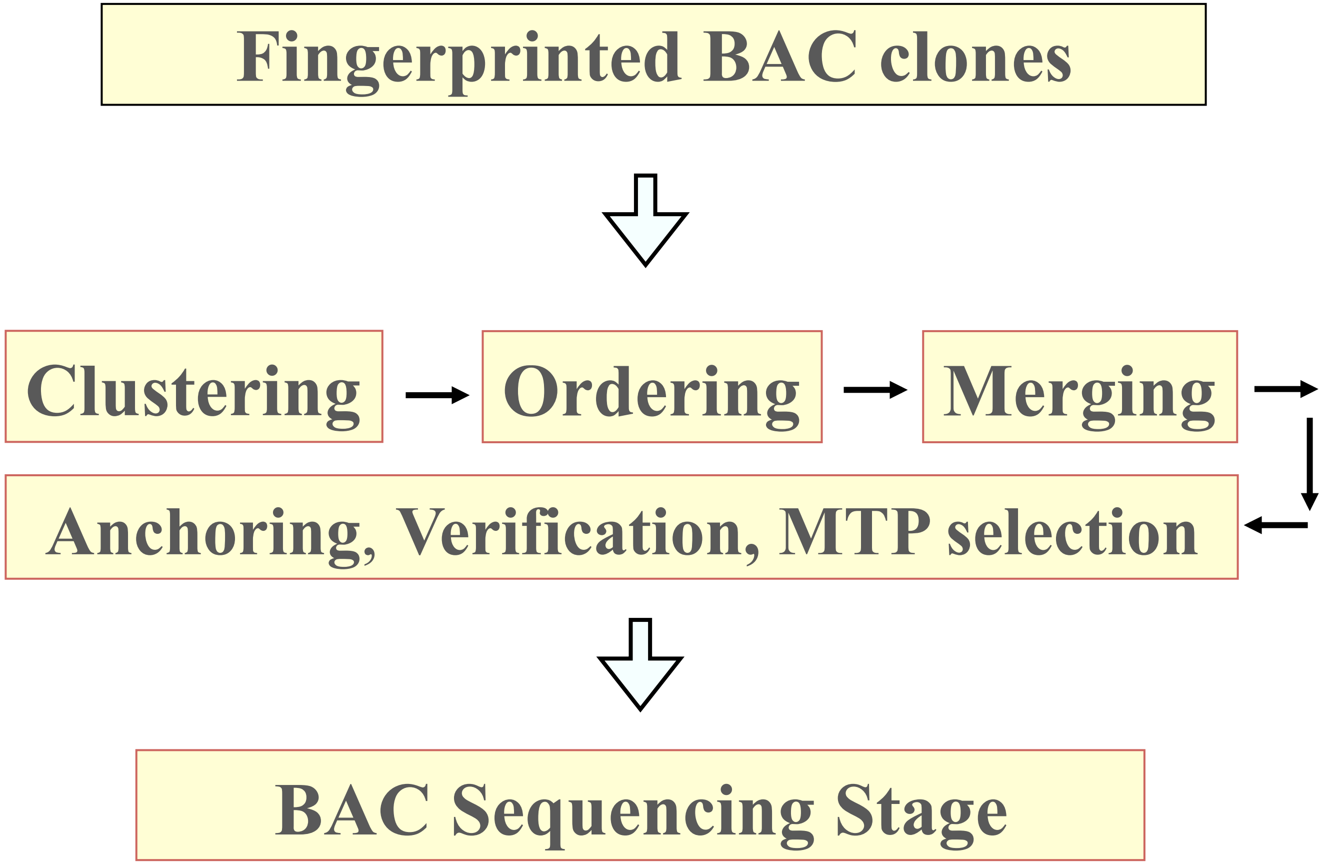
Using LTC Software for Physical Mapping and Assisting in Sequence Assembly

Z. Frenkel, V. Glikson, A. Korol
Institute of Evolution, University of Haifa

See also poster P1194

PAGXXIII, San Diego, January 2015

Physical mapping, anchoring, sequencing



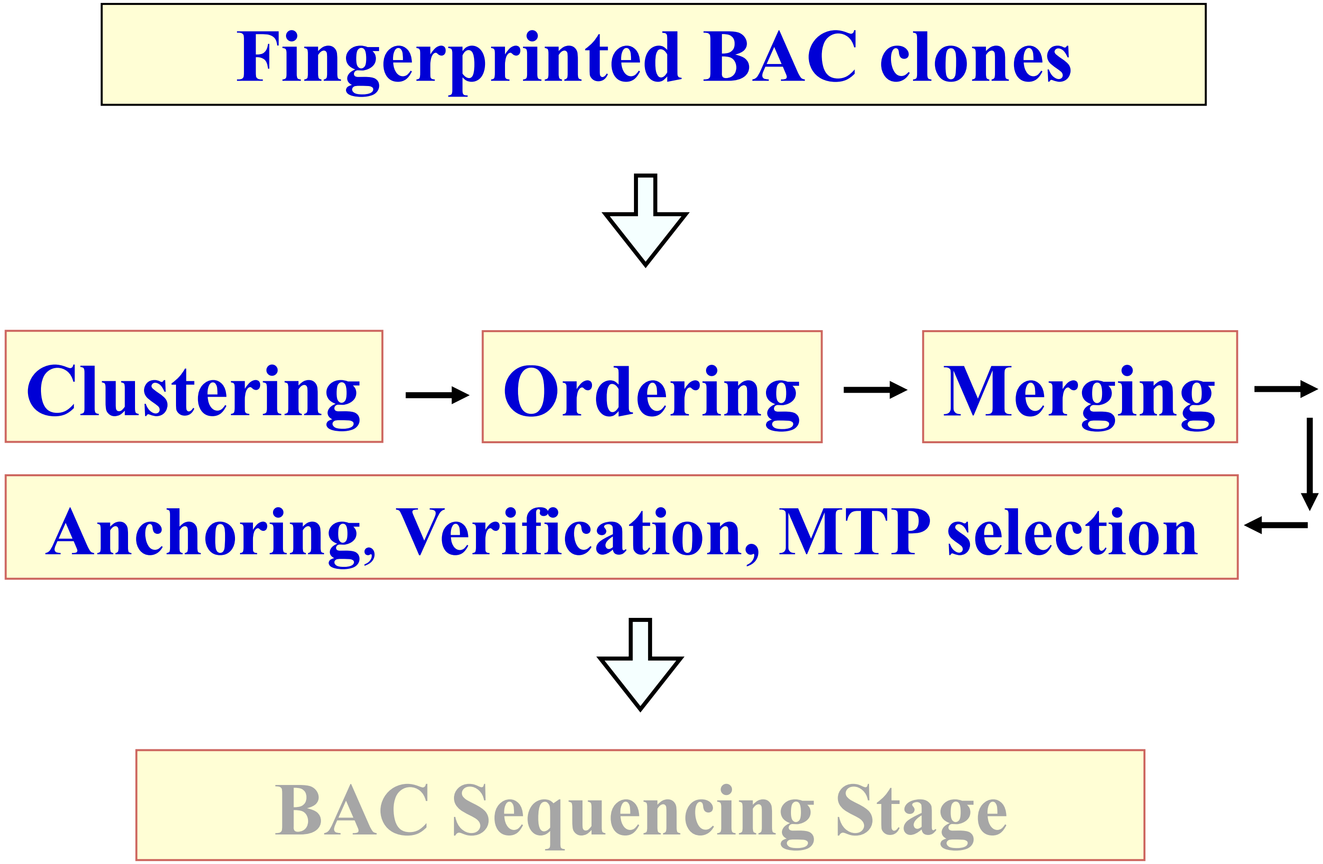
Fingerprinted BAC clones

Clustering → Ordering → Merging

Anchoring, Verification, MTP selection

BAC Sequencing Stage

Physical mapping, anchoring, sequencing



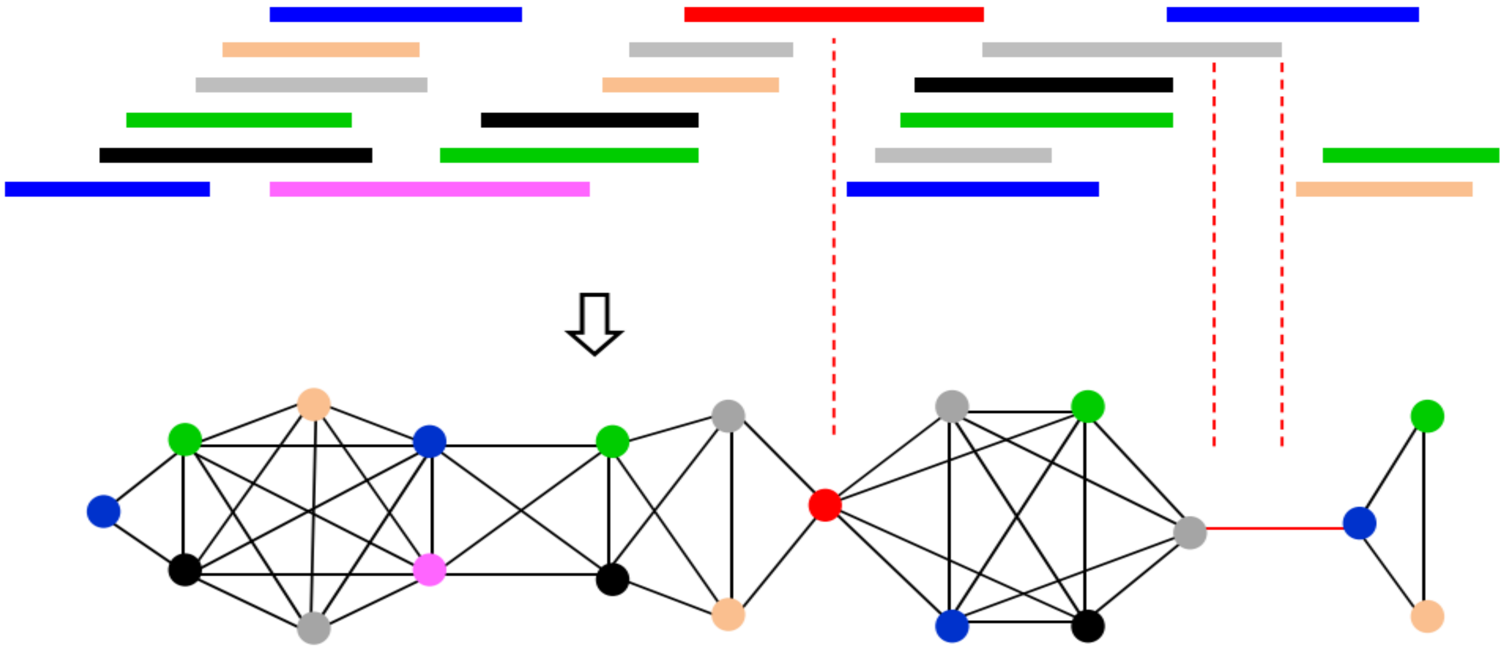
Fingerprinted BAC clones

Clustering → Ordering → Merging

Anchoring, Verification, MTP selection

BAC Sequencing Stage

Network representation of significant clone overlaps

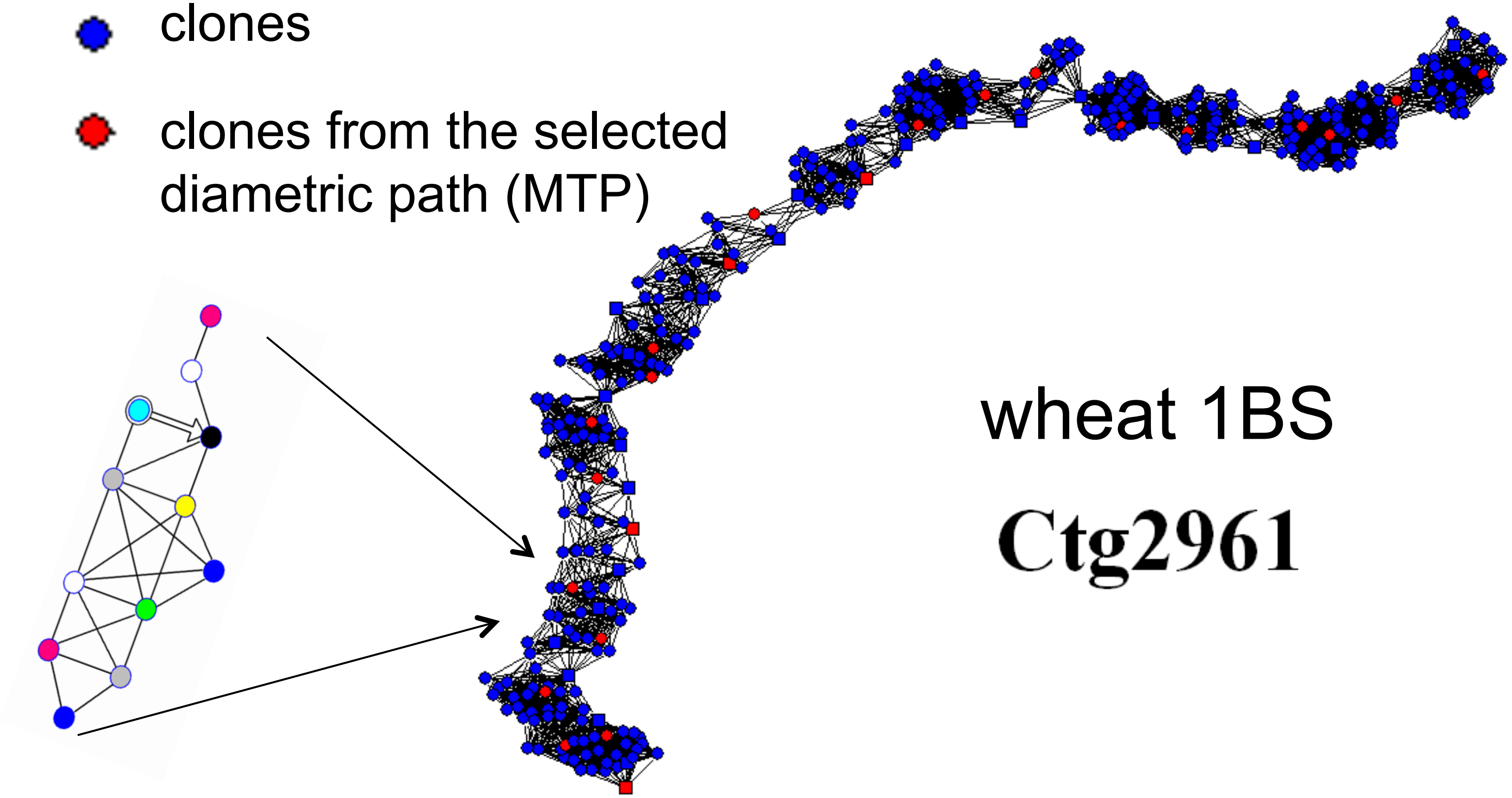


Unproven clone

Unproven overlap

Network representation of significant clone overlaps

- 🔵 clones
- 🔴 clones from the selected diametric path (MTP)

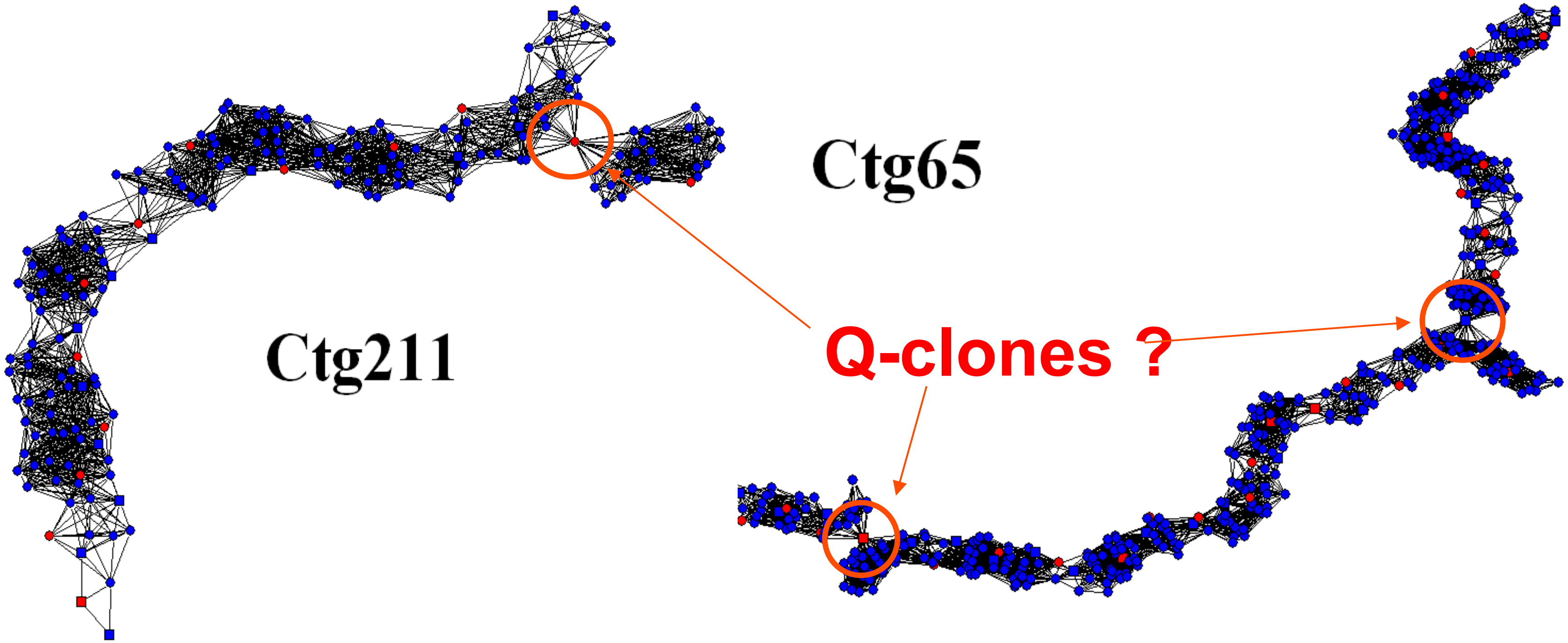


wheat 1BS

Ctg2961

Testing FPC contig quality by using LTC

Some FPC contigs have non-linear topological structure inconsistent with chromosome linear

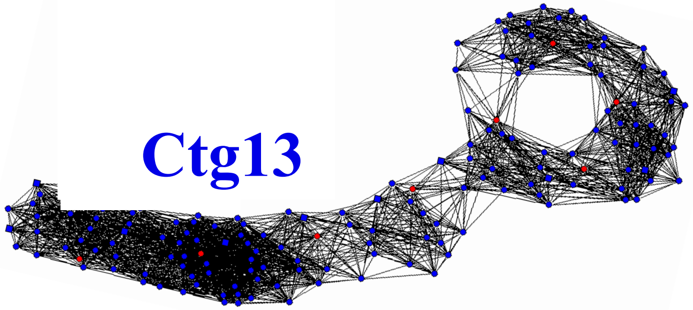


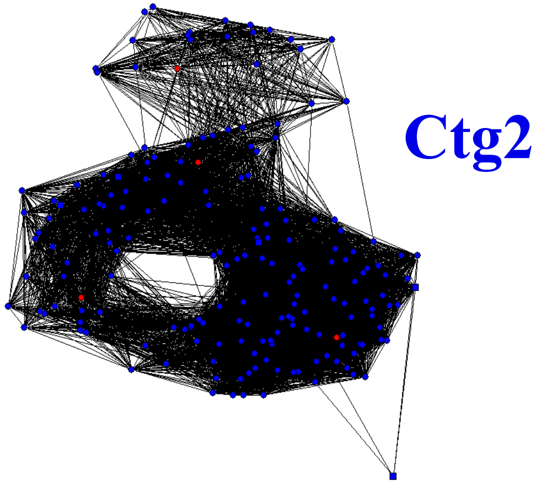
Q-clones ?

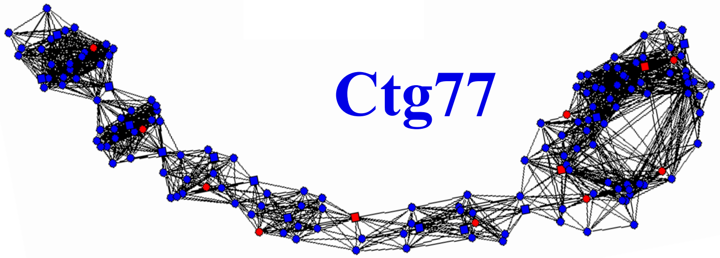
Vertices represent the clones; **edges** represent the significant overlaps (with cutoff 1e-25 Sulston score)

Testing FPC contig quality by using LTC

FPC contigs with non-linear topology and even cycles







Edges represent significant overlaps (with cutoff 1e-25 Sulston score). Increasing the stringency up to e-75 **does not help in non-trivial linearization!**

A big cluster with highly overlapped clones

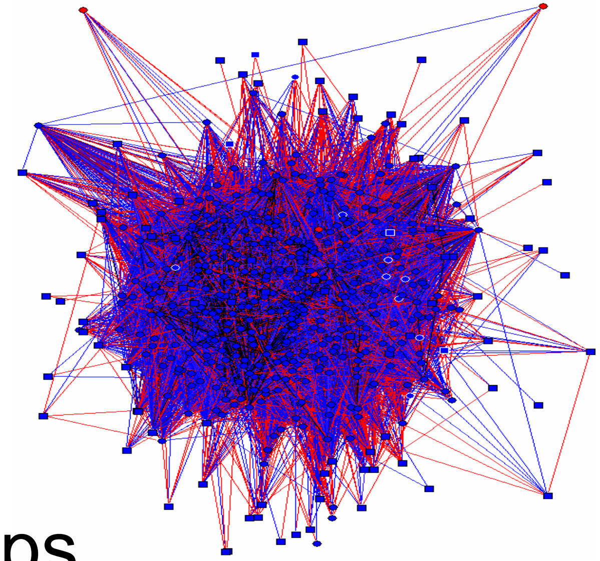
1BS: after exclusion of Q-clones at cutoff 10^{-30} we obtained cluster of 2110 clones.

⇩

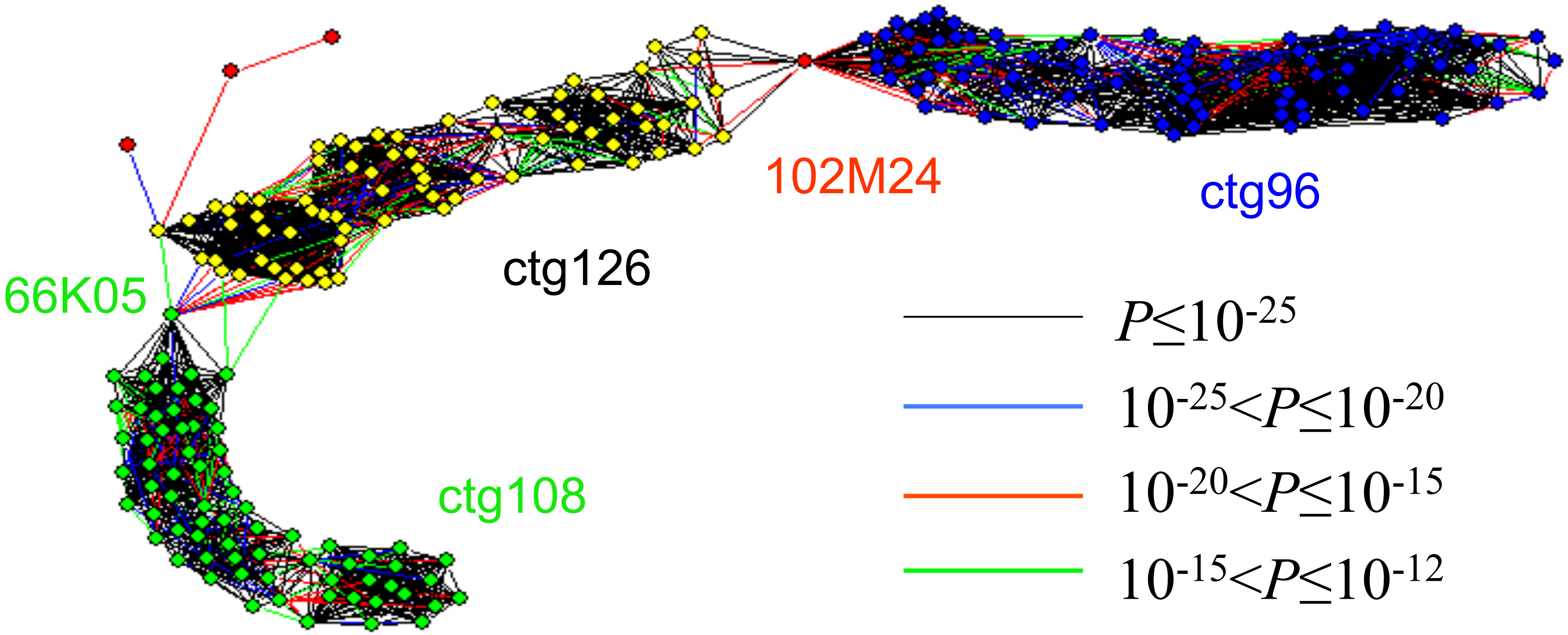
Manual separation of several linear parts

⇩

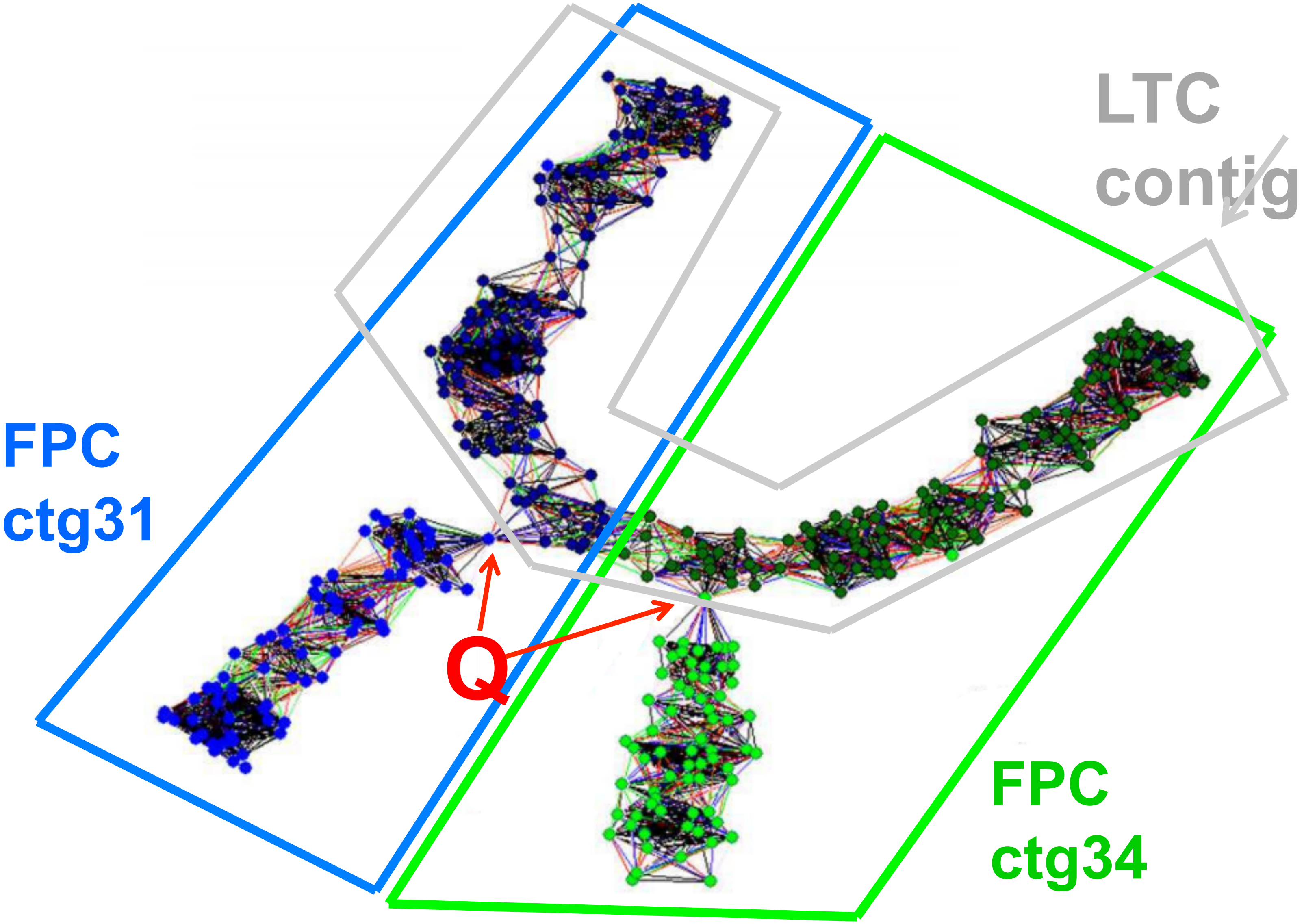
Cluster with 1218 clones,
~870 highly significant overlaps (cutoff 10^{-50}) per clone



End-to-end contig merging



LTC scaffolds vs. FPC contigs


LTC contig

FPC ctg31

Q

FPC ctg34

LTC scaffolds vs. FPC contigs

- Visual and analytical control of the net of significant clone overlaps
- Coordinating of scaffolding with anchoring

➔ Long well anchored physical scaffolds

Example: wheat 1BS (314 Mb, HICF, x15, ~50,000

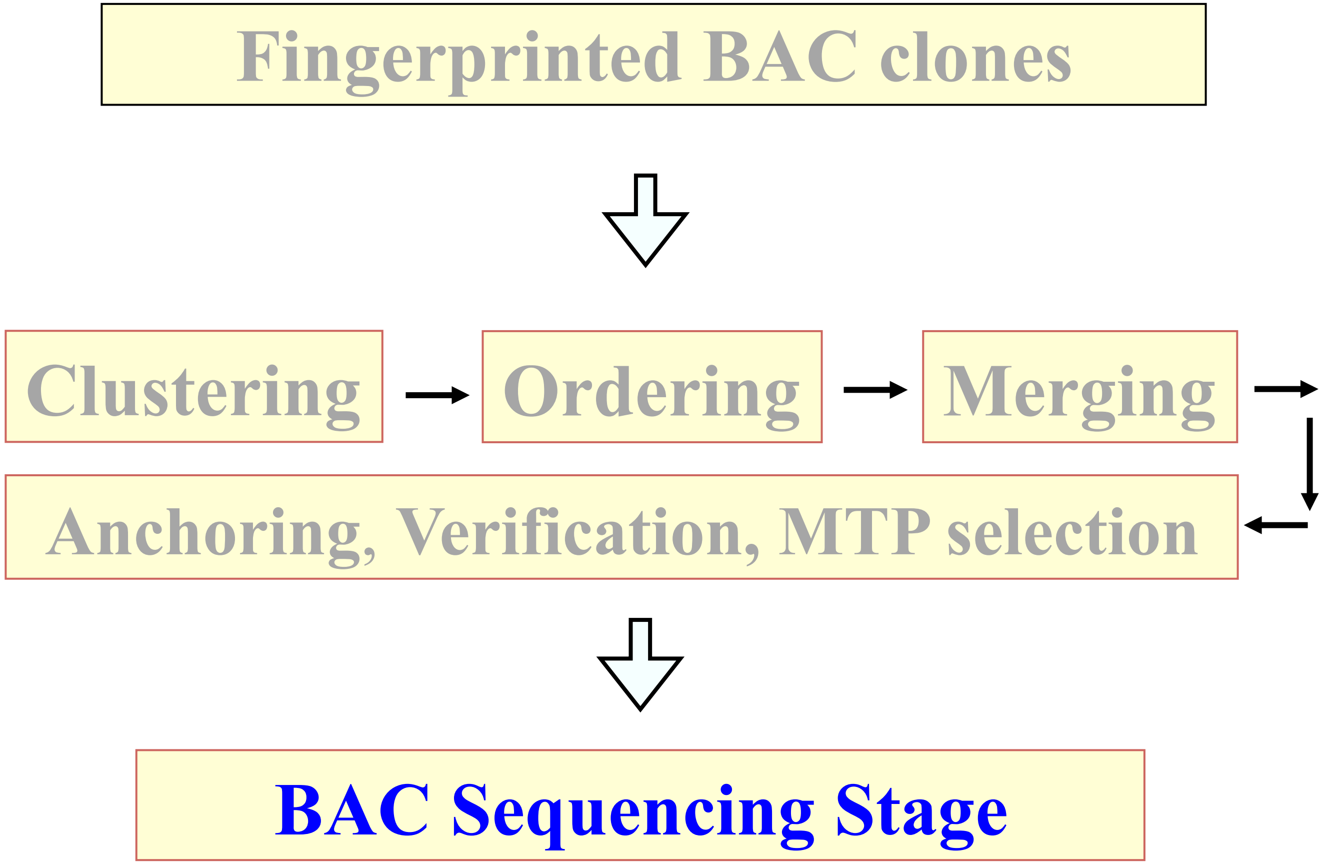
	FPC	LTC contigs	LTC scaffolds
Clones in contigs (≥6)	34,104	33,846	34,027
Longest contig (Mb)	4.7	**7.0**	**20.9**
N50 (Mb)	1.0	**2.4**	**8.5**
L50 (contigs)	81	**35**	**11**

Raats et al. *Genome Biology* *2013,* ***14:R138***

Selecting clones for sequencing

- Possibility to give priority to clones previously selected for anchoring or for BAC-end sequencing
- Higher (**more reliable**) overlaps of neighbor clones to avoid non-significant overlaps at sequence level in highly repeated genomes ➔ **less gaps**
- Double coverage at the **contig ends** (where ordering is less reliable)
- Whenever possible, trying to avoid selection of Q-clones and false clone overlaps in MTP ➔ **more reliable MTP**
- Supplementing the list of MTP clones by **candidate** “bridges” for end-to-end merging ➔ **longer contigs**

Physical mapping, anchoring, sequencing



Fingerprinted BAC clones

Clustering

Ordering

Merging

Anchoring, Verification, MTP selection

BAC Sequencing Stage

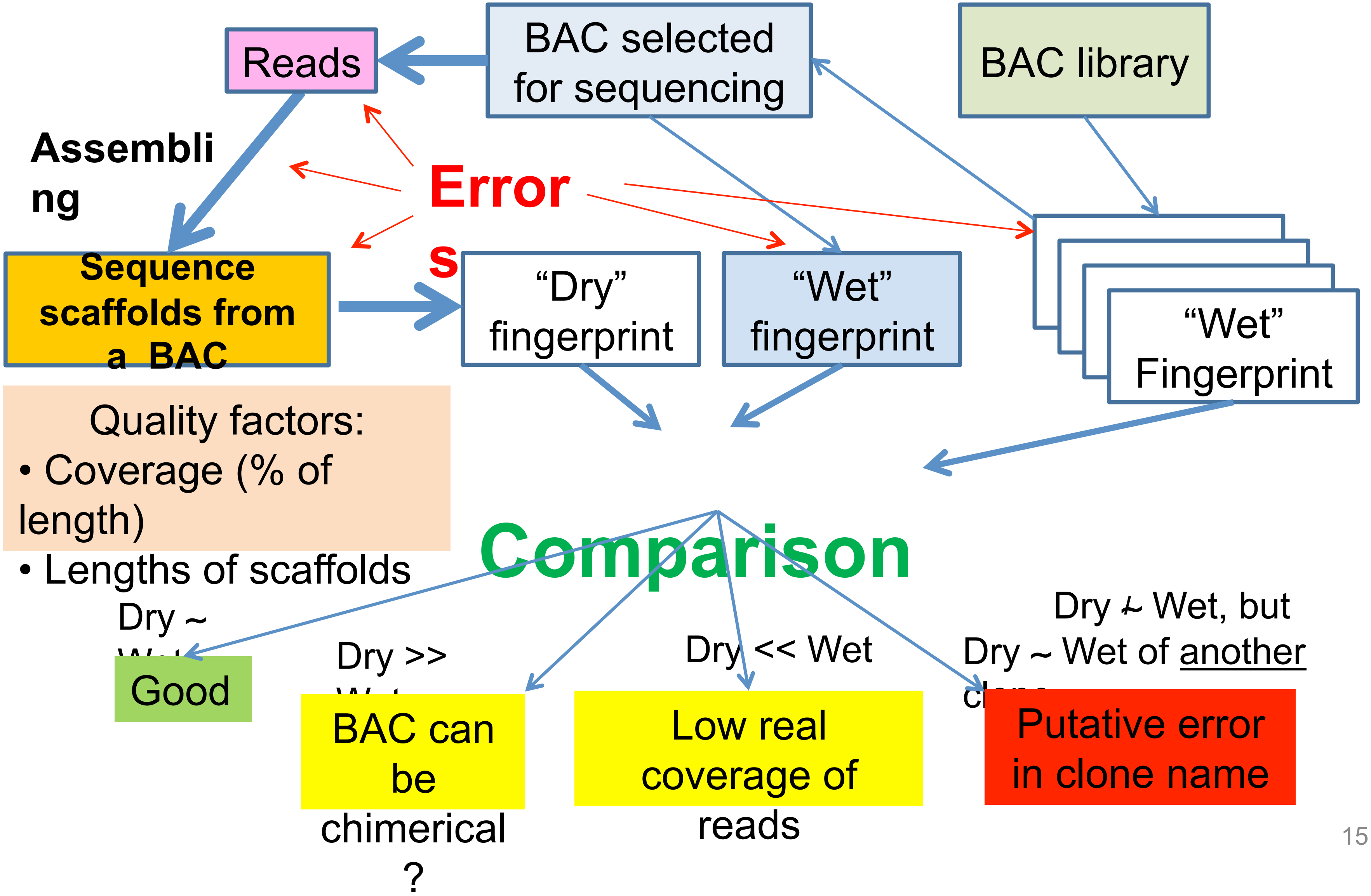
LTC assistance at sequencing stage

- Constructing long reliable physical contigs
- MTP selection
- Gap repairing

……

- **Gap repairing**
- **Controlling sequence quality**
- **Ordering of sequence contigs**
- **Facilitating of the anchoring**

Testing sequence quality



Reads

BAC selected for sequencing

BAC library

Assembling

Errors

Sequence scaffolds from a BAC

“Dry” fingerprint

“Wet” fingerprint

“Wet” Fingerprint

Quality factors:

- Coverage (% of length)
- Lengths of scaffolds

Comparison

Dry ~ Wet

Good

Dry >> Wet

BAC can be chimerical ?

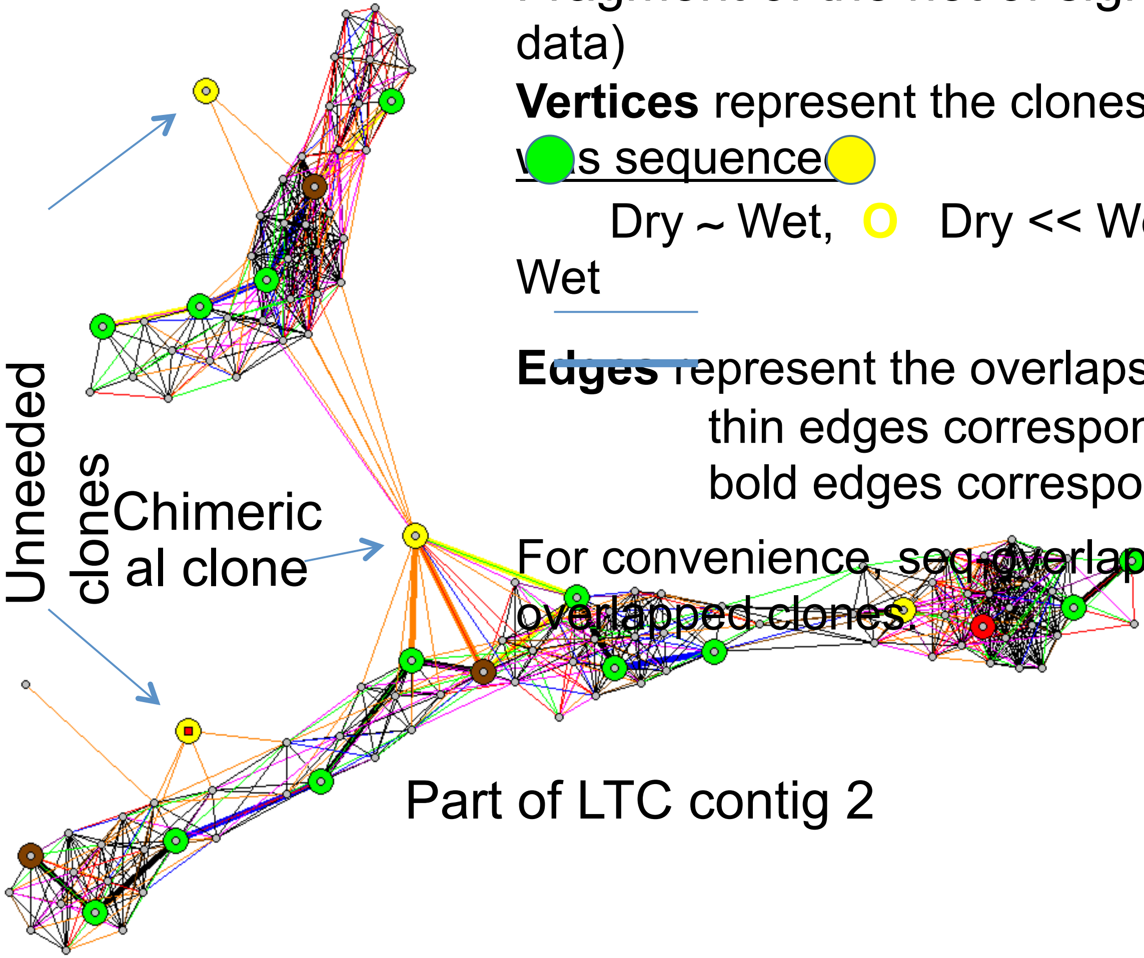
Dry << Wet

Low real coverage of reads

Dry ≁ Wet, but Dry ~ Wet of another clone

Putative error in clone name

LTC control of MTP clone-overlaps at sequence level



Fragment of the net of significant clone overlaps (7BS data)

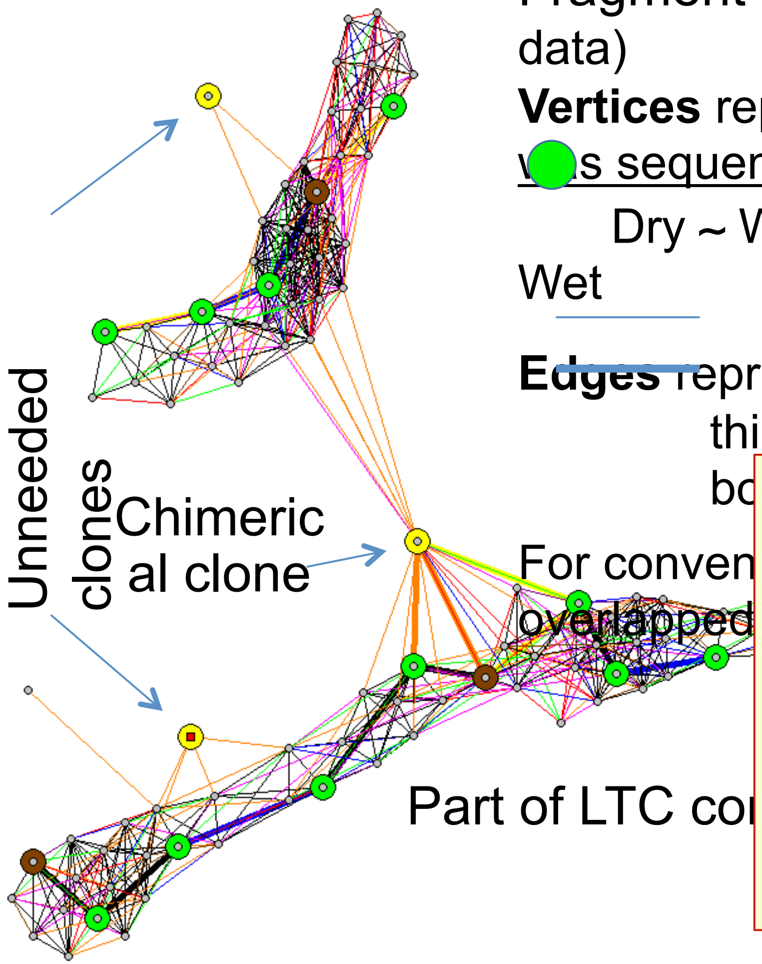
Vertices represent the clones: Disk indicates that the clone was sequenced 🟢 🟡 ● 🔴 /

Dry ~ Wet, O Dry << Wet, Dry >> Wet, Dry ~ Wet

Edges represent the overlaps, color reflects significance:
thin edges correspond to HICF-based overlaps
bold edges correspond to seq-based overlaps

For convenience, seq-overlaps are shown only for HICF-overlapped clones.

LTC control of MTP clone-overlaps at sequence level



Fragment of the net of significant clone overlaps (7BS data)

Vertices represent the clones: Disk indicates that the clone was sequenced ● ● ● ● /
Dry ~ Wet, O Dry << Wet, Dry >> Wet, Dry ~ Wet

Edges represent the overlaps, color reflects significance:
thin edges correspond to HICF-based overlaps
bo s

For conven CF-
overlapped

For chromosomal arms, where sequencing is already done, it is high time to employ this simple tactic for verification/testing of

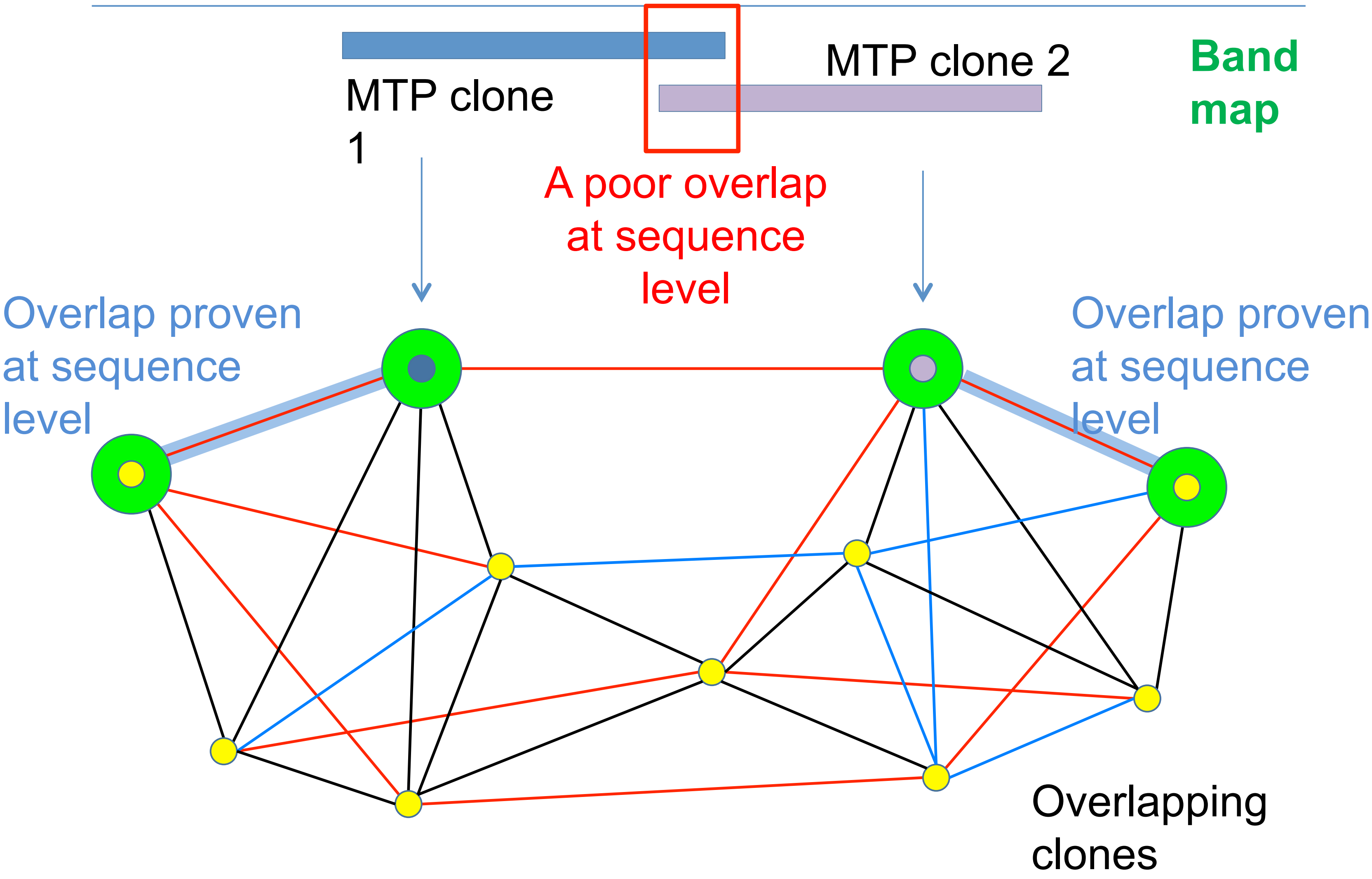
Curing of the detected sequence gaps

- **Check the physical contig:** a gap can be a result of error(s) in physical contig assembly
- **Check overlaps** in the fingerprints
- **Check sequence quality**: coverage, length and correspondence between wet and dry fingerprints
- **Add clones** to connect the sides of the gap via significant fingerprint-based overlaps
- If well sequenced clones appeared to overlap on fingerprint but not sequence level, try to **increase cutoff at the fingerprint level**

Curing of the detected sequence gaps

- **Check the physical contig:** a gap can be a result of error(s) in physical contig assembly
- **Check overlaps** in the fingerprints
- **Check sequence quality**: coverage, length and correspondence between wet and dry fingerprints
- **Add clones** to connect the sides of the gap via significant fingerprint-based overlaps
- If well sequenced clones appeared to overlap on fingerprint but not sequence level, try to **increase cutoff at the fingerprint level**

Example of gap repairing



MTP clone 1

MTP clone 2

Band map

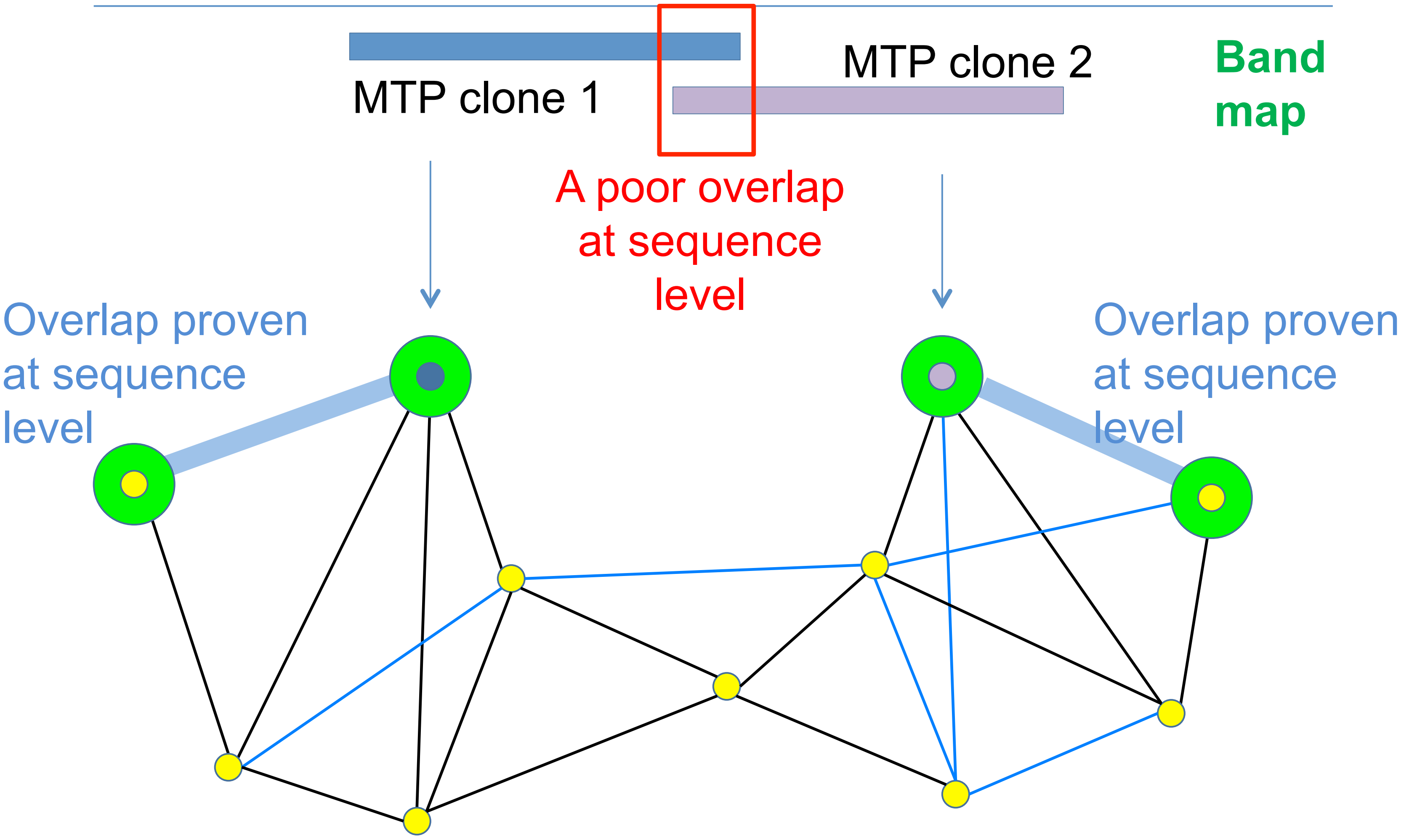
A poor overlap at sequence level

Overlap proven at sequence level

Overlap proven at sequence level

Overlapping clones

Example of gap repairing



MTP clone 1

MTP clone 2

Band map

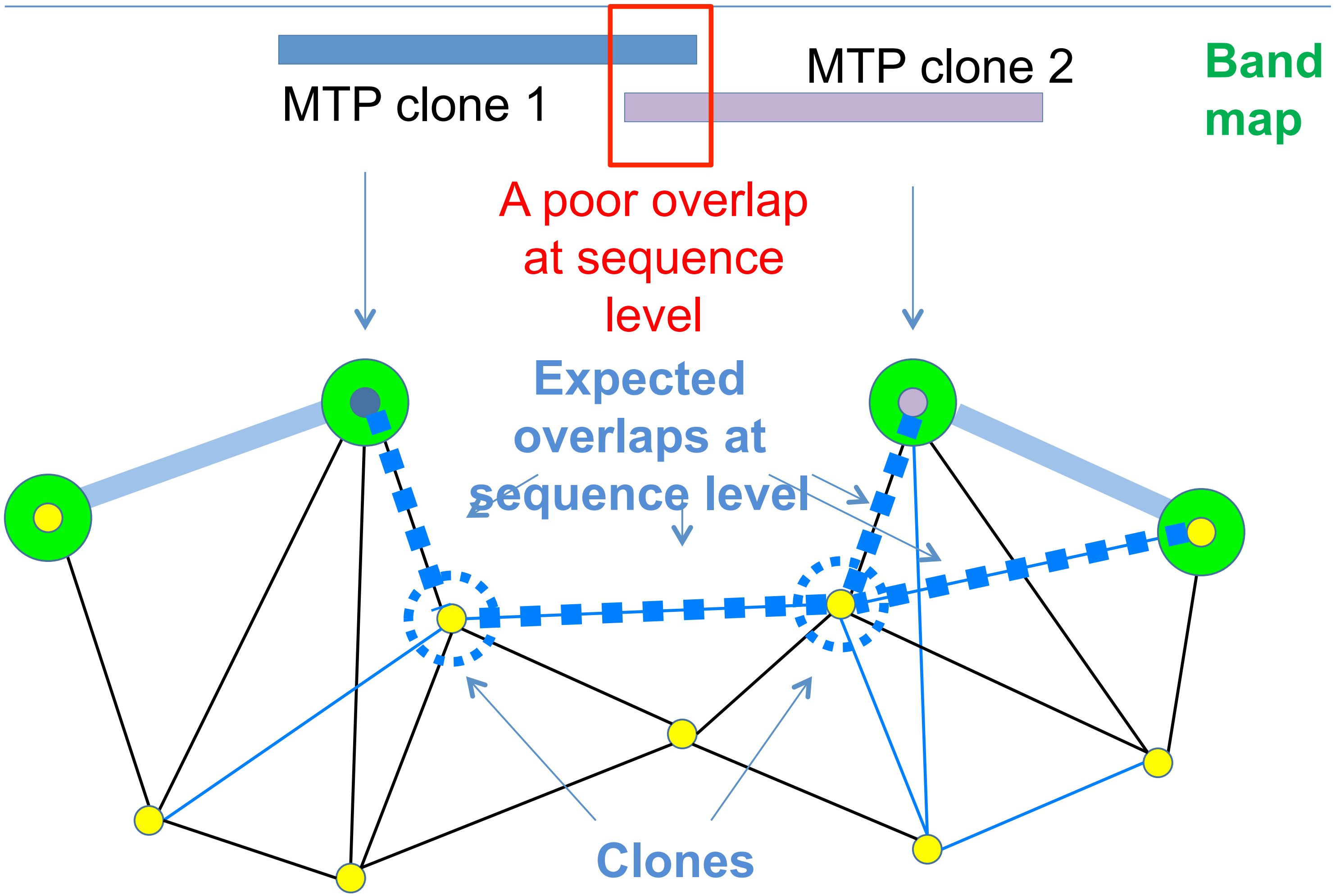
A poor overlap at sequence level

Overlap proven at sequence level

Overlap proven at sequence level

Clone overlaps detected at increased cutoff stringency

Example of gap repairing



MTP clone 1

MTP clone 2

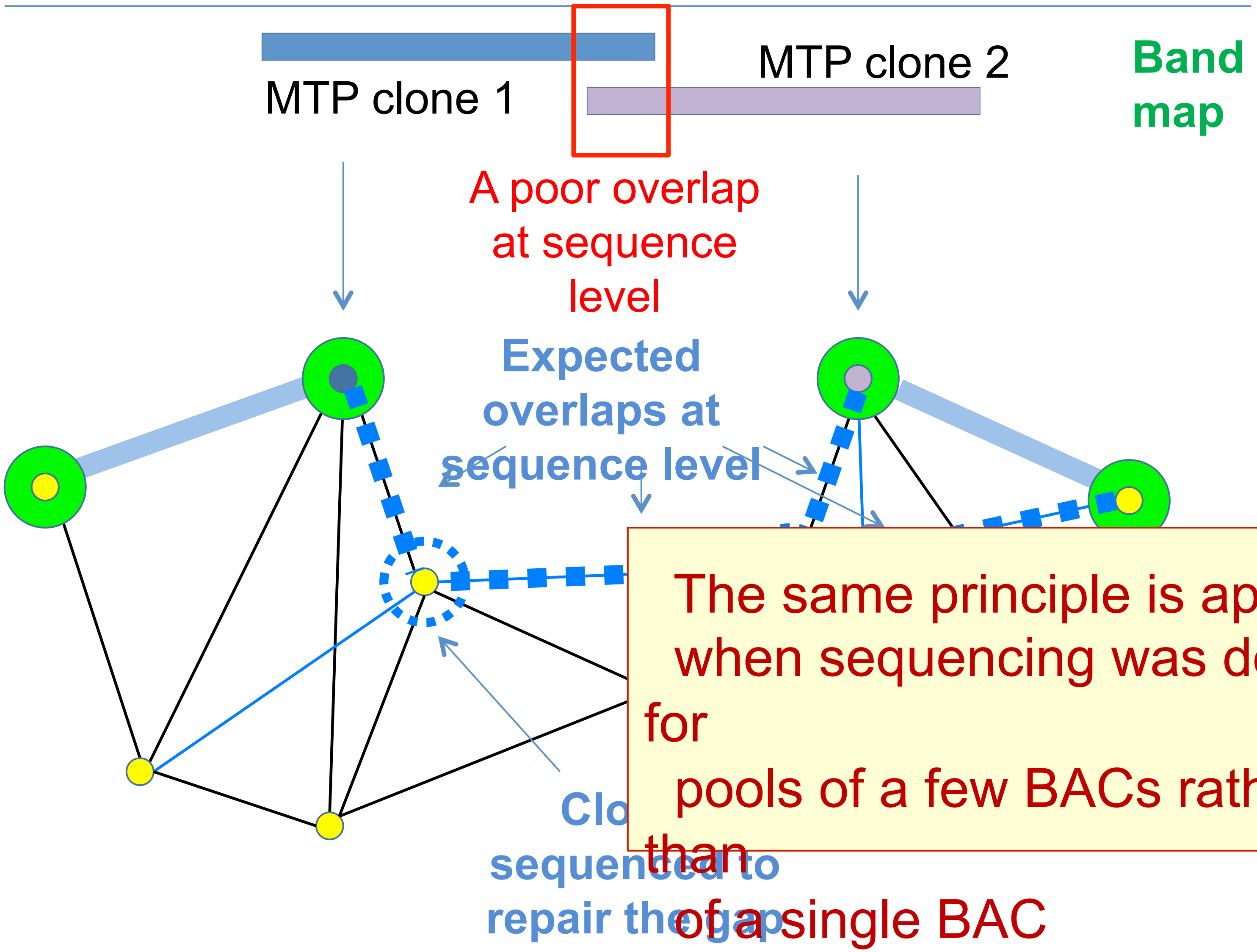
Band map

A poor overlap at sequence level

Expected overlaps at sequence level

Clones sequenced to repair the gap

Example of gap repairing



MTP clone 1

MTP clone 2

Band map

A poor overlap at sequence level

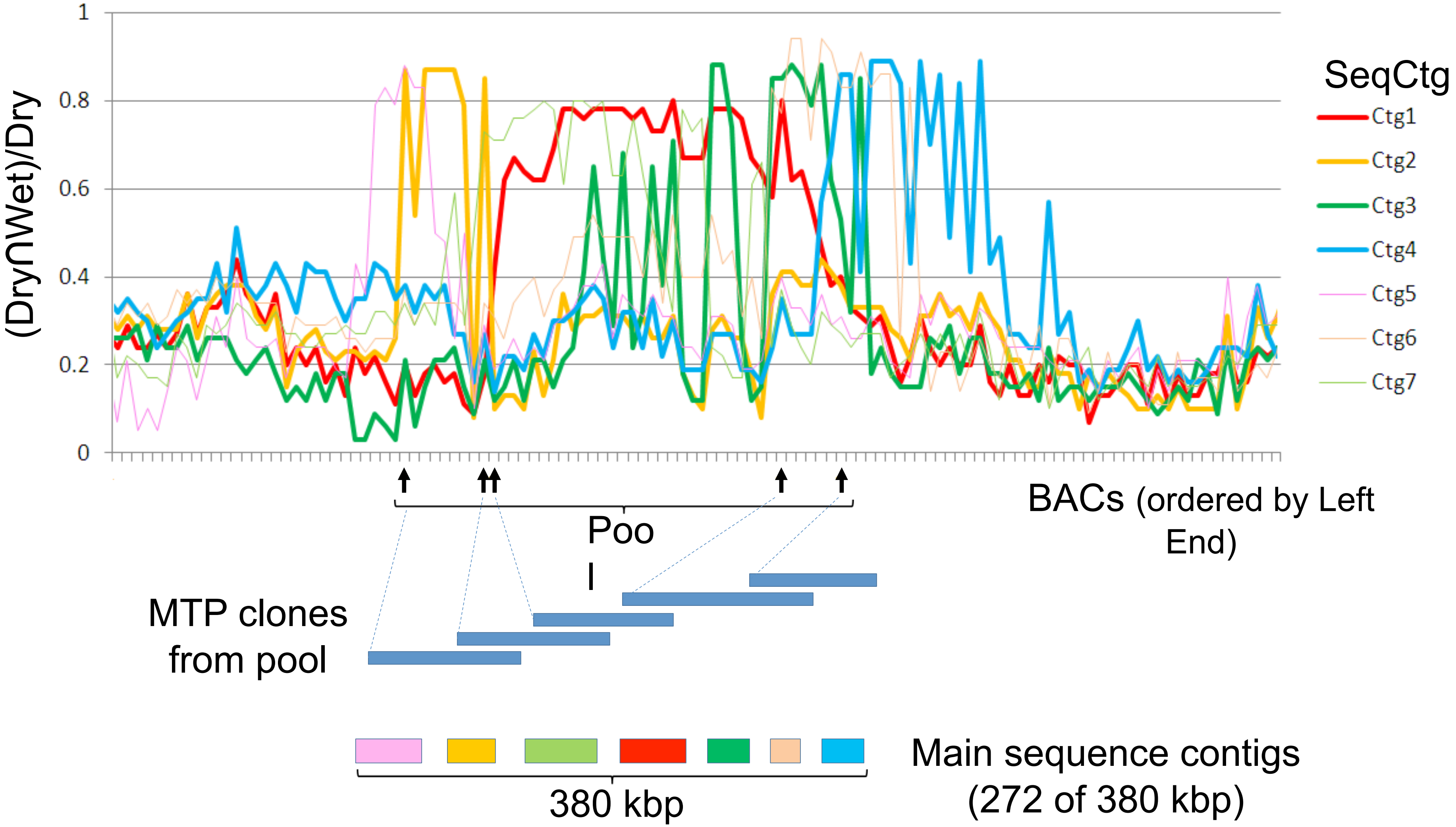
Expected overlaps at sequence level

Clo

sequenced to repair the gap

The same principle is applied when sequencing was done for pools of a few BACs rather than of a single BAC

Ordering and orientation of sequence contigs within pool



BACs (ordered by Left End)

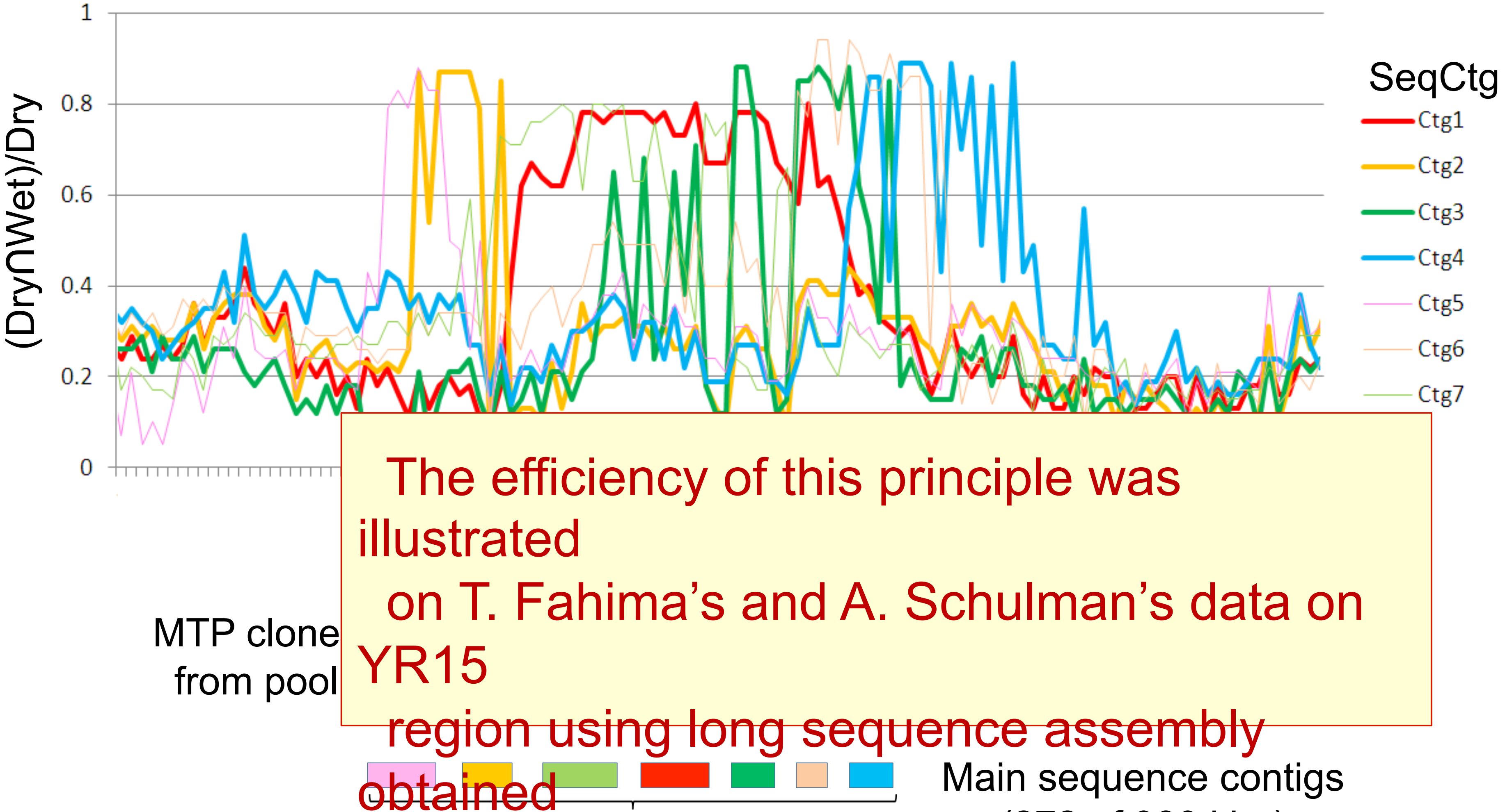
Pool

MTP clones from pool

380 kbp

Main sequence contigs (272 of 380 kbp)

Ordering and orientation of sequence contigs within pool



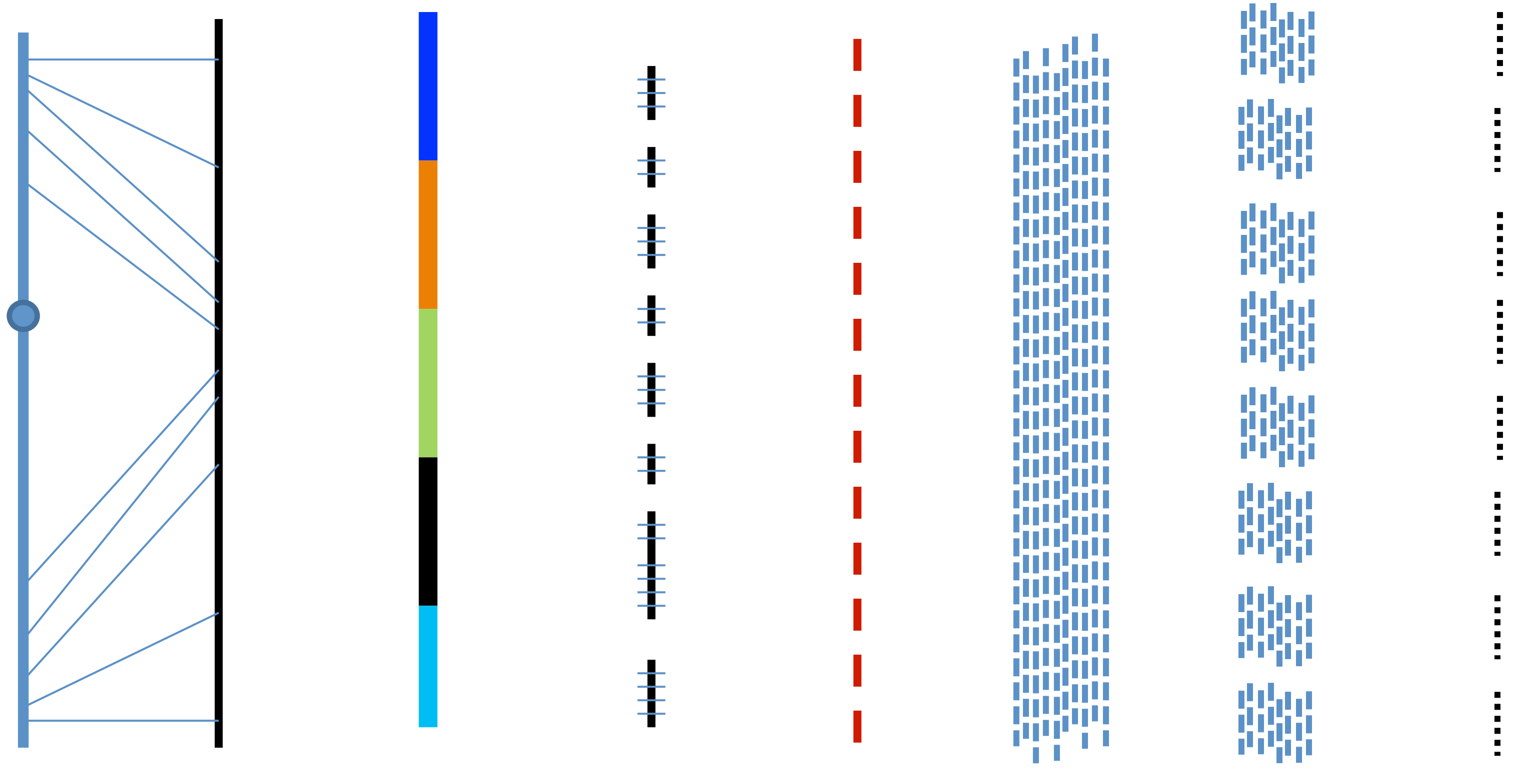
The efficiency of this principle was illustrated on T. Fahima’s and A. Schulman’s data on YR15 region using long sequence assembly obtained by PacBio sequencing

MTP clone
from pool

380 kbp

Main sequence contigs
(272 of 380 kbp)

Physical mapping, anchoring, sequencing



Chromosome

Genetic map

Deletion map

RH map

LD blocks

BACs

Physical map

Sequence contigs of MTP

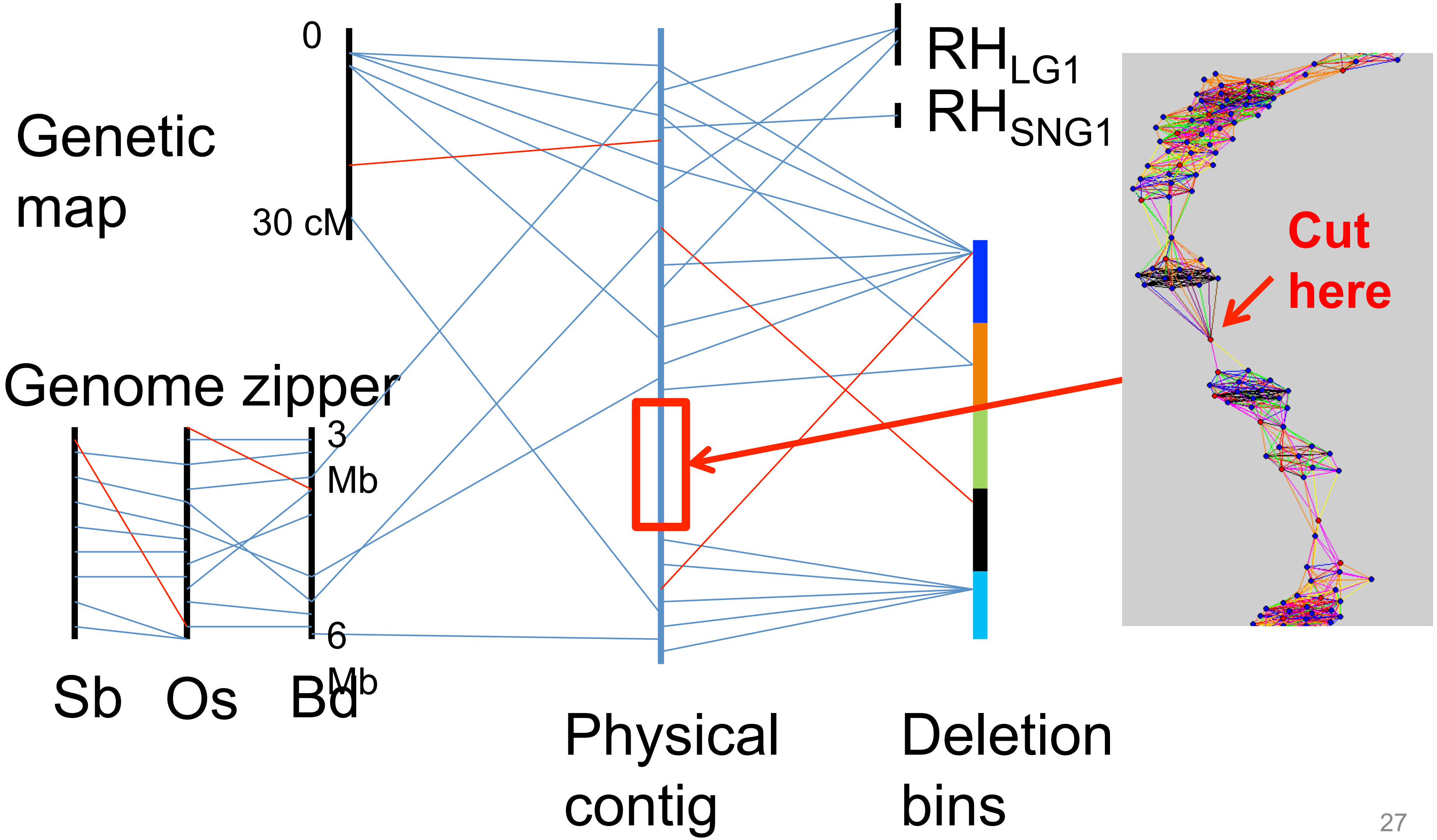
Low resolution

Local partial ordering

Local ordering

- **Unknown local orders**
- **Parasite sequences**
- **Misassembles**

Current challenge: Enabling for cross-talks



Further extension of LTC approach

Physical mapping

Parallel clone overlaps

→ detection of problematic clones and clone overlaps

Linear structure of the net of significant clone overlaps

→ Avoiding contradictions of contig topology with chromosome linear structure

➔ Long reliable contigs

➔ Simpler anchoring

Genetic/RH mapping

Parallel marker linkages

→ detection of problematic markers and pseudo-linkages

Linear structure of the net of tight linkages

→ Avoiding contradictions of linkage group topology with chromosome linear structure

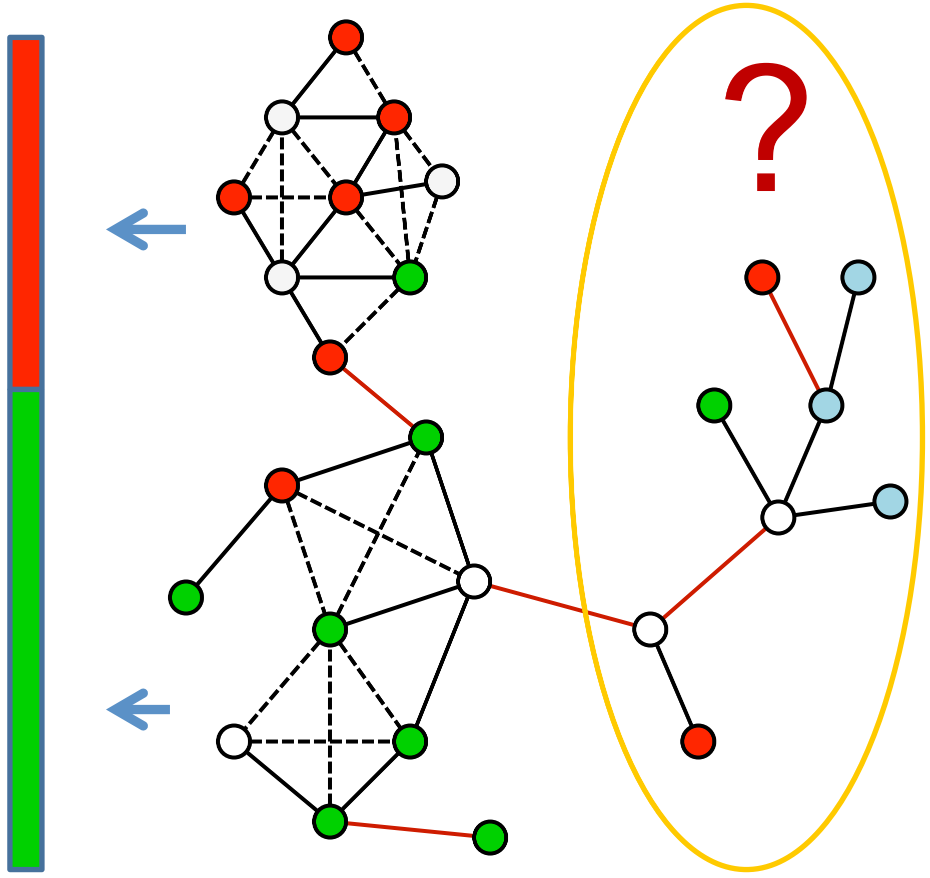
➔ Large reliable linkage groups

➔ Simpler ordering

False significant linkages in **RH**-mapping

Problems:

- Metric
- Filtration of markers
- Non-uniform coverage
- Many false linkages
- Ordering within “LG”
- Ordering of “LGs”



Minimal Spanning Tree (solid)
complemented by tight linkages
(dotted)

LTC for: anchoring ← → map editing

Provide marker anchoring results in the table:

marker → map_1 map_2 map_3 … map_M

Physical map: clones in contig, coordinates in contig

Deletion map: arm, bin name

Genetic / RH map: linkage group, coordinate in LG

Genome Zipper: best position on *Brachypodium* only, rice only, sorghum only, barley only

Integral map: orders rather than numerical positions

→ Sort all markers within contigs by coordinate

LTC for: anchoring ← → map editing

Provide marker anchoring results in the table:

marker → map$_1$ map$_2$ map$_3$ … map$_M$

Physical map: clones in contig, coordinates in contig

Deletion map: arm, bin n

Genetic/RH map: linkage in LG

Genome Zipper: best pos only, rice only, sorghum on

Integral map: orders rath ns

→ Sort all markers within coordinate

Using the 90K chip simplifies anchoring to Genome Zipper and to high quality genetic maps; these markers also can be mapped to deletion bins and used in RH-mapping (already done for 7B and

Conclusions

Although the results of physical mapping are central for BAC-sequencing, they can also be useful at sequence assembly and anchoring stages. In particular, LTC enables

- Testing the assembled sequence quality
- Controlling of MTP clone-overlaps at sequence level
- Curing the gaps in sequence contigs
- Ordering and orientation of sequence contigs assembled from pooled DNA of several BAC clones

Some of the approaches developed in LTC for physical mapping proved helpful in other aspects of genome mapping: building dense linkage maps and RH-mapping

Prospects: the assembly and anchoring of sequence contigs can be further improved by enabling coordinated/iterative analysis (cross-talk) of sequence contigs with all sources of positional information: physical contigs, deletion bins, genetic maps, RH-maps, LD data, and genome zippers

Acknowledgements

INRA

- Etienne Paux

BAYER

- Catherine Feuillet

MultiQTL

- Zeev Frenkel
- Vladimir Glikson
- Yefim Ronin
- Dina Minkov

MURDOCH
UNIVERSITY
PERTH, WESTERN AUSTRALIA

- Rudi Appels

IWGSC

- Kellye Eversole
- Jane Rogers

Triticeae Genome

BARD

MATIMOP מתימ"פ
Israeli Industry Center For R&D

Israeli Industry Center for R &D

Institute of Experimental Botany

- Miroslav Valarik

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP
MDCCCLIX

Norwegian University of Life Sciences

- Simen Sandve
- Bujie Zhan
- Tatiana Belova
- Odd-Arne Olsen

אוניברסיטת חיפה
University of Haifa

- Zeev Frenkel
- David Mester
- Dina Raats
- Tami Krugman
- Tzion

אוניברסיטת תל-אביב
TEL AVIV UNIVERSITY

- Assaf Destelfeld
- Raz Avni