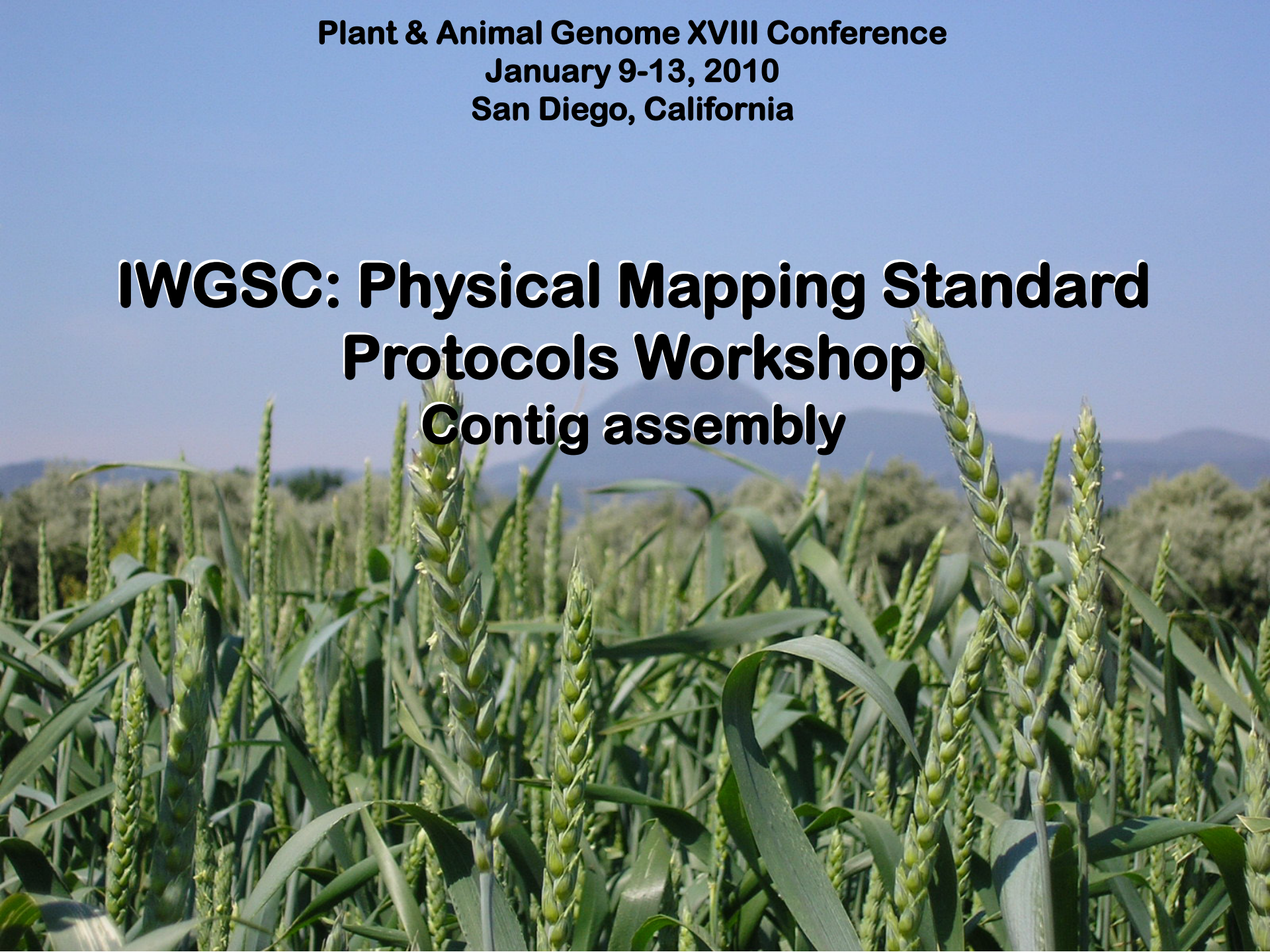
Plant & Animal Genome XVIII Conference
January 9-13, 2010
San Diego, California



IWGSC: Physical Mapping Standard Protocols Workshop

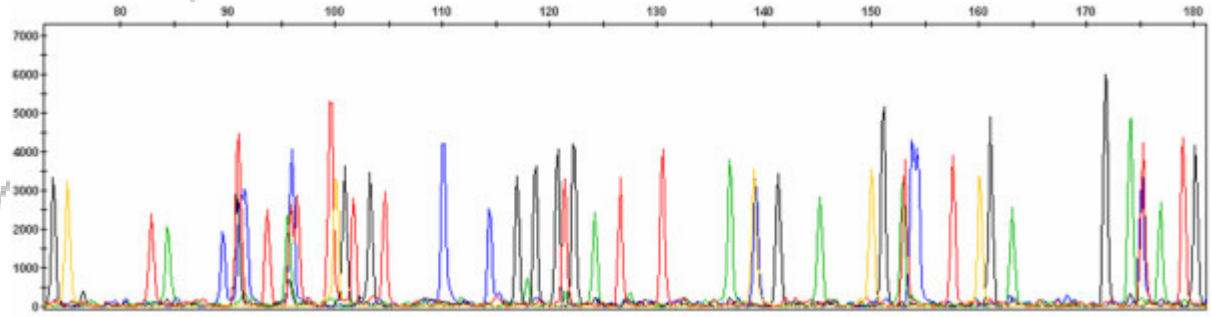
Contig assembly

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).



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

Cleaning fingerprints using FPB

Automated FingerPrint Background removal: FPB

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

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

Background removal

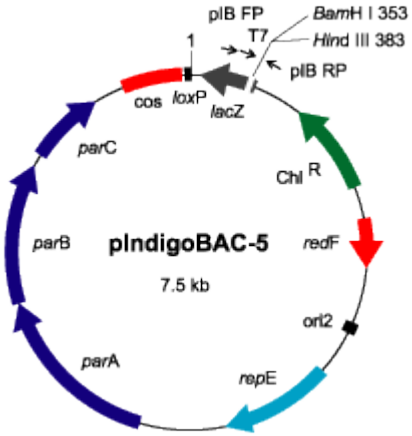
- ✓ low signal peak produced by the machine;
- ✓ partial digestion related peak;
- ✓ star activity by-product;
- ✓ *E. coli* genomic DNA band;

Pre-processing

- ✓ vector band;
- ✓ out of size standard range band (with unreliable sizing);
- ✓ wide area peak (unreliable, resulting from co-migrating fragments).

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

Vector bands



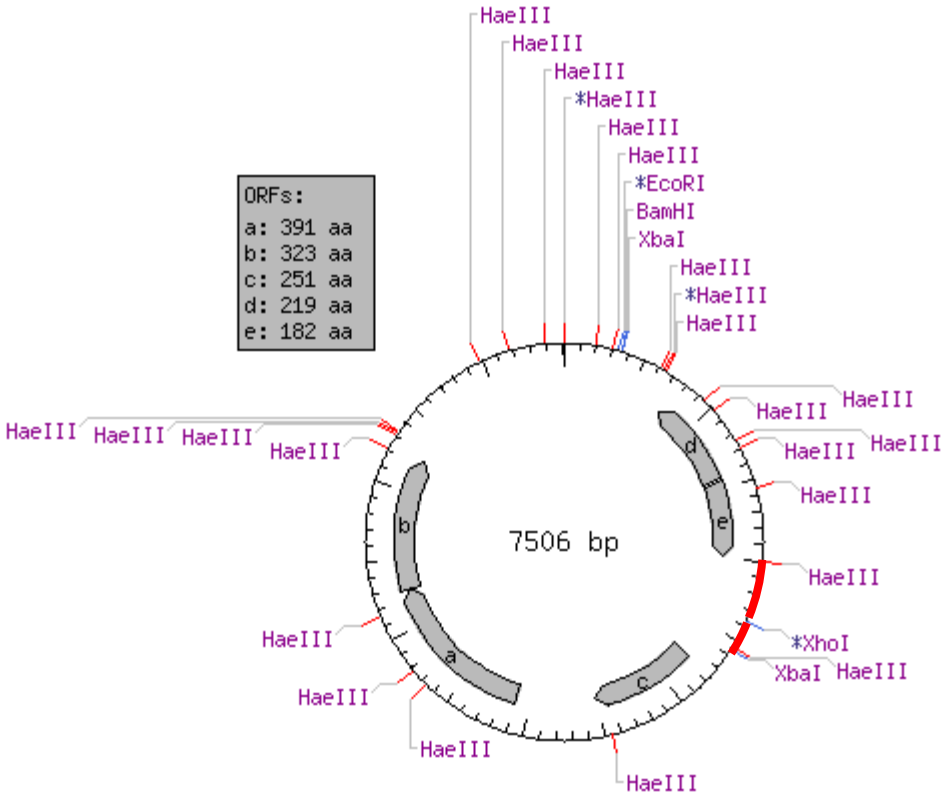
*Bam*HI

*Eco*RI

*Xba*I

*Xho*I

*Hae*III

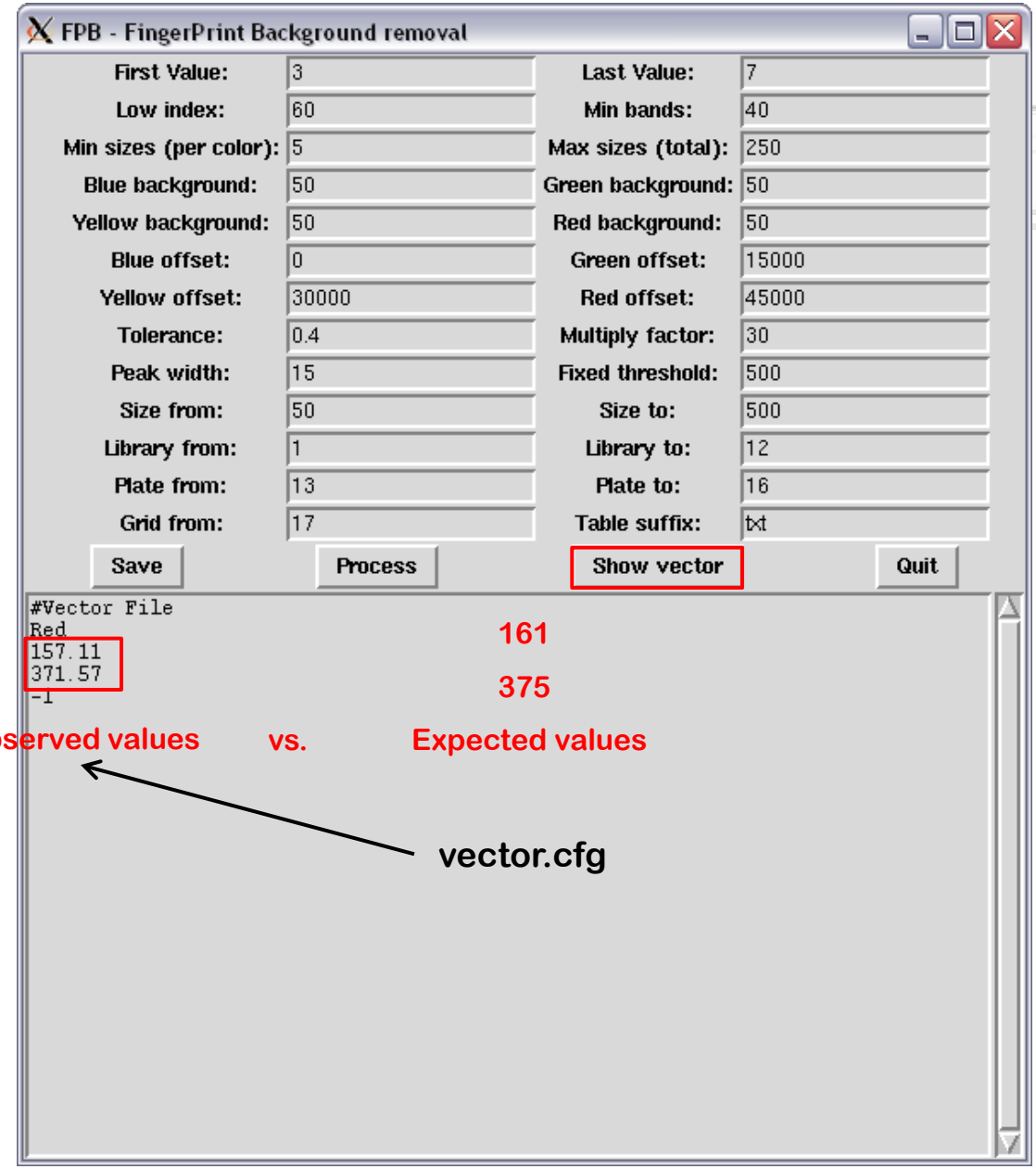


Two red fragments (*Xho*I):161 & 375 bp

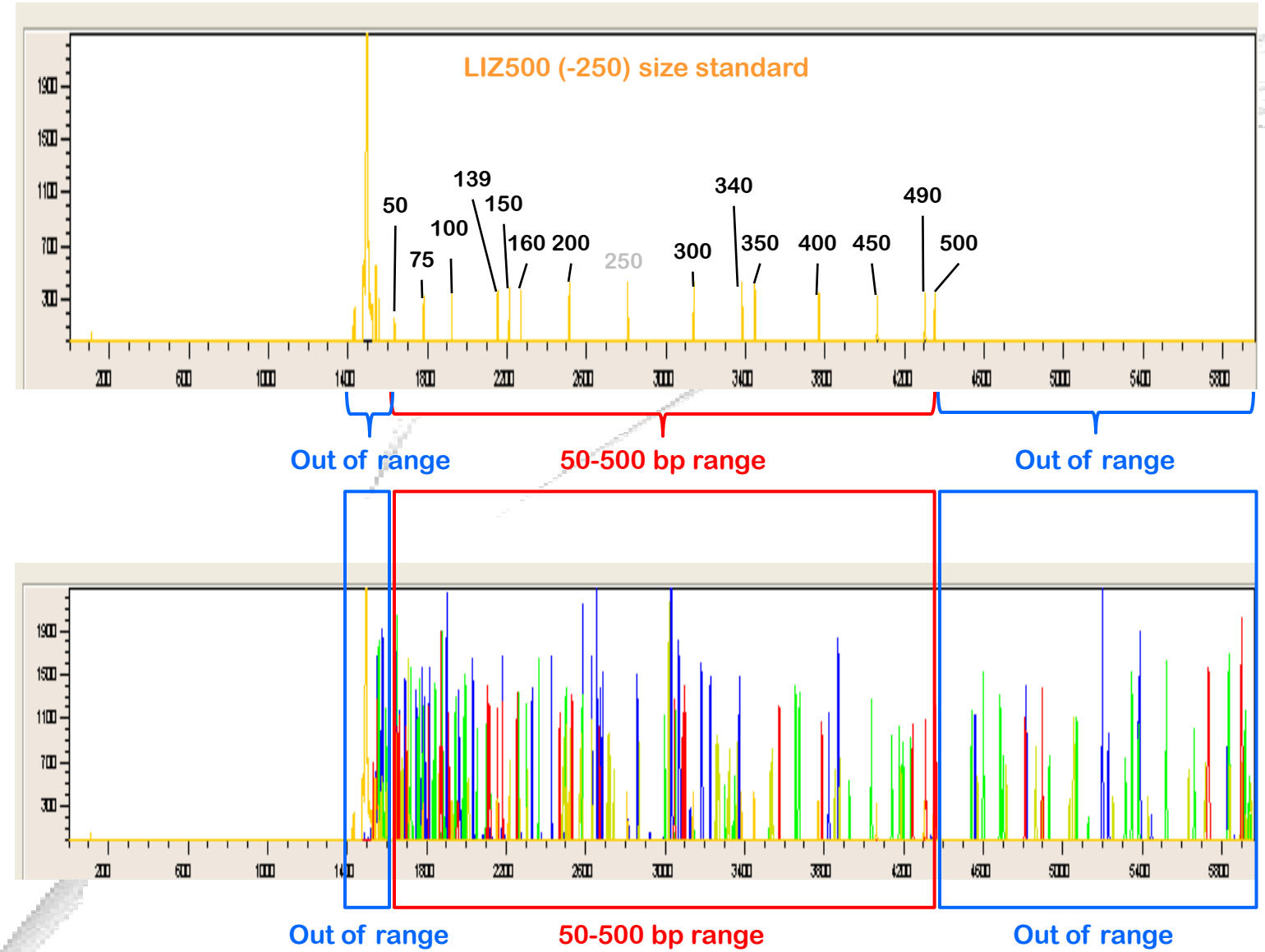
common to all fingerprints

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

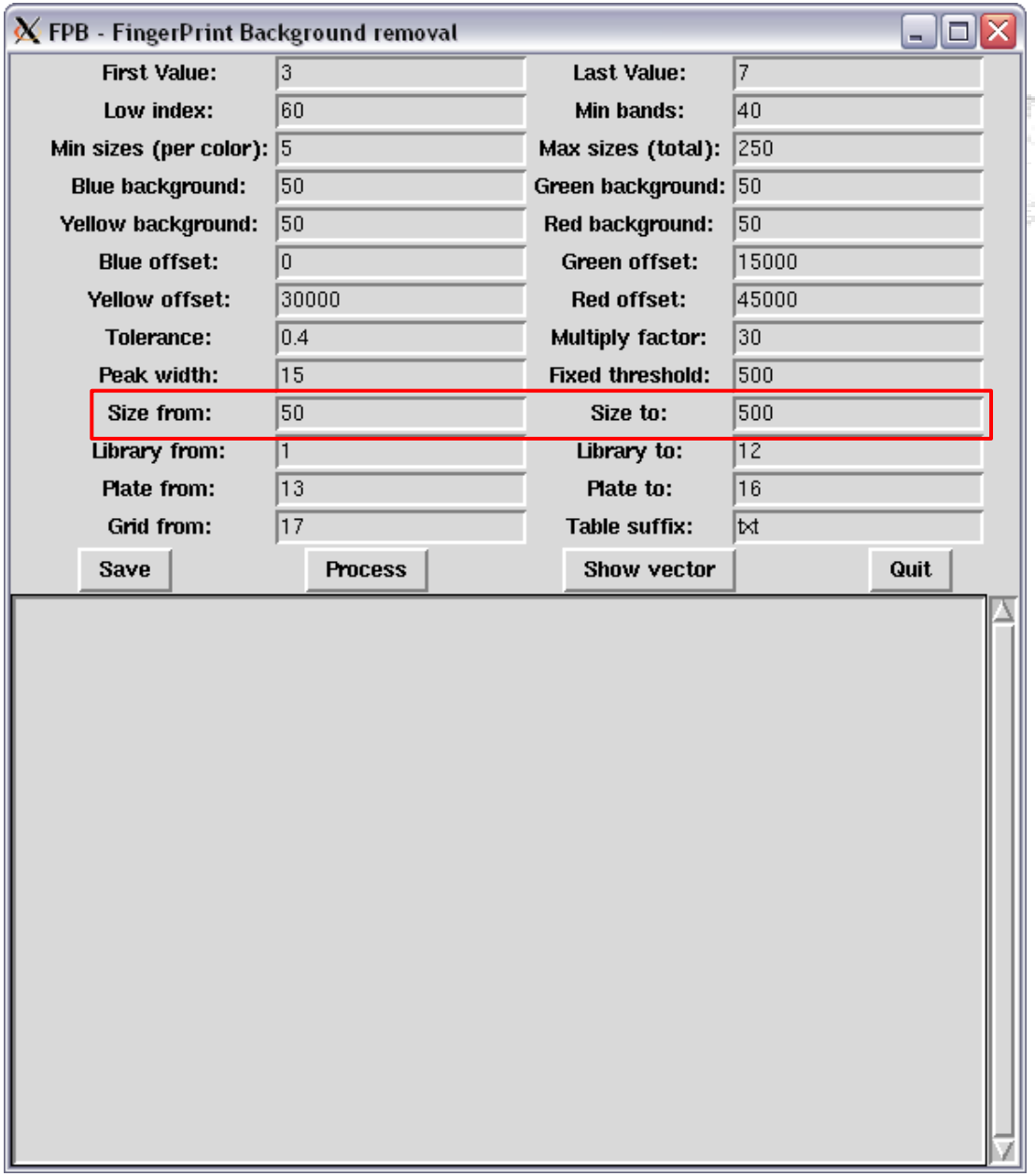
Removing vector bands



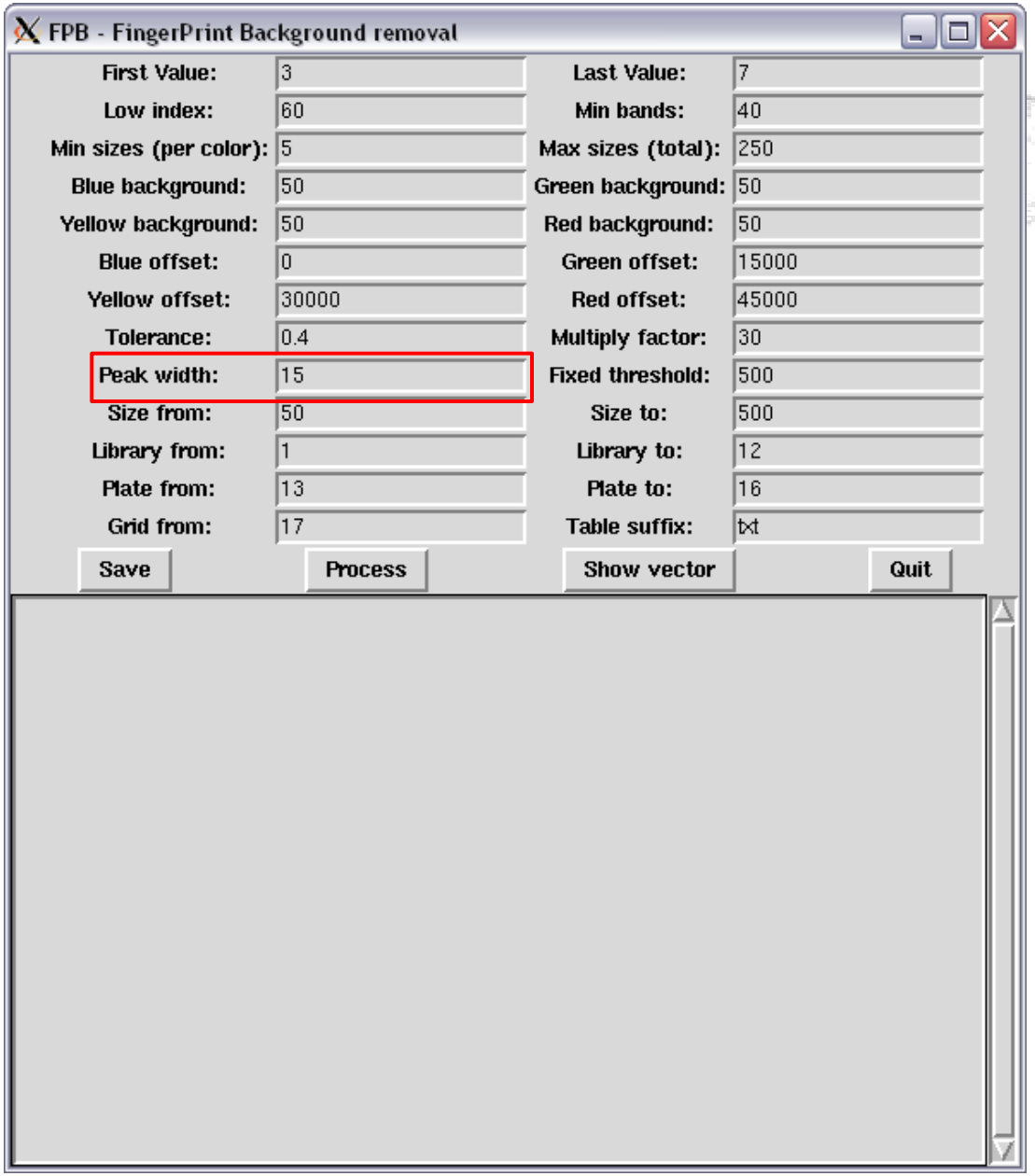
“Out of range” bands



Removing “out of range” bands

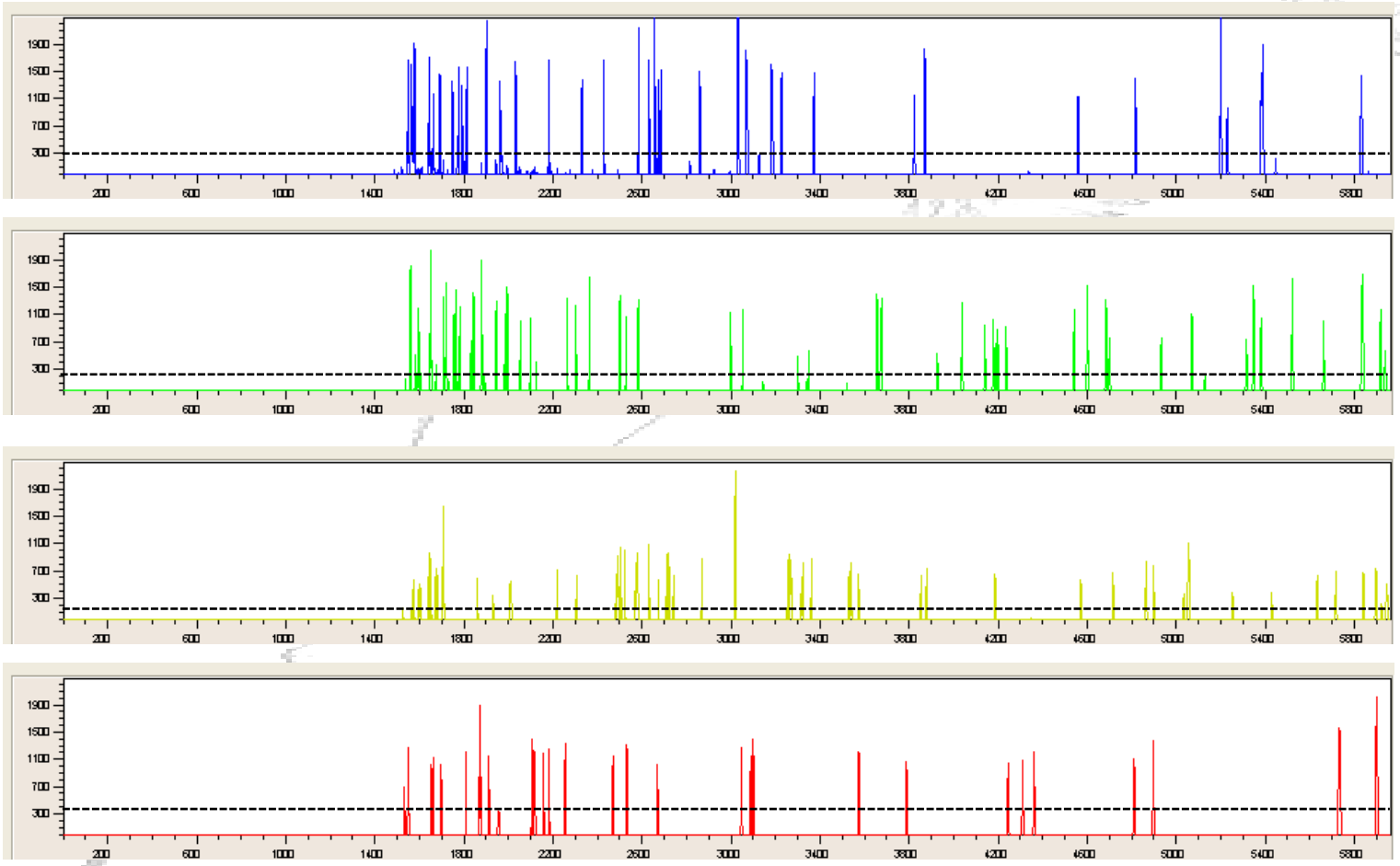


Removing wide peaks



True signal *vs.* background

Calculation of the background threshold for each dye



Removal of all peaks below the threshold

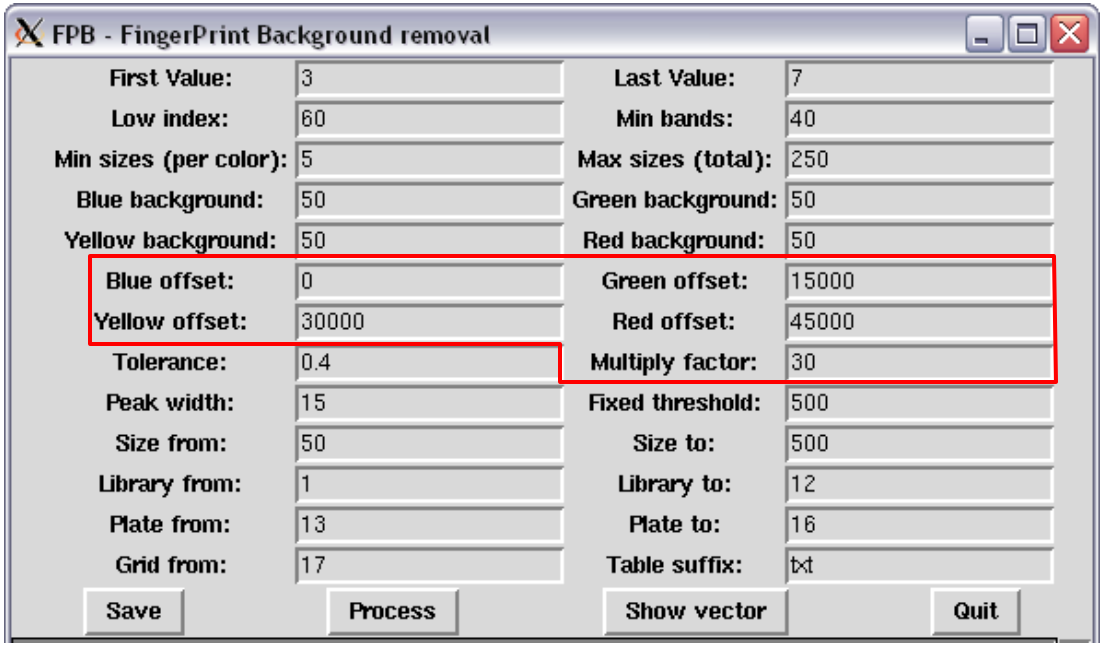
(adapted from Scalabrin *et al.*, *BMC Bioinformatics*, 2009)

Multiplication factor & color shift

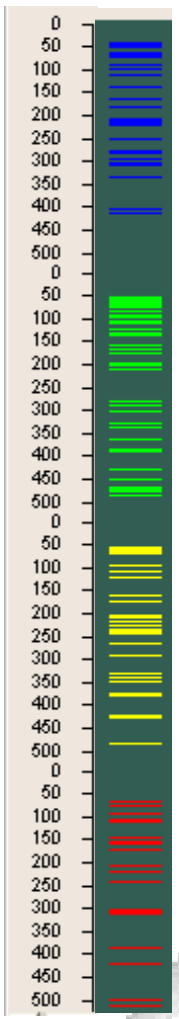
FPC does not accept color labels or fractional sizes, so the fragments must be manipulated before being loaded into FPC.

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).



Multiplication factor & color shift

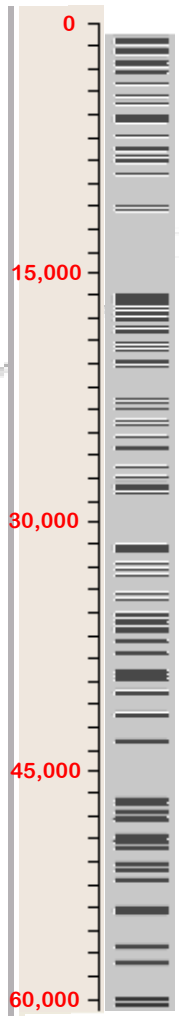


BamHI
Blue bands, from 50 to 500 bp

EcoRI
Green bands, from 50 to 500 bp

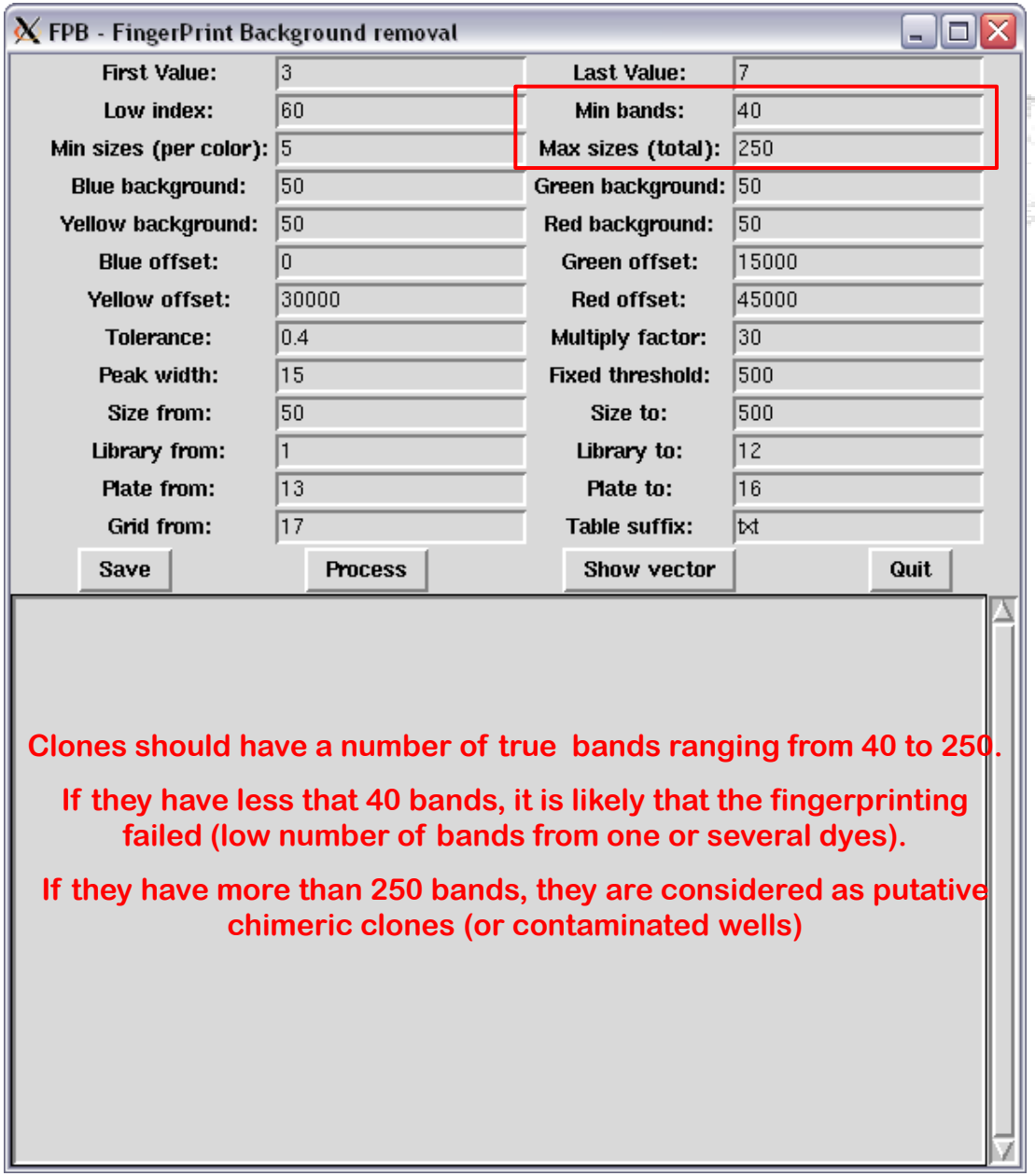
XbaI
Yellow bands, from 50 to 500 bp

XhoI
Red bands, from 50 to 500 bp



Complete fingerprint
'Black' bands, from 0 to 60,000
(with 4 gaps)

Removing low quality fingerprints



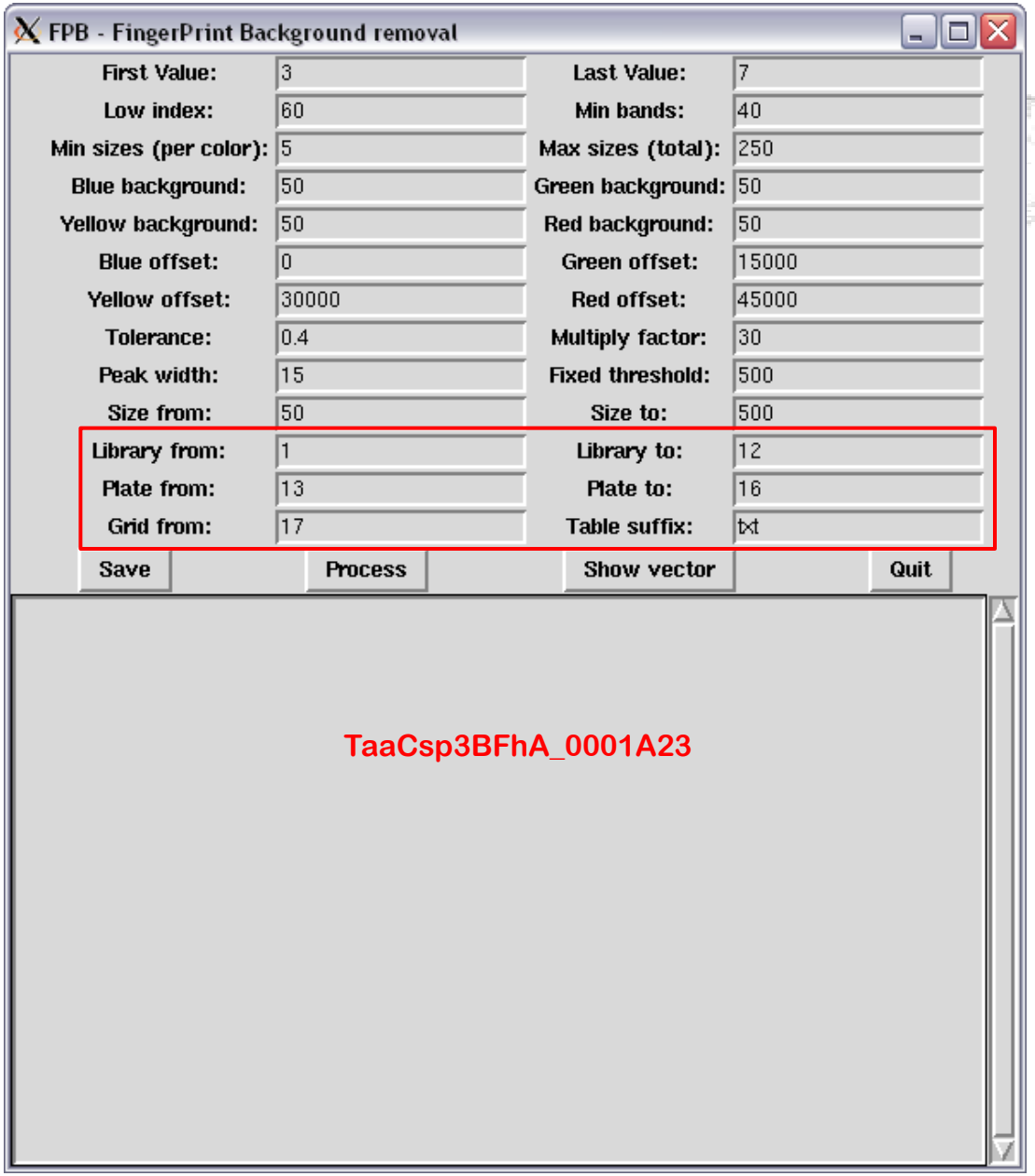
International naming convention (IWGSC)

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 *Hin*dIII, 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).

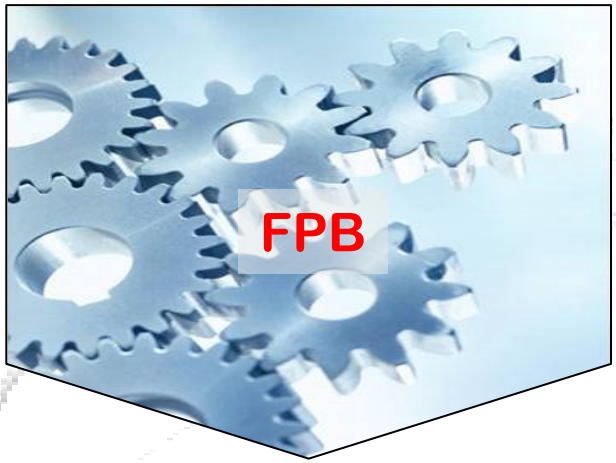
http://www.wheatgenome.org/pdf/Triticeae_Annotation_Group_Report_2007.pdf

Setting up clone name in FPB



FPB output

GeneMapper .txt files



FPB .sizes files

✓FPC-compatible

✓Background-free

✓Vector-free

✓Ranging from 50 to 500 bp…

⇩

Genoprofiler

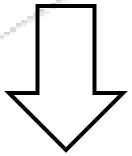
Editing fingerprints
2- Genoprofiler

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

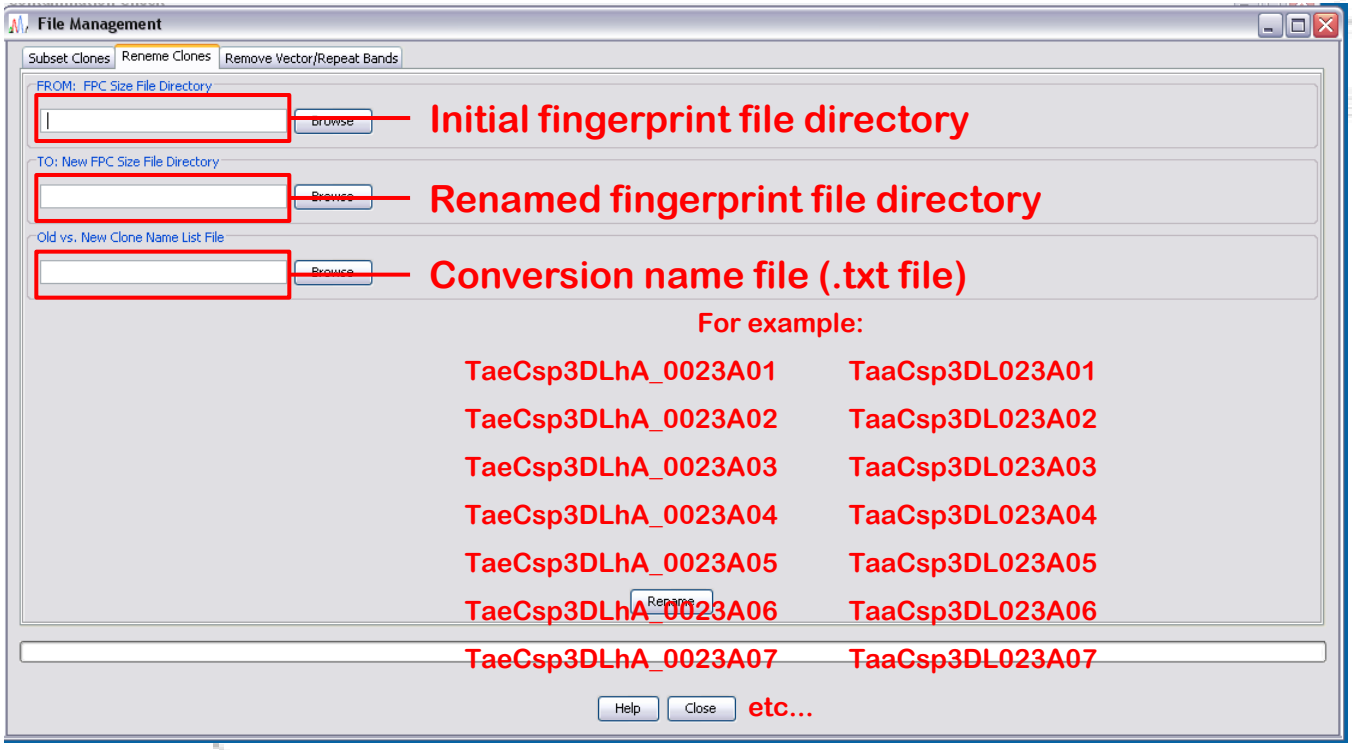


TaaCsp3BF001A23

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*

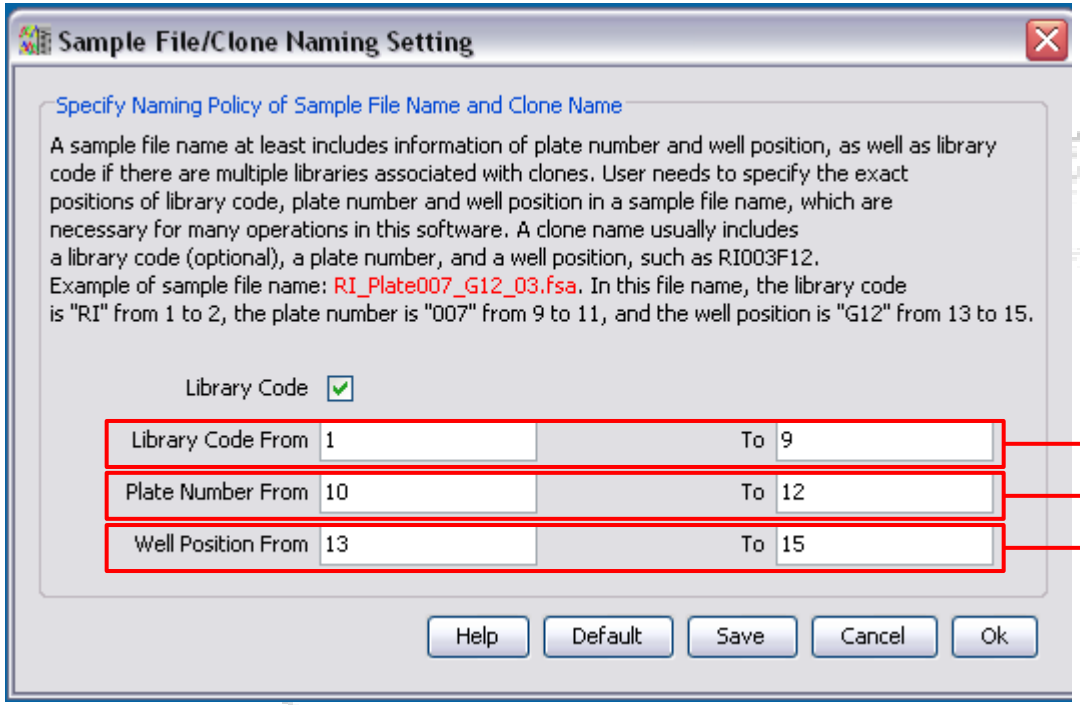
TaaCsp3BFhA_0001A01	TaaCsp3B001A01
TaaCsp3BFhA_0001A02	TaaCsp3B001A02
TaaCsp3BFhA_0001A03	TaaCsp3B001A03

…

Command line:

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

Configuring Genoprofiler

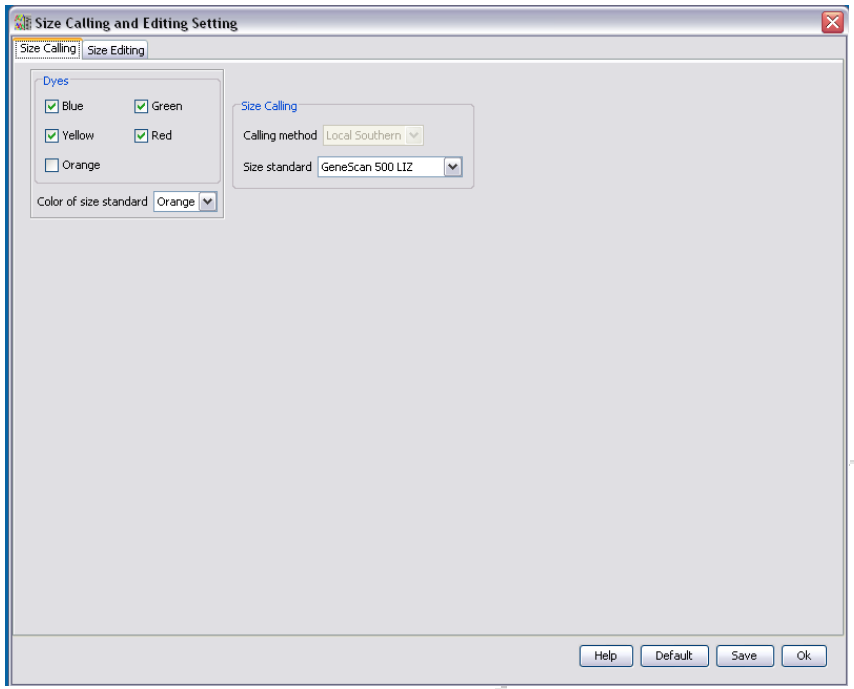


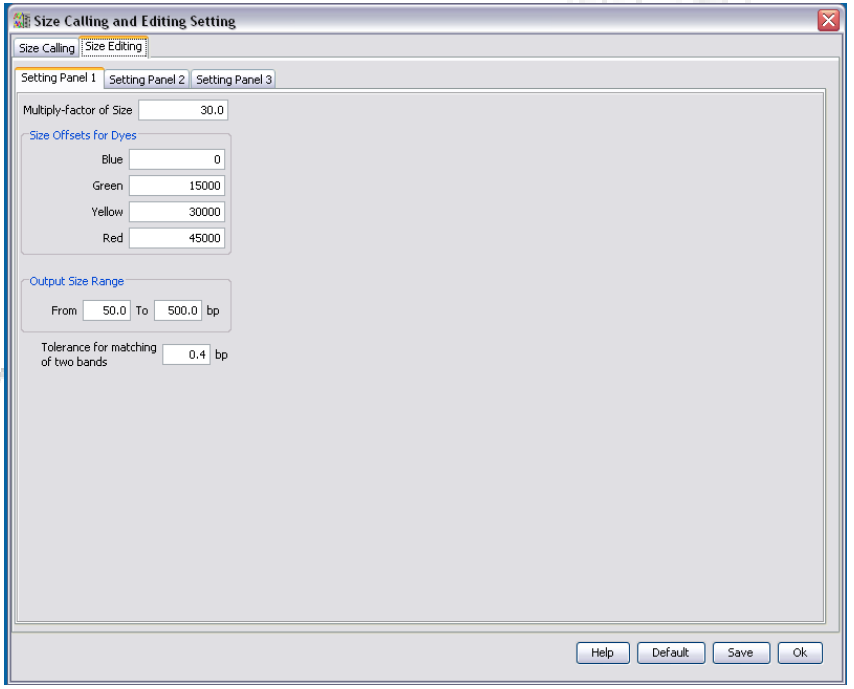
TaaCsp3DL

023

A01

Configuring Genoprofiler





Sources of DNA contamination

✓ Chloroplastic DNA contamination

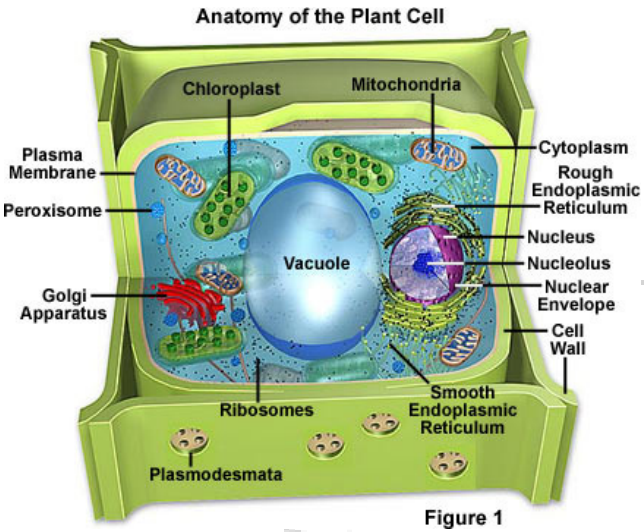
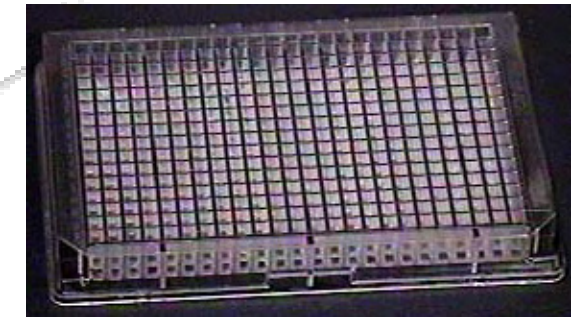
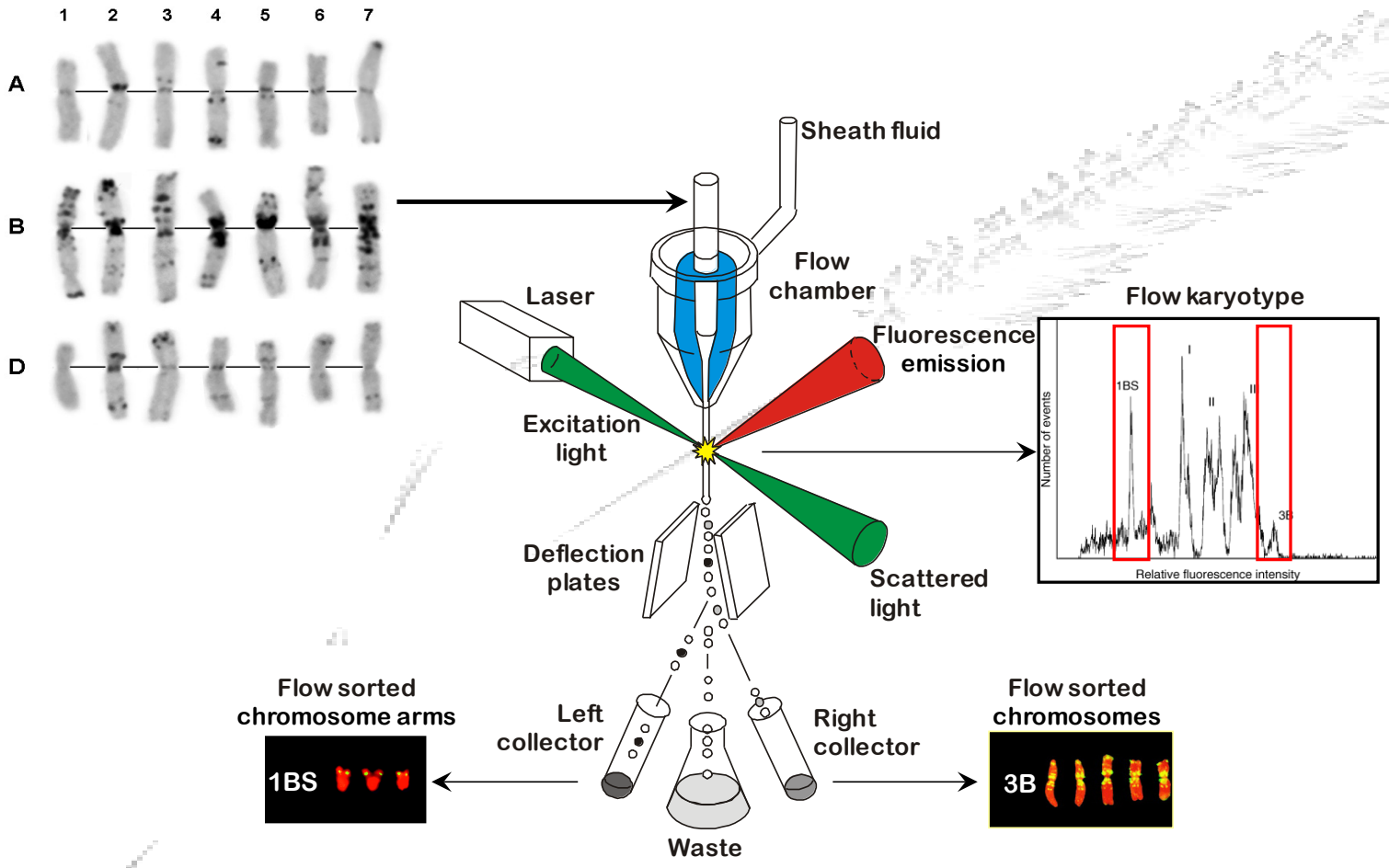


Figure 1

✓ Well-to-well contamination



Chloroplast DNA contamination



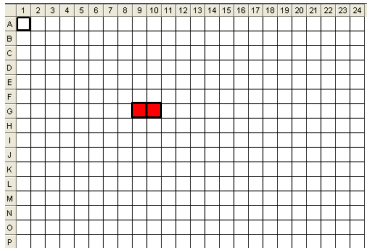
No chloroplast DNA contamination since chromosomes are flow-sorted and not simply extracted

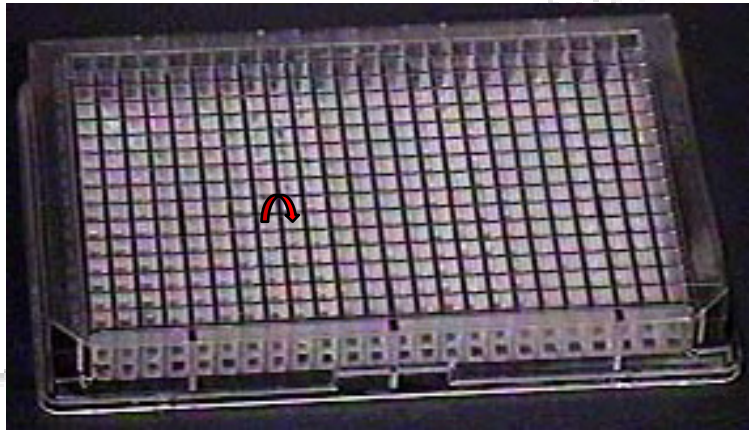
(kindly of J. Dolezel)

Well-to-well contamination

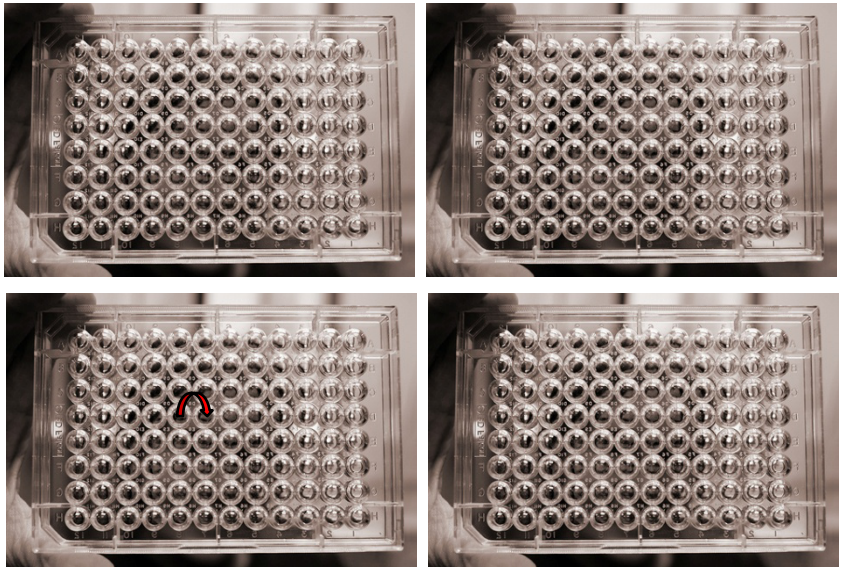
✓ Well-to-well contamination in 384-well plate format

➔ Adjacent wells showing similar profiles



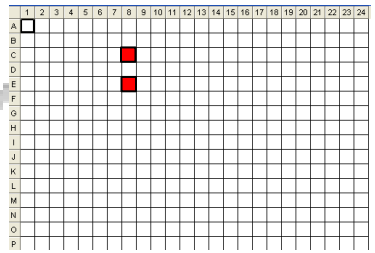


Splitting of 384-well plate into four 96-well plate during DNA extraction process.



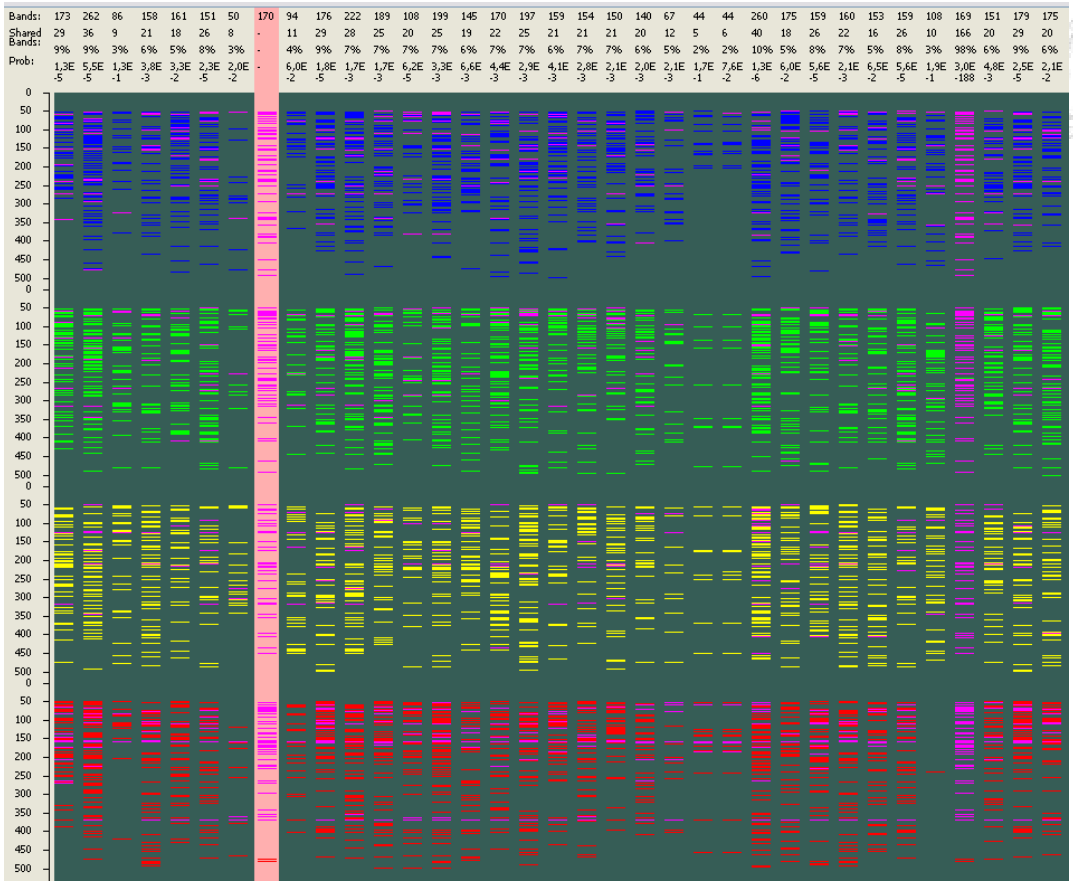
✓ Well-to-well contamination in 96-well plate format

➔ Non-adjacent wells showing similar profiles



‘One-to-one’ contamination

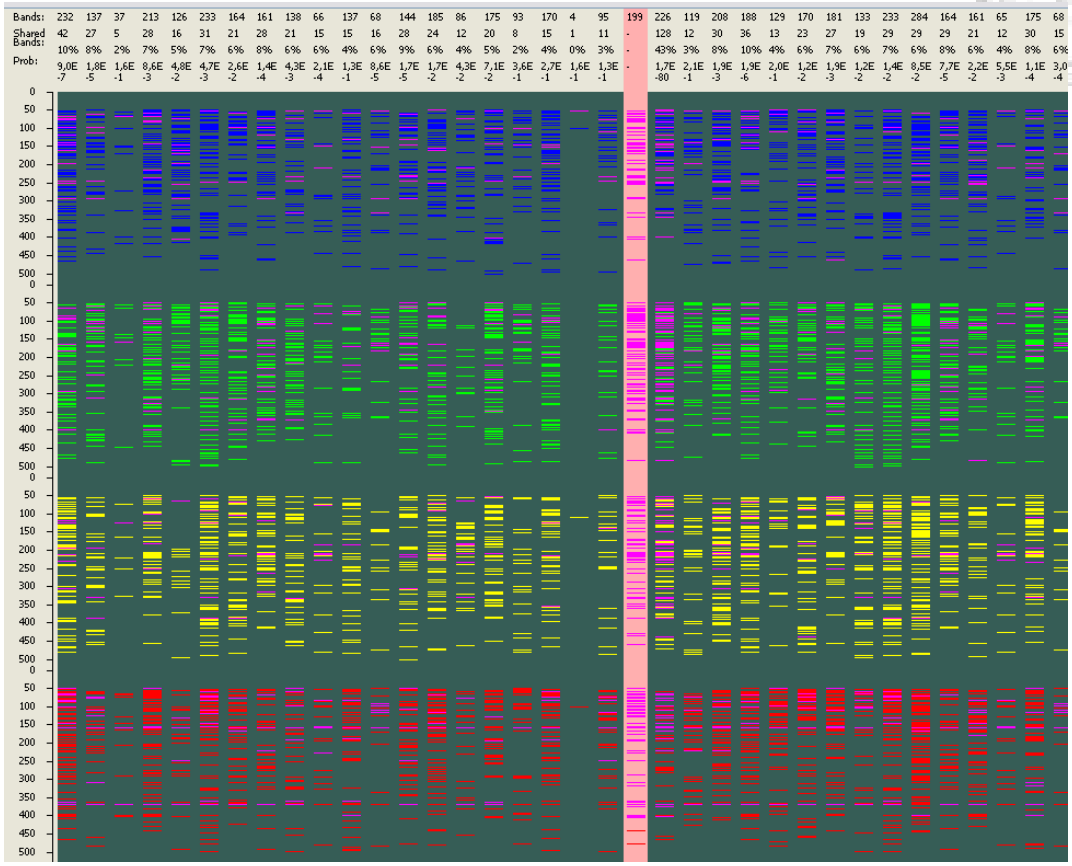
Two adjacent wells contain the same clone B1



80-100% identity of fingerprints

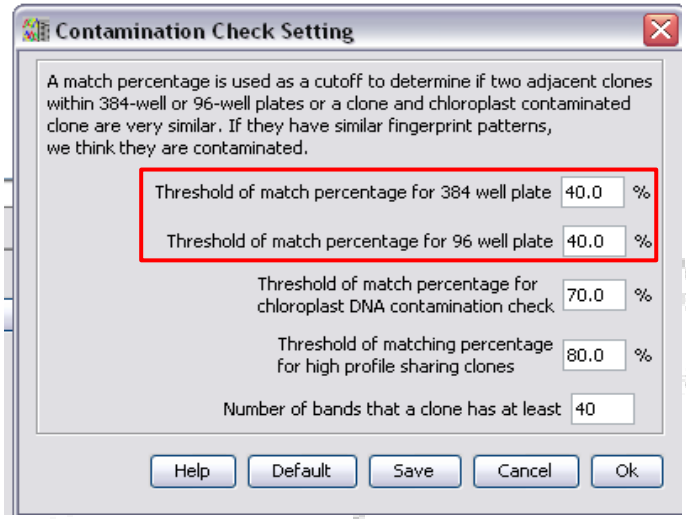
‘One-to-two’ contamination

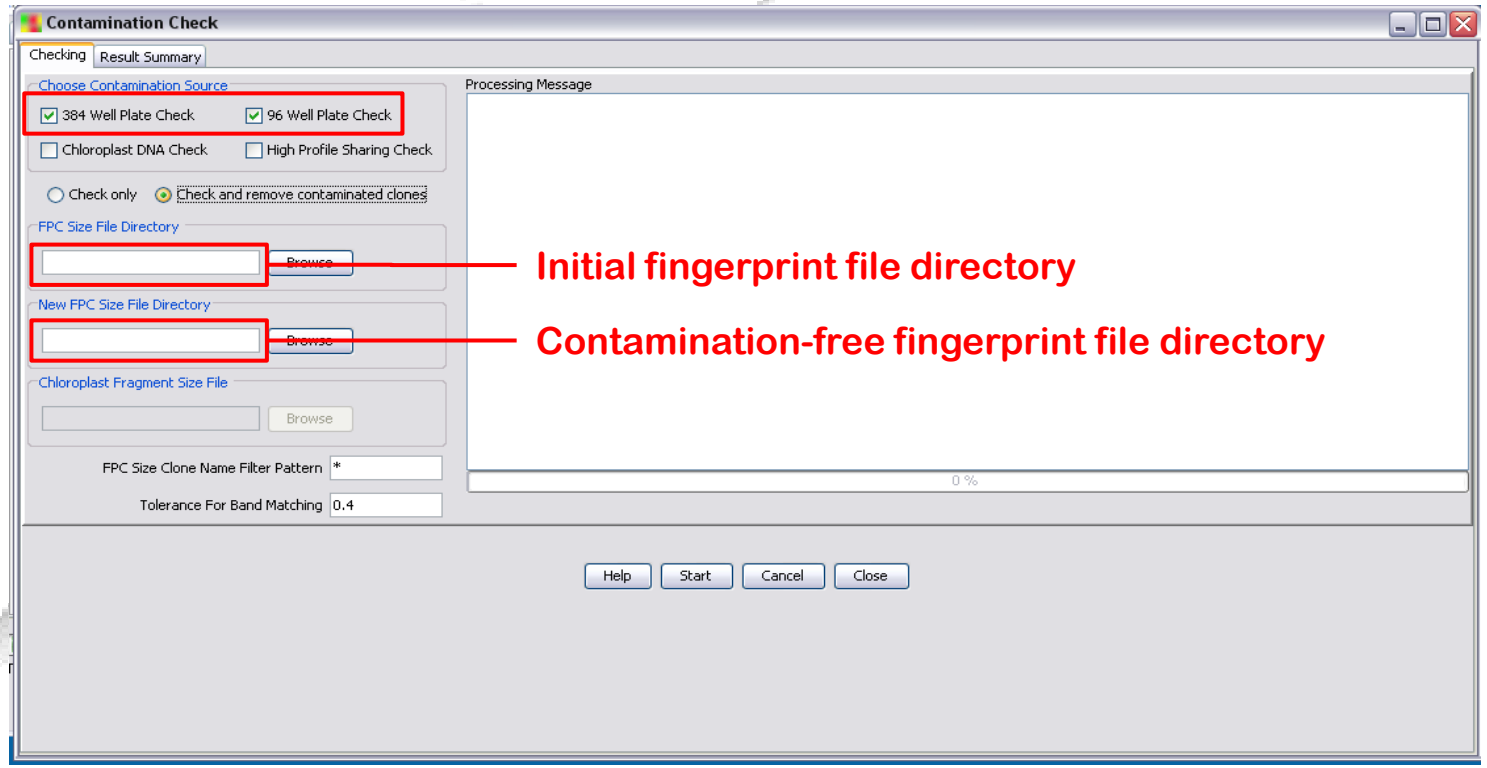
One well contains one clone B1 and the adjacent one contains the same clone B1 and another one B2



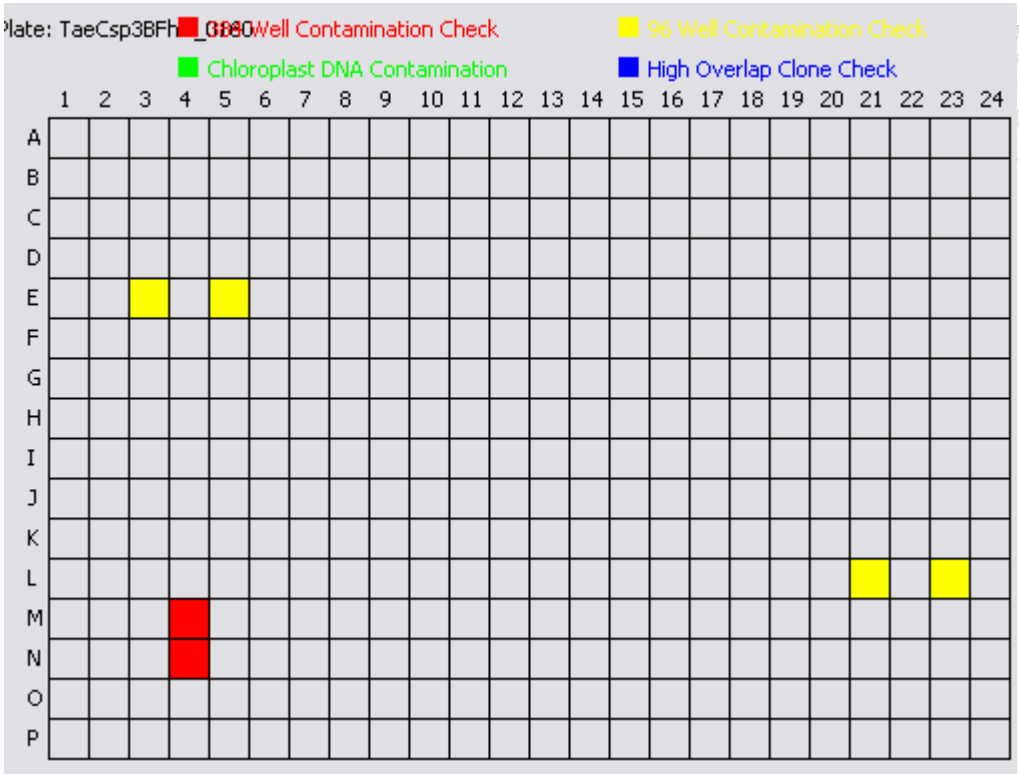
35-50% identity of fingerprints:
one of the well displays two merged fingerprints

Contamination removal using Genoprofiler

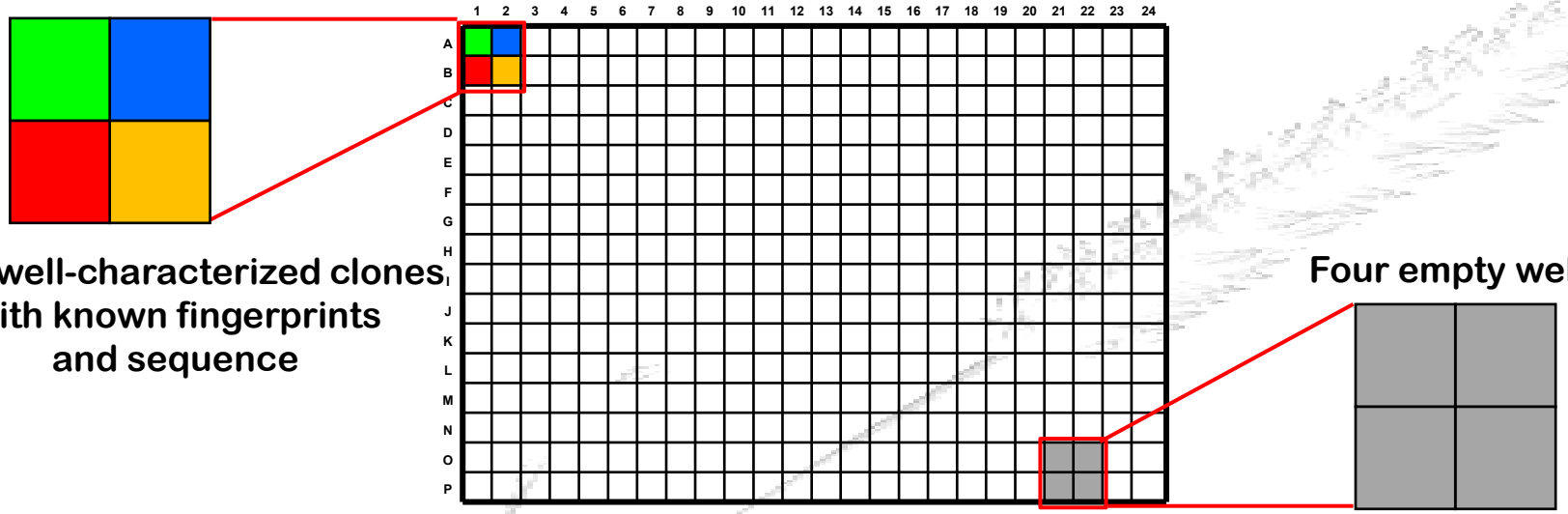




Contamination removal using Genoprofiler



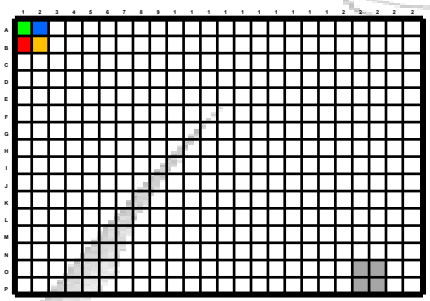
Control clones for quality check

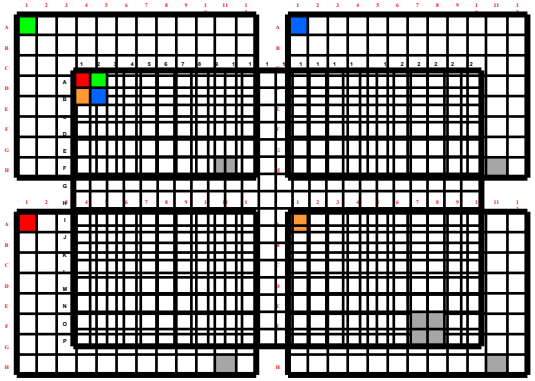


Four well-characterized clones with known fingerprints and sequence

Four empty wells

✓ Control of plate rotation or inversion



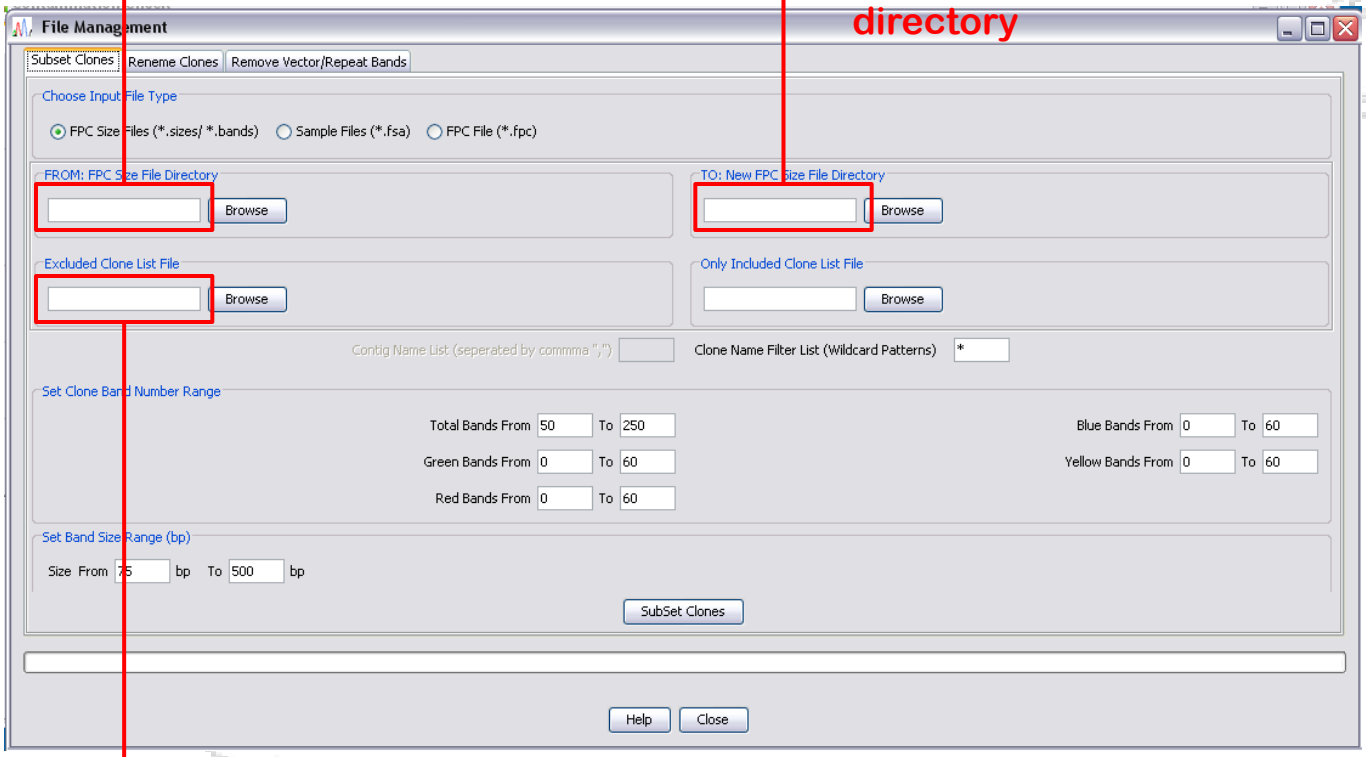


✓ Calculation of contamination rate

Removing control clones using Genoprofiler

Input fingerprint file directory

Output fingerprint file directory



List of excluded clones (.txt file)

For example:

TaeCsp3DL023A01

TaeCsp3DL023A02

TaeCsp3DL023B01

TaeCsp3DL023B02

TaeCsp3DL023O21

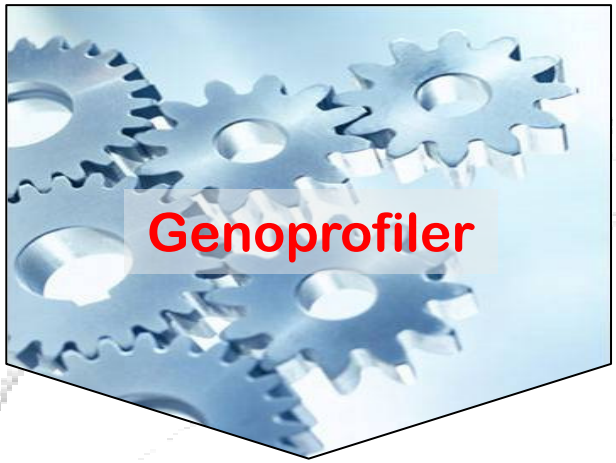
TaeCsp3DL023O22

TaeCsp3DL023P21

TaeCsp3DL023P22

Genoprofiler output

FPB .sizes files



Genoprofiler.sizes files

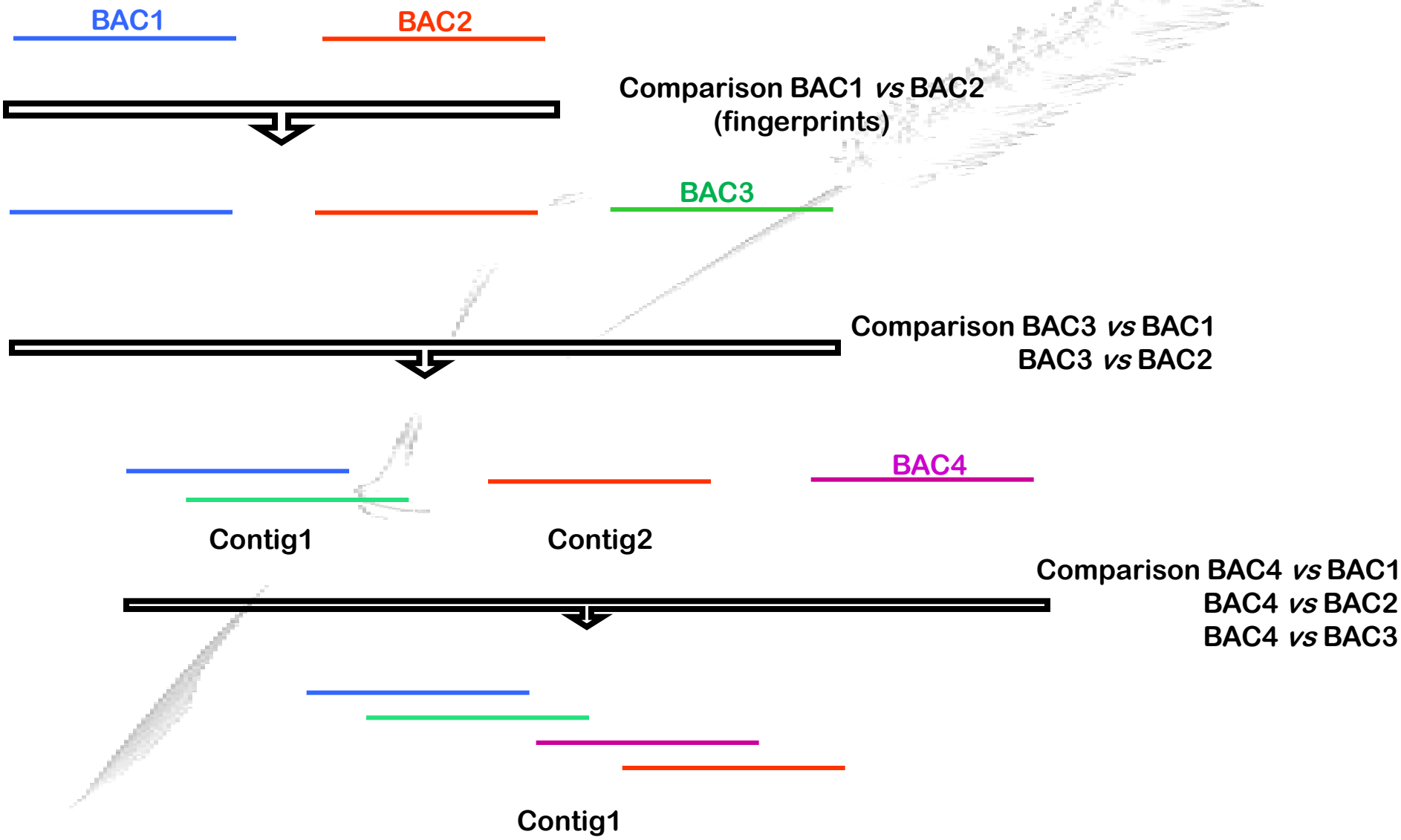
✓Contamination-free

✓Control clone-free…

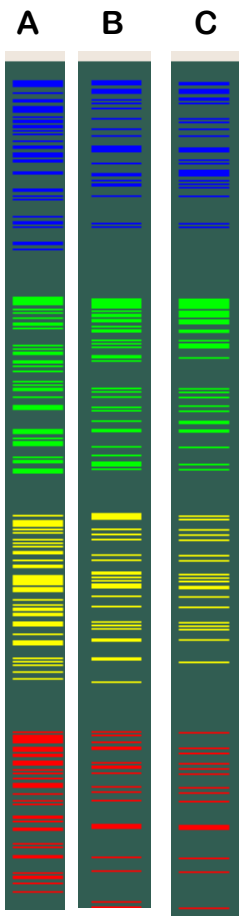
FPC

Contig assembly
1- Overview

Pairwise comparison and contig assembly



Overlap calculation: the Sulston score



FingerPrinted Contigs (FPC)

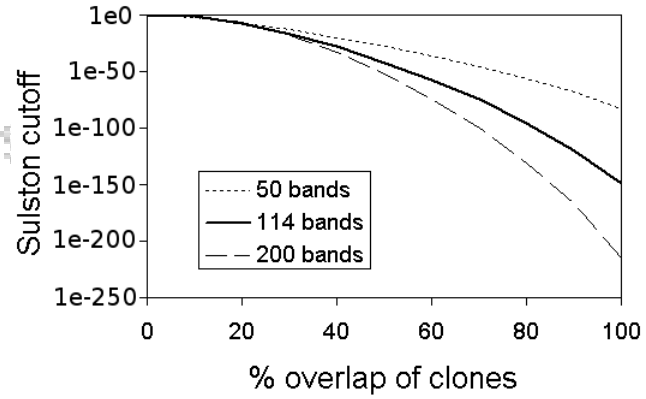
$$\sum_{m=M}^{nL}\left[\binom{nL}{m}\left((1-p)^{m}p^{nL-m}\right)\right]$$

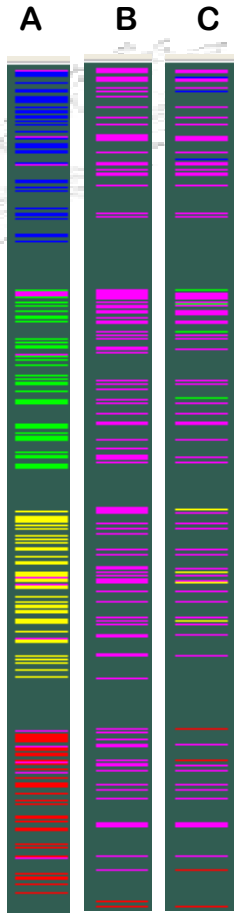
Tolerance for two bands to be identical

Number of possible values for bands

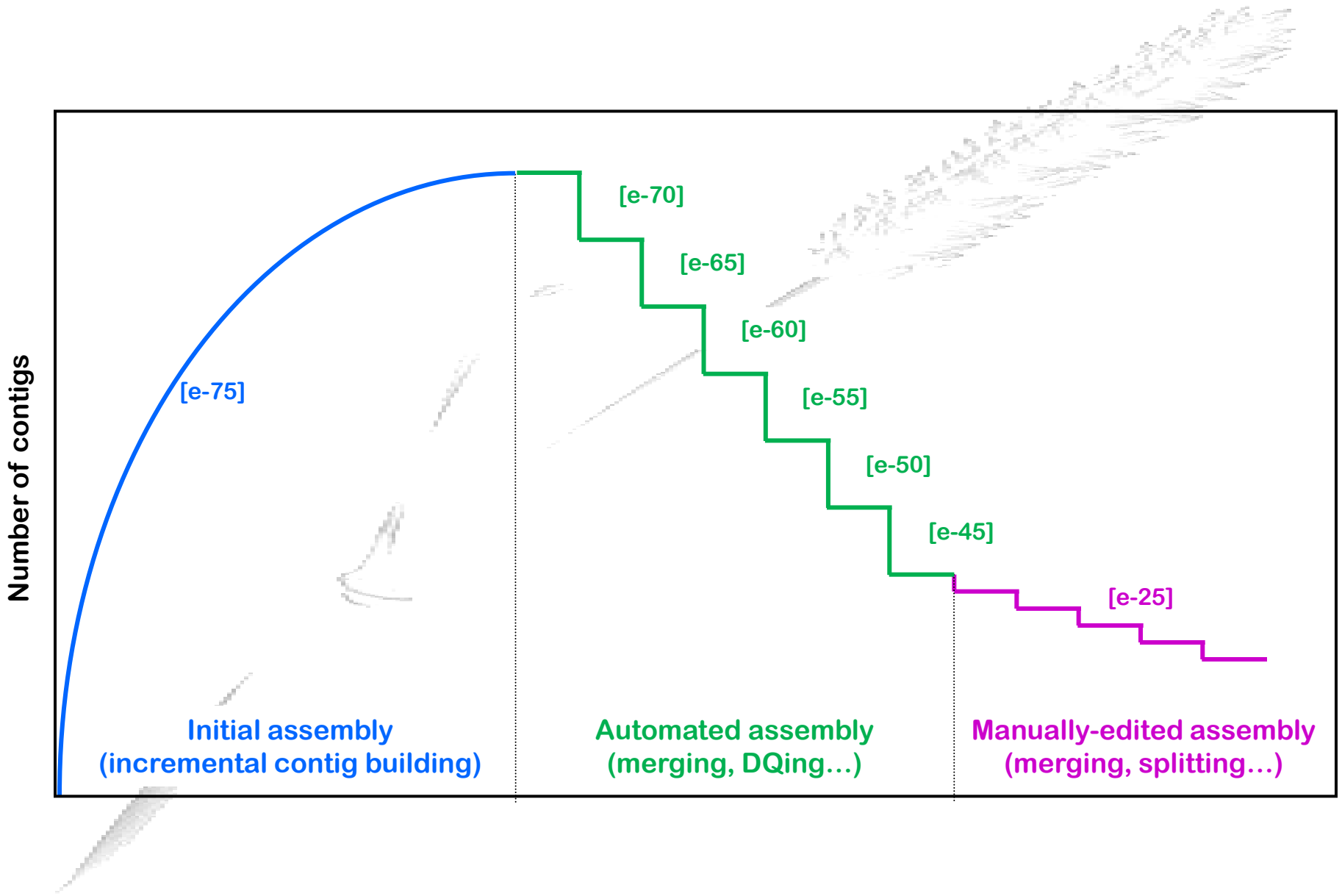
Number of bands for two clones

Number of shared bands





Assembly of the physical map



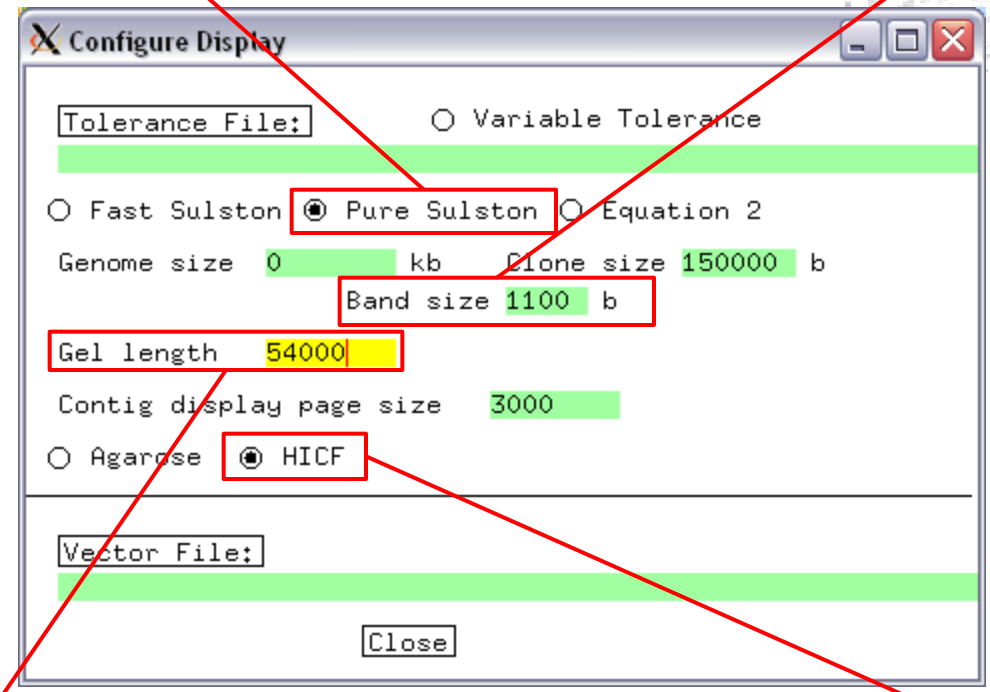
Contig assembly
2- FPC overview

Contig assembly
3- Initial assembly

Configuring FPC: configure window

For clones larger than 100 bands

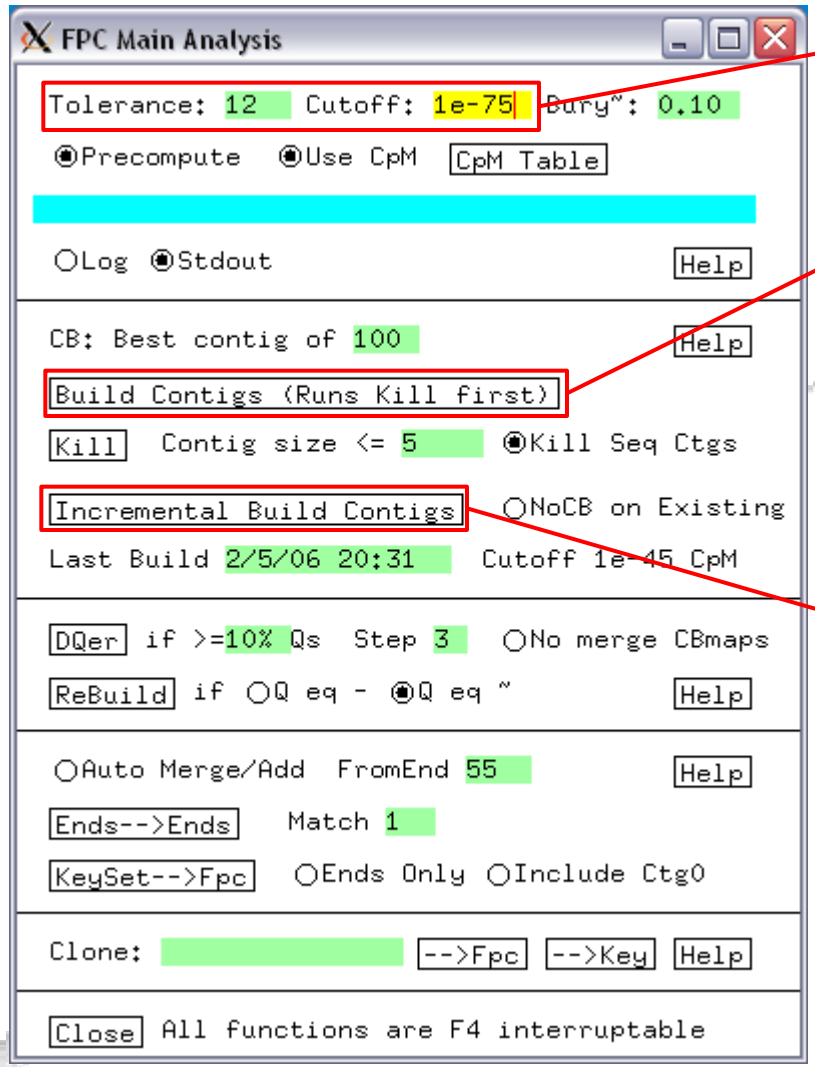
Average band size
(based on fingerprints and sequences)



Number of possible values for one band:
(15,000 – 1500) x 4 = 54,000

SNaPshot labelling & capillary sequencer

Building contigs

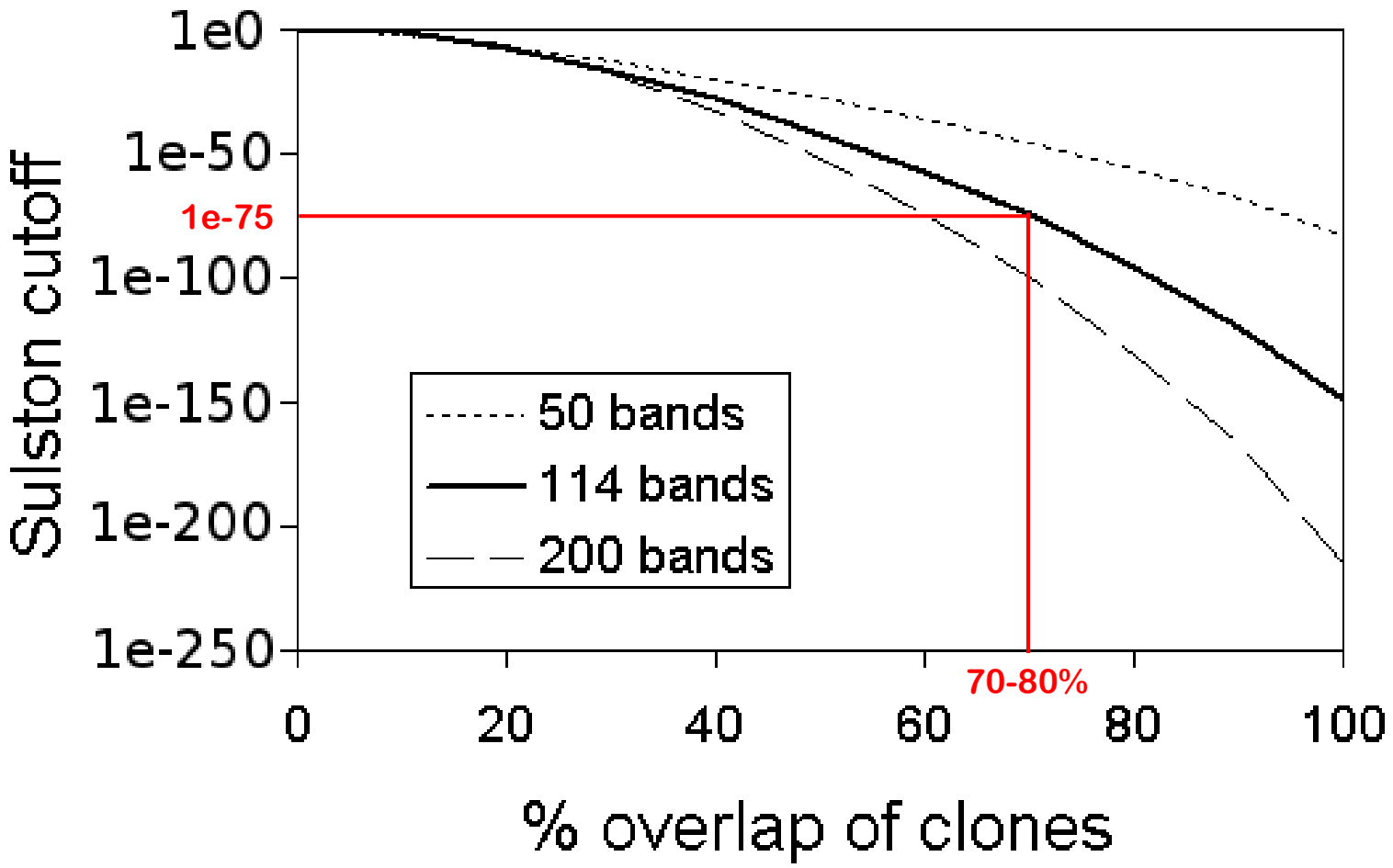


Start at very high stringency (1e-75)

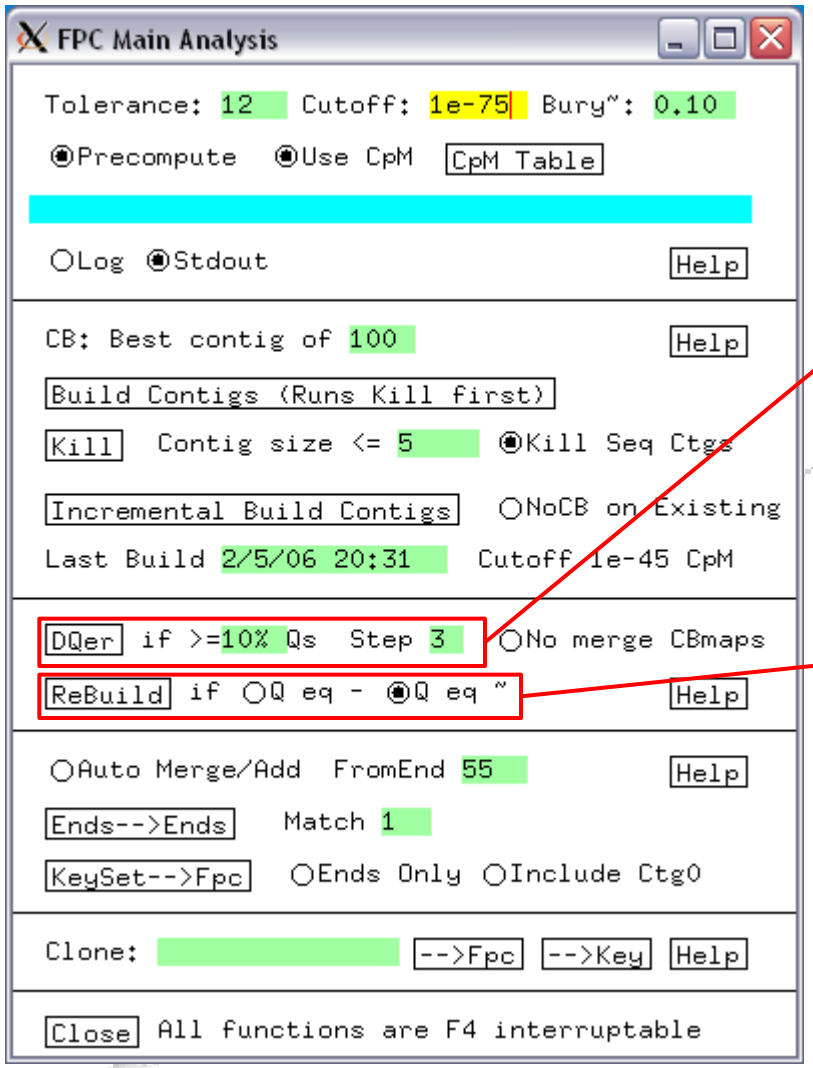
Start a new assembly

Compute newly added fingerprints

Sulston score overlap



DQing contigs



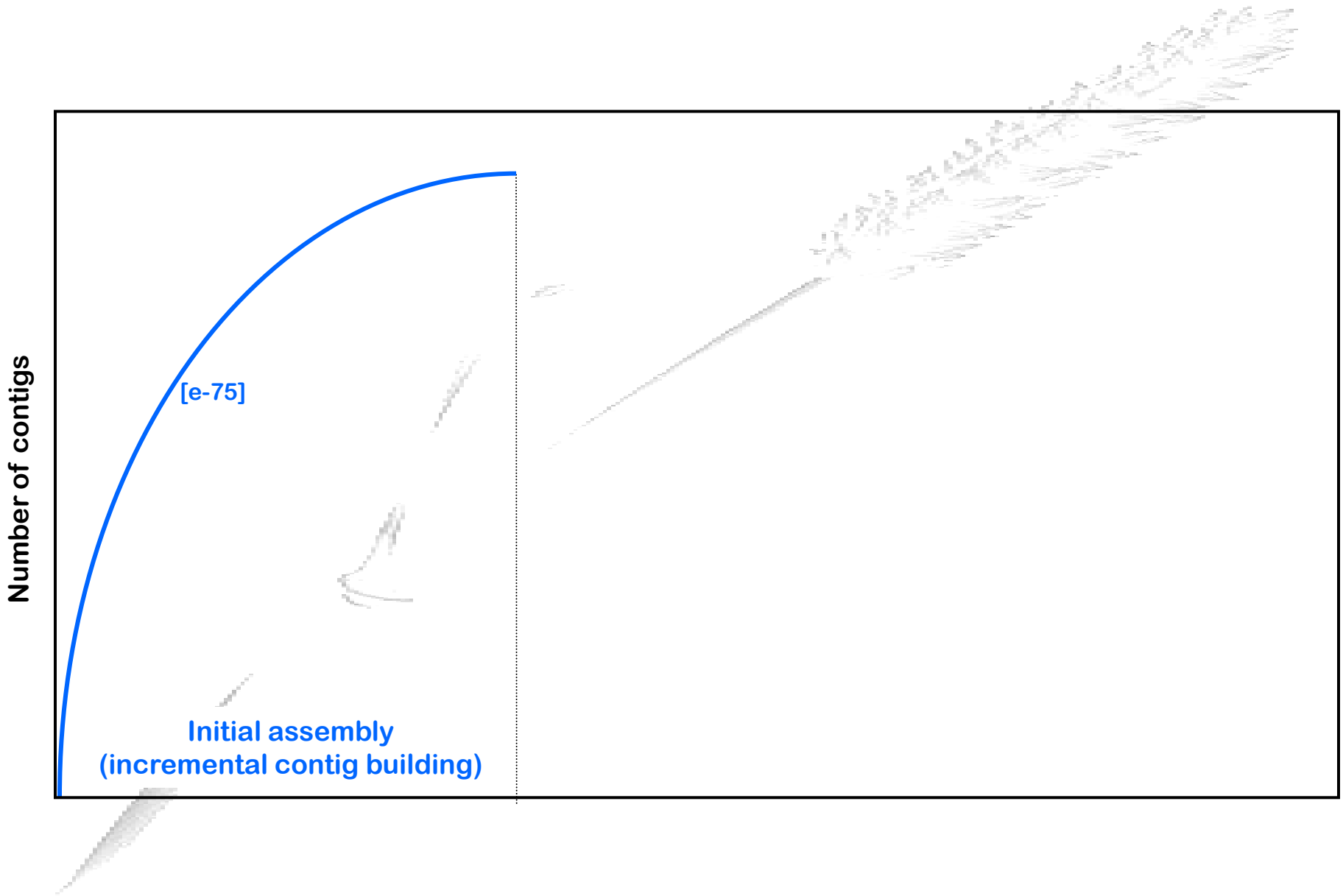
1- DQer

- ✓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

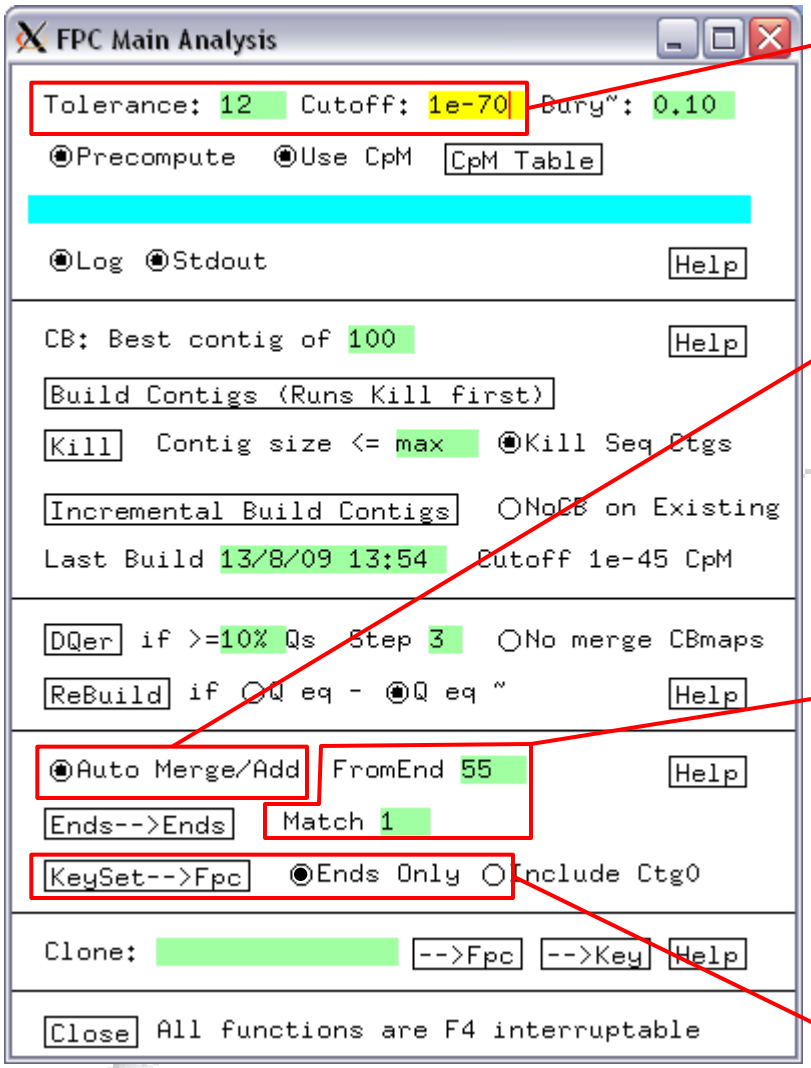
3- If necessary, perform a new DQer step, starting at 1e-84, followed by Rebuild…

Assembly of the physical map



Contig assembly
4- Automated assembly

Single-to-end merging



Decrease the stringency stepwise

(1e-70, 1e-65, 1e-55, 1e-50, e-45)

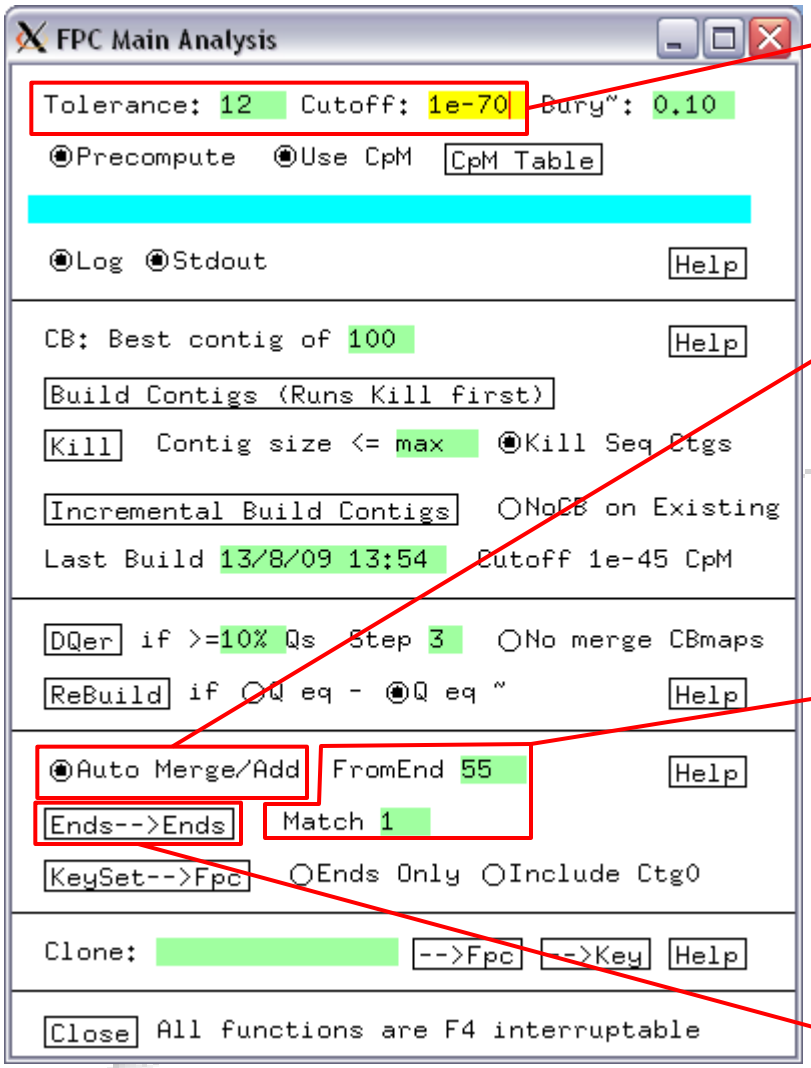
Select Automerge for automatic merging

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



Decrease the stringency stepwise

(1e-70, 1e-65, 1e-55, 1e-50, e-45)

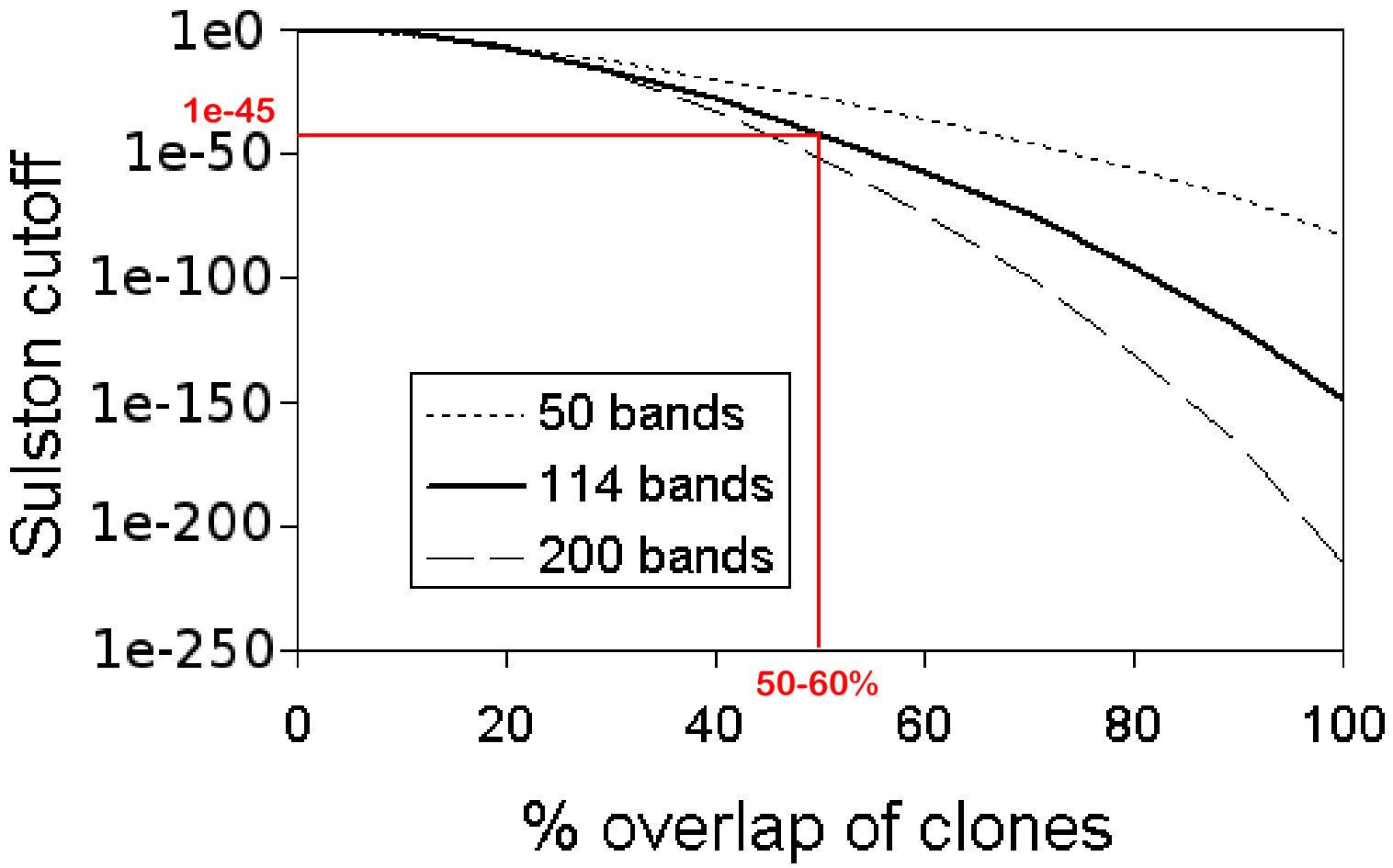
Select Automerge for automatic merging

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)

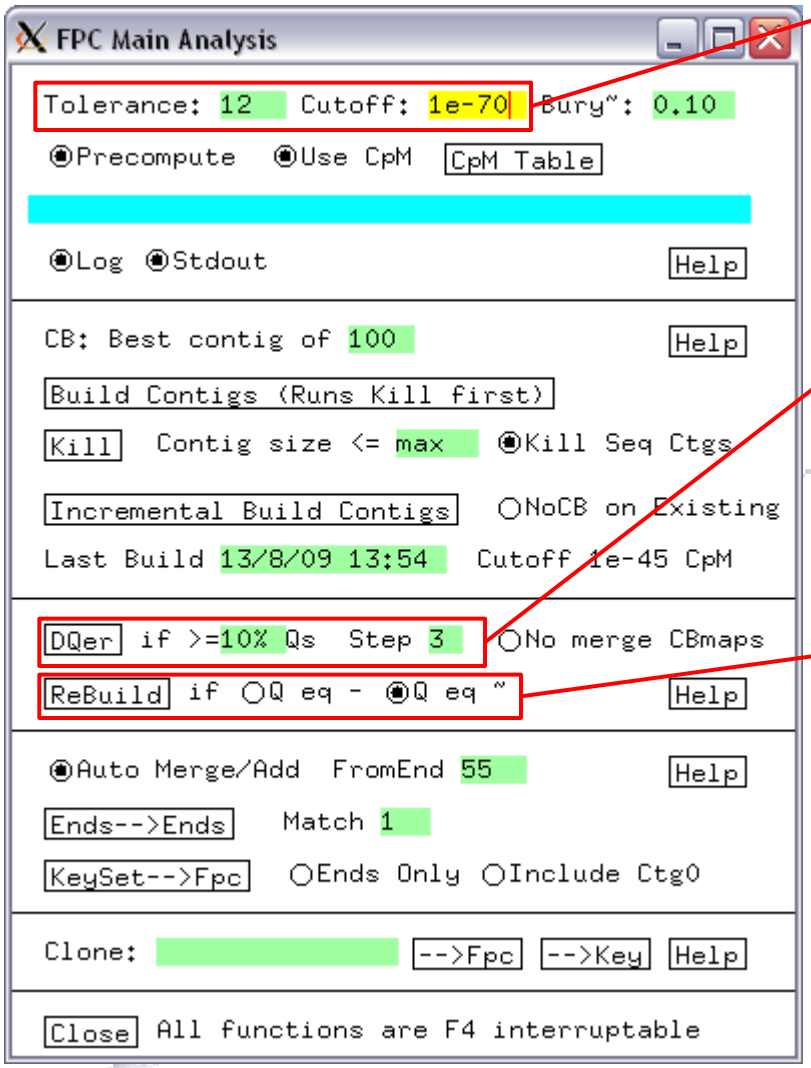
Match tells the number of clones from one contig that have to match with another contig for merging

Perform end-to-end merging

Sulston score overlap



DQing contigs



1- Rebuild contigs at merging stringency

2- DQer

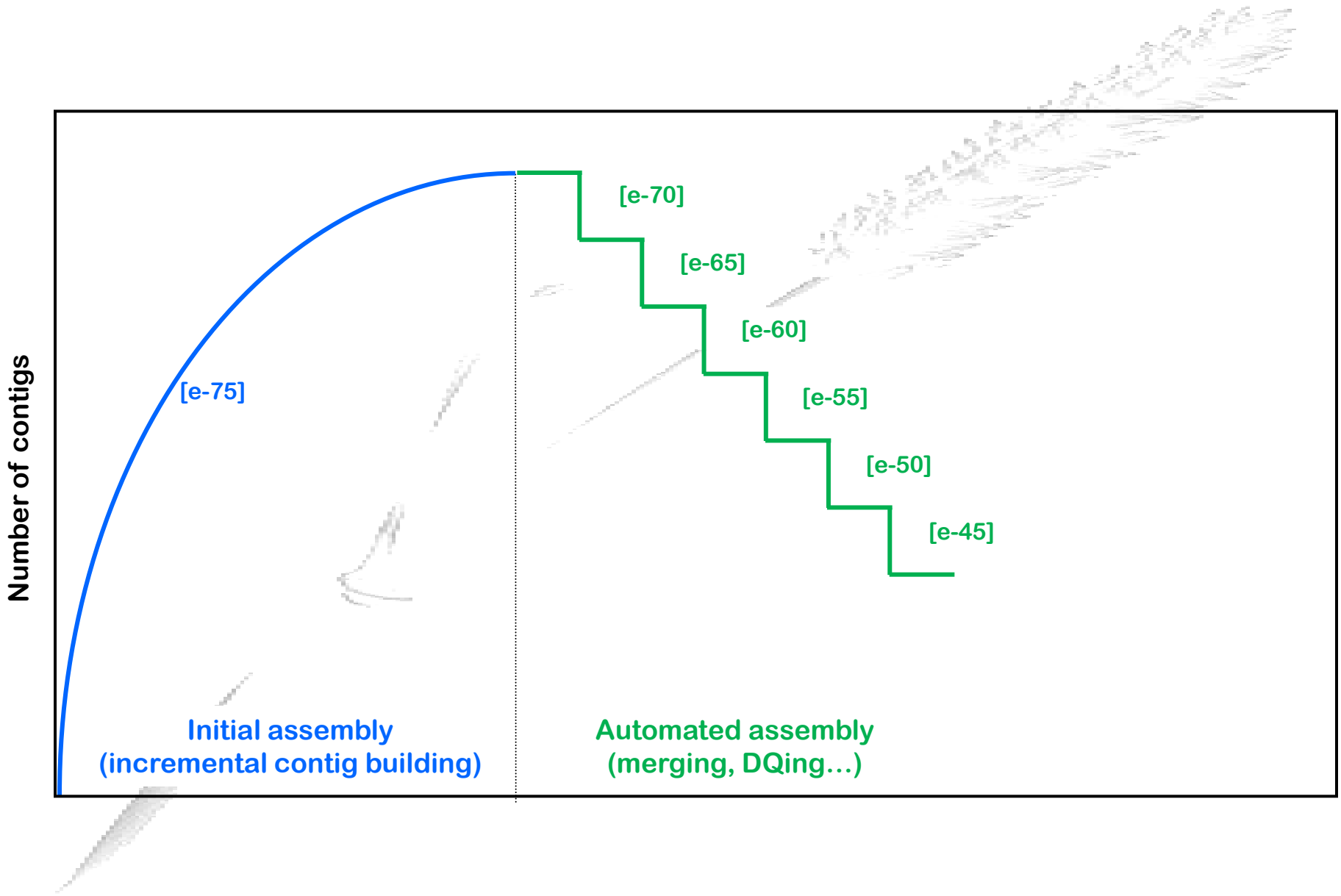
- ✓decreasing the cut-off to remove Qs
- ✓only for contigs having more than 10% Qs
- ✓Three times

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…

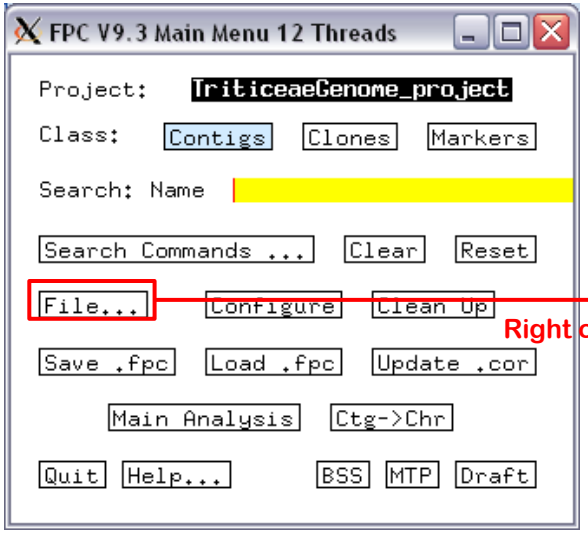
5- Perform single-to-end and end-to-end merging until 1e-45

Assembly of the physical map

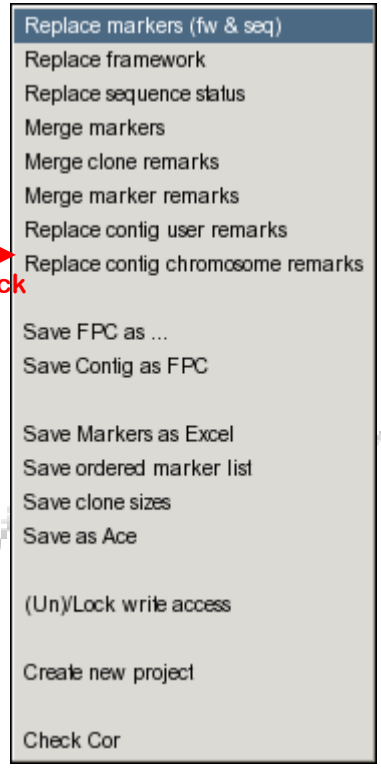


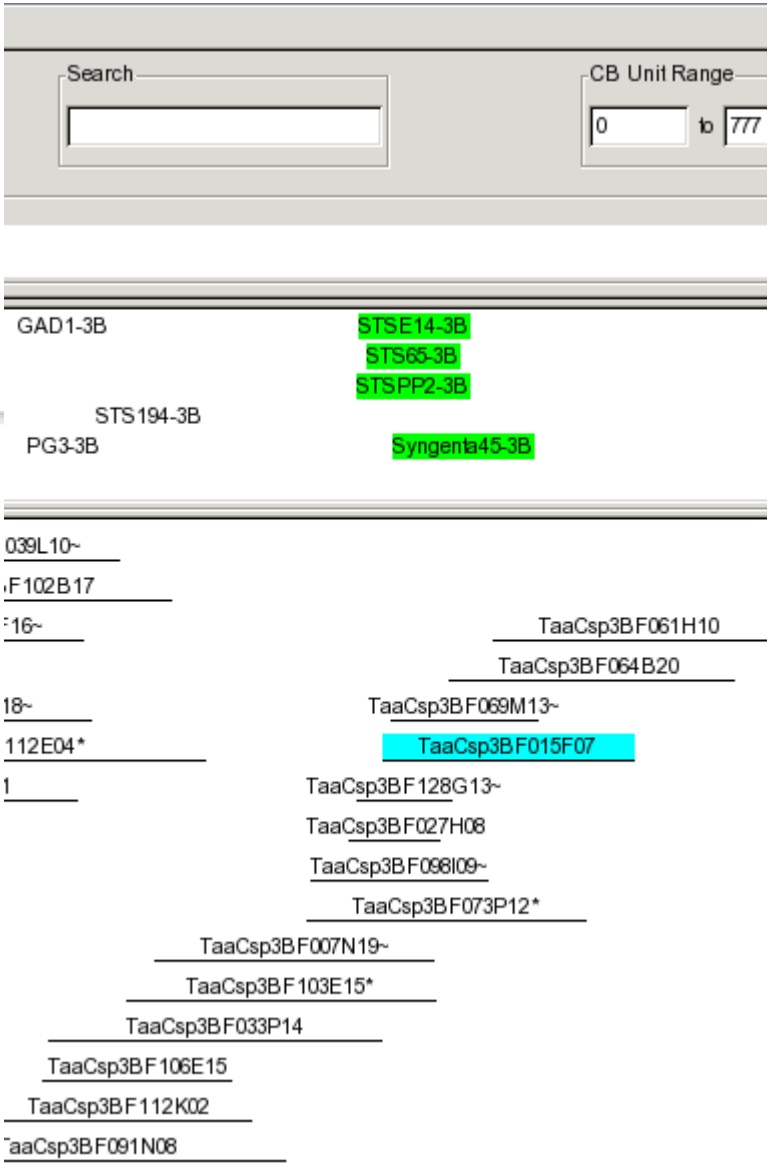
Contig assembly
5- Manually-edited assembly

Adding markers



Right click





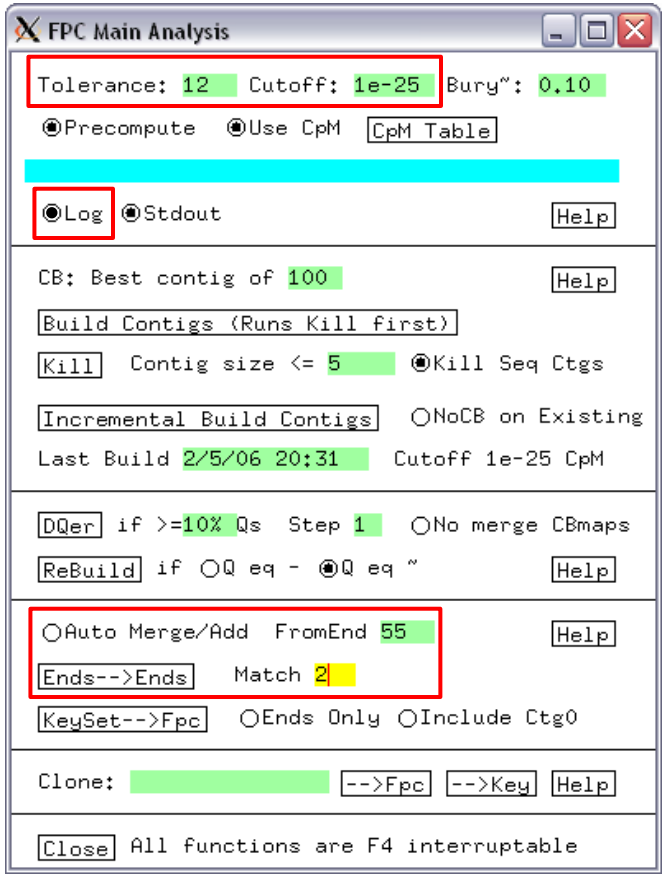
Marker.ace file

```
Clone    :    "TaaCsp3BF015F07"
Positive_STS     "STS65-3B"
Positive_STS     "STSPP2-3B"
Positive_STS     "STSE14-3B"
Positive_STS     "Syngenta45-3B"

Clone    :    "TaaCsp3BF015F08"
```

Files/

Looking for small overlaps

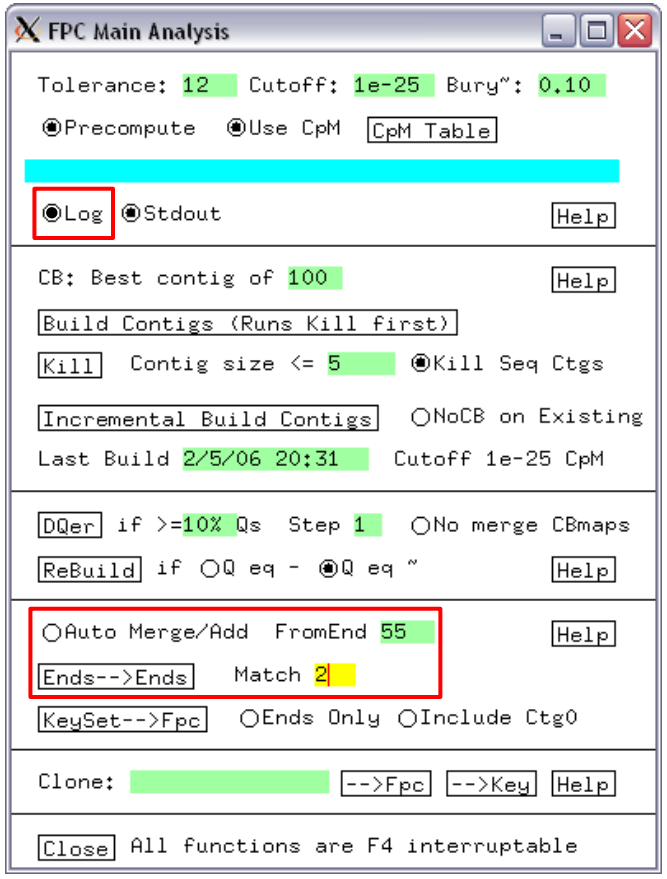


✓ **Stdout (screen)**

✓ **.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




```
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
```

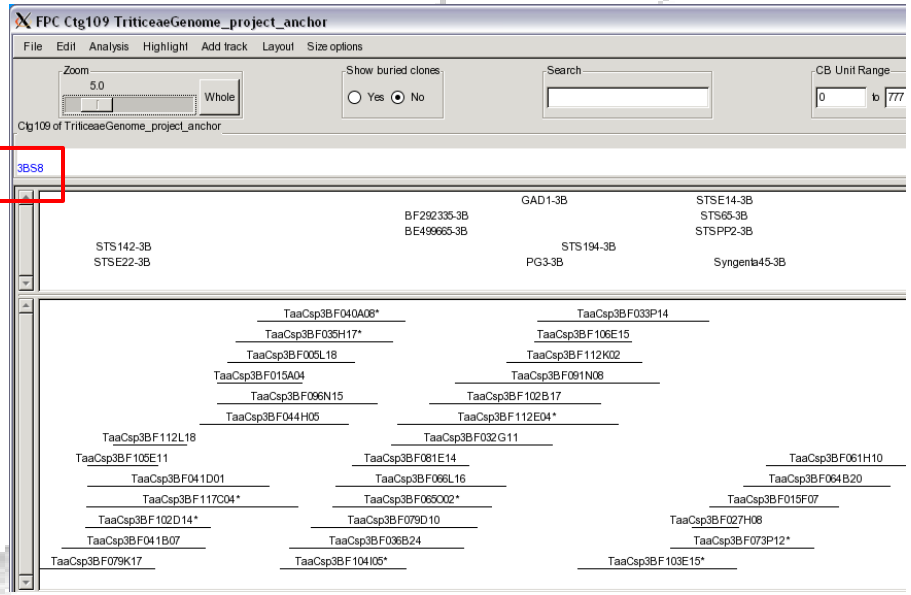
⇩

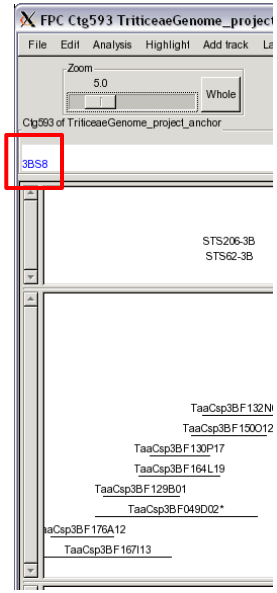
Perform merging
(unless mapping data are conflicting)

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

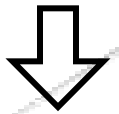
Check mapping data
& perform merging if mapping data are consistent





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 TaaCsp3BF007O23 165b  Ctg742  L TaaCsp3BF152K02  74b  Match  42 4e-29
Match: 7R 742L cutoff:4e-29
```

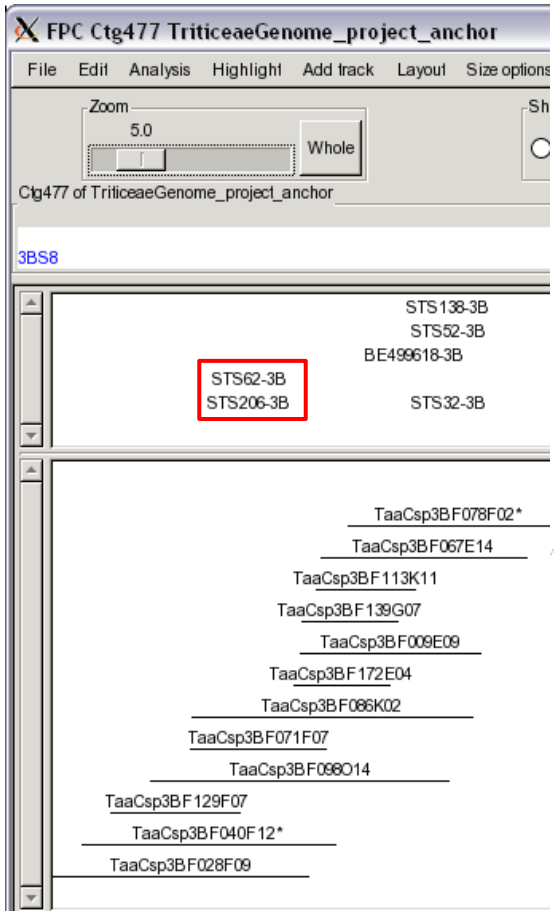


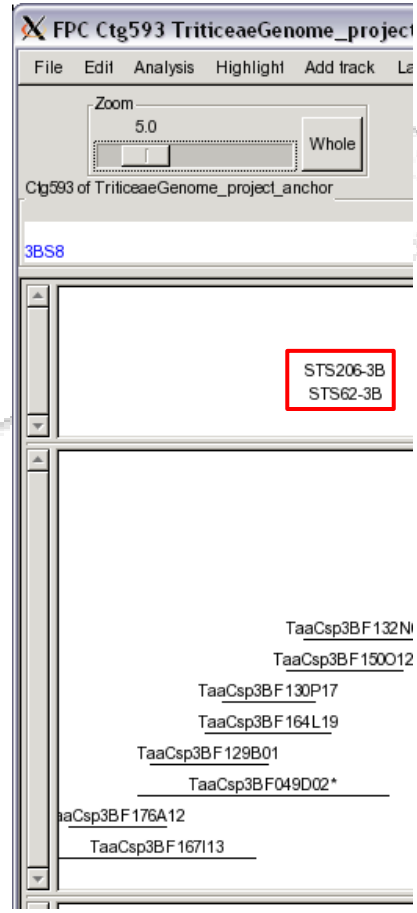
Check manually

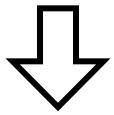
✓ **Small contig included into the others**

✓ **Chimeric clones…**

No match but shared markers

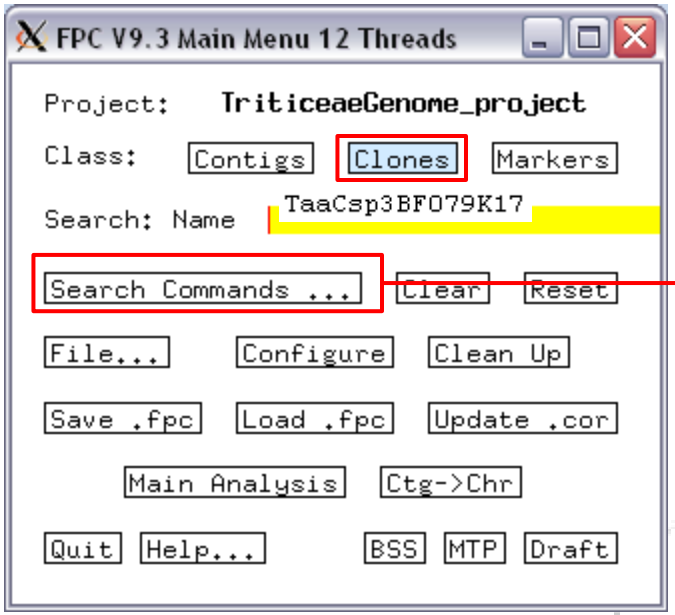


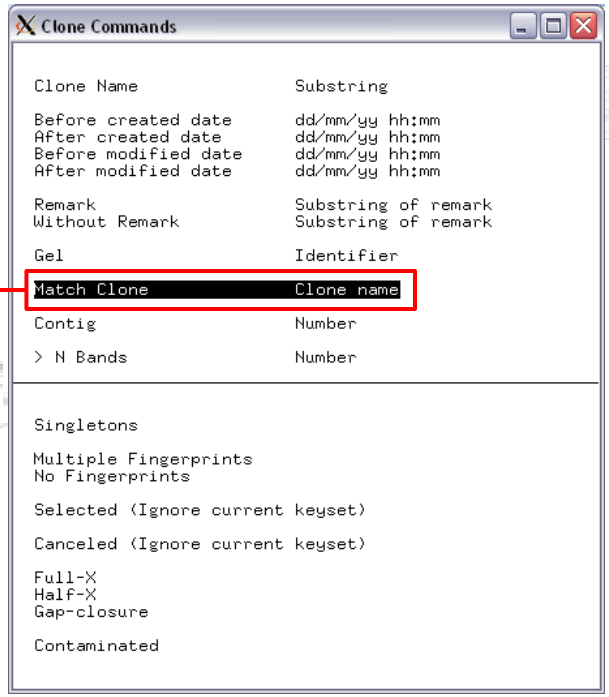


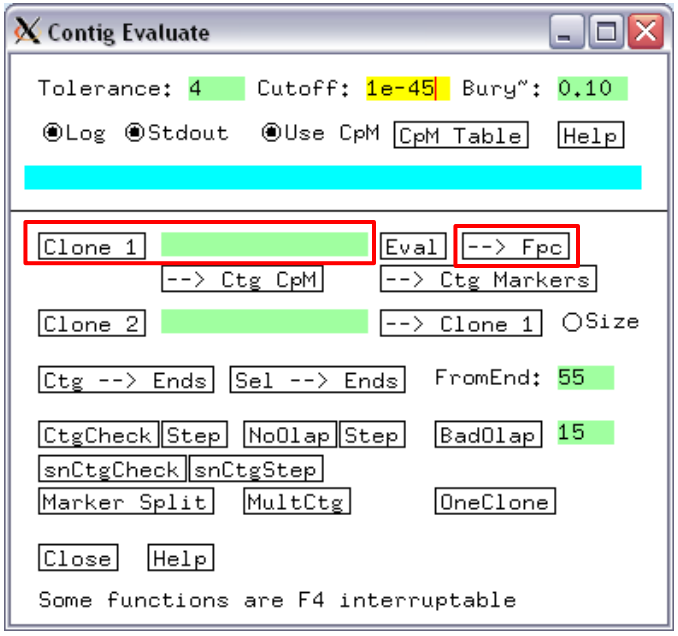


Perform merging
(if marker data are reliable)

Looking for small overlaps

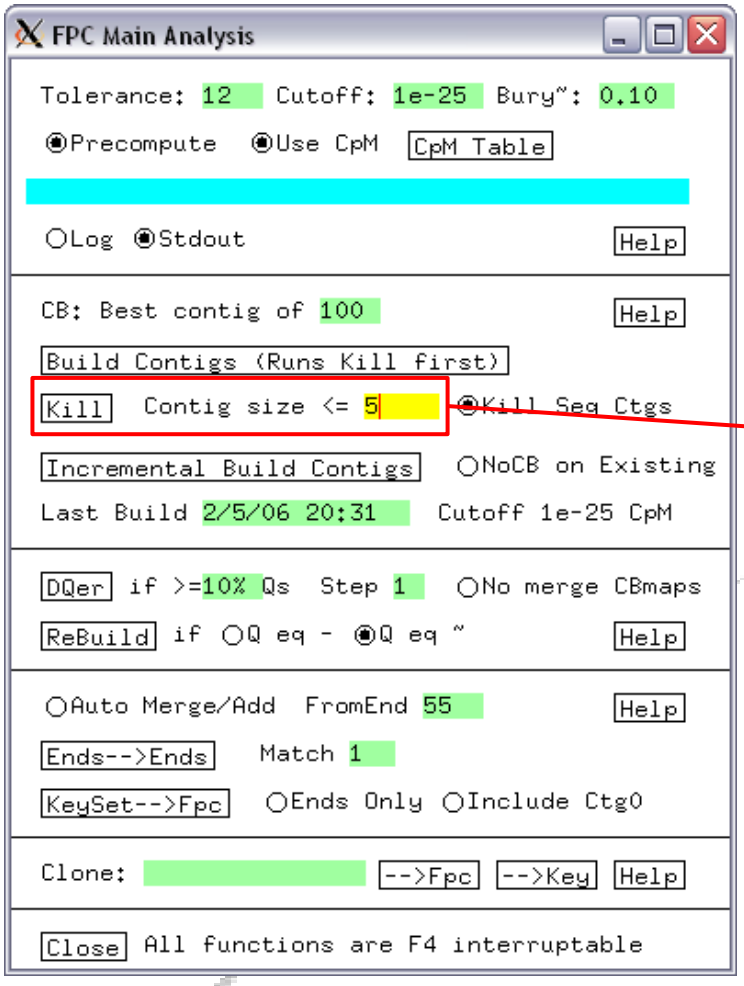






Useful to check MTP results when clones belong to 2 different contigs.

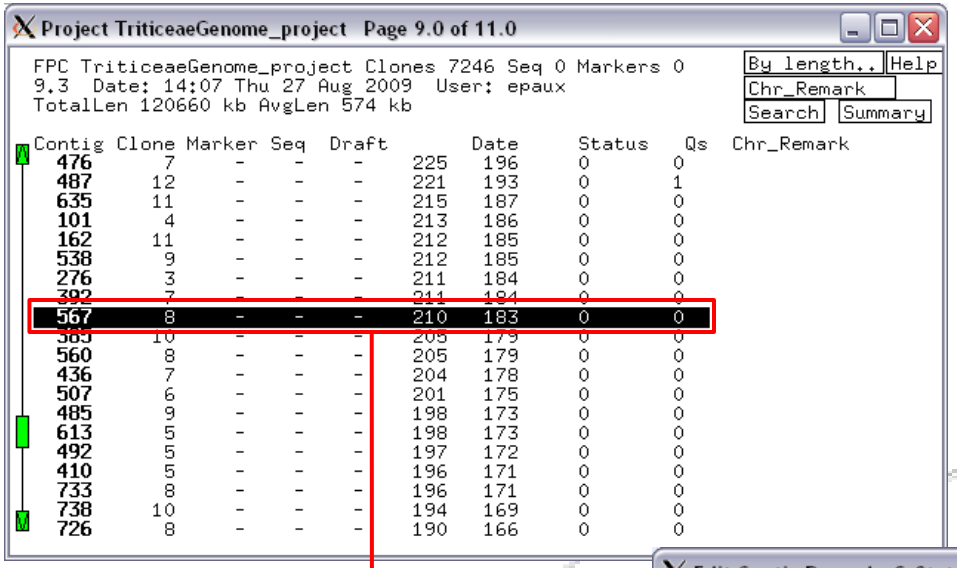
Killing small contigs



Kill contigs containing less than 6 clones

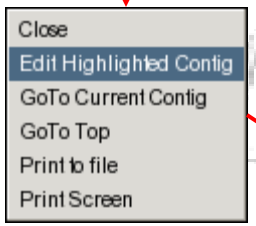
('max' to kill all the contigs)

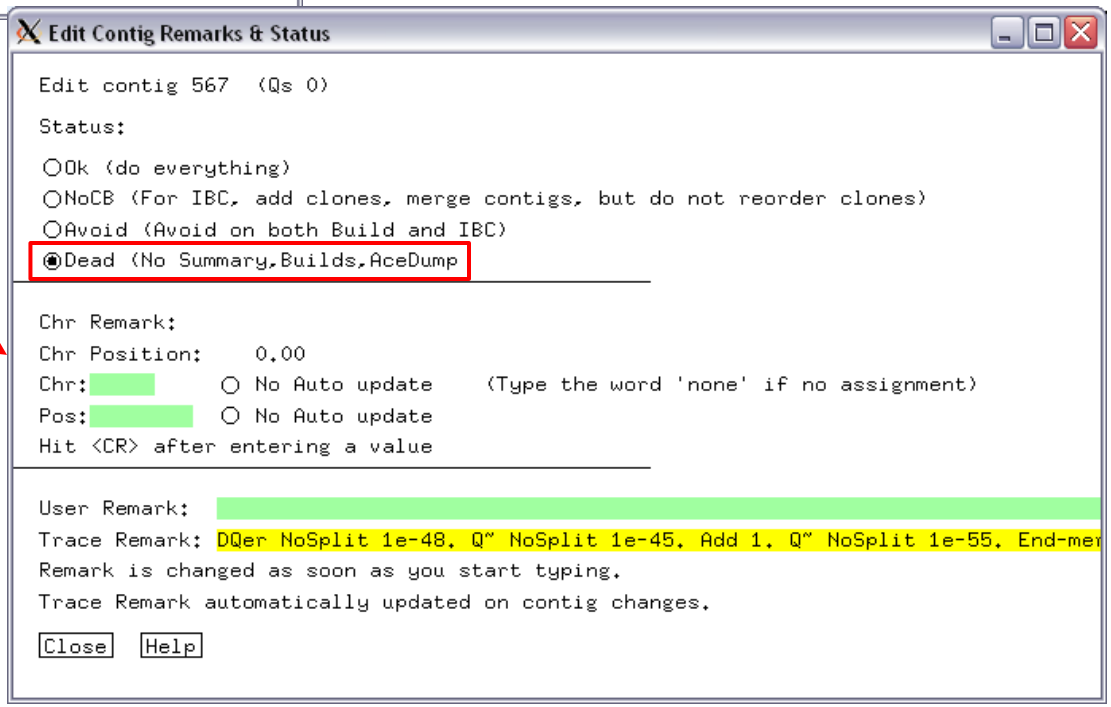
Killing small contigs



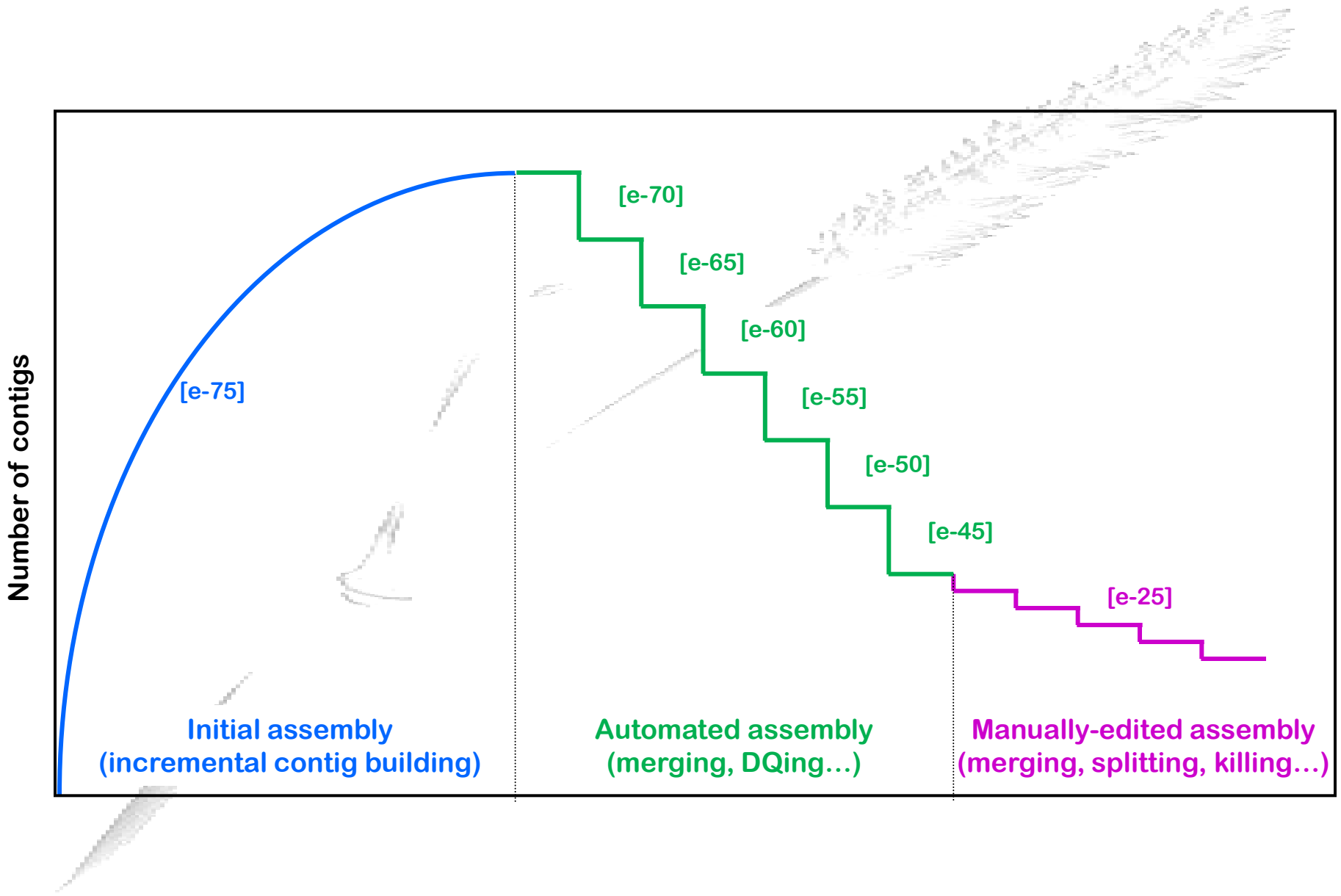
Contigs smaller than 300 kb

Right click





Assembly of the physical map

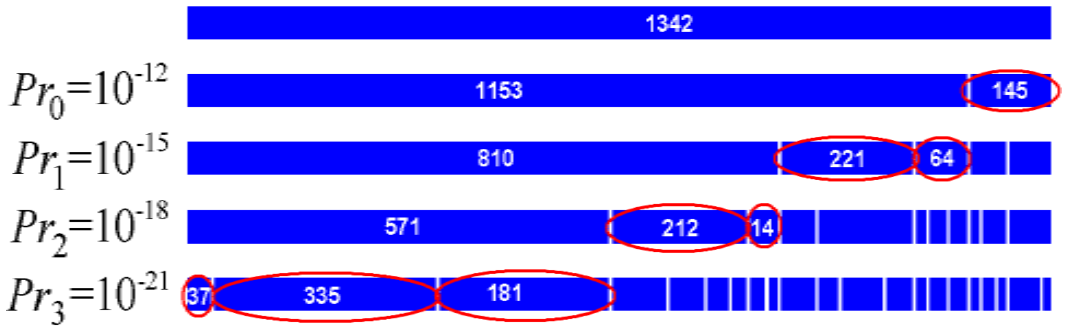


Contig assembly
6- LTC: Linear Topology Contig

LTC program

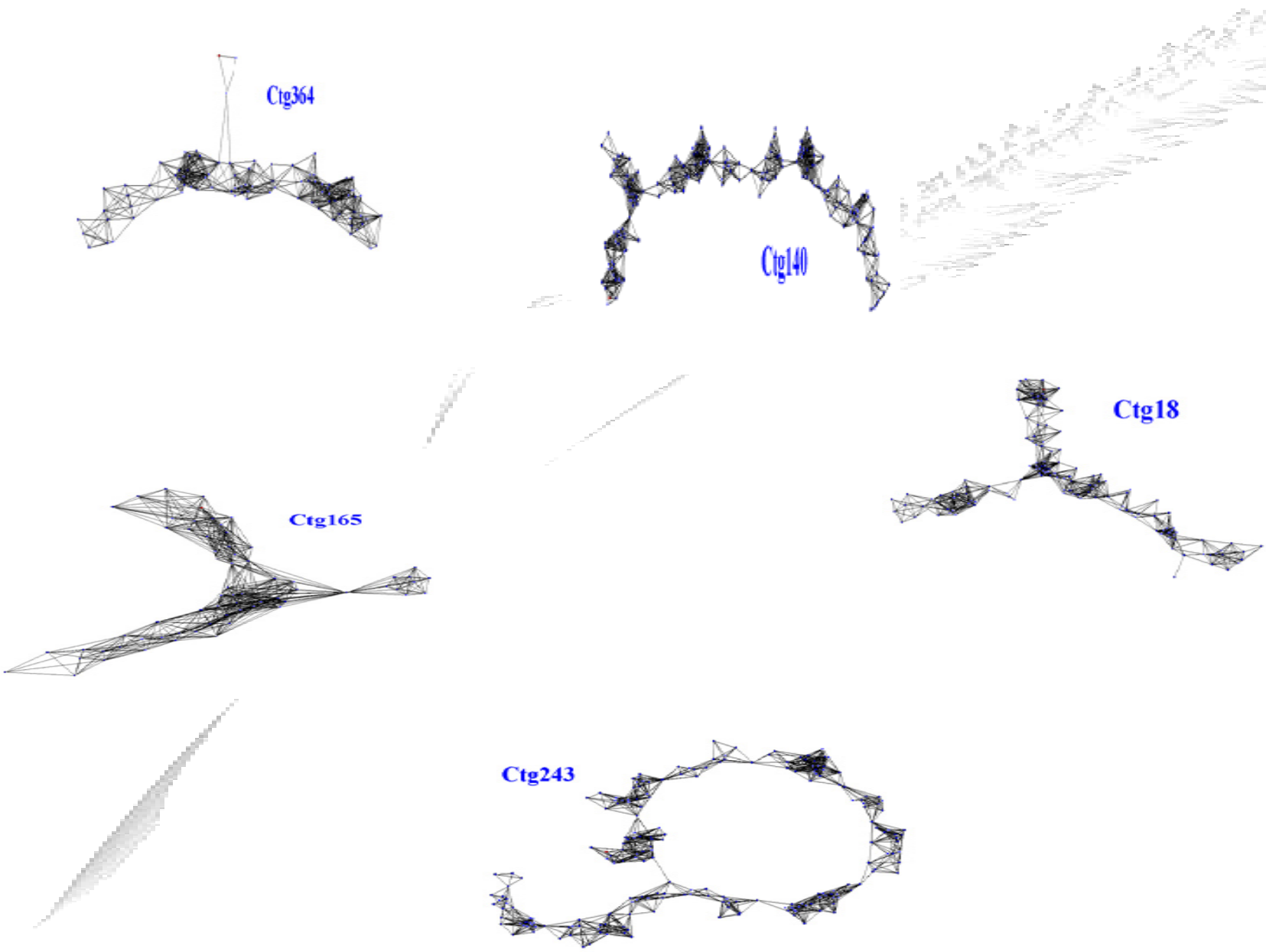
Frenkel Z, Paux E, Mester D, Feuillet C and Korol A (2009) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *Manuscript in prep.*

- ✓ 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.



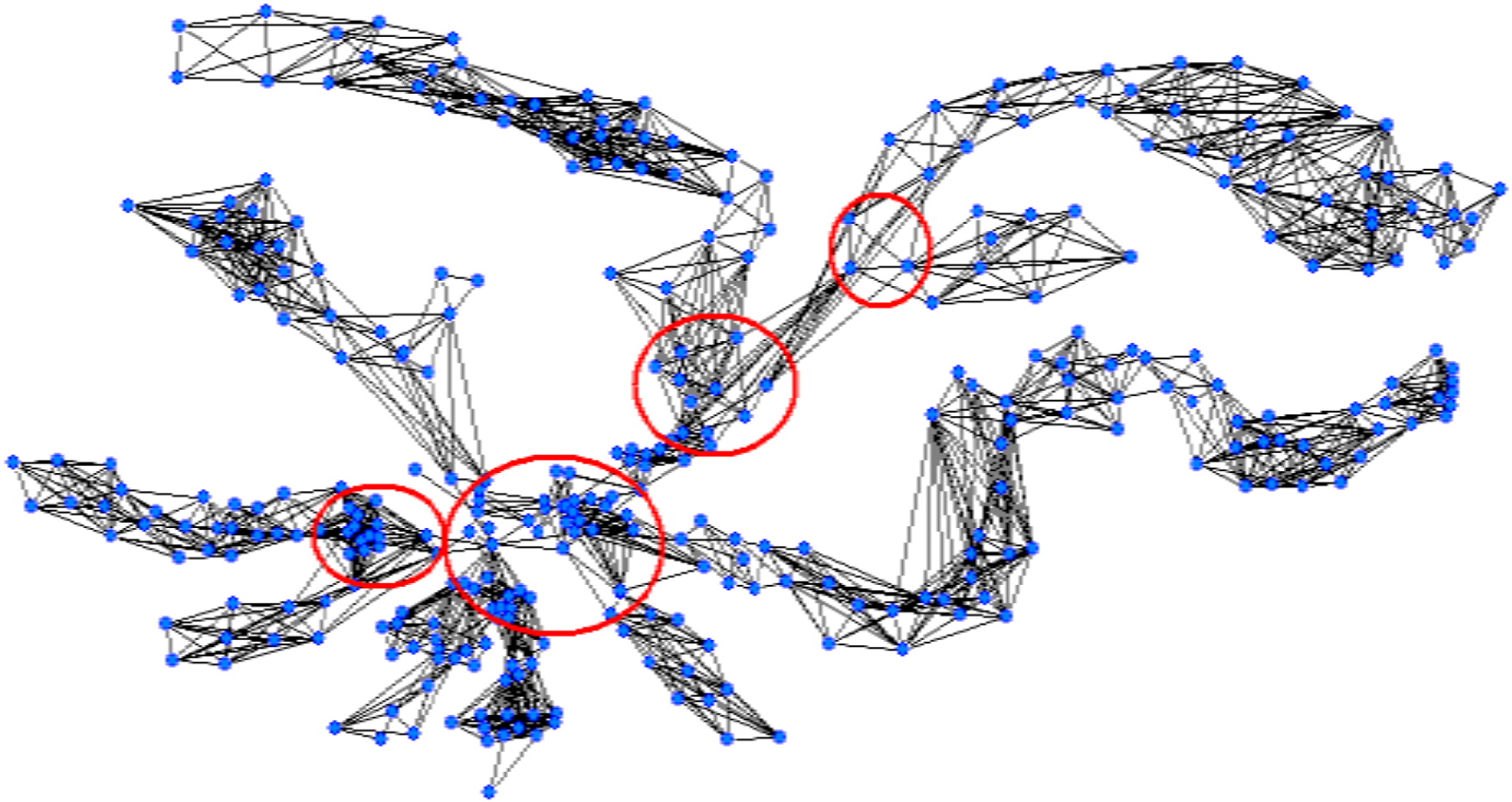
(kindly of A. Korol)

Examples of non linear topology contigs



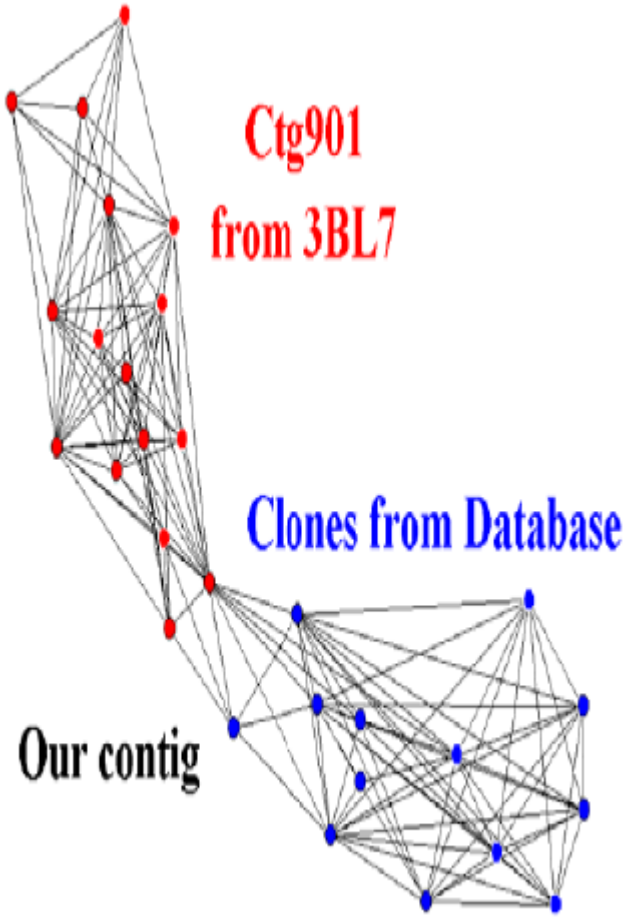
(kindly of A. Korol)

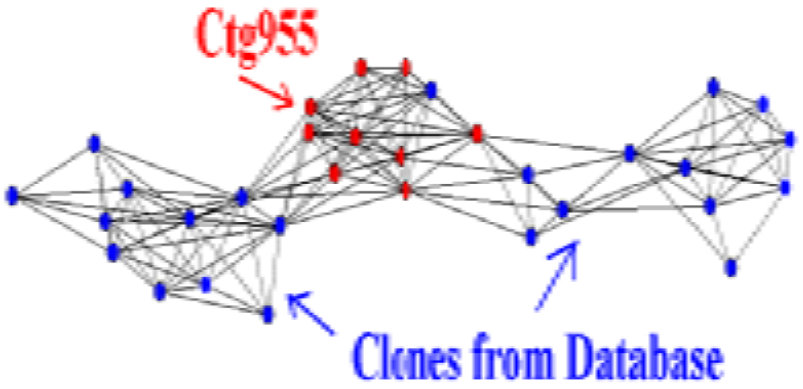
“Linearization” by removing clones in cluster branching

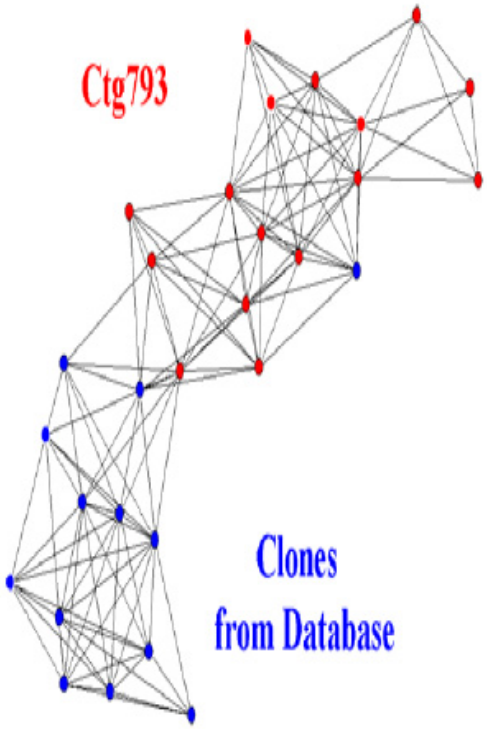


(kindly of A. Korol)

Examples of contig elongation

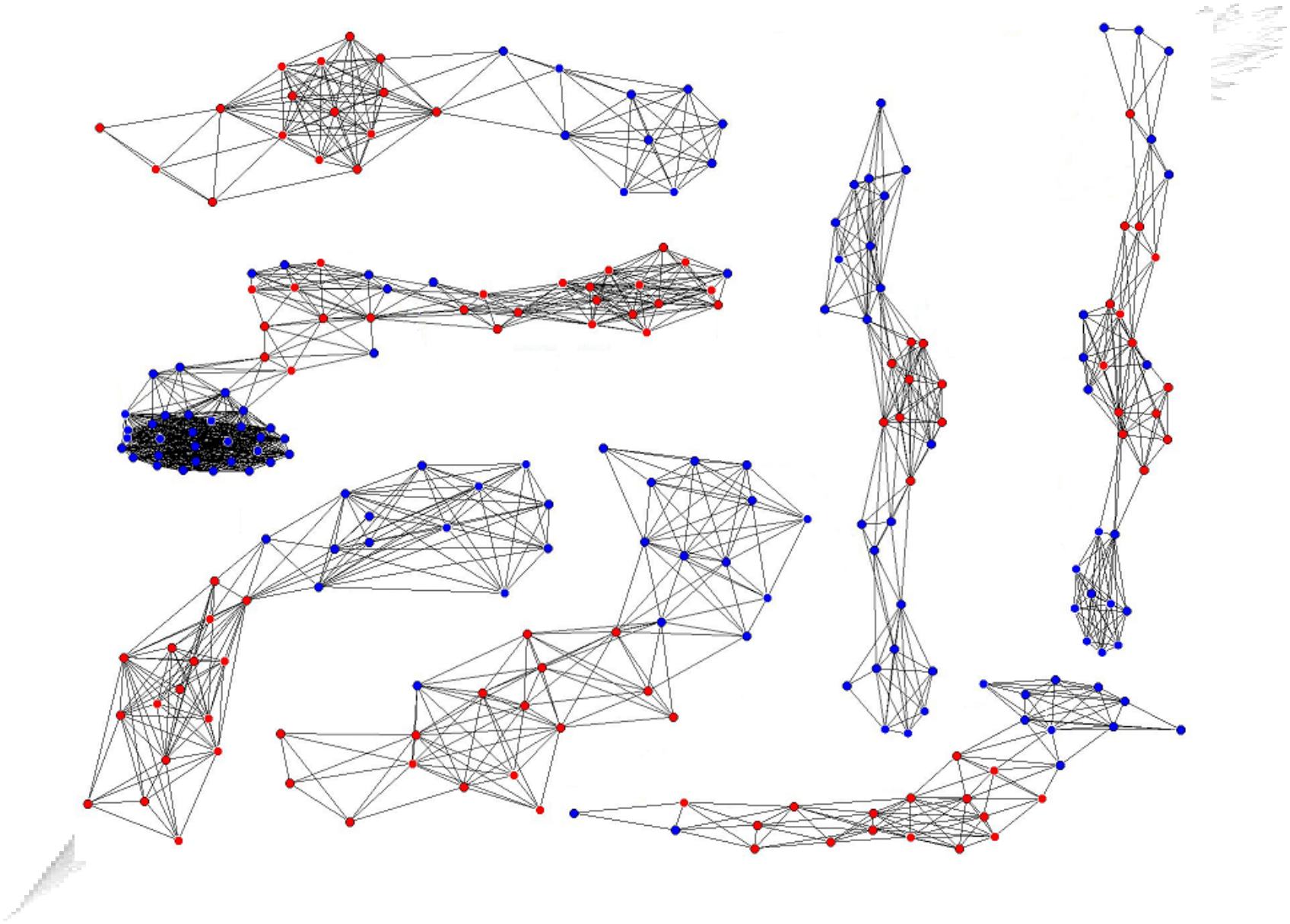






(kindly of A. Korol)

Examples of *de novo* assembled contigs



(kindly of A. Korol)