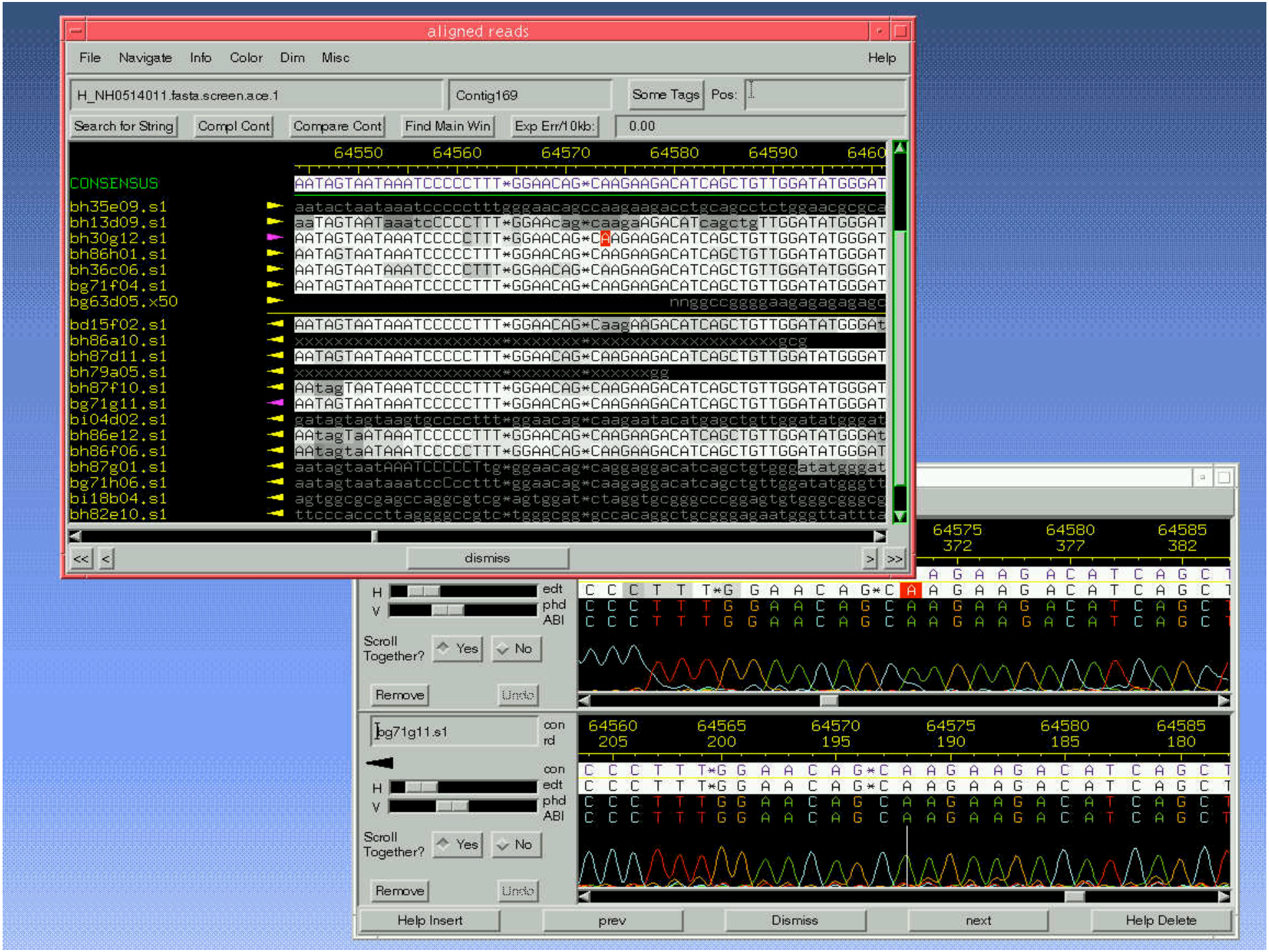
Finishing Genomic Sequence

Melissa de la Bastide

Cold Spring Harbor Laboratory

What is Finishing?

- Interactive sequence inspection
- Directed sequencing
- Assembly verification


aligned reads
File Navigate Info Color Dim Misc Help
H_NH0514011.fasta.screen.ace.1
Contig169
Some Tags
Pos:
Search for String
Compl Cont
Compare Cont
Find Main Win
Exp Err/10kb:
0.00
64550 64560 64570 64580 64590 6460
CONSENSUS
bh35e09.s1
bh13d09.s1
bh30g12.s1
bh86h01.s1
bh36c06.s1
bg71f04.s1
bg63d05.x50
bd15f02.s1
bh86a10.s1
bh87d11.s1
bh79a05.s1
bh87f10.s1
bg71g11.s1
bi04d02.s1
bh86e12.s1
bh86f06.s1
bh87g01.s1
bg71h06.s1
bi18b04.s1
bh82e10.s1
<< < dismiss > >>
64575 64580 64585
372 377 382
H
V
edt
phd
ABI
Scroll Together?
Yes
No
Remove
Undo
bg71g11.s1
con
rd
64560 64565 64570 64575 64580 64585
205 200 195 190 185 180
Help Insert
prev
Dismiss
next
Help Delete

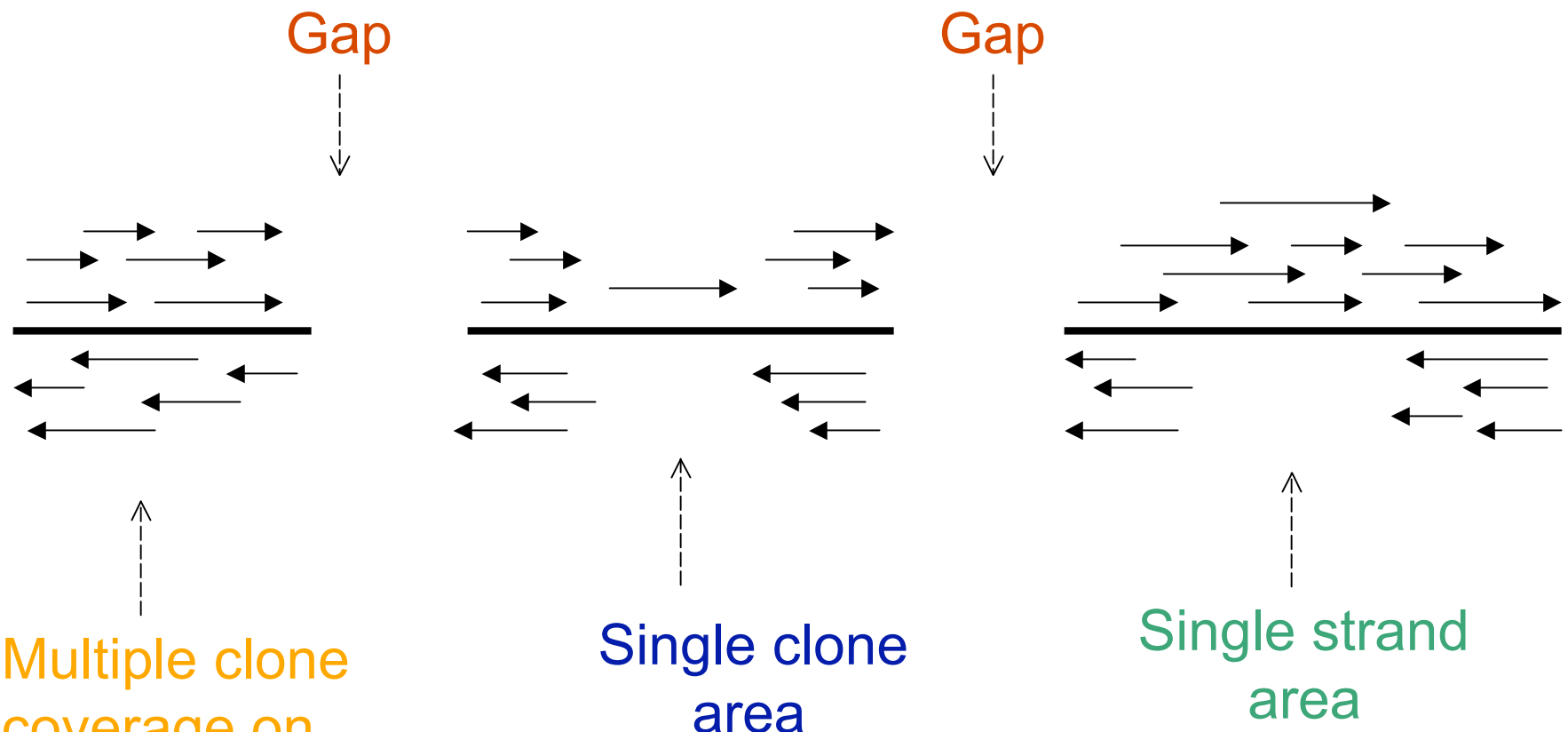
Why finish sequence?

- gene/pseudo-gene difference can be a little as a single base
- assembly algorithms alone are not sufficient to correctly assemble sequence

Goals of Finishing

- Resolve sequence ambiguities and discrepancies, such that the error rate is less than one in 10,000 bases.
- Provide “double-stranded” coverage for every base:
  - minimum of two different clones
  - two strands
  - two different chemistries
  - or quality >phred30
- Achieve contiguity.
- Delineate vector/insert junctions.

Types of problems



Background

- Shotgun to 8X coverage
- Use double-stranded subcloning vectors
- Use dye terminator chemistries

Comparison of Sequencing Chemistries

Dye Primers

- Advantages - Even peak heights, no base dropouts.
- Disadvantages - Compressions, must use universal primers.

Dye Terminators

- Advantages - Resolve compressions, work on poor templates, can use custom primers.
- Disadvantage – Base dropouts (especially G).

Basic Finishing Strategy

Step One: Pre-Finishing

- CONSED Autofinish
- SWIFT software

CONSED Autofinish

David Gordon, University of Washington

- aims for a pre-set target number of errors per Megabase
- pre-set cost parameters guide number and type of reactions called
- automatic custom oligo selection

TKFinish

Gabor Marth, Washington University, St. Louis

- automatically selects “long” dye primer, dye terminator and reverse sequencing reactions
- cannot select custom oligos
- no way to modify reaction selection

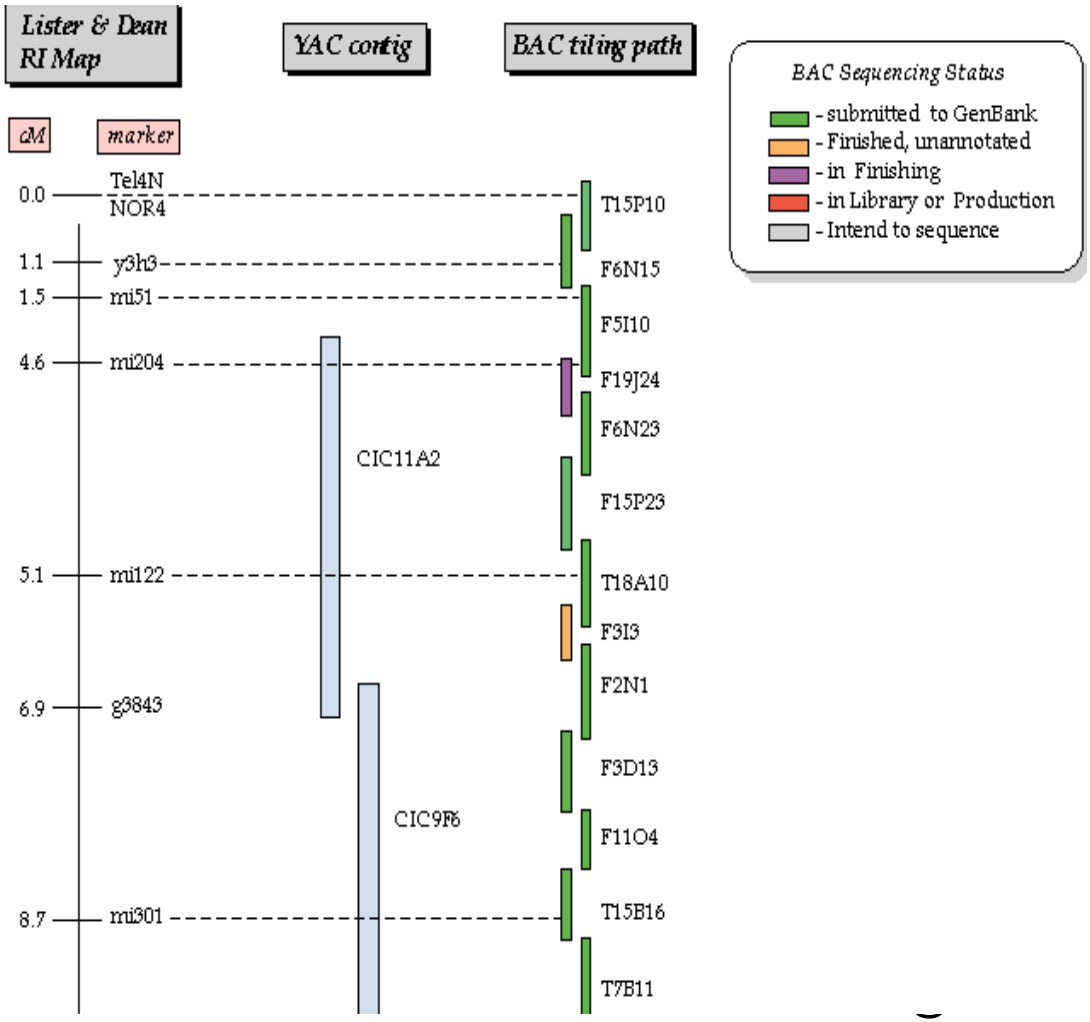
np_edit and nd_edit

Sanger Center, UK

- automatically applies edits to the assembly using information from the shotgun primary data
- tags edits for human assessment
- does not suggest finishing reactions

Basic Finishing Strategy

Step Two: Determine Map Location



- Delineate overlaps with adjacent clones to minimize finishing redundancy
- Note genetic markers

Basic Finishing Strategy

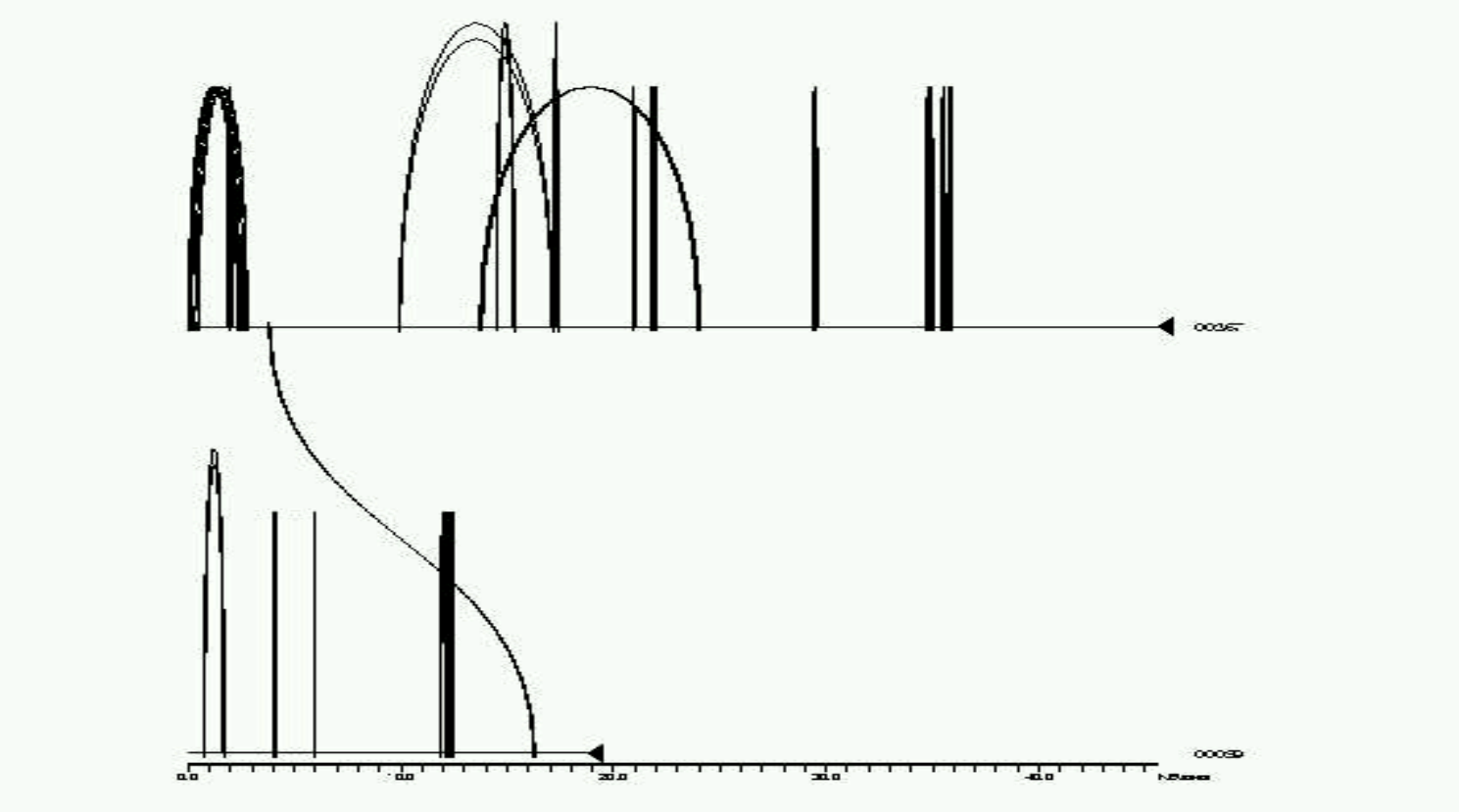
Step Three: Examine Contig Ends

- BAC/vector cloning junctions must contain a restriction site
- Noting the vector junctions is a first step in ordering contigs

Basic Finishing Strategy

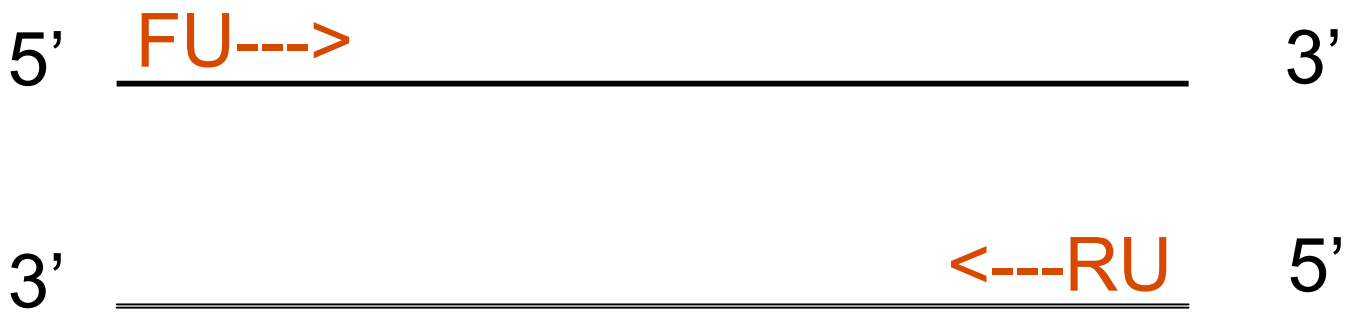
Step Four: Assembly Overview

- Printrepeats (J. Parsons)



Basic Finishing Strategy
Step Five: Assembly Inspection

- check forward-reverse sequence pairs to order and orient contigs



Basic Finishing Strategy
Step Six: First Pass Edit

- custom oligos or transposons are chosen to extend sequence into gaps
- initial attempt to correct sequence discrepancies (editing)
- directed sequencing reactions (alternate chemistries, etc) are chosen to resolve problem sequence

Aggressive Finishing Approaches

- multiple (2-5) sequencing reactions are chosen to resolve problems
- multiple custom oligos, templates for transposons or sequencing chemistries are attempted for areas requiring coverage
- new sequence data is intermittently assembled to reduce cycle time if reactions fail

Basic Finishing Strategy
Step Seven: Additional Editing

- data from previous passes is used to correct sequence ambiguities
- additional custom oligos are chosen for subclone “walking” and PCR
- alternate techniques are employed to attack problems that were not resolved by initial methods


Contig Editor: -3322 PCR1.x5.seq
C: 75
Q: -1
Insert
Edit Modes
Cutoffs
Undo
Next Problem
Commands
Settings
Quit
Help
-3407 za76b07.x3.seq
-3300 p2arr.x9.seq
3497 za77d11.x3.seq
3462 za77a04.x3.seq
3476 za77e10.x3.seq
3475 za77e09.x3.seq
3449 za76f02.x3.seq
3471 za77c11.x3.seq
3453 za76f11.x3.seq
3424 za77d07.x3.seq
3460 za76h09.x3.seq
CONSENSUS
3495 za76f12.x3.seq
3467 za77b07.x3.seq
3494 za77f10.x3.seq
3299 p2ar2.x9.seq
Info
Diff
Quit

Basic Finishing Strategy
Step Eight: Final Edit

- clone is inspected for contiguity, sequence quality, coverage, and proper resolution of repetitive sequence
- the assembly's predicted "in silico" restriction fragment sizes must agree with clone fingerprint data

Finishers' Clone Submission Checklist

Finisher:
Clone Name:
Organism:
Chromosome:

Date assigned:
Date completed:
Estimated clone size:
Actual Clone Size:

Vector type:
Restriction site:

Neighbors/overlaps (note extent of overlap and whether data was stolen):

Is the entire clone finished (if not, delineate the boundaries):

Is overall error rate less than 1 in 10,000 bases?

Are there any extra contigs greater than 2kb?

Has the clone been screened for E coli contamination/transposons?

Have you edited single stranded/single chemistry regions at phred30 quality?

Have you eliminated single clone areas?

Are there any regions covered only by PCR?

Do any areas need annotation (unresolved repeats, low quality regions, etc)? If so, list the locations and reasons:

Does the mapsort agree with the restriction digest? List enzymes used, and corresponding mapsort/digest band sizes, including vector bands! Attach mapsort and digest data.

Other remarks:

Quality Assurance of Finished Sequence

- clone reassembly (PHRAP)
- overlap discrepancy analysis
- cumulative PHRAP assembly scores

Finishing Problem Sequence

I. Physical Gaps

- little or no representation of an area of sequence
- caused by cloning bias, too little shotgun, repeats, etc.

Physical Gap Resolution

- PCR
  - try different enzymes/buffers
  - add DMSO
  - vary oligo position and length
  - alter cycling conditions
- different template types
- subclone restriction fragments
- oligo screen for subclones that may bridge the gap

Finishing Problem Sequence

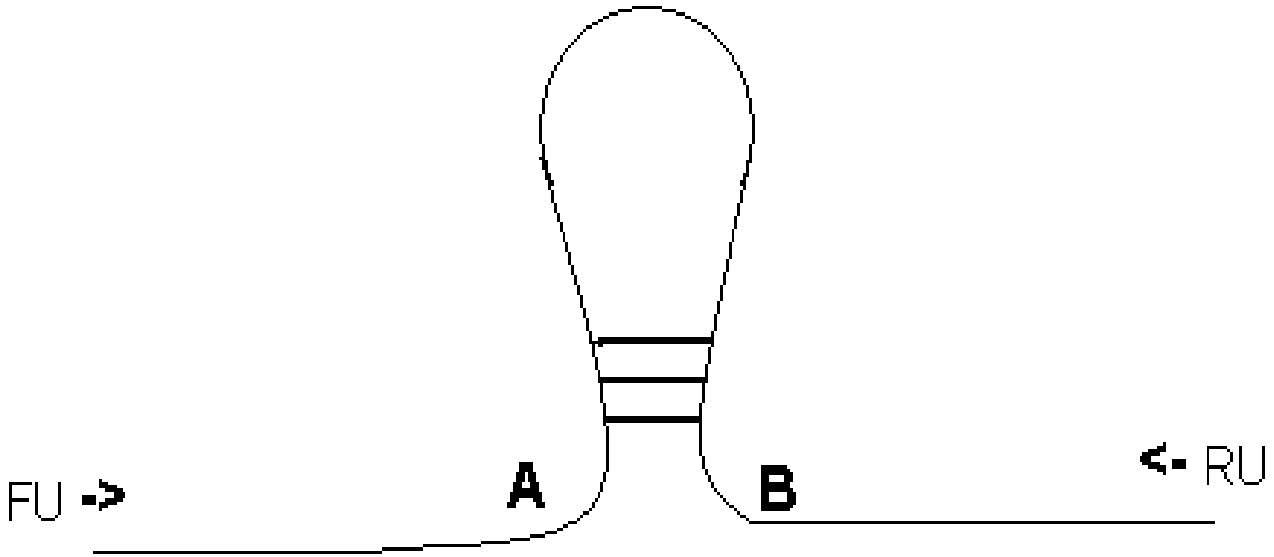
II. Sequence Gaps

- templates that span the gap are available, but sequencing reactions are inhibited
- caused by compressions, mono/polynucleotide runs, secondary structure, etc.

Sequence Gap Resolution

- compressions
  - sequence both strands
  - terminator sequencing chemistry
- mono/polynucleotide runs
  - dye primer chemistry
  - use subclones if possible (PCR may slip)
  - in-vitro transposons/shatter libraries
  - thermofidelase at high temperature
- secondary structure
  - dGTP chemistry
  - in-vitro transposons/shatter libraries (SILs)
  - TA subcloning

Compression



Compressions are the result of the extension product folding back upon itself, thereby altering migration of the fragment through the sequencing matrix. Base spacing is noticeably altered.


Contig Editor: -5891 ay30c01r001.x9.s
C: 75
Q: -1
Insert
Edit Modes
Cutoffs
Undo
Next Problem
Commands
Settings
Quit
Help
826 ay54e09.s1
-1618 ay53h07.s1
-4233 ay41a01.s1
-1742 ay18f01.s1
-4232 ay32g01.s1
-4238 ay20c11.s1
-4236 ay78h11.x1
1261 ay68h09.x1
-4229 ay45e02.s1
-4235 ay64e04.x1
CONSENSUS
Lock
Quit
Help
4231 ay31c04.s1
4228 ay17b07.s1
585 ay66h03.x1
-2045 ay27b10.s1
Info
Diff
Quit


Contig Editor: -2069 bd05e08.s1
C: 75
Q: -1
Insert
Edit Modes
Cutoffs
Undo
Next Problem
Commands
Settings
Quit
Help
276 bd18c04.s1
3557 bd40f01.x1
1720 bd34h03.x1
1723 bd34h03.x3
1988 bd47f01.x1
-3757 bd45c04.s5.seq
CONSENSUS
traces
Lock
Quit
Help
3754 bd40f01.s5.seq
2005 bd08d03.s1
457 bd29a12.x3
Info
Diff
Quit
X
Y


Contig Editor: -2069 bd05e08.s1
< C: 75 > < Q: -1 > Insert Edit Modes Cutoffs Undo Next Problem Commands Settings Quit Help
<< < > >>
-1523 bc83h06.s1
-672 bc83h07.s1
-816 bd10d12.s1
731 bd31f10.x1
-813 bc83h07.x4
CONSENSUS -**-
traces
Lock Quit Help
739 bd31f10.x4
3744 bd39e0116.x5.seq
-1125 bd20f01.s1
-3934 bc83h07.x8.seq
Info
Diff
Quit
X Y

Finishing Problem Sequence

III. Repeats

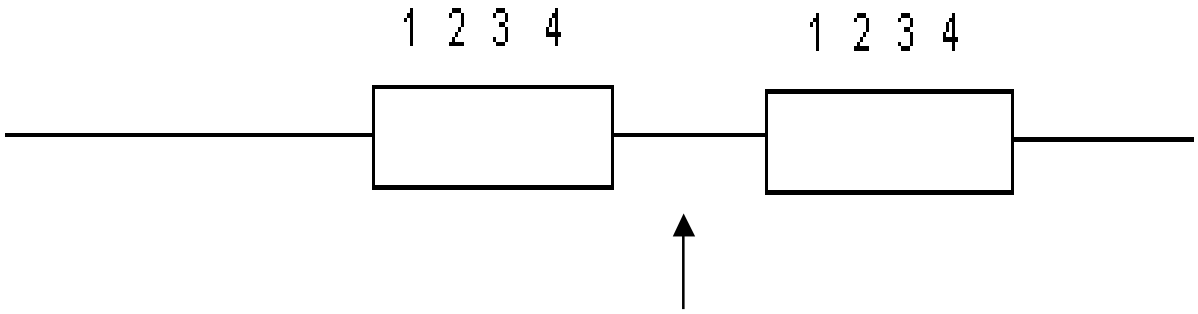
- often cause both physical and sequence gaps
- require careful inspection and interpretation of sequence data
- forward-reverse subclone sequence pairs and long read lengths are very useful in sorting repeats

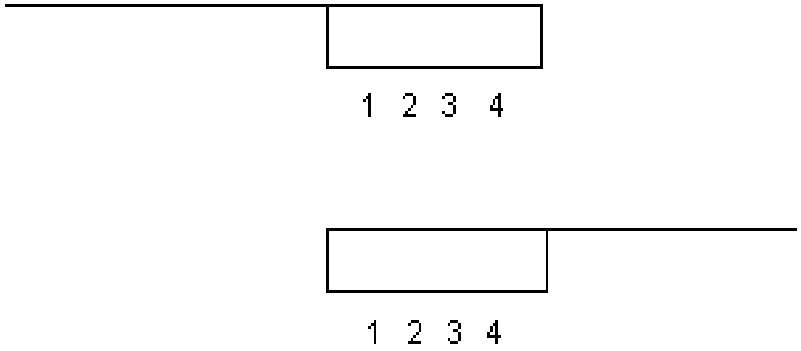
Direct Repeats

Can differ by as little as one base in several kilobases.

- try primer walking into unique sequence between repeats
- use double stranded templates for read pair information
- use restriction maps
- subclone restriction fragments, use in vitro transposons or SILs, etc.

Direct Repeats





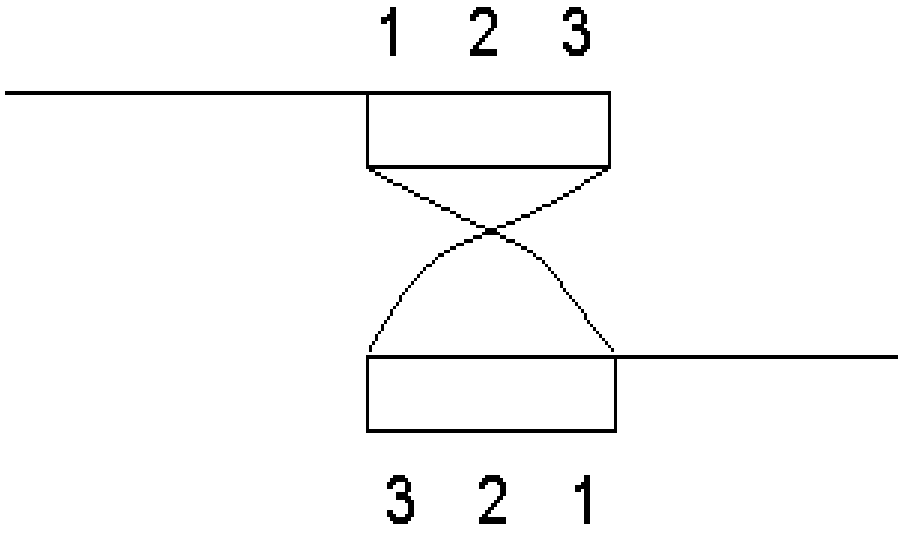
Possible Sequence Join

Inverted Repeats

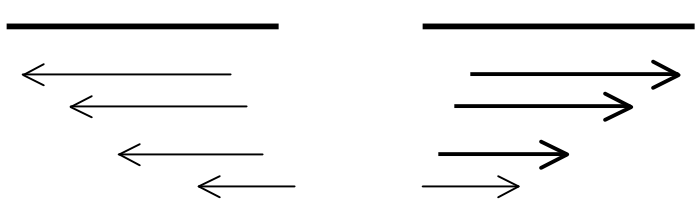
Often cause gaps as sequence proceeds uni-directionally away from the loop.

- PCR with oligos chosen in unique areas away from the gap
- double-stranded templates
- TA subclone
- subclone restriction fragments, use transposons or SILs, etc

Inverted Repeat



To identify inverted repeat causing a gap: no clones to walk on because all reads go away from the gap.

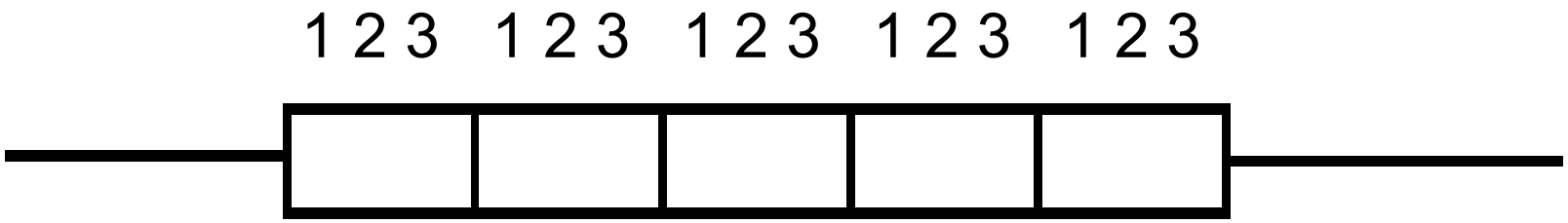


Tandem Repeats

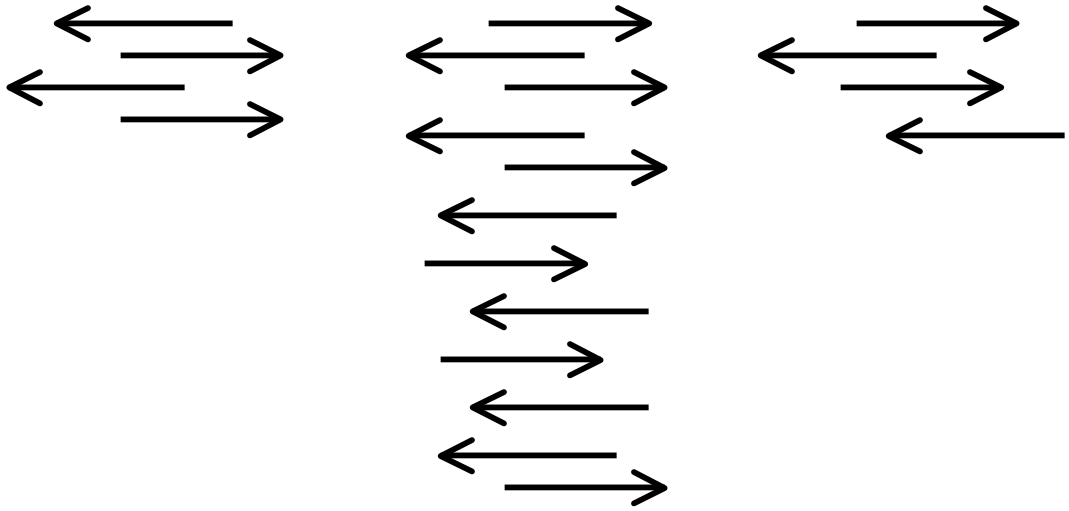
Repeat copies may be strung together with little or no unique sequence between them. It is often difficult to determine the number of copies of the repeat.

- sequence in both directions on subclones of varying insert sizes
- in-vitro transposons
- non-cycling sequenase
- may need to characterize the repeat units and analyze the size of the repeat with restriction fragment information to estimate the number of repeat copies

Tandem Repeats



Characteristic of a tandem repeat in the assembly: coverage tends to pile up in these regions.



Acknowledgements

W. Richard McCombie, Cold Spring Harbor Lab

Doug Johnson, Washington University, St. Louis