc ☐ Ends Only ☐ Include Ctg0

Clone: -->Fpc -->Key Help

Close All functions are F4 interruptable

Decrease the stringency stepwise
(1e-70, 1e-65, 1e-55, 1e-50, e-45)

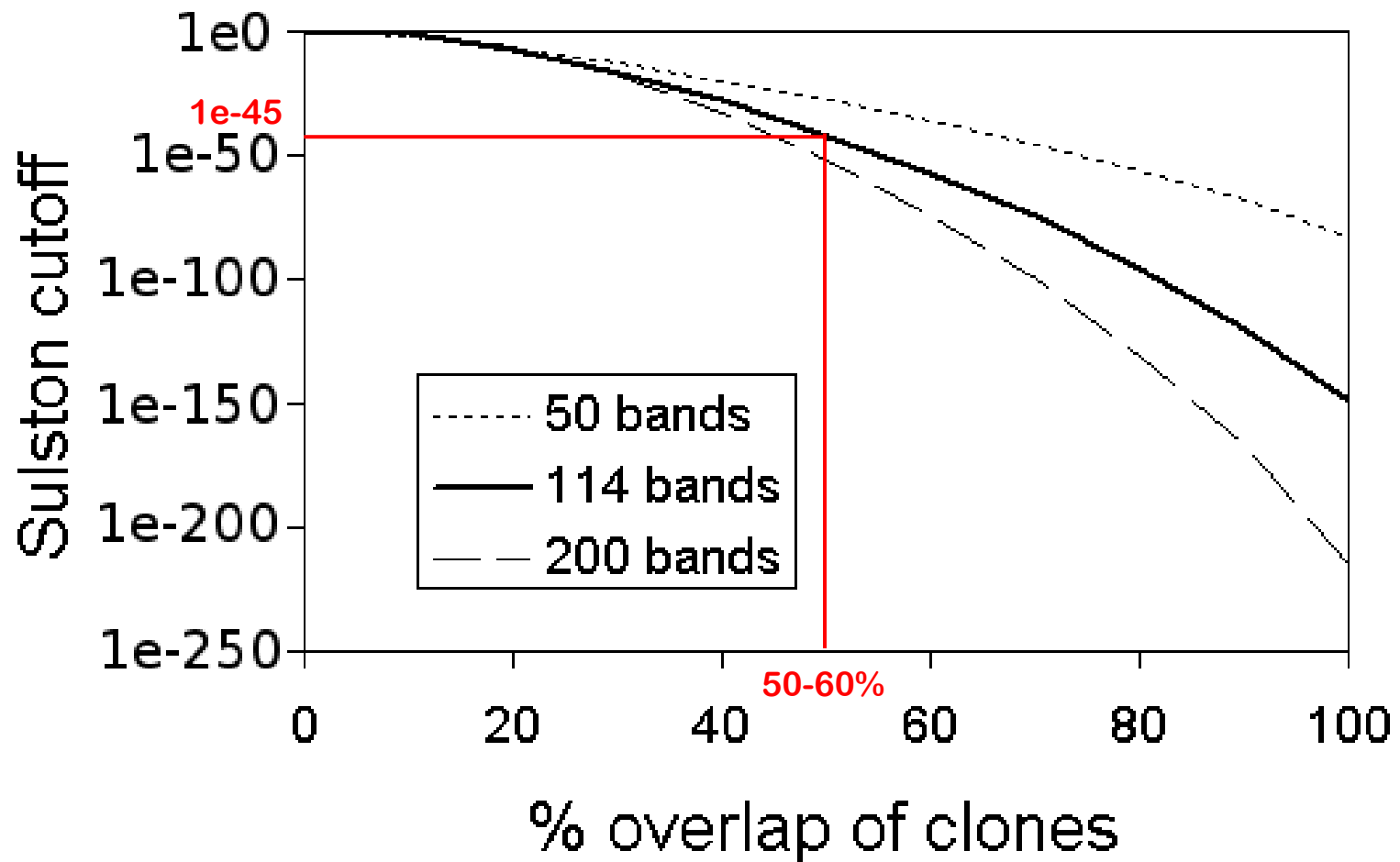
Select Automerge for automatic merging

FromEnd tells how close to the contig
end a clone must be in order to
count as an end-clone (1/2 the
number of bands in an average
clone)

Match tells the number of clones from
one contig that have to match with
another contig for merging

Perform end-to-end merging

Sulston score overlap



DQing contigs

The screenshot shows the 'FPC Main Analysis' window. Red boxes and arrows highlight specific settings: a box around 'Tolerance: 12 Cutoff: 1e-70 Bury~: 0.10' with an arrow pointing to '1- Rebuild contigs at merging stringency'; a box around 'DQer if >=10% Qs Step 3' with an arrow pointing to '2- DQer'; and a box around 'ReBuild if OQ eq - @Q eq ~' with an arrow pointing to '3- Rebuild modified contigs as the number of Qs is no longer reliable at merging stringency'. Other visible settings include 'Precompute', 'Use CpM', 'CpM Table', 'Log', 'Stdout', 'Help', 'CB: Best contig of 100', 'Build Contigs (Runs Kill first)', 'Kill', 'Contig size <= max', 'Kill Seq Ctgs', 'Incremental Build Contigs', 'NoCB on Existing', 'Last Build 13/8/09 13:54', 'Cutoff 1e-45 CpM', 'Auto Merge/Add', 'FromEnd 55', 'Ends-->Ends', 'Match 1', 'KeySet-->Fpc', 'Ends Only', 'Include Ctg0', 'Clone:', '-->Fpc', '-->Key', 'Help', and 'Close All functions are F4 interruptable'.

1- Rebuild contigs at merging stringency

2- DQer

✓ decreasing the cut-off to remove Qs

✓ only for contigs having more than 10% Qs

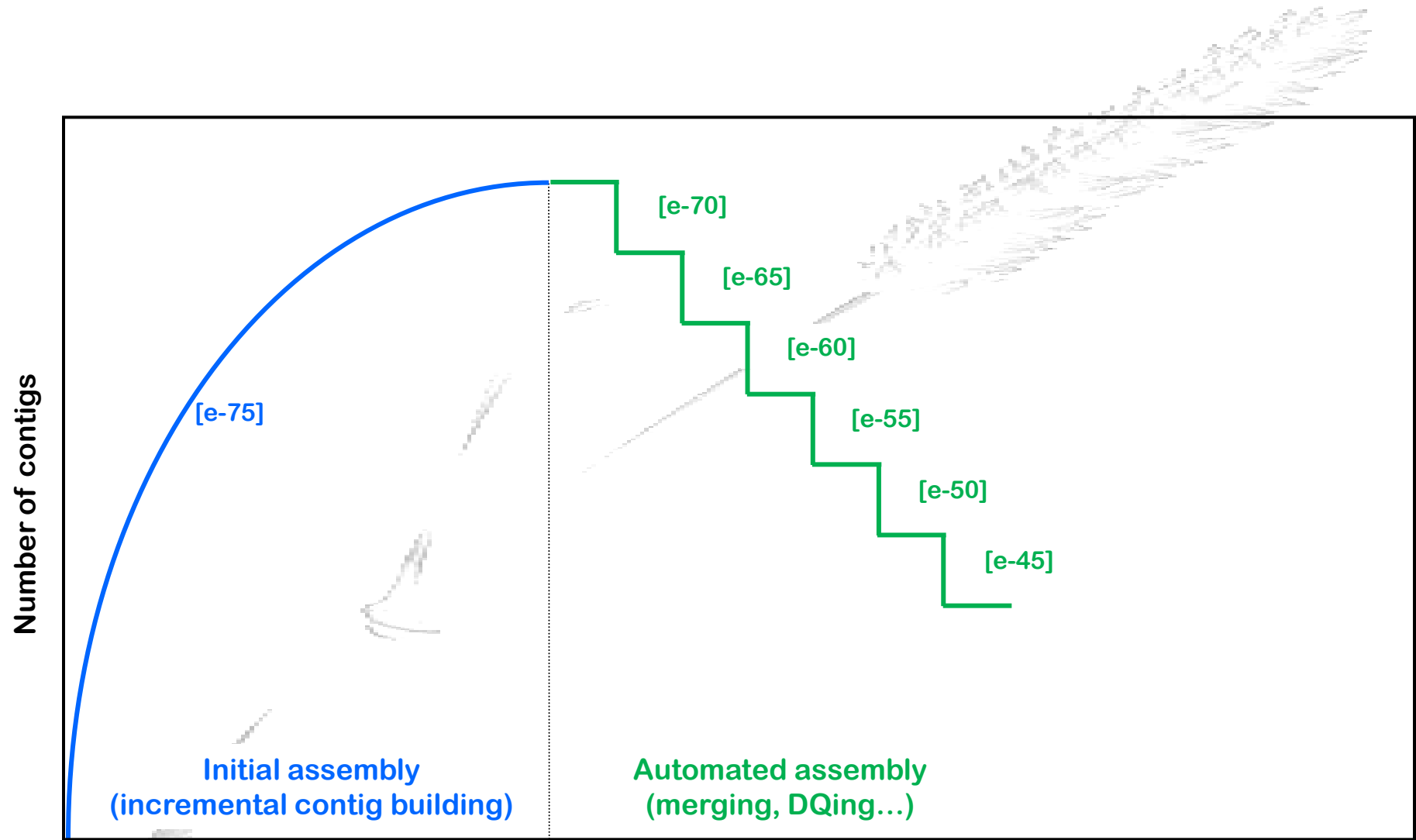
✓ Three times

3- Rebuild modified contigs as the number of Qs is no longer reliable at merging stringency

4- If necessary, perform a new DQer step,, followed by Rebuild...

5- Perform single-to-end and end-to-end merging until 1e-45

Assembly of the physical map



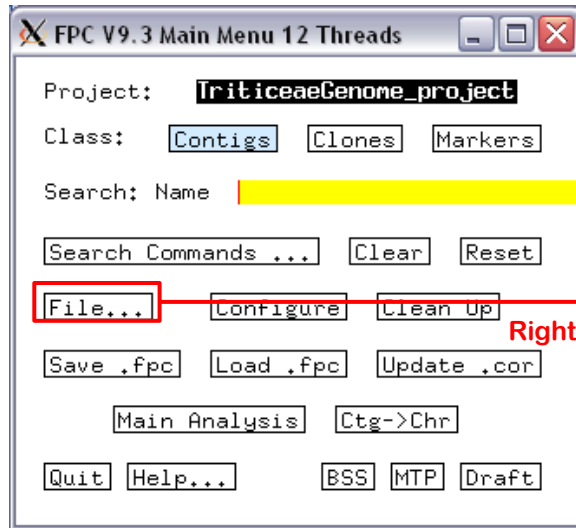


A background image of a lush green wheat field with several stalks in sharp focus in the foreground. The sky is a pale, clear blue.

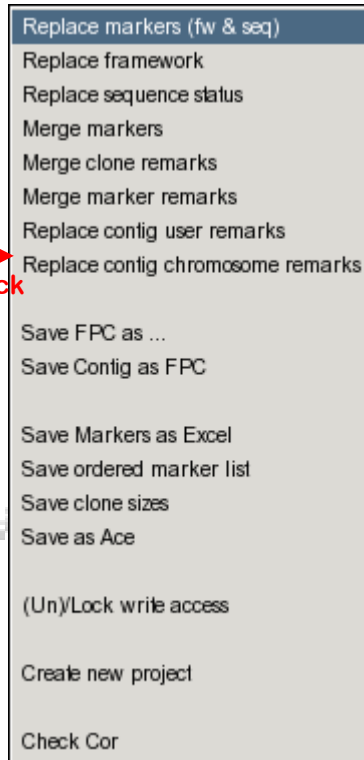
Contig assembly

5- Manually-edited assembly

Adding markers



Right click

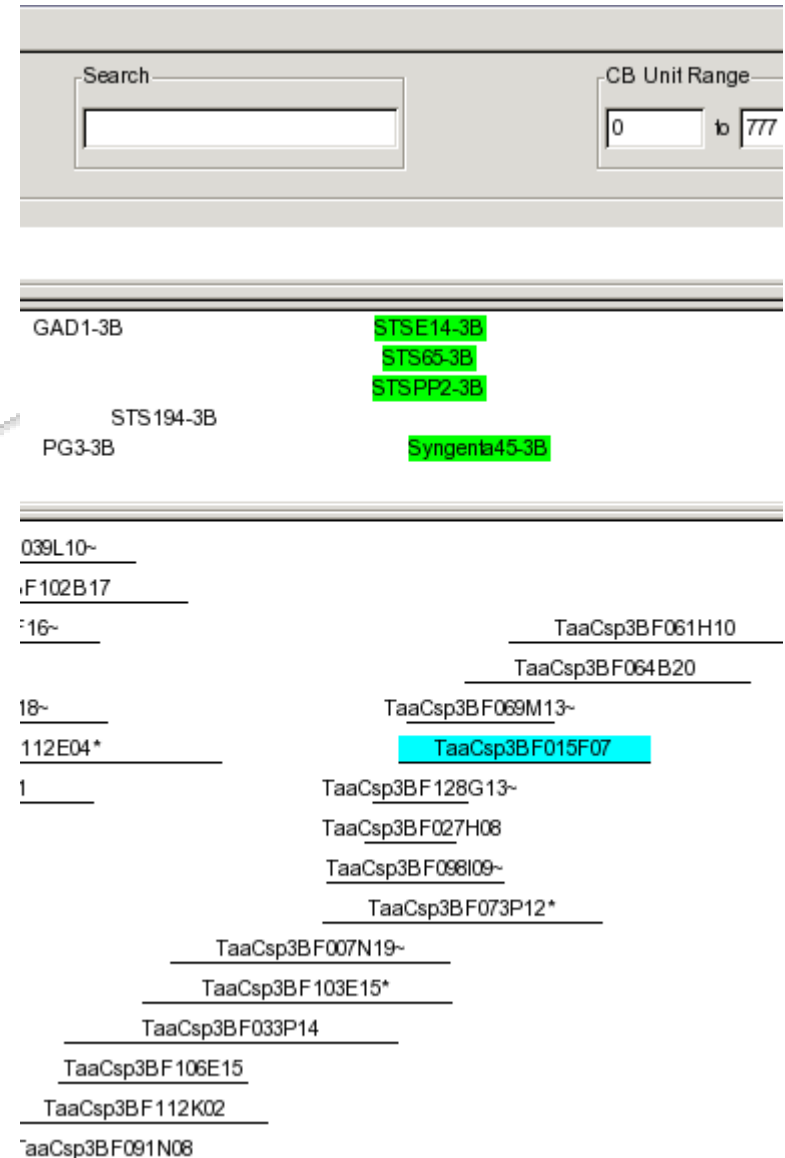


Marker.ace file

```
Clone : "TaaCsp3BF015F07"
Positive_STS "STS65-3B"
Positive_STS "STSP2-3B"
Positive_STS "STSE14-3B"
Positive_STS "Syngenta45-3B"

Clone : "TaaCsp3BF015F08"
```

Files/



Looking for small overlaps

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-25 Bury: 0.10

☐ Precompute ☒ Use CpM

☒ Log ☒ Stdout

CB: Best contig of 100

Contig size <= 5 ☒ Kill Seq Ctgs

☐ NoCB on Existing

Last Build 2/5/06 20:31 Cutoff 1e-25 CpM

if >= 10% Qs Step 1 ☐ No merge CBmaps

if ☐ Q eq - ☒ Q eq ~

☐ Auto Merge/Add FromEnd 55

Match 2

☐ Ends Only ☐ Include Ctg0

Clone:

All functions are F4 interruptable

✓ Stdout (screen)

```

CTG4  R  TaaCsp3BF047P12  94b  CTG16  L  TaaCsp3BF096G20  114b  Match  43  6e-31
CTG4  L  TaaCsp3BF062K17  117b  CTG535 B  TaaCsp3BF077D16  137b  Match  52  2e-32
CTG4  L  TaaCsp3BF062K17  117b  CTG535 L  TaaCsp3BF137B13   79b  Match  48  1e-41
CTG4  L  TaaCsp3BF062K17  117b  CTG535 L  TaaCsp3BF168D06   81b  Match  45  1e-36
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF037E18  128b  Match  50  9e-36
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF092H10  154b  Match  47  9e-29
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF098K16   62b  Match  40  5e-39
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF153J05   58b  Match  35  7e-33
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF037E18  128b  Match  49  8e-35
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF092H10  154b  Match  48  4e-30
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF098K16   62b  Match  40  3e-39
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF153J05   58b  Match  35  5e-33
Match: 5L 49R cutoff:9e-36
  
```

✓ .log file

```

CTG4  R  TaaCsp3BF047P12  94b  CTG16  L  TaaCsp3BF096G20  114b  Match  43  6e-31
CTG4  L  TaaCsp3BF062K17  117b  CTG535 B  TaaCsp3BF077D16  137b  Match  52  2e-32
CTG4  L  TaaCsp3BF062K17  117b  CTG535 L  TaaCsp3BF137B13   79b  Match  48  1e-41
CTG4  L  TaaCsp3BF062K17  117b  CTG535 L  TaaCsp3BF168D06   81b  Match  45  1e-36
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF037E18  128b  Match  50  9e-36
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF092H10  154b  Match  47  9e-29
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF098K16   62b  Match  40  5e-39
CTG5  L  TaaCsp3BF113B08  100b  CTG49  R  TaaCsp3BF153J05   58b  Match  35  7e-33
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF037E18  128b  Match  49  8e-35
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF092H10  154b  Match  48  4e-30
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF098K16   62b  Match  40  3e-39
CTG5  L  TaaCsp3BF140A15   99b  CTG49  R  TaaCsp3BF153J05   58b  Match  35  5e-33
Match: 5L 49R cutoff:9e-36
  
```

Match 2

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-25 Bury: 0.10

☐ Precompute ☒ Use CpM

☒ Log ☐ Stdout

CB: Best contig of 100

Contig size <= 5 ☒ Kill Seq Ctgs

☐ NoCB on Existing

Last Build 2/5/06 20:31 Cutoff 1e-25 CpM

if >=10% Qs Step 1 ☐ No merge CBmaps

if ☐ Q eq - ☒ Q eq ~

☐ Auto Merge/Add FromEnd 55

Match 2

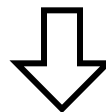
☐ Ends Only ☐ Include Ctg0

Clone:

All functions are F4 interruptable

Ctg4	R	TaaCsp3BF047P12	94b	Ctg16	L	TaaCsp3BF096G20	114b	Match	43	6e-31
Ctg4	L	TaaCsp3BF062K17	117b	Ctg535	B	TaaCsp3BF077D16	137b	Match	52	2e-32
Ctg4	L	TaaCsp3BF062K17	117b	Ctg535	L	TaaCsp3BF137B13	79b	Match	48	1e-41
Ctg4	L	TaaCsp3BF062K17	117b	Ctg535	L	TaaCsp3BF168D06	81b	Match	45	1e-36
Ctg5	L	TaaCsp3BF113B08	100b	Ctg49	R	TaaCsp3BF037E18	128b	Match	50	9e-36
Ctg5	L	TaaCsp3BF113B08	100b	Ctg49	R	TaaCsp3BF092H10	154b	Match	47	9e-29
Ctg5	L	TaaCsp3BF113B08	100b	Ctg49	R	TaaCsp3BF098K16	62b	Match	40	5e-39
Ctg5	L	TaaCsp3BF113B08	100b	Ctg49	R	TaaCsp3BF153J05	58b	Match	35	7e-33
Ctg5	L	TaaCsp3BF140A15	99b	Ctg49	R	TaaCsp3BF037E18	128b	Match	49	8e-35
Ctg5	L	TaaCsp3BF140A15	99b	Ctg49	R	TaaCsp3BF092H10	154b	Match	48	4e-30
Ctg5	L	TaaCsp3BF140A15	99b	Ctg49	R	TaaCsp3BF098K16	62b	Match	40	3e-39
Ctg5	L	TaaCsp3BF140A15	99b	Ctg49	R	TaaCsp3BF153J05	58b	Match	35	5e-33

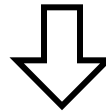
Match: 5L 49R cutoff: 9e-36



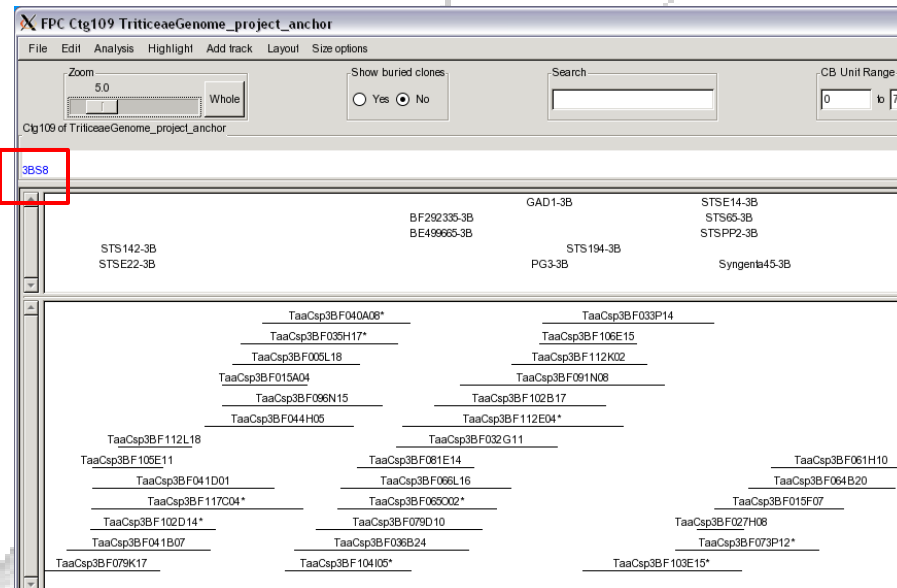
Perform merging
(unless mapping data are conflicting)

Match 1

Ctg107	R	TaaCsp3BF082P11	162b	Ctg215	L	TaaCsp3BF007K02	104b	Match	54	1e-34
Ctg109	L	TaaCsp3BF079K17	104b	Ctg593	L	TaaCsp3BF099P24	90b	Match	41	6e-31
Ctg109	L	TaaCsp3BF079K17	104b	Ctg593	L	TaaCsp3BF129B01	70b	Match	40	2e-35
Ctg109	L	TaaCsp3BF079K17	104b	Ctg593	L	TaaCsp3BF167I13	117b	Match	41	5e-26
Ctg112	L	TaaCsp3BF076D24	122b	Ctg440	B	TaaCsp3BF106C03	148b	Match	59	2e-37
Ctg112	L	TaaCsp3BF083A16	153b	Ctg440	B	TaaCsp3BF106C03	148b	Match	52	2e-24

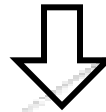


Check mapping data
& perform merging if mapping data are consistent



Conflicting results

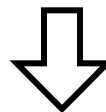
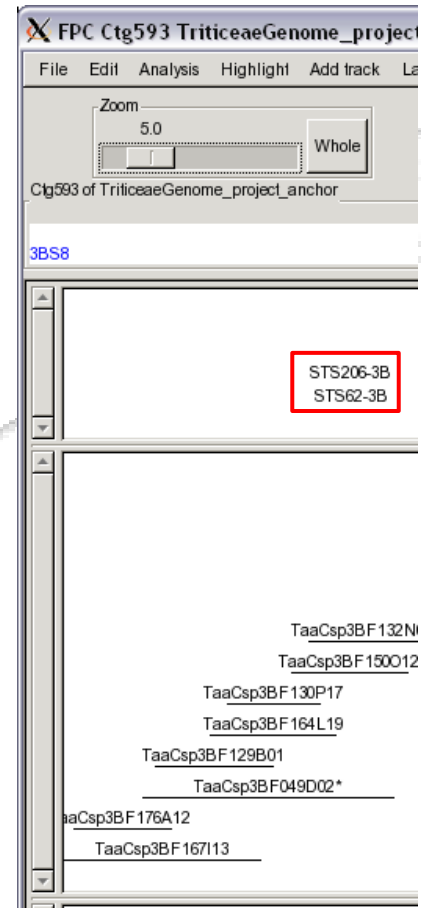
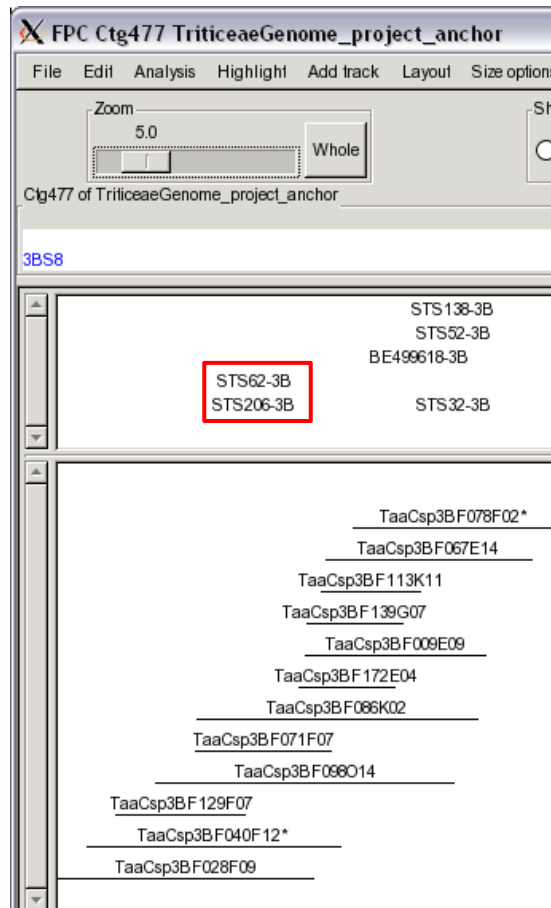
Ctg7	R	TaaCsp3BF155A17	64b	Ctg728	R	TaaCsp3BF146A02	90b	Match	36	9e-34
Ctg7	R	TaaCsp3BF155A17	64b	Ctg728	R	TaaCsp3BF147C07	87b	Match	40	1e-40
Ctg7	R	TaaCsp3BF155A17	64b	Ctg728	R	TaaCsp3BF149M19	72b	Match	41	2e-45
Ctg7	R	TaaCsp3BF166D21	64b	Ctg728	R	TaaCsp3BF060G18	194b	Match	40	2e-27
Ctg7	R	TaaCsp3BF166D21	64b	Ctg728	R	TaaCsp3BF109M02	99b	Match	39	7e-37
Ctg7	R	TaaCsp3BF166D21	64b	Ctg728	R	TaaCsp3BF149M19	72b	Match	29	1e-26
Match: 7R 728R cutoff:2e-45										
Ctg7	R	TaaCsp3BF002K10	178b	Ctg742	L	TaaCsp3BF014L24	147b	Match	59	7e-28
Ctg7	R	TaaCsp3BF002K10	178b	Ctg742	L	TaaCsp3BF055P11	132b	Match	60	6e-32
Ctg7	R	TaaCsp3BF002K10	178b	Ctg742	L	TaaCsp3BF127A14	61b	Match	38	2e-27
Ctg7	R	TaaCsp3BF002K10	178b	Ctg742	L	TaaCsp3BF152K02	74b	Match	44	3e-30
Ctg7	R	TaaCsp3BF007023	165b	Ctg742	L	TaaCsp3BF152K02	74b	Match	42	4e-29
Match: 7R 742L cutoff:4e-29										



Check manually

- ✓ **Small contig included into the others**
- ✓ **Chimeric clones...**

No match but shared markers



**Perform merging
(if marker data are reliable)**

Looking for small overlaps

FPC V9.3 Main Menu 12 Threads

Project: **TriticeaeGenome_project**

Class:

Search: Name **TaaCsp3BF079K17**

Clone Commands

Clone Name	Substring
Before created date	dd/mm/yy hh:mm
After created date	dd/mm/yy hh:mm
Before modified date	dd/mm/yy hh:mm
After modified date	dd/mm/yy hh:mm
Remark	Substring of remark
Without Remark	Substring of remark
Gel	Identifier
<input checked="" type="button" value="Match Clone"/> <input type="button" value="Clone name"/>	
Contig	Number
> N Bands	Number

Singletons

Multiple Fingerprints
No Fingerprints

Selected (Ignore current keyset)
Canceled (Ignore current keyset)

Full-X
Half-X
Gap-closure
Contaminated

Contig Evaluate

Tolerance: 4 Cutoff: **1e-45** Bury~: 0.10

☒ Log ☒ Stdout ☒ Use CpM

Some functions are F4 interruptable

Useful to check MTP results when clones belong to 2 different contigs.

Killing small contigs

FPC Main Analysis

Tolerance: 12 Cutoff: 1e-25 Bury~: 0.10

☒Precompute ☒Use CpM

☐Log ☒Stdout

CB: Best contig of 100

Contig size <= 5 ☒Kill Seq Ctgs

☐NoCB on Existing

Last Build 2/5/06 20:31 Cutoff 1e-25 CpM

if >=10% Qs Step 1 ☐No merge CBmaps

if ☐Q eq - ☒Q eq ~

☐Auto Merge/Add FromEnd 55

Match 1

☐Ends Only ☐Include Ctg0

Clone:

All functions are F4 interruptable

Kill contigs containing less than 6 clones
(‘max’ to kill all the contigs)

Killing small contigs

Project TriticeaeGenome_project Page 9.0 of 11.0

FPC TriticeaeGenome_project Clones 7246 Seq 0 Markers 0
9.3 Date: 14:07 Thu 27 Aug 2009 User: epaux
TotalLen 120660 kb AvgLen 574 kb

By length.. Help
Chr_Remark
Search Summary

Contig	Clone	Marker	Seq	Draft	Date	Status	Qs	Chr_Remark
476	7	-	-	-	225 196	0	0	
487	12	-	-	-	221 193	0	1	
635	11	-	-	-	215 187	0	0	
101	4	-	-	-	213 186	0	0	
162	11	-	-	-	212 185	0	0	
538	9	-	-	-	212 185	0	0	
276	3	-	-	-	211 184	0	0	
392	7	-	-	-	211 184	0	0	
567	8	-	-	-	210 183	0	0	
383	10	-	-	-	209 179	0	0	
560	8	-	-	-	205 179	0	0	
436	7	-	-	-	204 178	0	0	
507	6	-	-	-	201 175	0	0	
485	9	-	-	-	198 173	0	0	
613	5	-	-	-	198 173	0	0	
492	5	-	-	-	197 172	0	0	
410	5	-	-	-	196 171	0	0	
733	8	-	-	-	196 171	0	0	
738	10	-	-	-	194 169	0	0	
726	8	-	-	-	190 166	0	0	

Contigs smaller than 300 kb

Right click

- Close
- Edit Highlighted Contig
- GoTo Current Contig
- GoTo Top
- Print to file
- PrintScreen

Edit Contig Remarks & Status

Edit contig 567 (Qs 0)

Status:

☐ Ok (do everything)

☐ NoCB (For IBC, add clones, merge contigs, but do not reorder clones)

☐ Avoid (Avoid on both Build and IBC)

☒ Dead (No Summary,Builds,AceDump)

Chr Remark:

Chr Position: 0.00

Chr: ☐ No Auto update (Type the word 'none' if no assignment)

Pos: ☐ No Auto update

Hit <CR> after entering a value

User Remark:

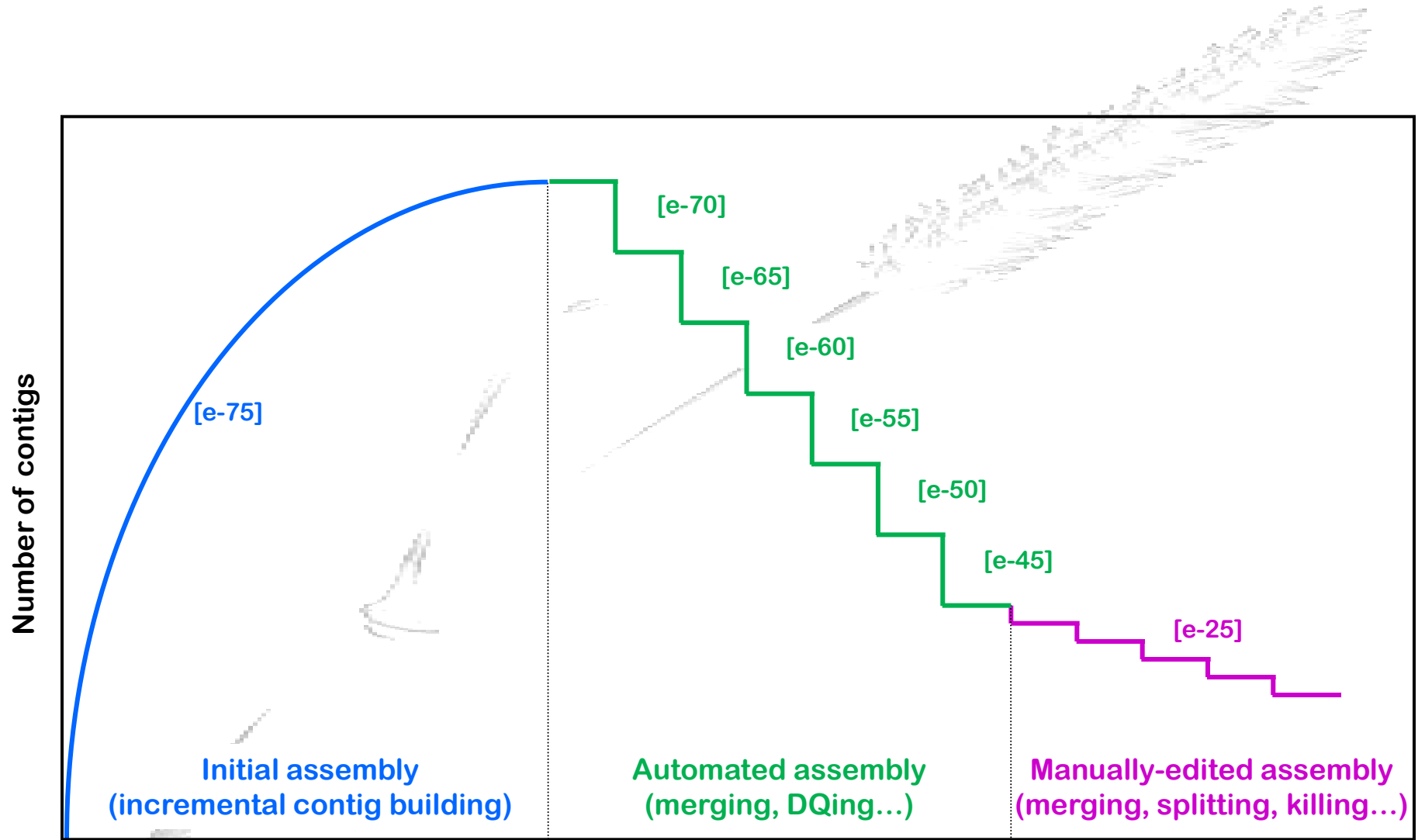
Trace Remark: DQer NoSplit 1e-48, Q~ NoSplit 1e-45, Add 1, Q~ NoSplit 1e-55, End-mer

Remark is changed as soon as you start typing.

Trace Remark automatically updated on contig changes.

Close Help

Assembly of the physical map



A background image of a lush green wheat field with several stalks in sharp focus in the foreground, reaching towards a bright, clear sky. The text is overlaid on the upper half of the image.

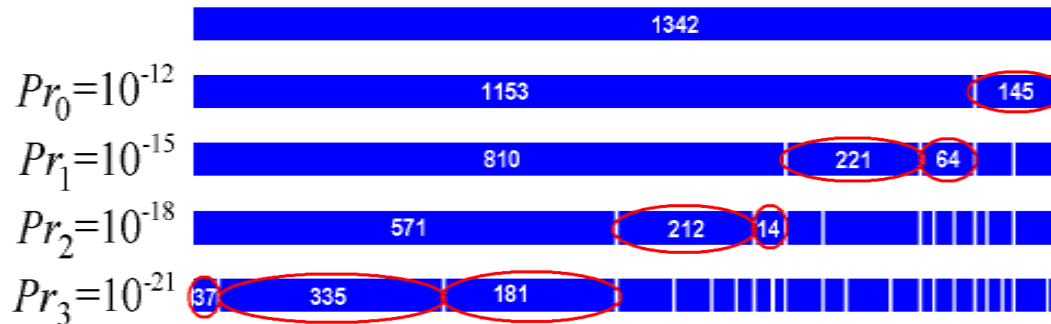
Contig assembly

6- LTC: Linear Topology Contig

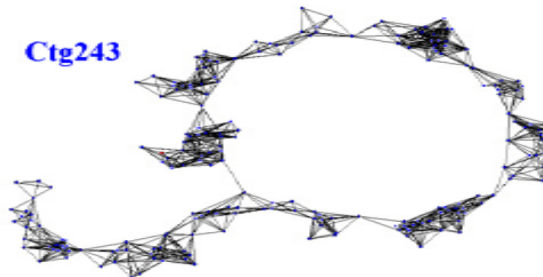
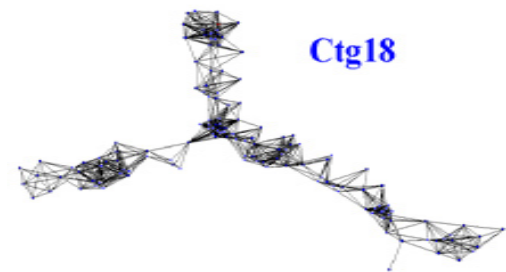
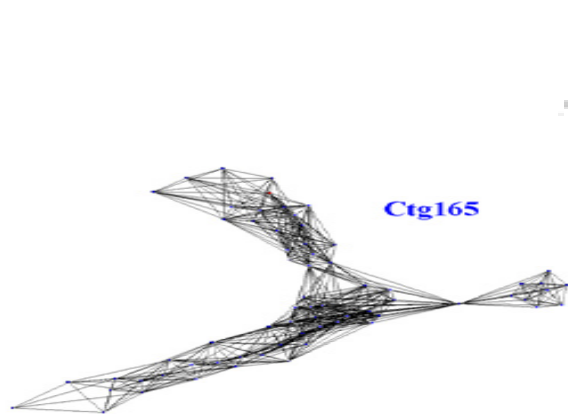
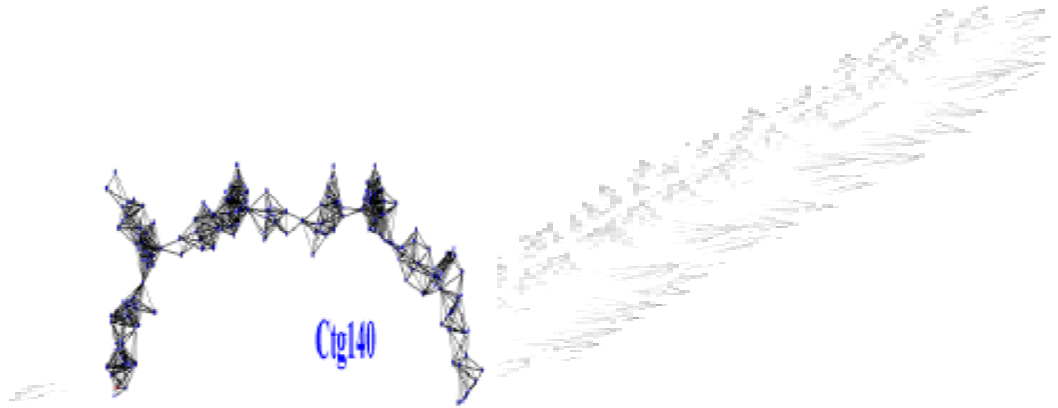
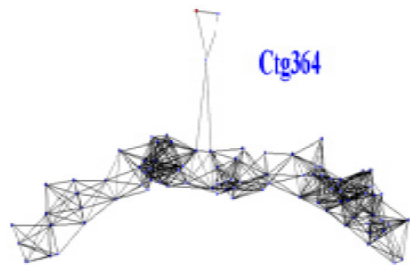
LTC program

Frenkel Z, Paux E, Mester D, Feuillet C and Korol A (2009) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *Manuscript in prep.*

- ✓ LTC program starts clustering with a relatively relaxed cutoff and uses the topology of significant clone overlapping to obtain longer contigs with realistic (linear) structure.
- ✓ In each cluster, clones are ordered based on a global optimization procedure and clones that disturb the order stability (assessed by re-sampling analysis) are excluded from the contig.
- ✓ Ordered contigs are then merged upon a relaxed cutoff into longer contigs using for control of the contig topology the network representation of the significant clone overlaps.

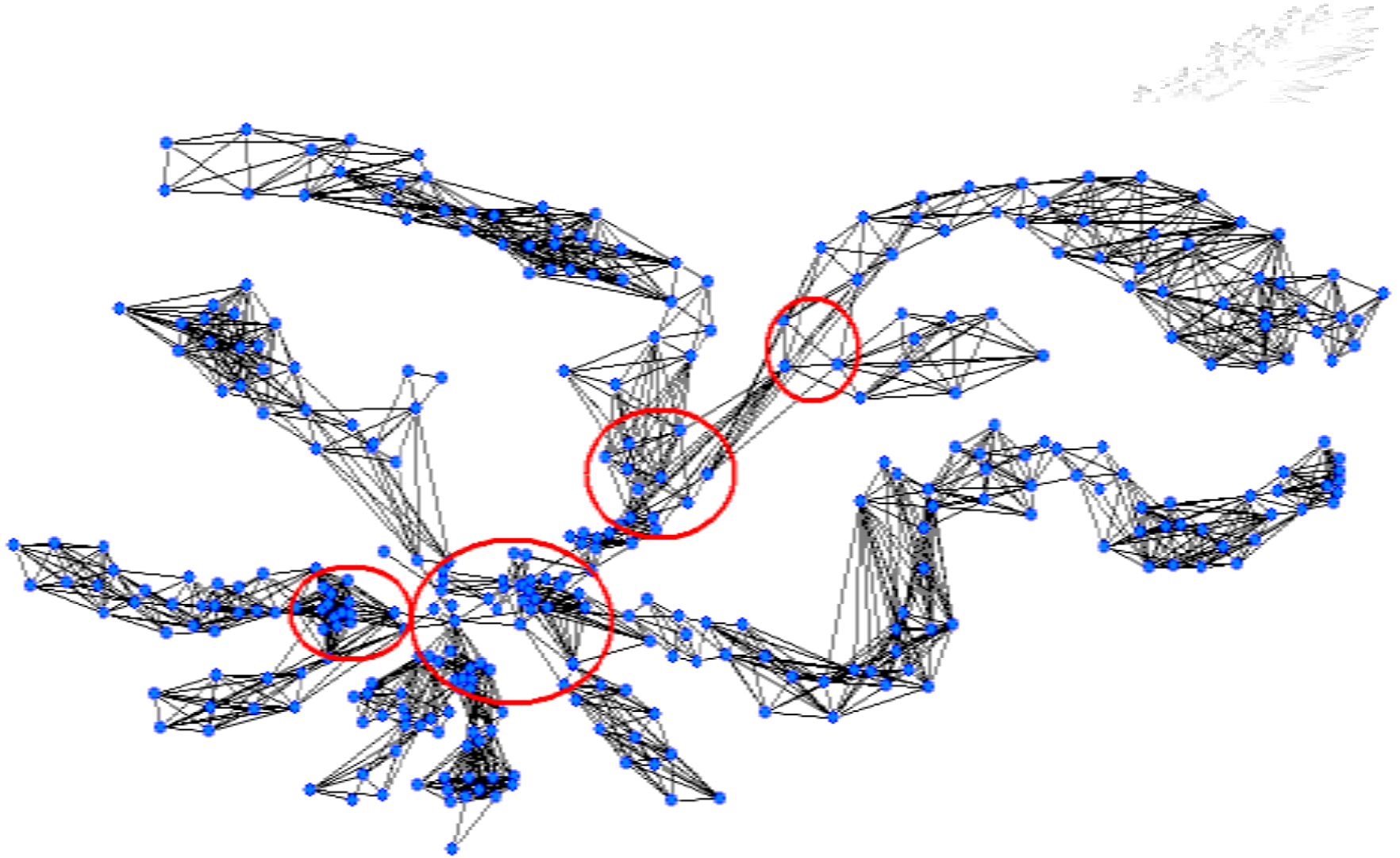


Examples of non linear topology contigs

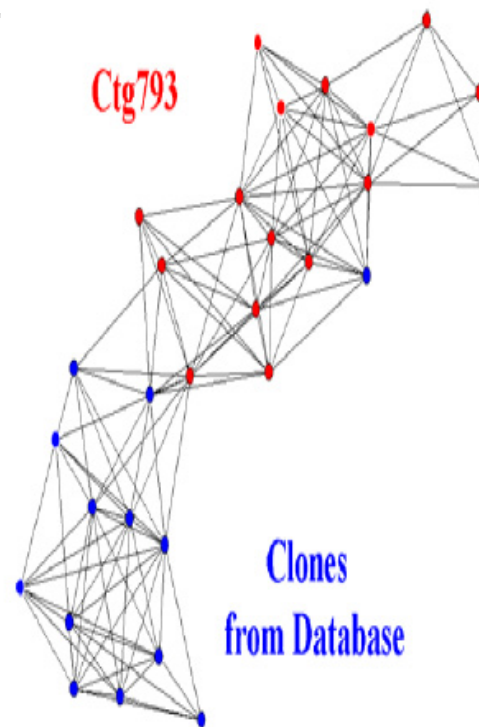
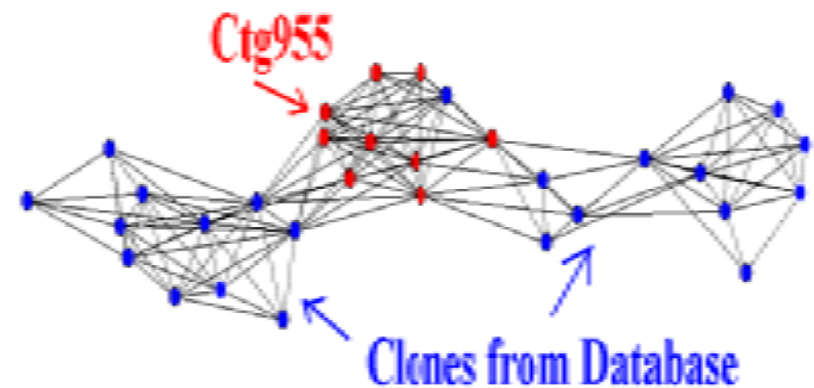
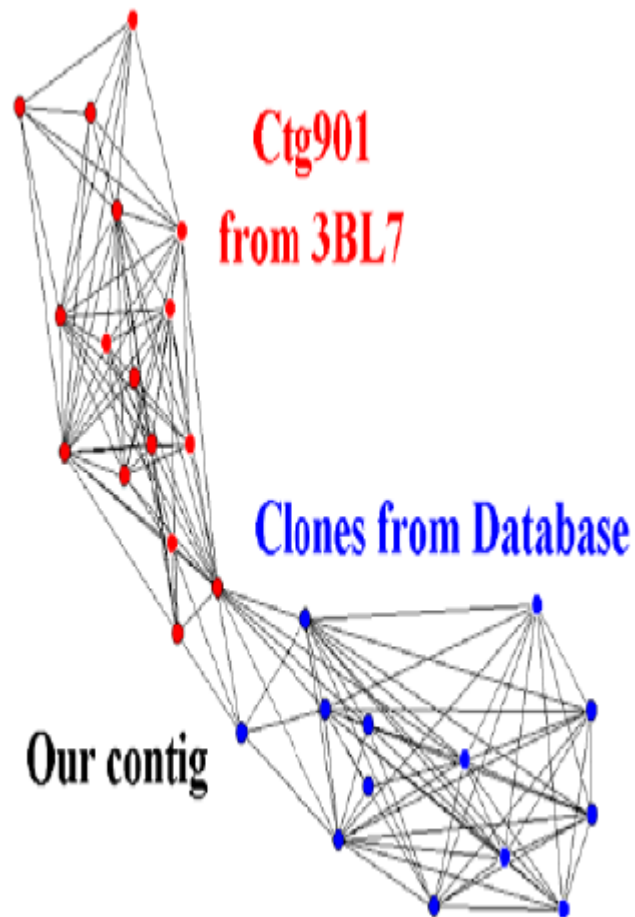


(kindly of A. Korol)

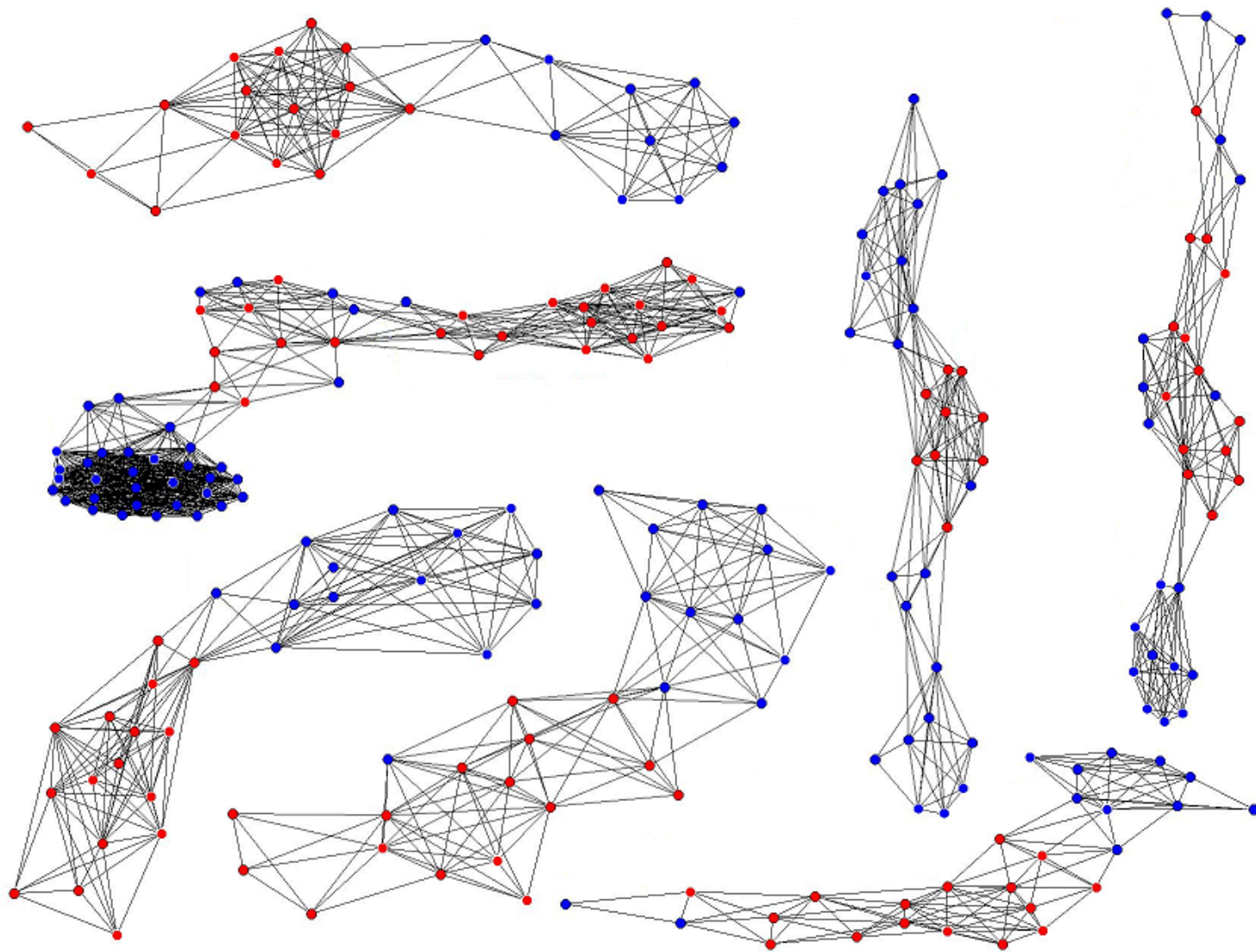
“Linearization” by removing clones in cluster branching



Examples of contig elongation



Examples of *de novo* assembled contigs



(kindly of A. Korol)