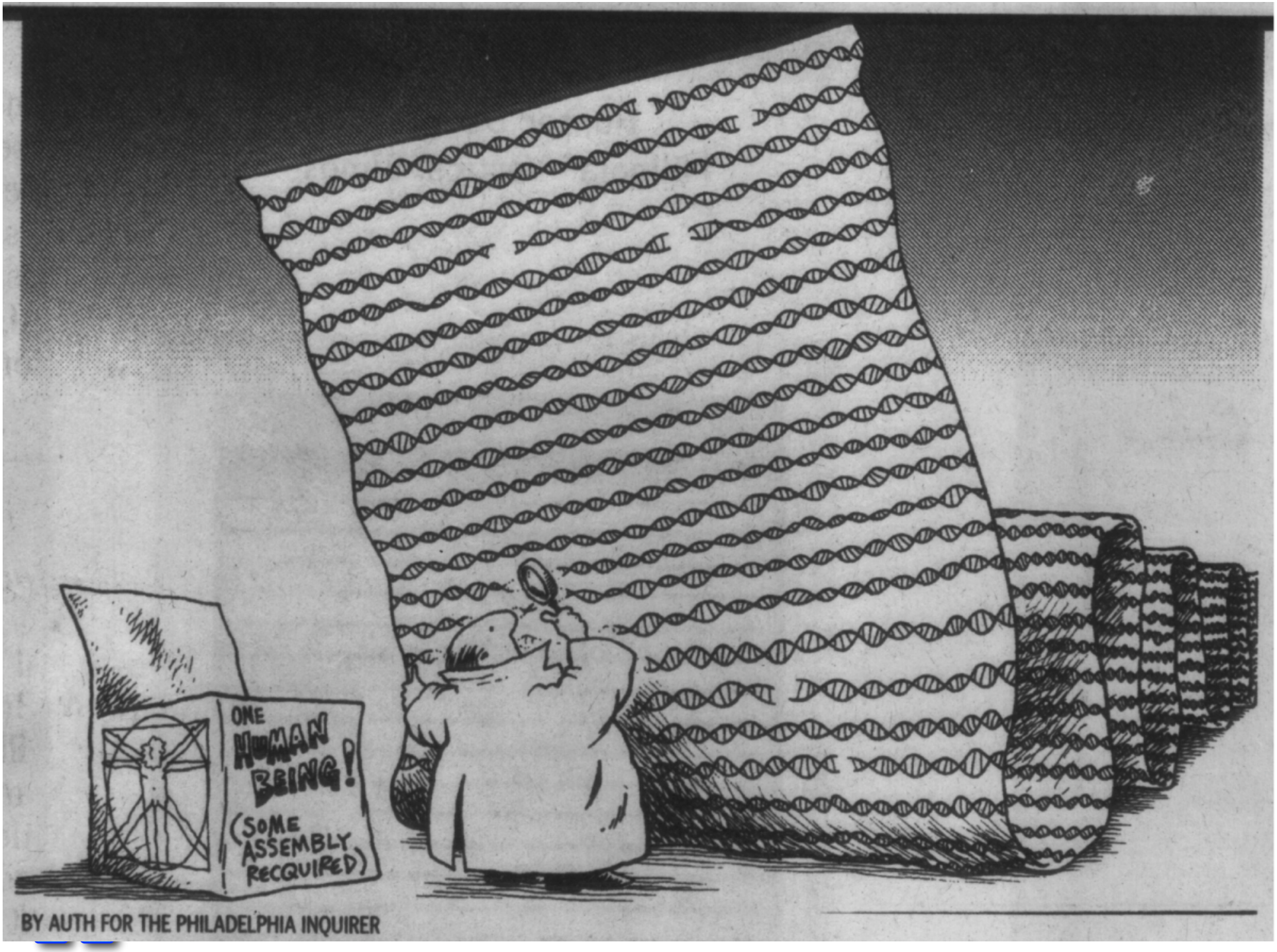
From Fragments to Finished Genome

Overview of Sequencing, Assembly, and Closure at TIGR

Luke Tallon

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

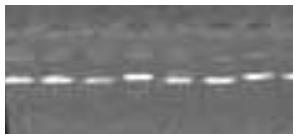

ONE
HUMAN
BEING!
(SOME
ASSEMBLY
RECQUIRED)
BY AUTH FOR THE PHILADELPHIA INQUIRER

Genome Sequencing Process

Library Construction

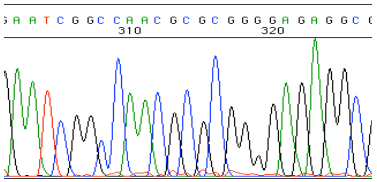
Clone Picking

Template Preparation

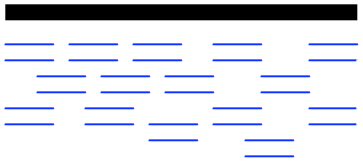


Sequencing Reactions

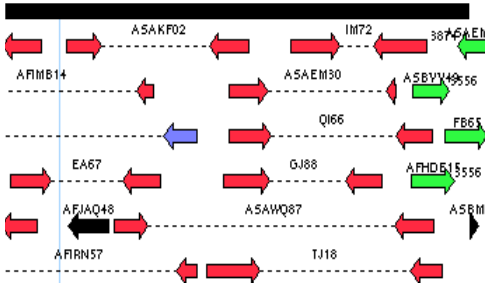
Electrophoresis and Base Calling



Genome Assembly

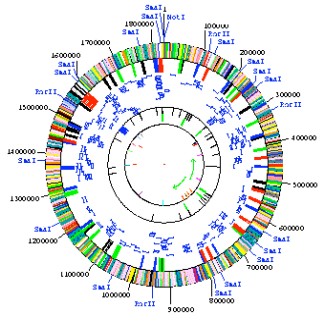


Genome Closure
Order Contigs
Close Gaps



Identify Repeats
Finish the Genome

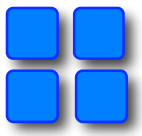
Annotation



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Library Construction

- Break the genomic DNA from the organism into many small fragments
- The fragments are cloned into plasmid vectors
- This collection of fragments is called a library



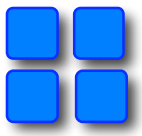
Library Requirements

1. Free vector should be at low or undetectable level.

2. No chimeric clones. Chimeras occur two or more random fragments from separate parts of the genome recombine and end up next to each other.

3. The majority of the inserts should be of relatively uniform size.

4. Libraries need to be random and cover the whole genome.

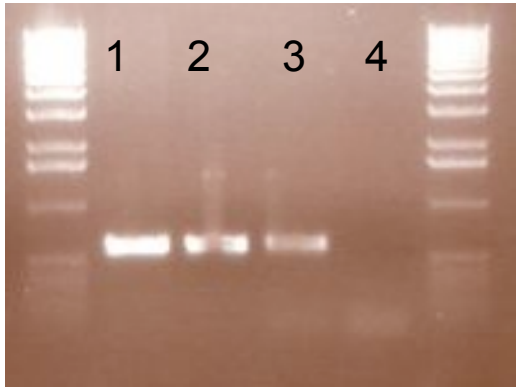


TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

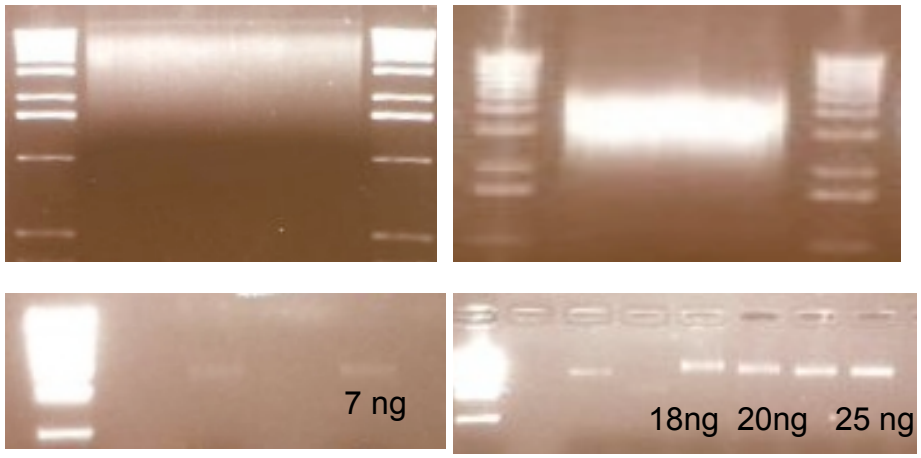
DNA Shearing with HydroShear

HydroShear Assembly Cleaning QC



PCR Products After 4 Wash Cycles
1 BAC DNA, 2 sheared BAC DNA,
3 sheared water, 4 sterile water

Nebulized DNA vs HydroShear sheared DNA
Size Selection Gels



Quant Gels Prior to Vector Ligation

BACs assembled and checked for contaminant sequences

XSCQB – 3 seqs
(2- *Lotus japonicus, 1- Homo sapiens)*
XSCQC – 0
XSCQD – 0
XSCQE – 0

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

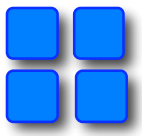
Types of vectors used in genomic library construction at TIGR

1. pHOS2 plasmid is a modified pBR192. It is low copy number: 20 copies/ cell, used for small insert library (2-4 Kb) and medium insert library (10-12 Kb).

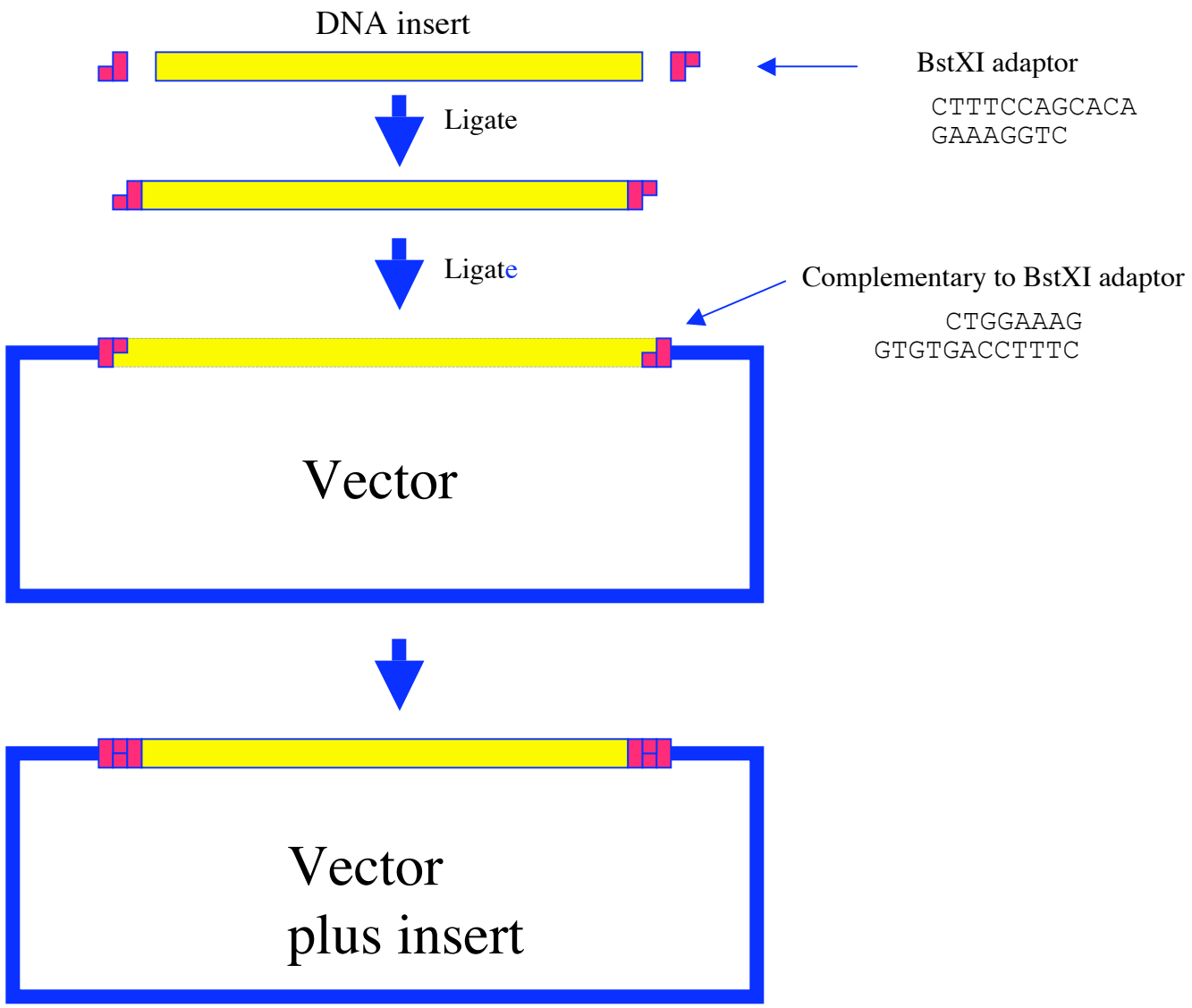
2. BACs (Bacteria Artificial Chromosome) are used for positional cloning, physical mapping and genomic sequencing of large DNAs. They can accomodate up to 350 Kb inserts. One copy/cell.

3. Fosmids are cosmids that have the F plasmid origin of replication and cos sites, the DNA should be cut into 40 Kb pieces and packaged into the phage head which is then transfected into E.coli to produce colonies. One copy/cell.

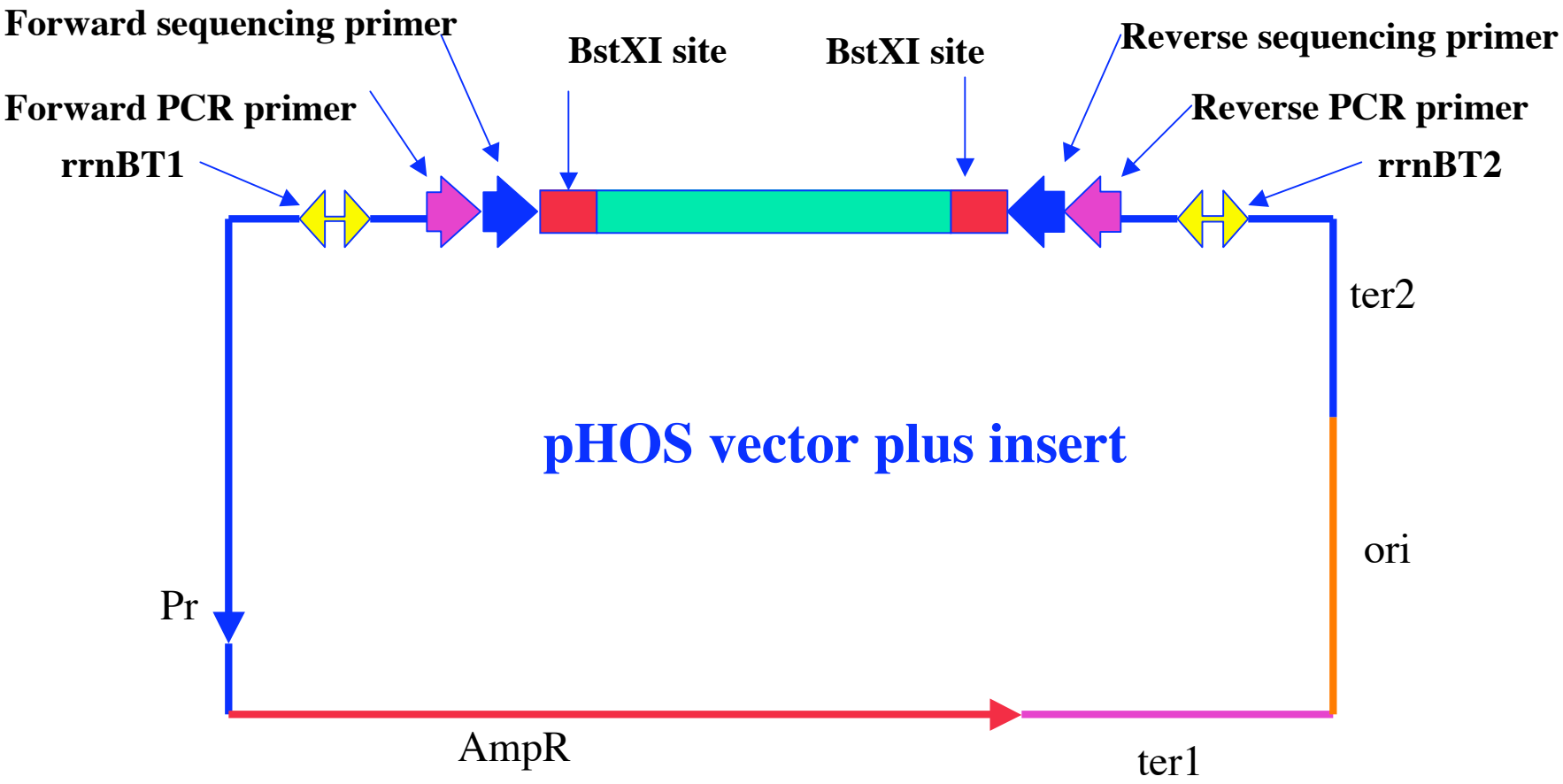
4. Linking libraries.



BstXI adaptor cloning system



Design Features of the BstXI Adaptor Cloning System



pHOS Vector Features

- The sequencing primer sites immediately flank the cloning site to avoid excessive re-sequencing of vector DNA.

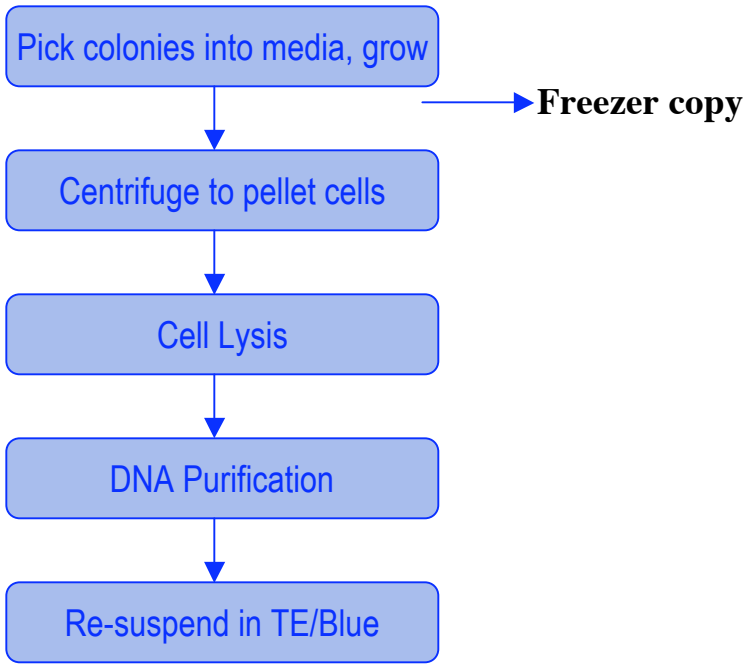
- PCR primer sites are located immediately outside of the Sequencing primer sites to allow PCR amplification for template preparation.

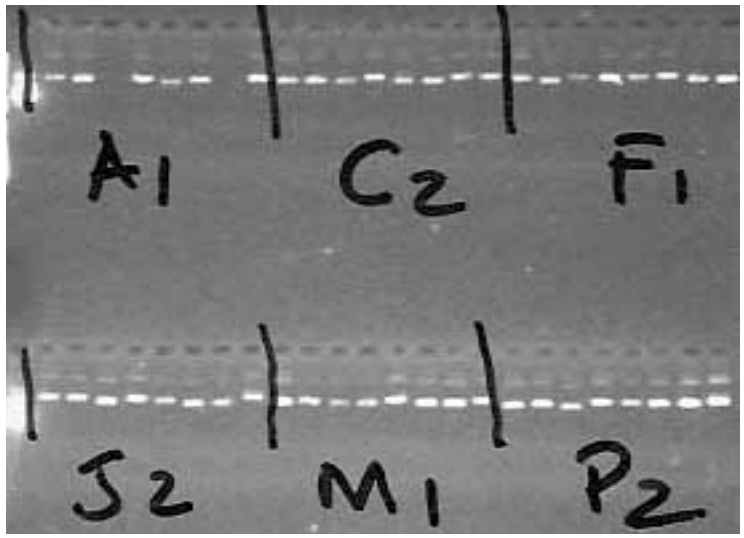
- The vector continues to have the regular, good vector features such as a gene for replication (origin of replication), antibiotic resistance and multiple cloning sites where we can cut the circular plasmid and insert desired features.

Template Production

- Transform clones into *E. coli*
- Plate and pick colonies to isolate clones
- Replicate and isolate plasmid for sequencing
- 384 well HT plate processing

Alkaline Lysis 384-well DNA Isolation Method





Average concentration
30 - 40 ng/μL

Template Production Laboratory

Current Capacity: 22,000,000 plasmids/year
Expansion Capacity: 54,000,000 plasmids/year

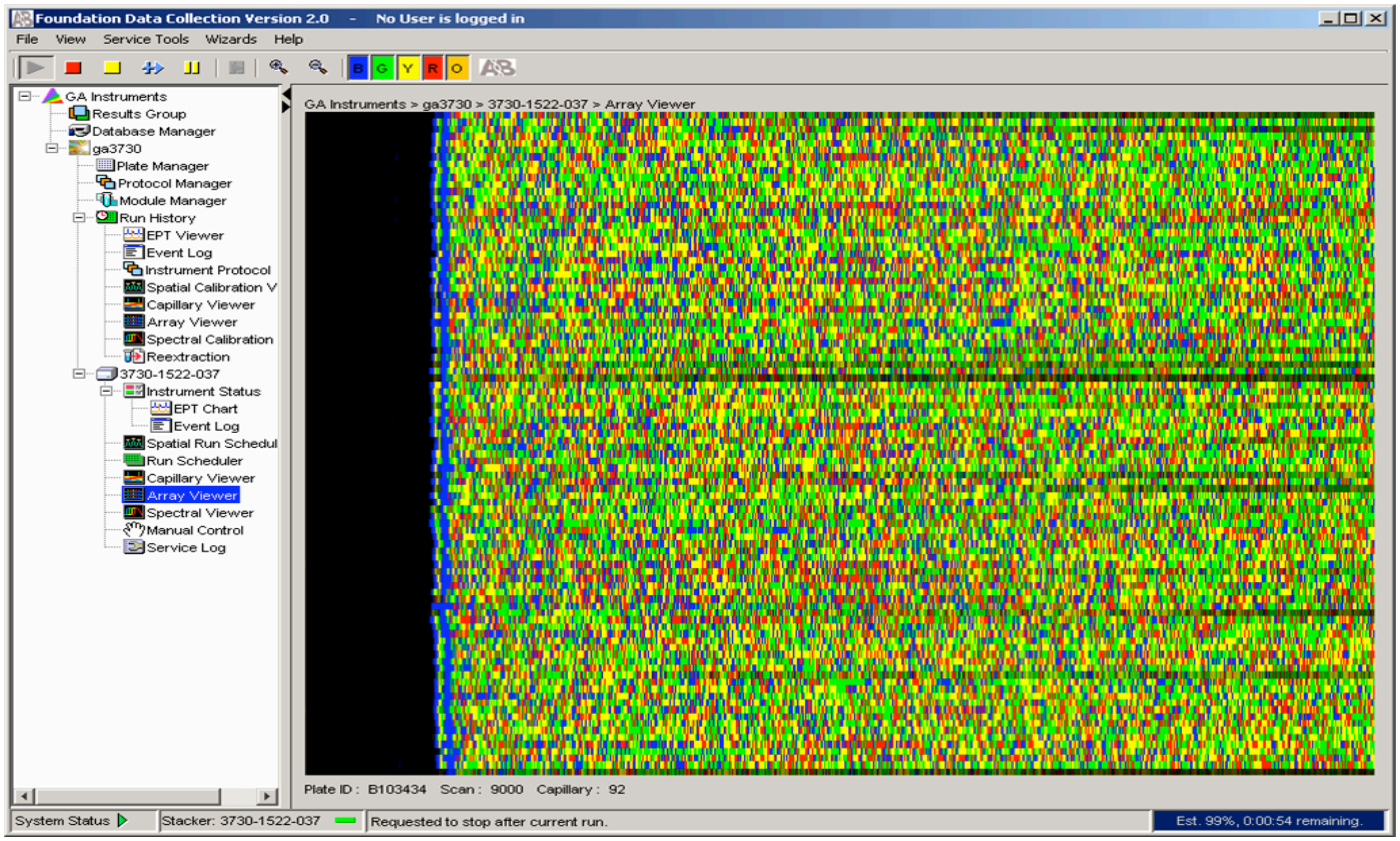
Sequence Production

- Sequencing Reactions and Precipitations
- Samples are loaded into the capillaries on a sequencing machine
- Data Collection
- Data Analysis

Sequence Production Laboratory

Current Capacity: 40,000,000 sequences/year
Expansion Capacity: 100,000,000 sequences/year

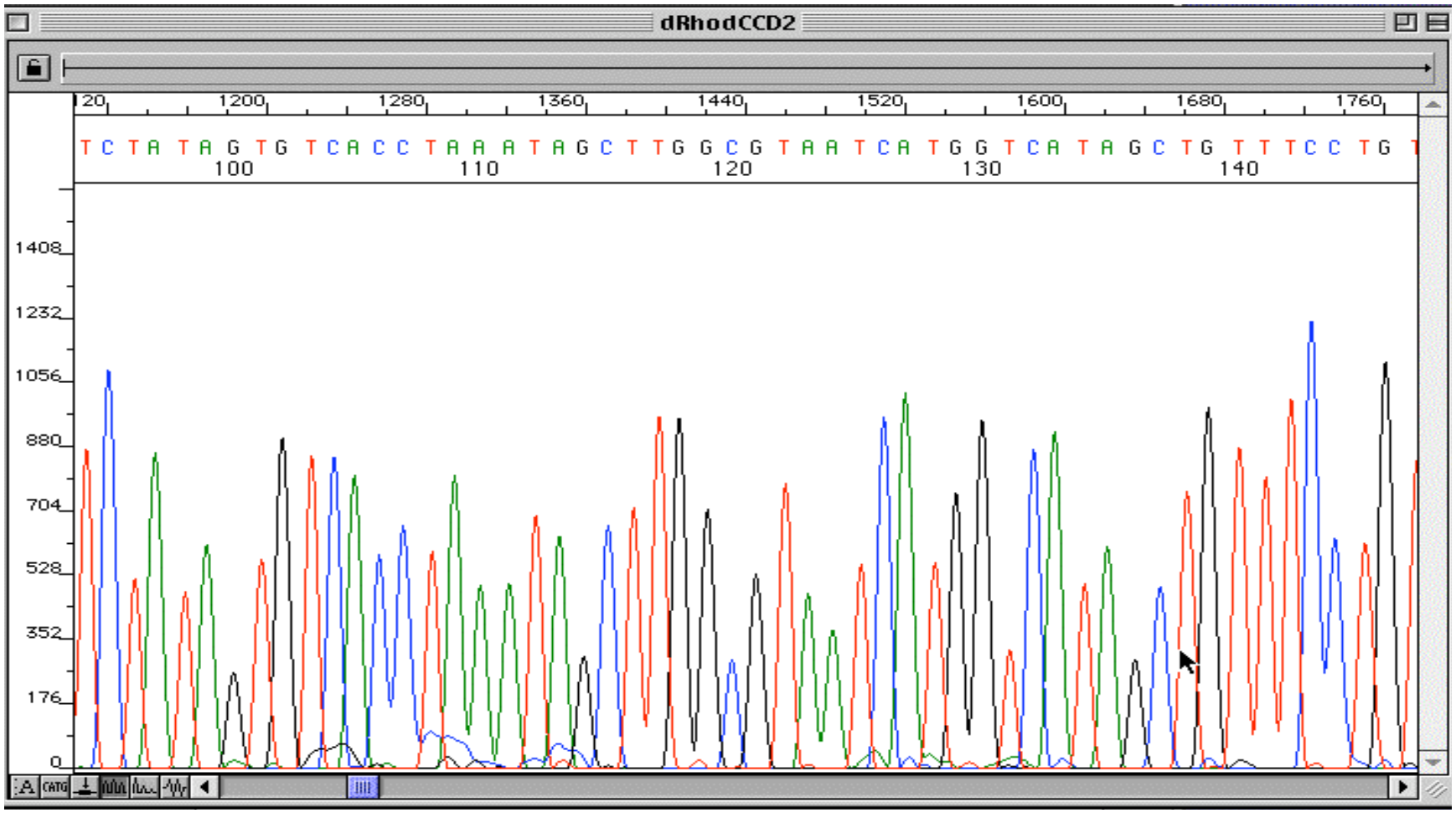
Capillary array view



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Sample Electropherogram



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Basecalling & Quality Assignments

Warner Brothers, Inc.

Phred & TraceTuner

- **Read DNA sequencer traces**
- **Call bases**
- **Assign base quality values**
- **Write basecalls and quality values to output files.**

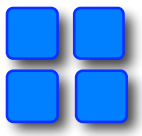
Lowering the error rate, averaging 40% - 50% fewer errors than ABI software independent of position in read, machine running conditions, or sequencing chemistry

What are *phred* quality values?

The quality value *q* assigned to a base call is defined as:

$$q = -10 \times \log_{10}(p)$$

where *p* is the estimated error probability for that base-call.



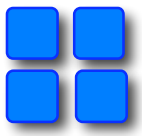
TIGR

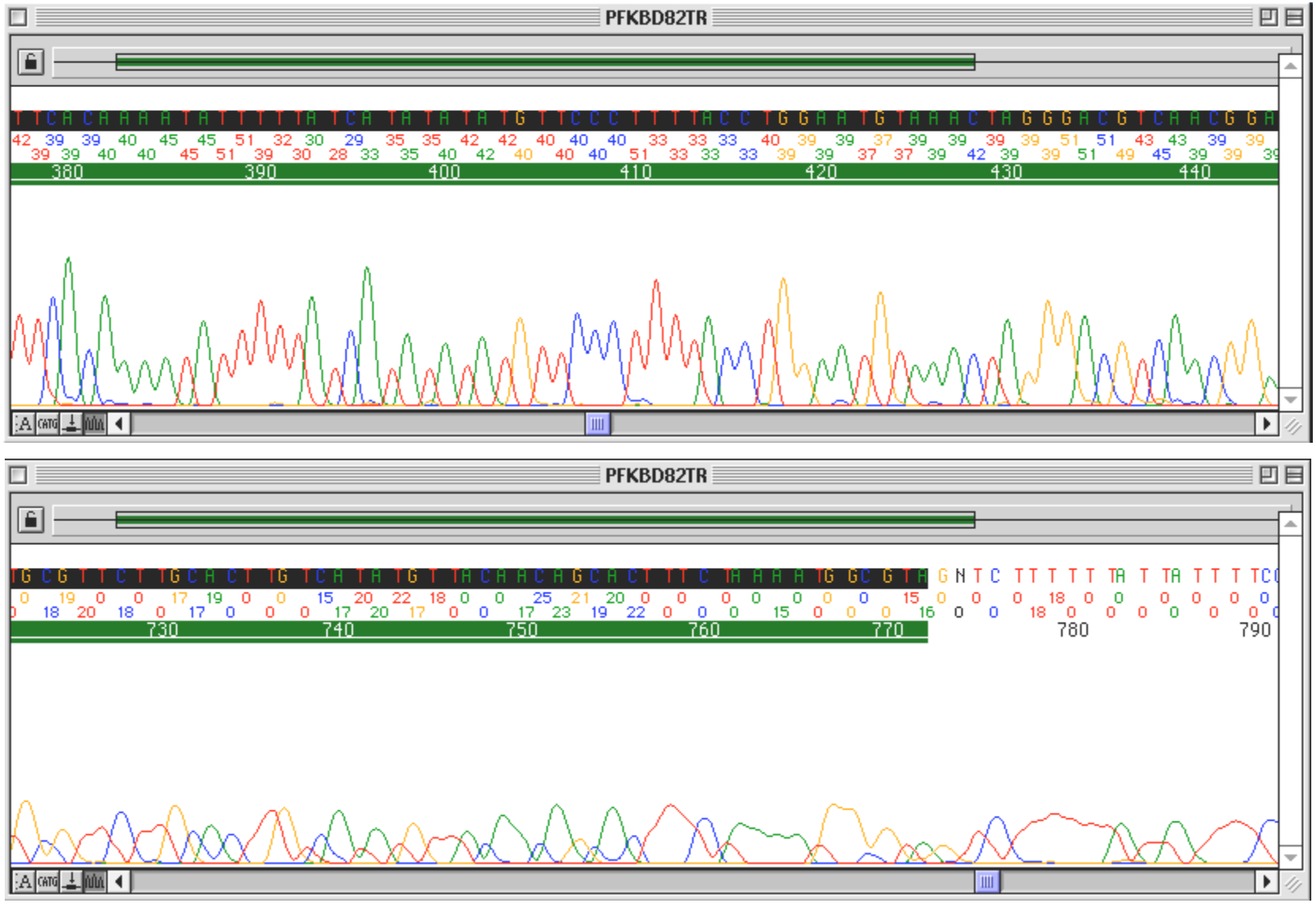
THE INSTITUTE FOR GENOMIC RESEARCH

OR

A base-call having a probability of 1/1000 of being incorrect is assigned a quality value of 30.

Probability	Quality Value
1/100	20
1/10	10



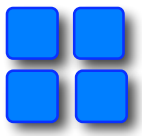


TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

QC Tools

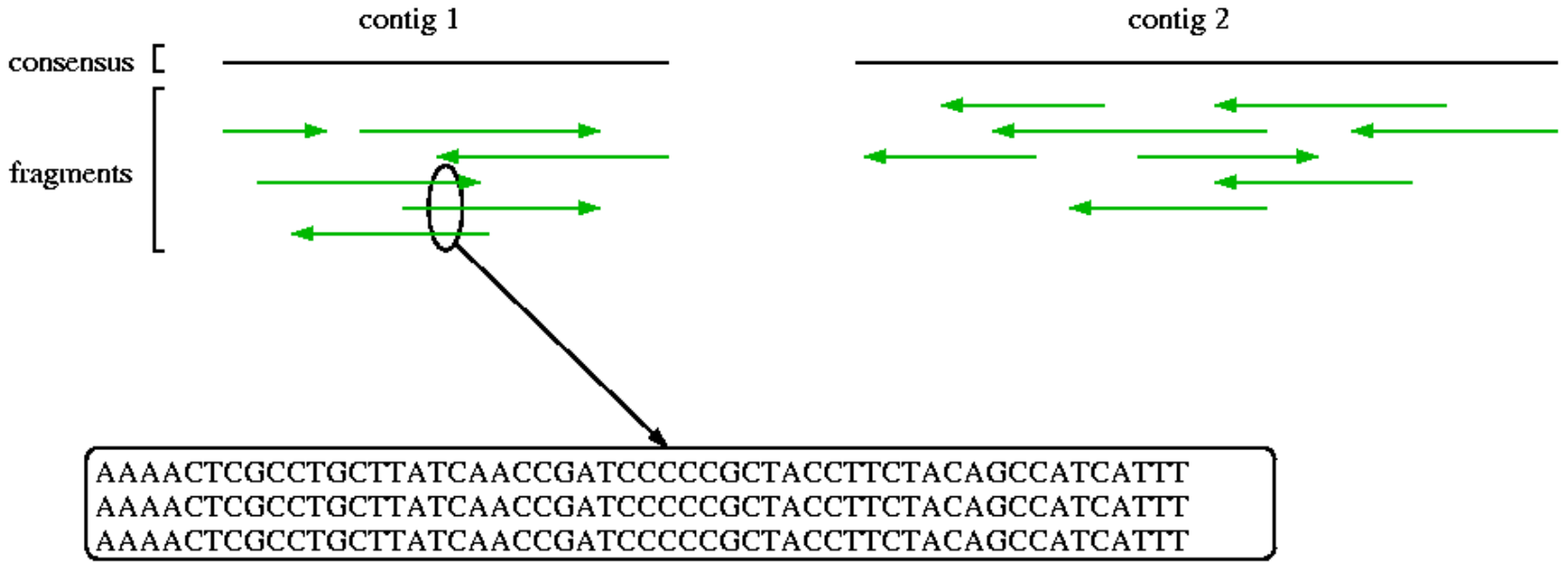
1. Web based document control system including SOP's and Process Control Forms
2. Receipt testing and lot control of sequencing reagents
3. QC audits of reagents and protocols used by production teams
4. Record of operators performing each step of the process
5. Equipment calibration and preventive maintenance program
6. QC/R&D group and QC representative on each team



TIGR

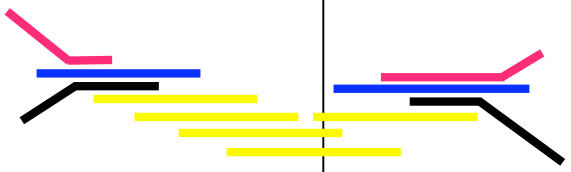
THE INSTITUTE FOR GENOMIC RESEARCH

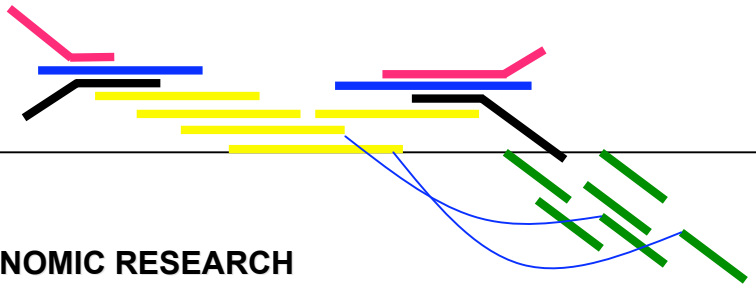
Assembling the fragments



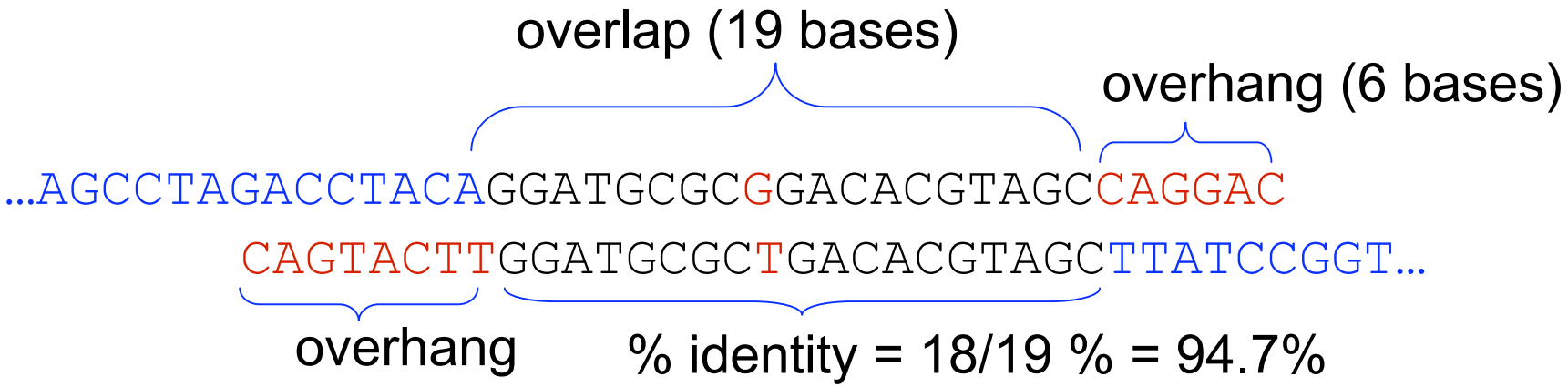
Celera Assembler

- creates high confidence “uniquely assembleable contigs” = unitigs
- marks those that appear repetitive w.r.t. arrival-rate statistics (surrogates and degenerates)
- uses insert (clone-mate) information to build contigs & mark ambiguous unitigs (surrogates)



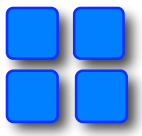


Merging two sequences



overlap - region of similarity between regions
overhang - un-aligned ends of the sequences

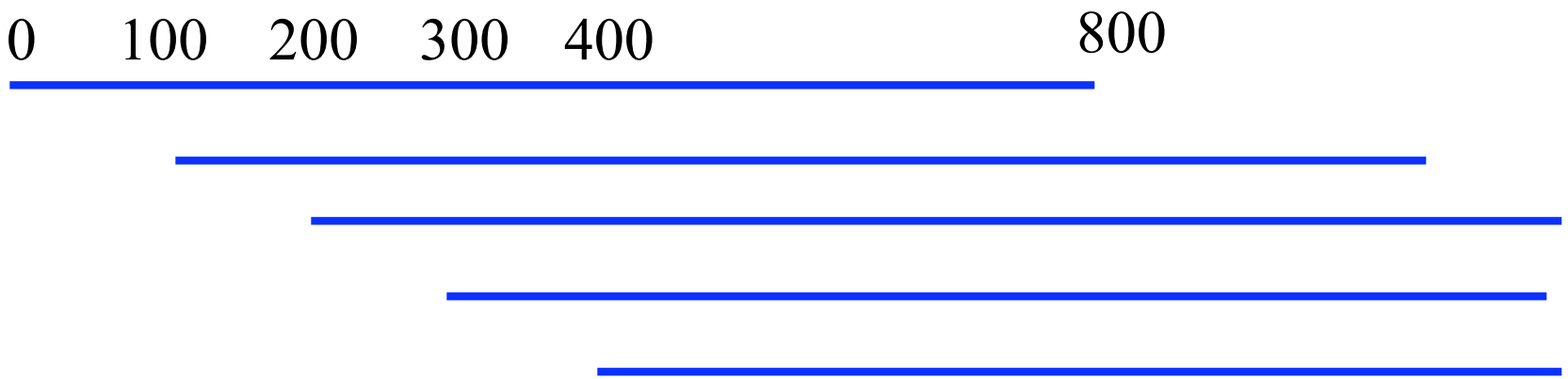
The assembler screens merges based on:
- length of overlap
- % identity in overlap region
- maximum overhang size.



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Arrival rate statistics

8x coverage, 800bp reads - a read every 100bp

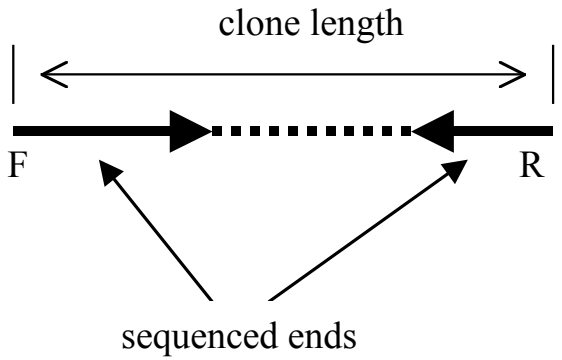


A-stat = log (p(single copy) / p(two-copy))
p(single copy) = ((pF/G)^k/k!) exp(-pF/G)
p(two-copy) = ((2pF/G)^k/k!) exp(-2pF/G)

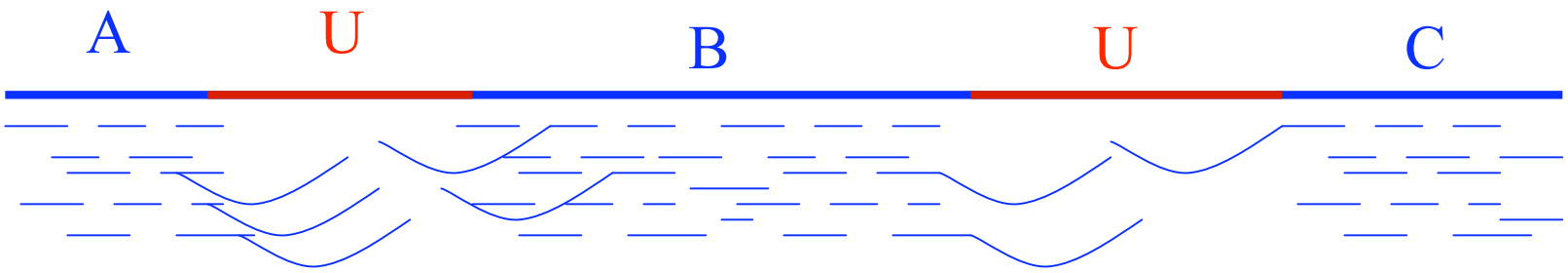
F - # fragments
G - genome size
p/k - average arrival rate

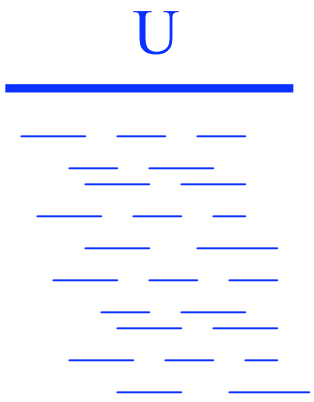
Forward-reverse constraints

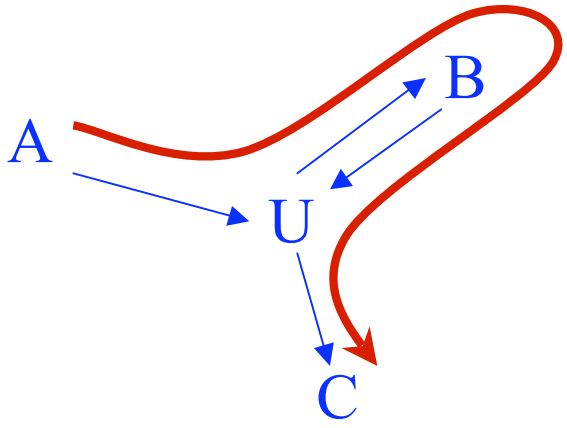
- The sequenced ends are facing towards each other
- The distance between the two fragments is known (within certain experimental error)



Surrogate Unitigs

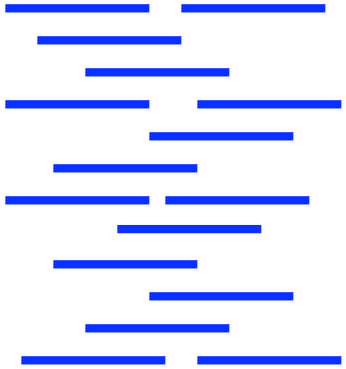






Degenerate contigs

- Too deep



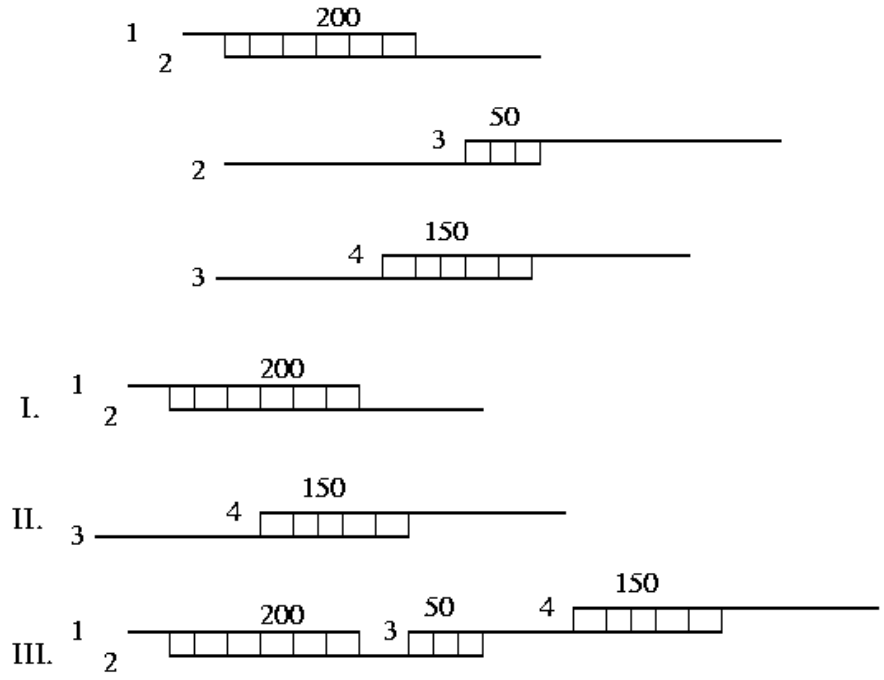
- Too few reads

(NOT IN SCAFFOLDS)

TIGR Assembler

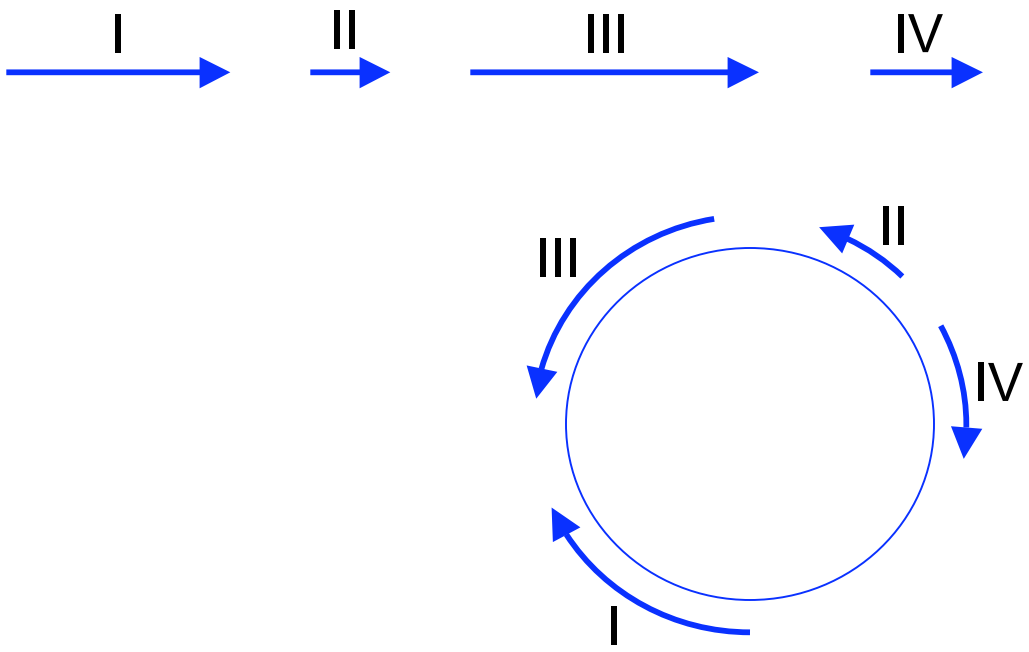
Greedy

- Build a rough map of fragment overlaps
- Pick the largest scoring overlap
- Merge the two fragments
- Repeat until no more merges can be done

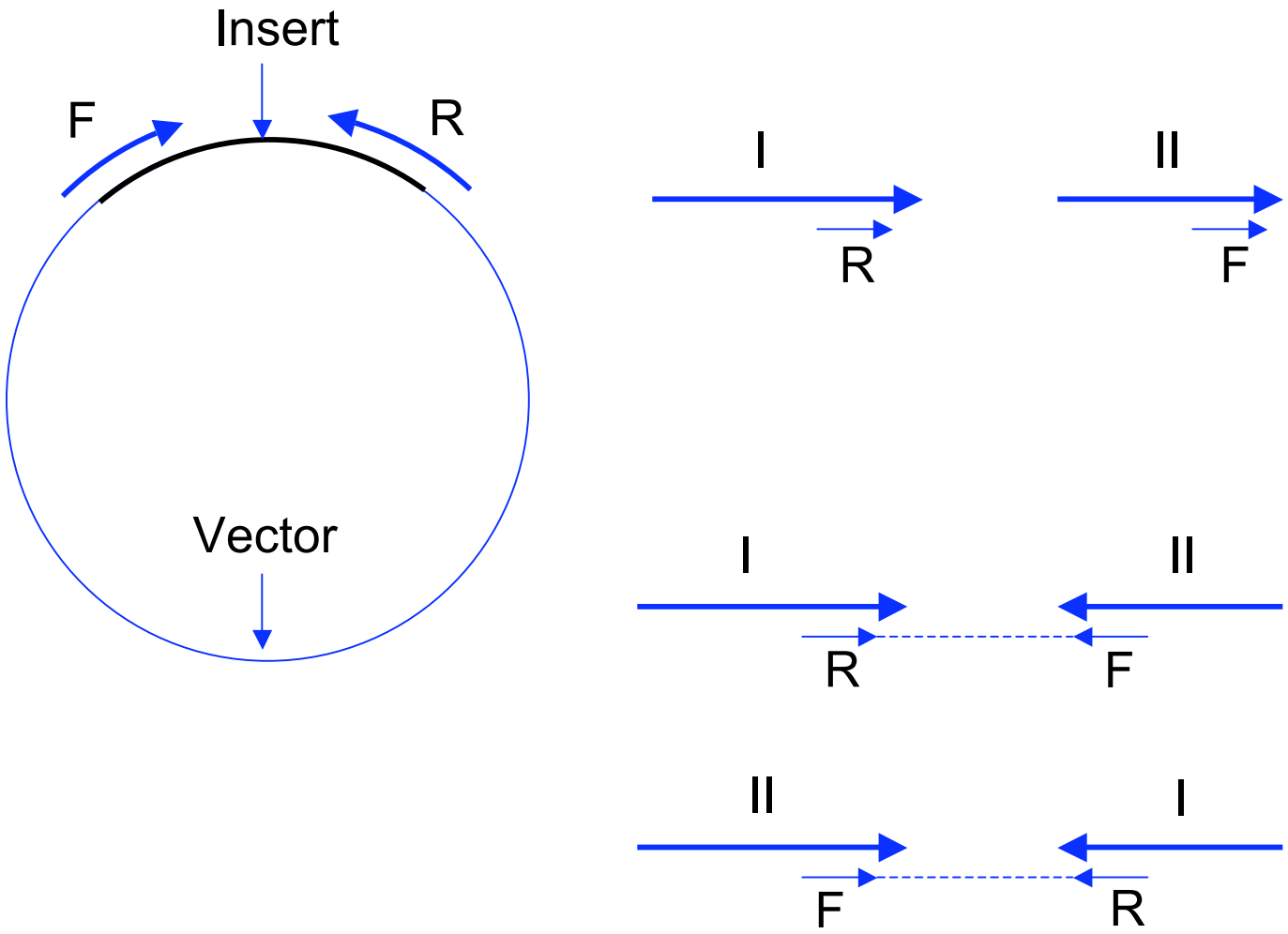


Scaffolding

- Given a set of non-overlapping contigs order and orient them along a chromosome

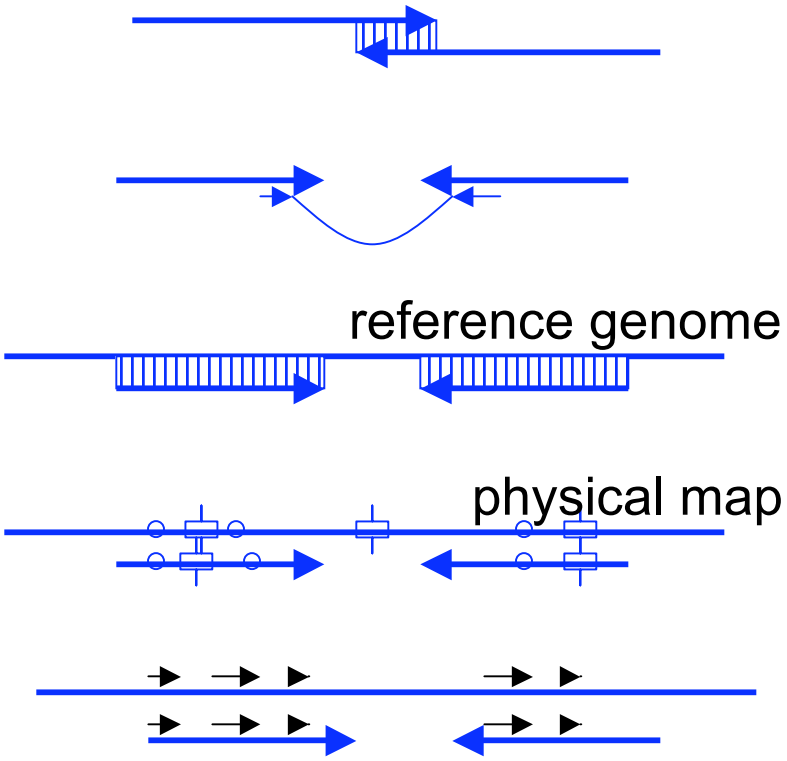


Clone-mates

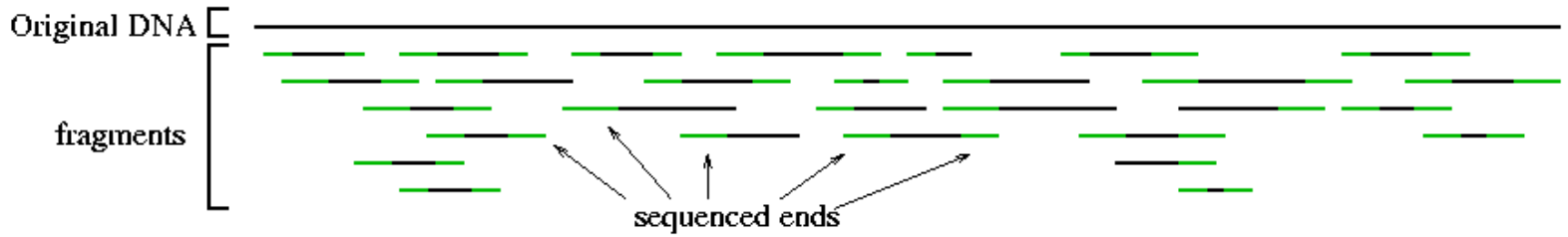


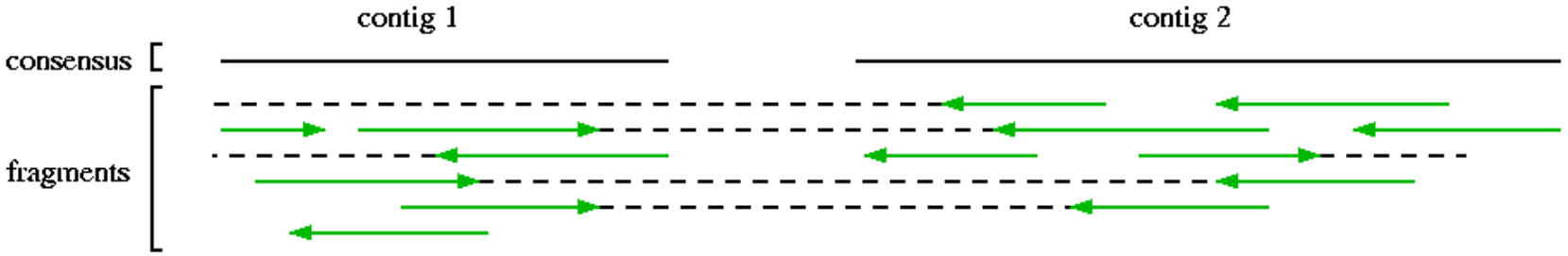
Linking information

- Overlaps
- Mate-pair links
- Similarity links
- Physical markers
- Gene synteny

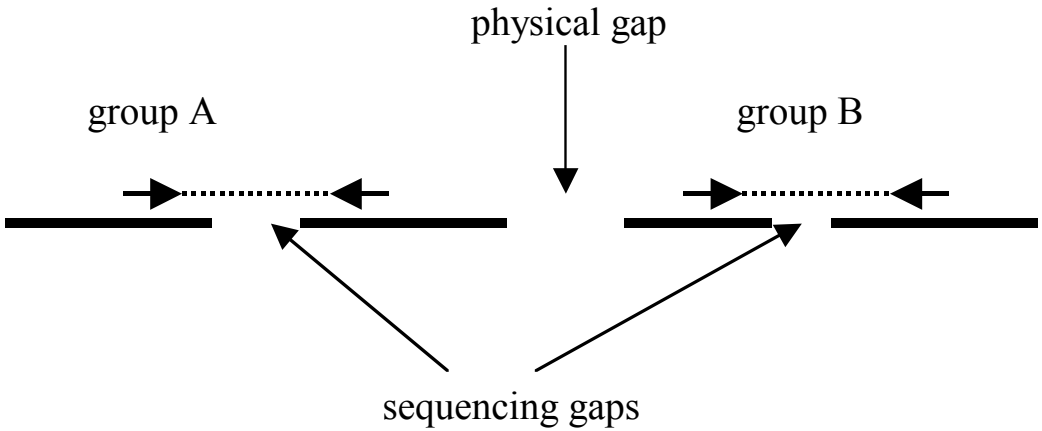


Grouping the contigs





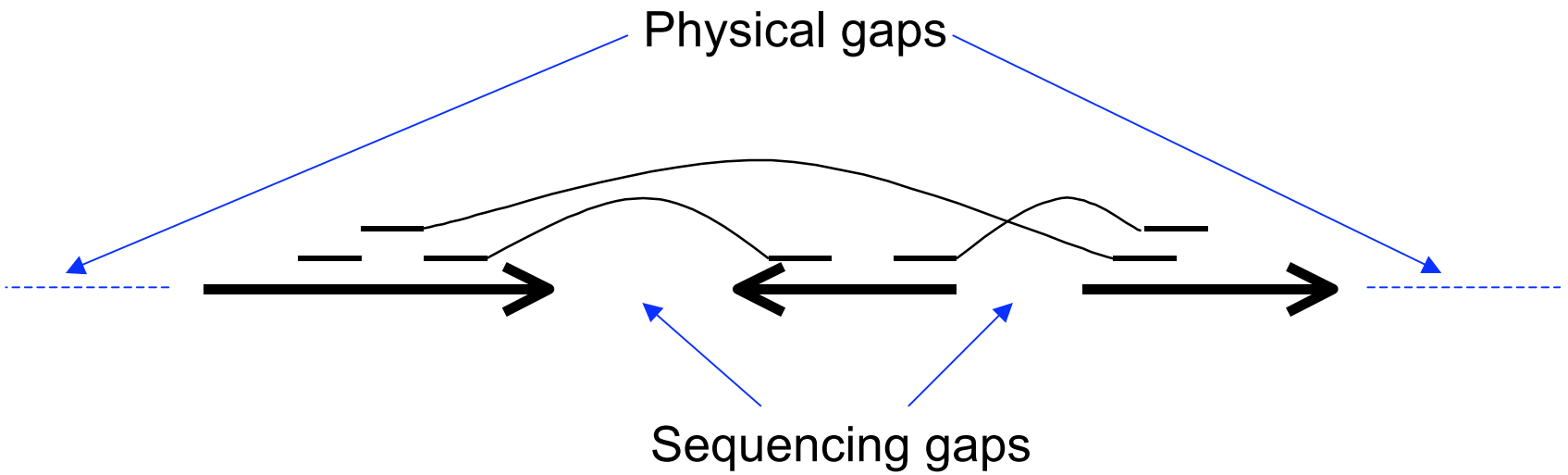
Assembly gaps



sequencing gap - we know the order and orientation of the contigs and have at least one clone spanning the gap

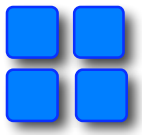
physical gap - no information known about the adjacent contigs, nor about the DNA spanning the gap

Scaffolder output

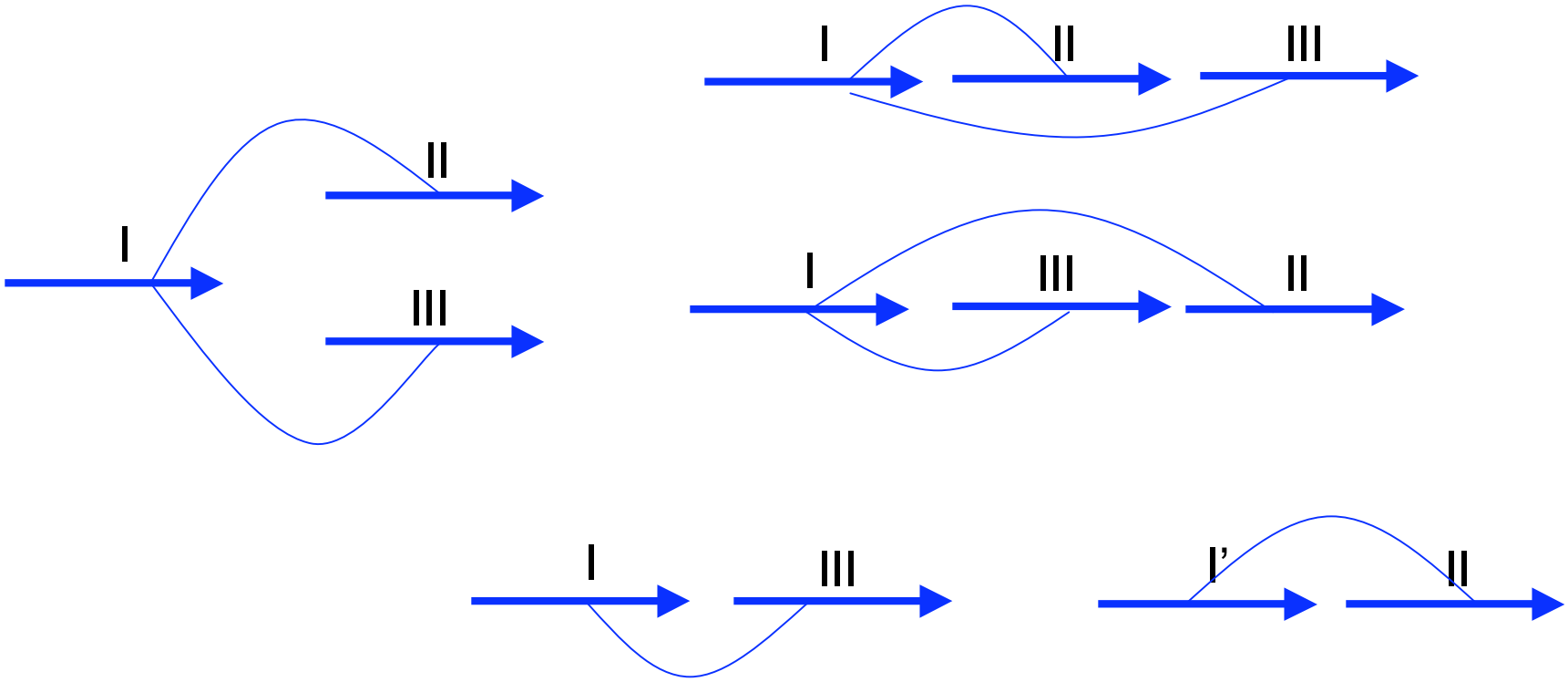


Problems with the data

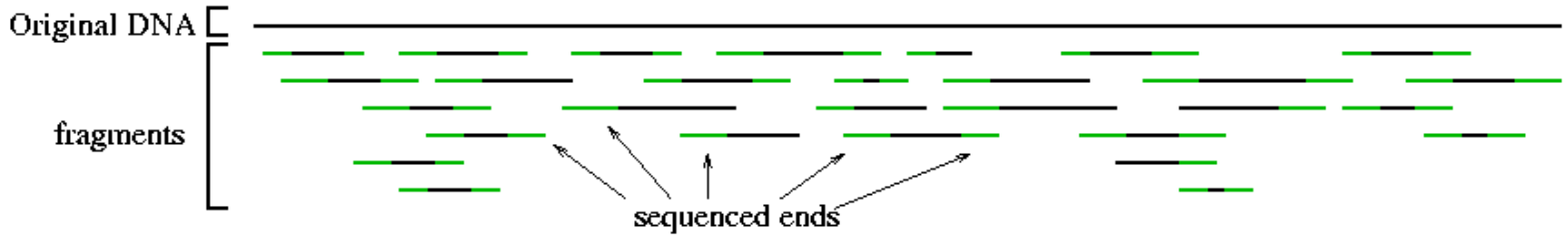
- Incorrect sizing of inserts
  - cut from gel – sizing is subjective
  - error increases with size
- Chimeras (ends belong to different inserts)
  - biological reasons (esp. for large sized inserts)
  - sample tracking (human error)
- Software must handle a certain error rate.

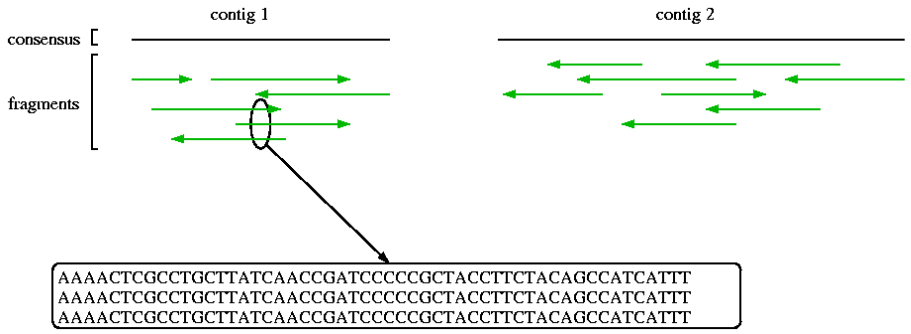


Ambiguous scaffold



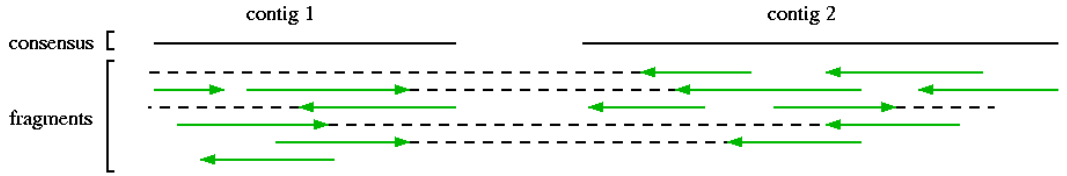
Unifying view of assembly

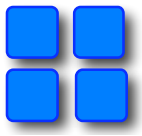




Assembly

Scaffolding





TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

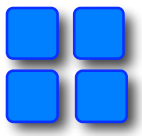
BAC-by-BAC Sequencing and Closure

Why?

- Aids complete closure by avoiding genomic repeat problems
- More streamlined collaboration with other sequencing and closure centers
- Allows for data release and annotation of completed sequence throughout the project

Why Not?

- Complete closure not required
- Small genome size
- Time/cost concerns

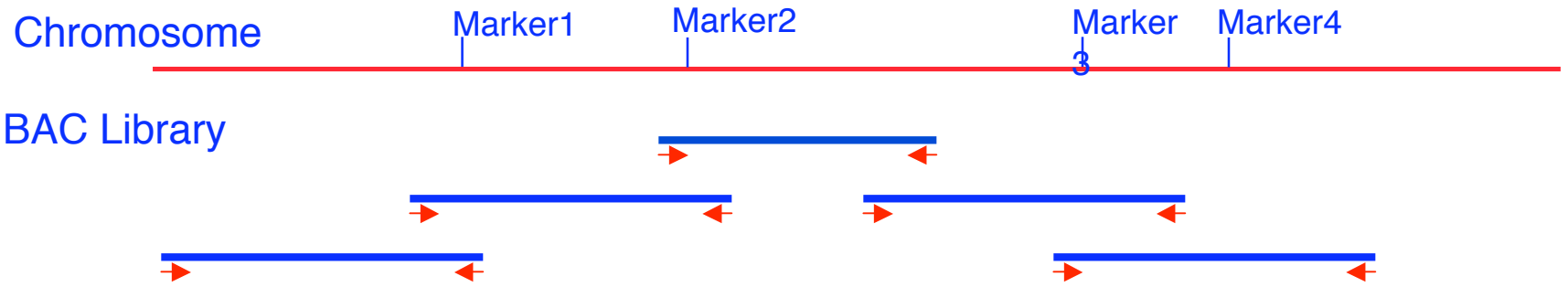


TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

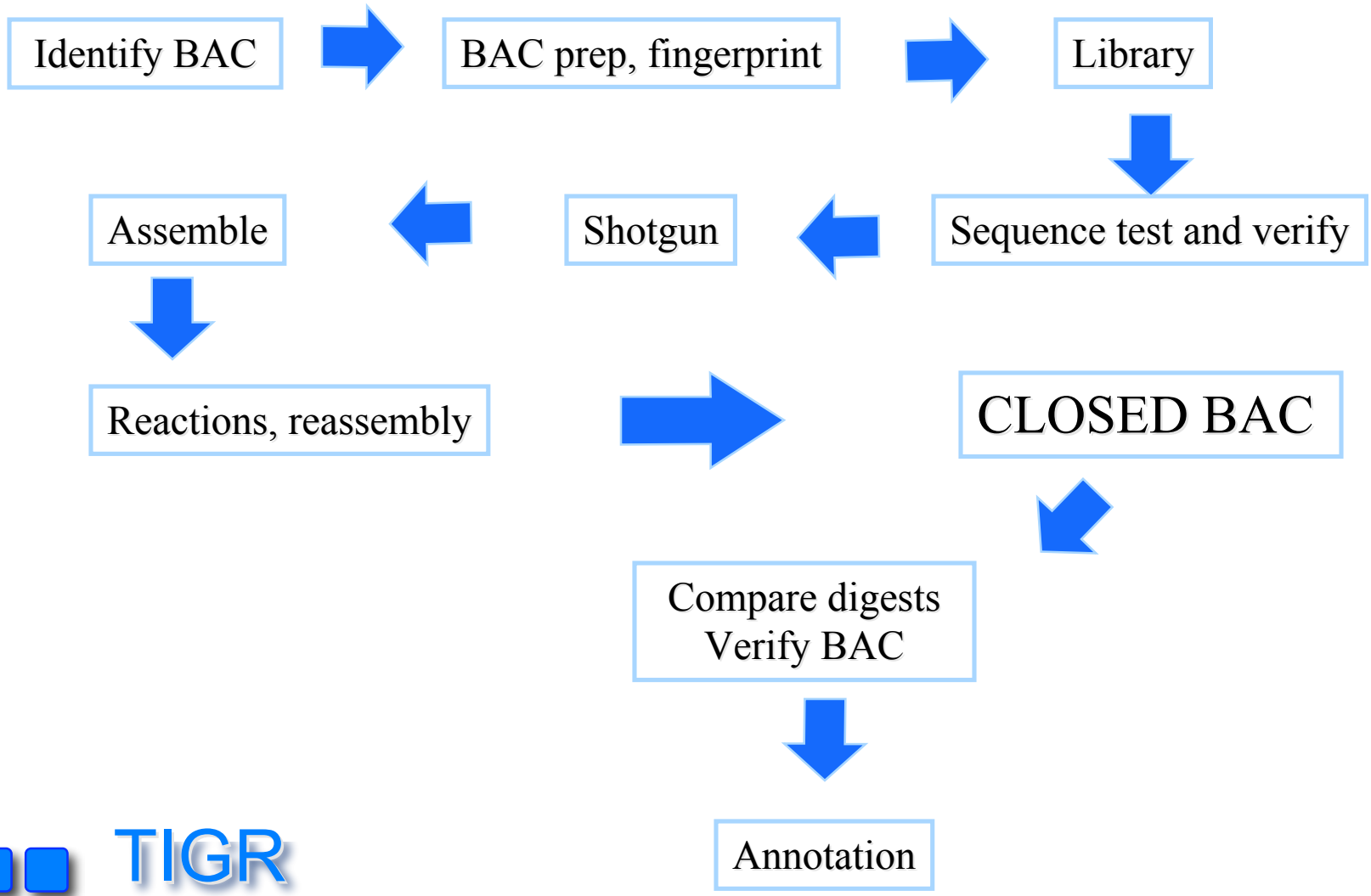
BAC-by-BAC Sequencing and Closure

- Generate BAC-end sequence data
- Choose and sequence “seed” BACs
- Select minimally overlapping “tiling” BACs for contig extension



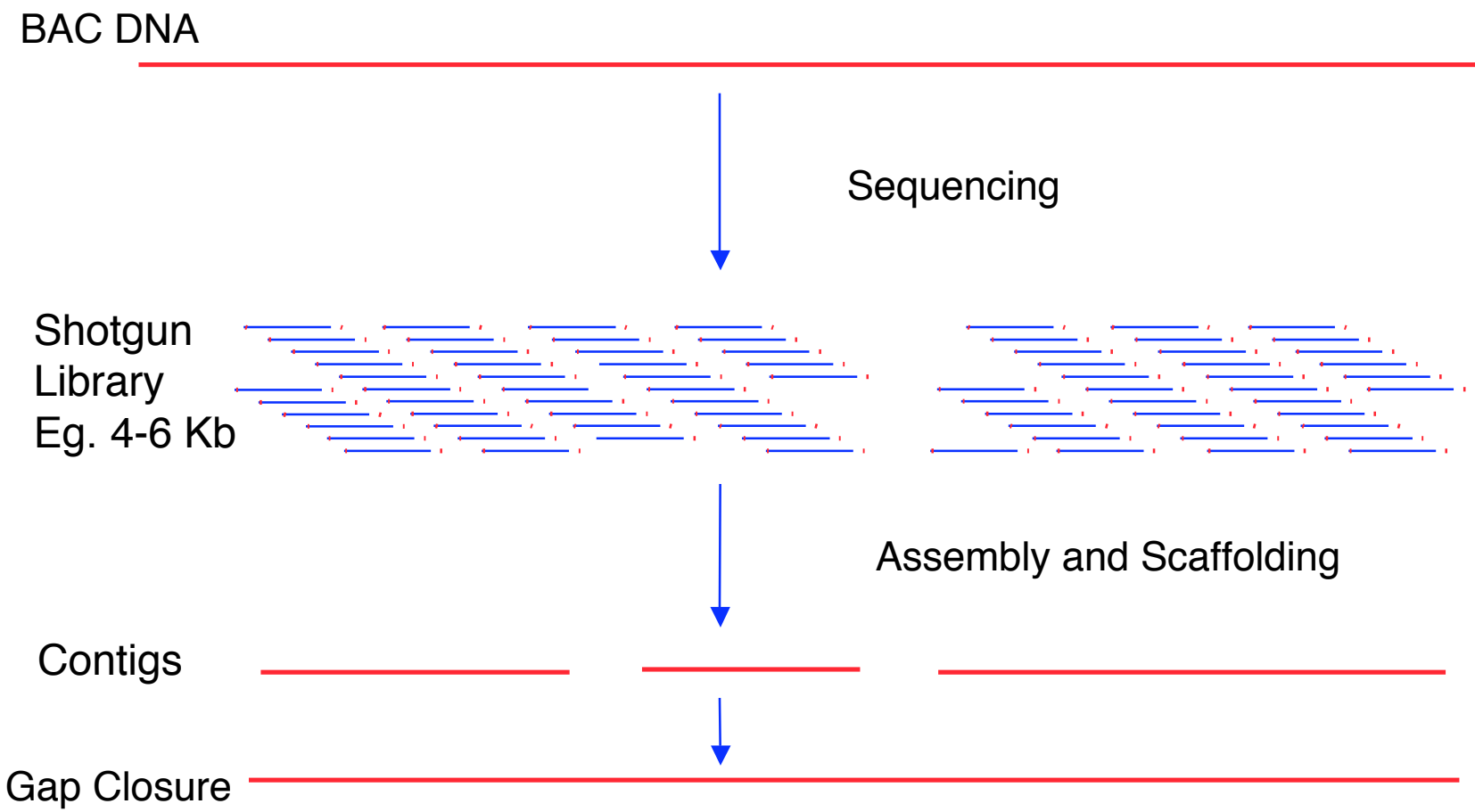
Select BACs based on BAC-end overlaps, physical mapping data, etc.

Life of a BAC at TIGR



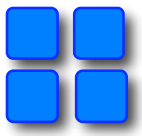
TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

BAC Shotgun Sequencing



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Genome Closure/Finishing

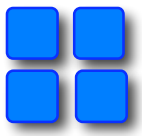


TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

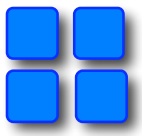
Why Completeness is Important

- Improves characterization of genome features
  - Gene order, replication origins
- Better comparative genomics
  - Genome duplications, inversions
- Determination of presence and absence of particular genes and features is less subjective
- Missing sequence might be important (e.g., centromere)
- Allows researchers to focus on biology not sequencing
- Facilitates large scale correlation studies
- Controls for contamination

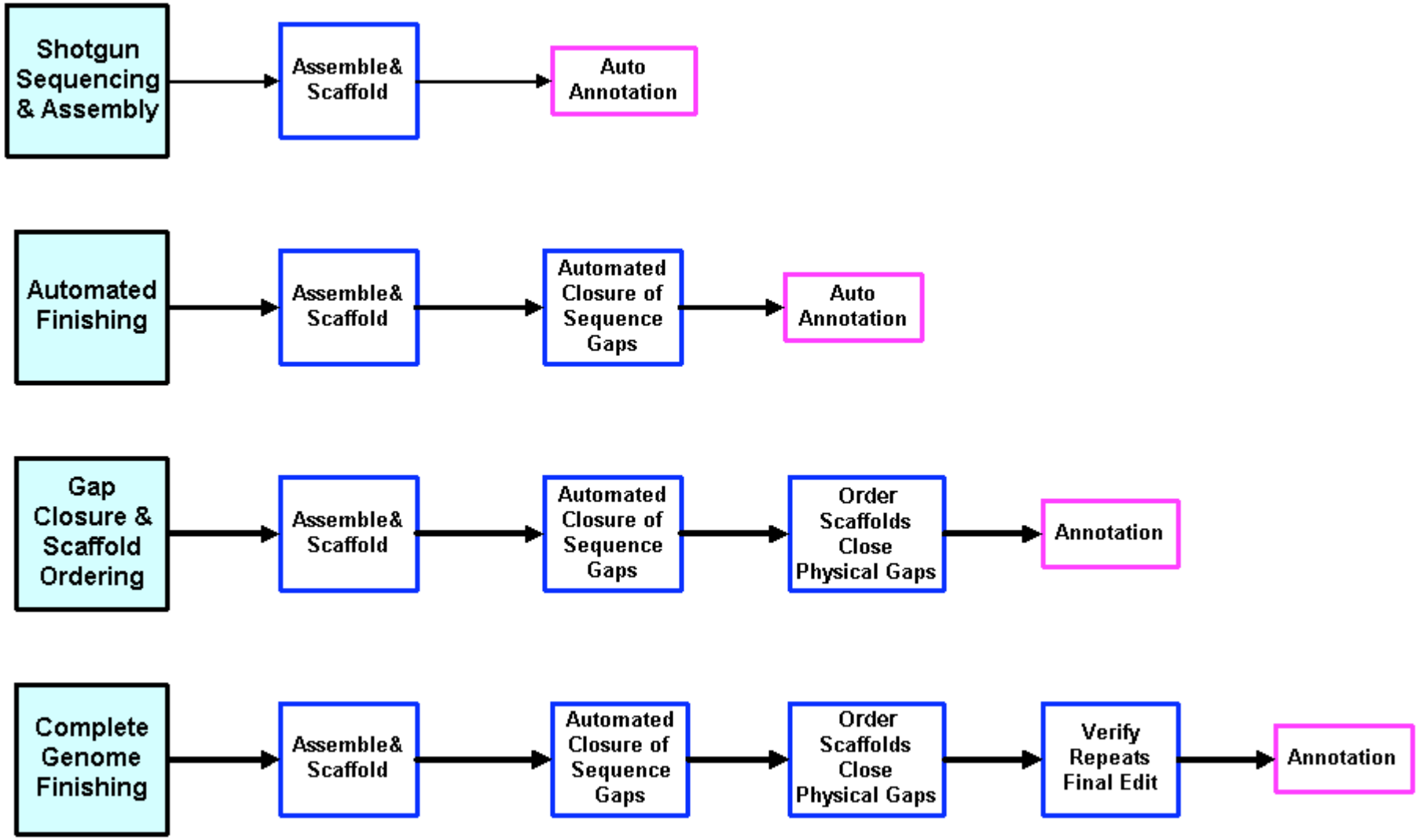


What Is Closure?

- Obtaining sequence that was not obtained during random sequencing which resulted in:
    - Sequencing Gaps
    - Physical Ends
- Confirming the integrity of assemblies
    - Repeats and misassemblies
    - Verification of Clone Coverage
- Confirming the base sequence of the consensus
    - Editing
    - Verification of Sequence Coverage



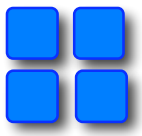
Levels of Genome Closure



TIGR Finishing Criteria

A genome is considered finished and ready if it satisfies the following criteria:

1. A continuous consensus of DNA sequence
2. No ambiguous consensus basepairs
3. At least 2X sequence coverage over the entire genome.
    a) both strands of one clone are sequenced
    b) two different clones sequenced
    c) Same clone sequenced with dual chemistry
4. At least 2X clone coverage over the entire genome
5. Complete confidence in all repetitive areas

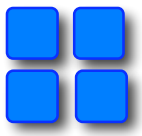


TIGR

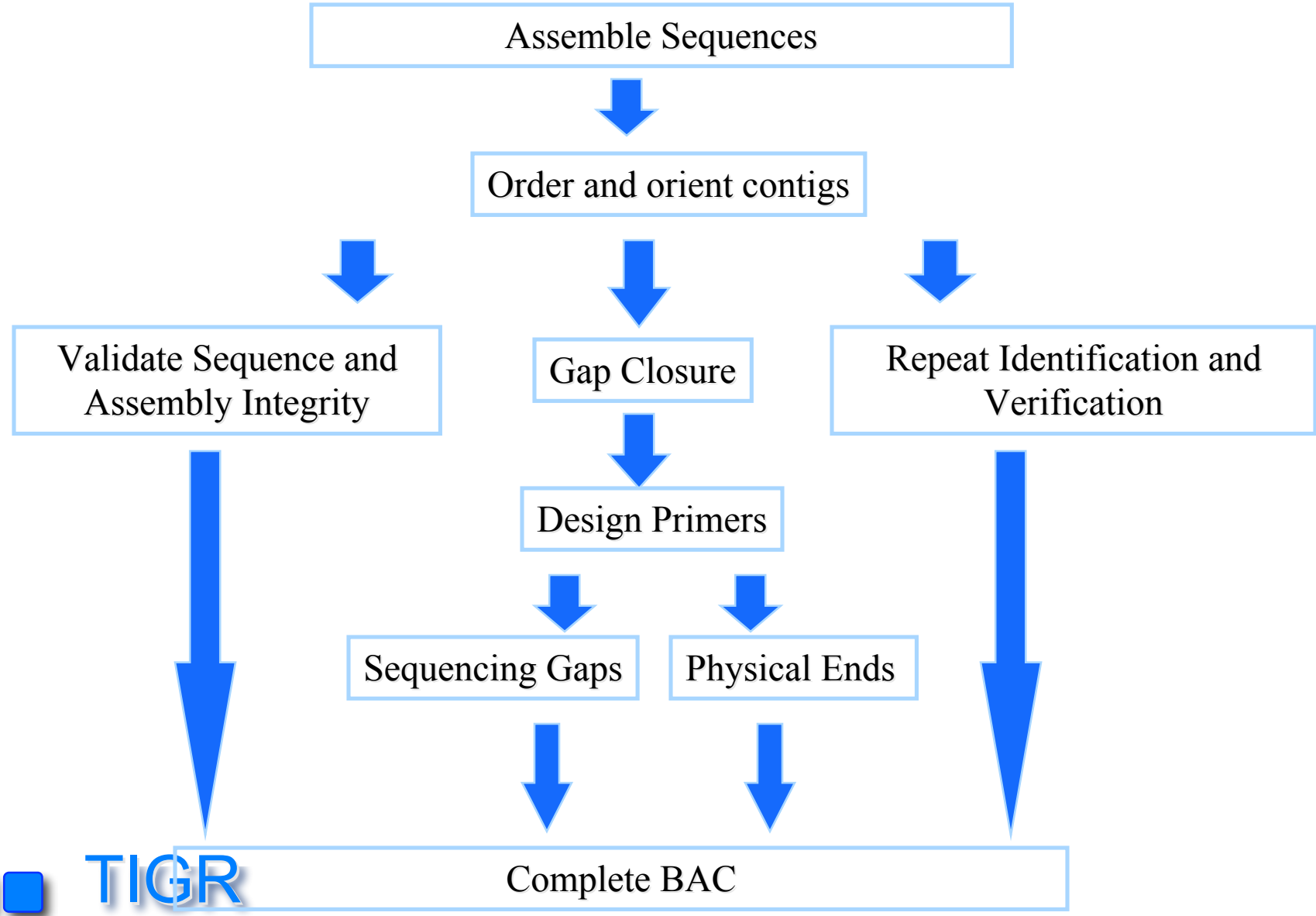
THE INSTITUTE FOR GENOMIC RESEARCH

Causes for gaps

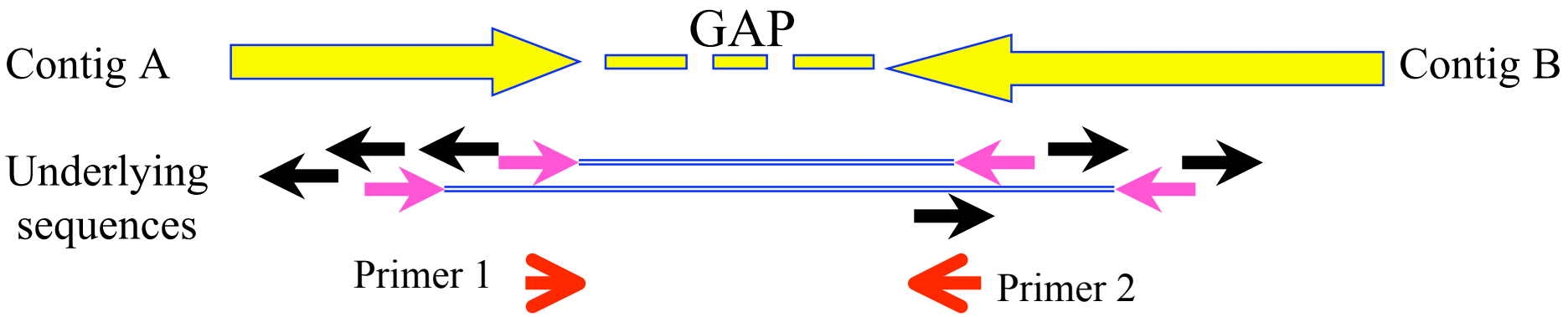
- Non-random shotgun library
  - Toxicity of genes or promoters in *E. coli*
  - Genomic DNA difficult to clone (capsular polysaccharides)
  - Unstable regions (low complexity)
- Sequencing problems
  - Hard stops
    - Secondary structures
    - Very high or low GC content
    - Small unit tandem repeats
  - Loss of signal
    - Homopolymeric tracts
    - Very high or low GC content



Genome Closure Steps



Sequencing Gaps



To close Sequencing gaps:

A. Resequence short reads at contig ends

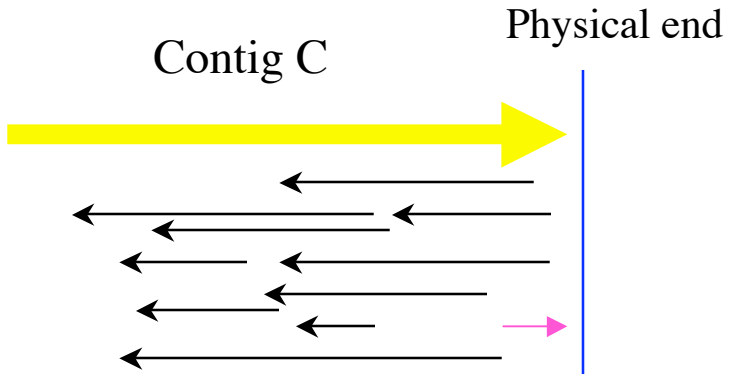
B. Perform sequencing reactions:

1. Design primers at contig ends
2. Do sequencing reactions with:

Insert 1 + primer 1	Insert 1 + primer 2
Insert 2 + primer 1	Insert 2 + primer 2

Physical Ends

No known links exist; there is no template available for sequencing.



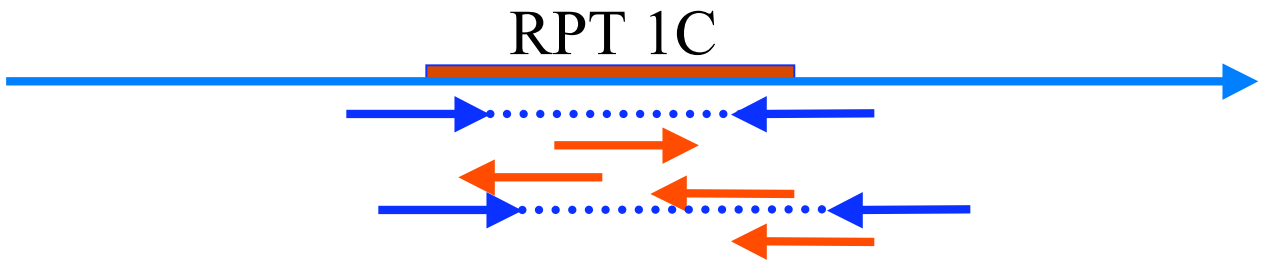
To link and close Physical ends use PCR:

1. Design primers at contig ends
2. If PCR products are obtained, sequence PCR products completely

Repetitive Areas

- Repetitive areas are regions of high similarity within the genome/BAC.
- Sequences in these areas may be misassembled by the Assembler.
- Verification of the sequence of repetitive areas:
    - A. Identify potential repetitive areas, using *repeatFinder* and other tools.
    - B. Classify repeats based on length, copy number, % similarity, structure and complexity.
    - C. If repeats are misassembled, transpose spanning clones or obtain PCR products and sequence to verify assembly.
- Misassemblies can occur in two ways:

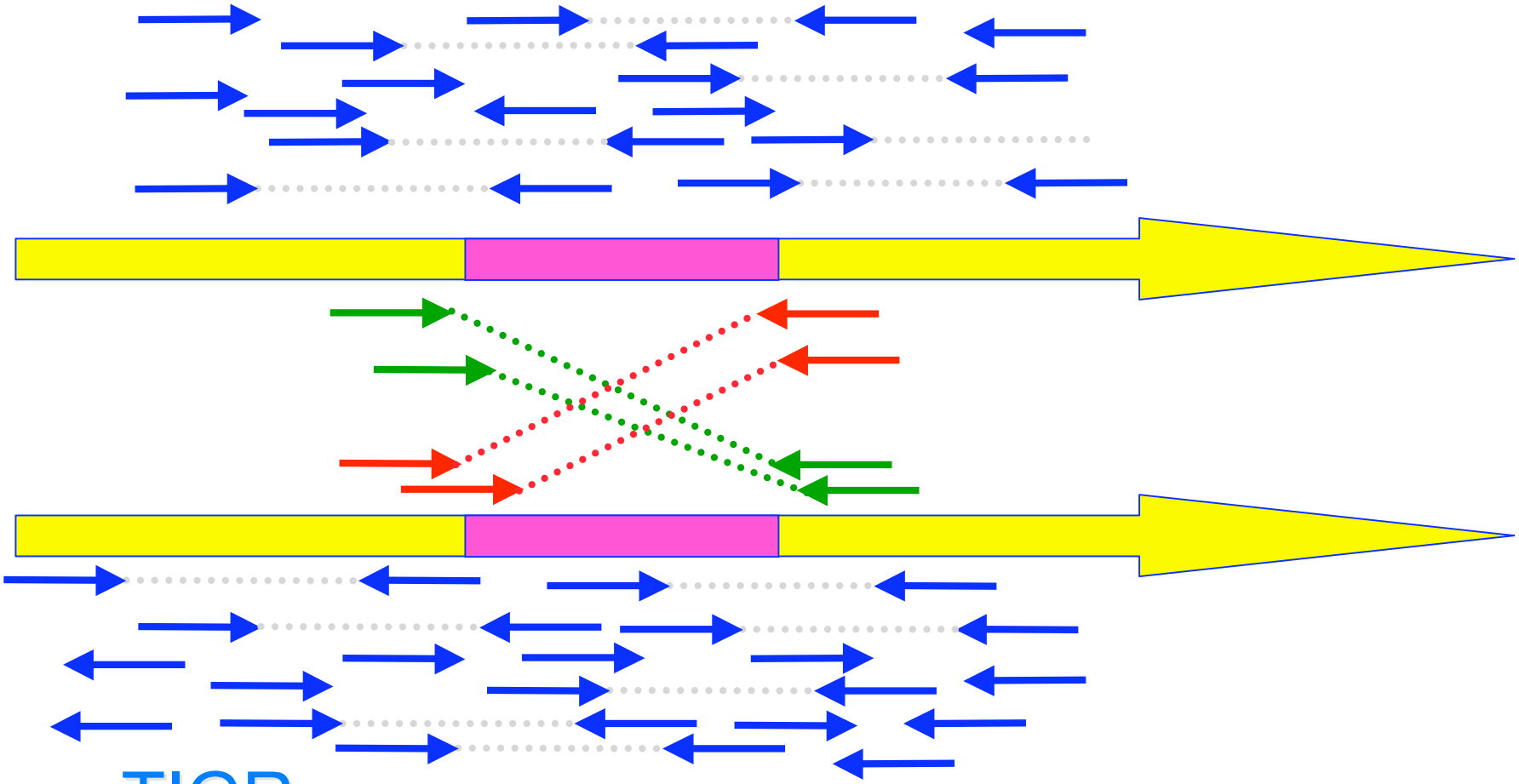
The base sequence of the repeat may be misassembled:



- Repeat is spanned by linking clones
- Underlying sequences are unlinked clones, with mates in other assemblies (branched clones) or mates that did not sequence (dead end clones).
- Results in incorrect consensus sequence

Mis-assembled repeat

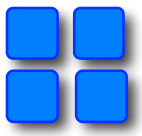
Clones link different repeat flanks



Identifying repetitive areas

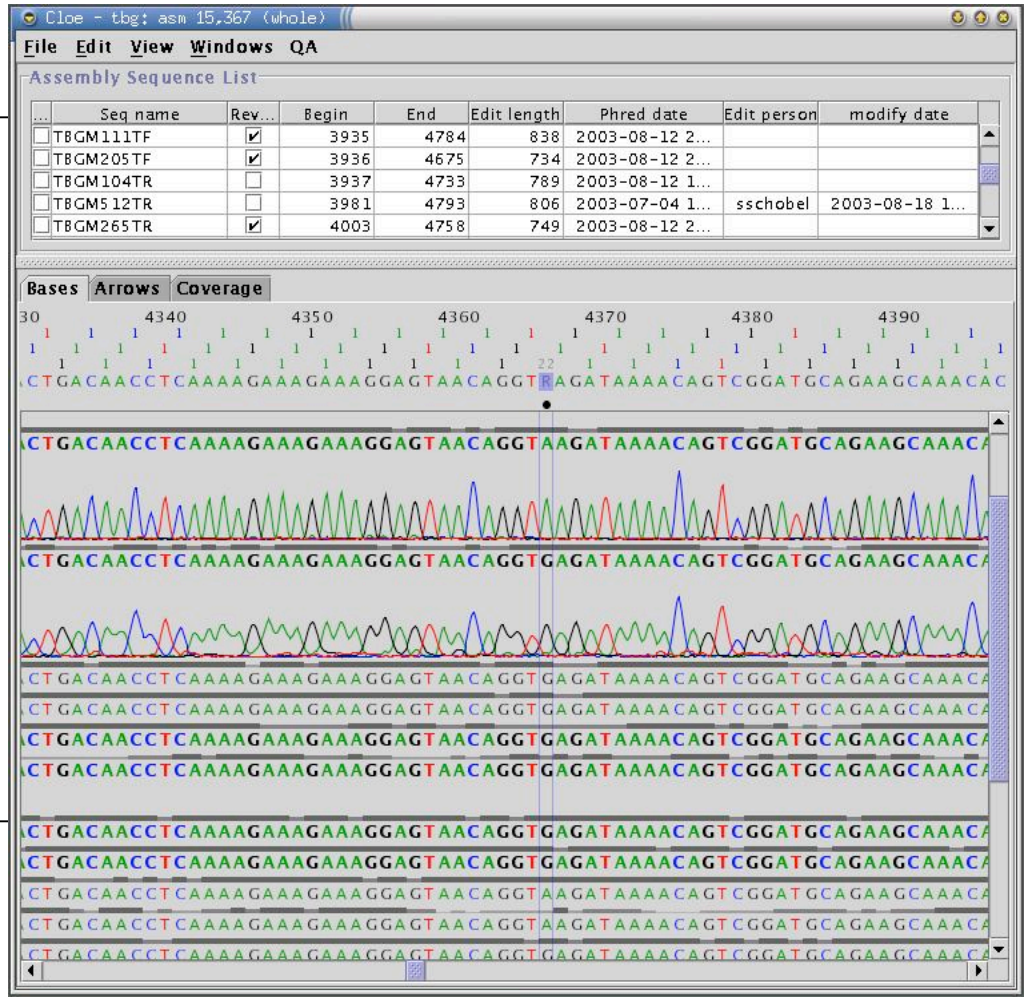
Repetitive areas are identified by the program *repeatFinder*
ftp://ftp.tigr.org/pub/software/repeatFinder/

- Analyzes sets of contigs for repeats greater than 50 bp and then groups these repeats into classes
- *repeatFinder* calls *REPuter*



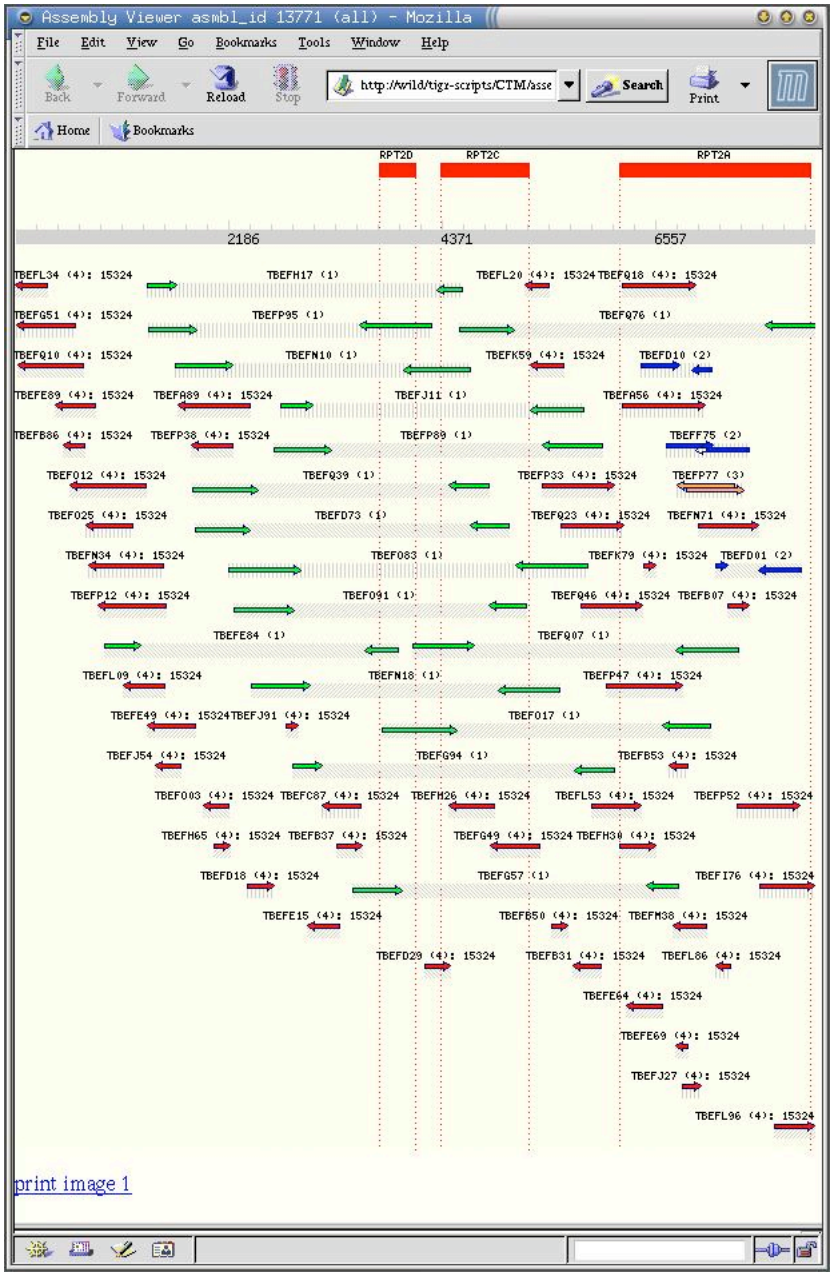
Sequence editor

- In this example there is an obvious discrepancy between the base calls of several of the underlying clones in this region.



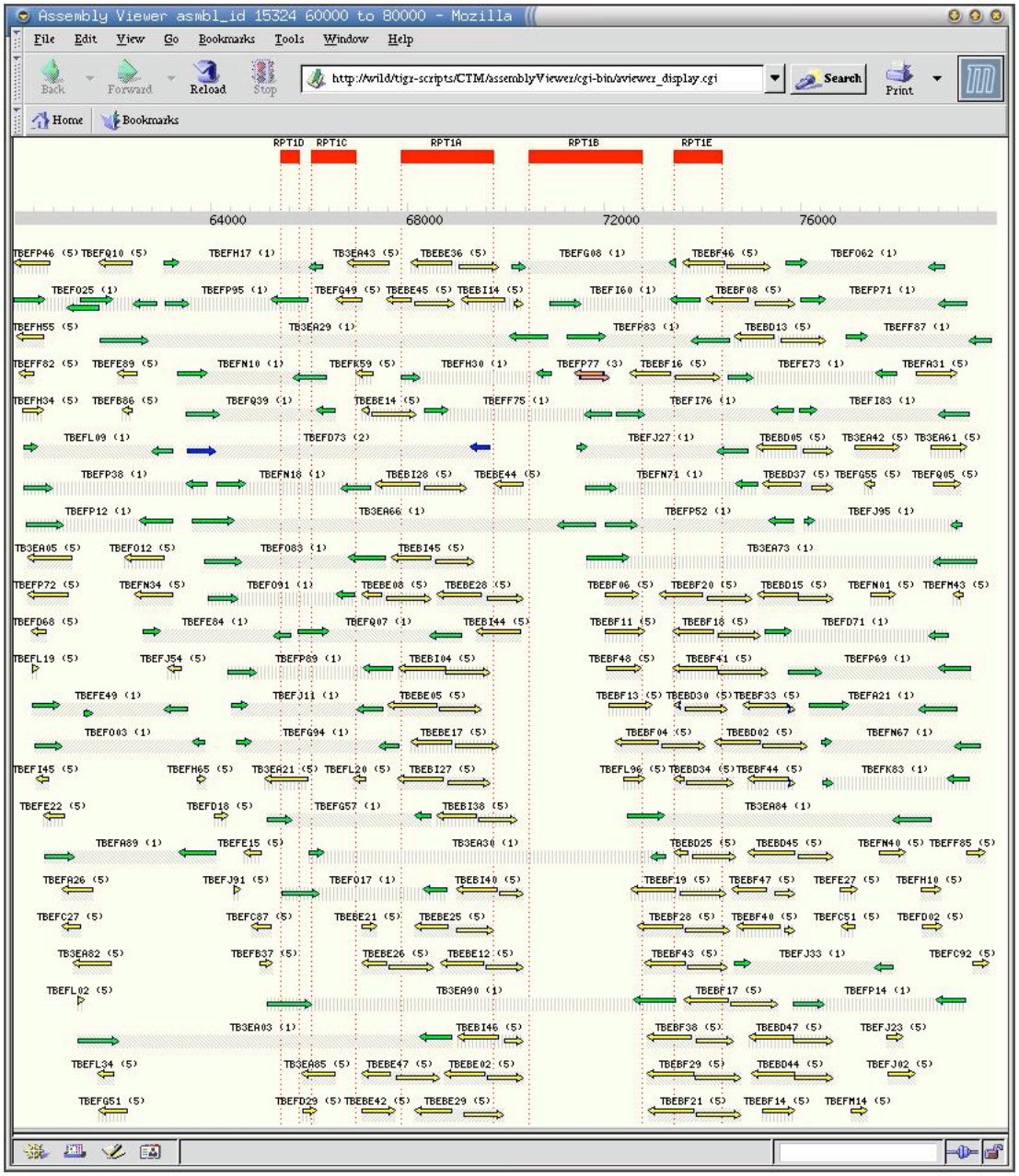
Assembly Viewer

- Several clones appear to be misassembled
- There are both obvious size violations and orientation issues



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

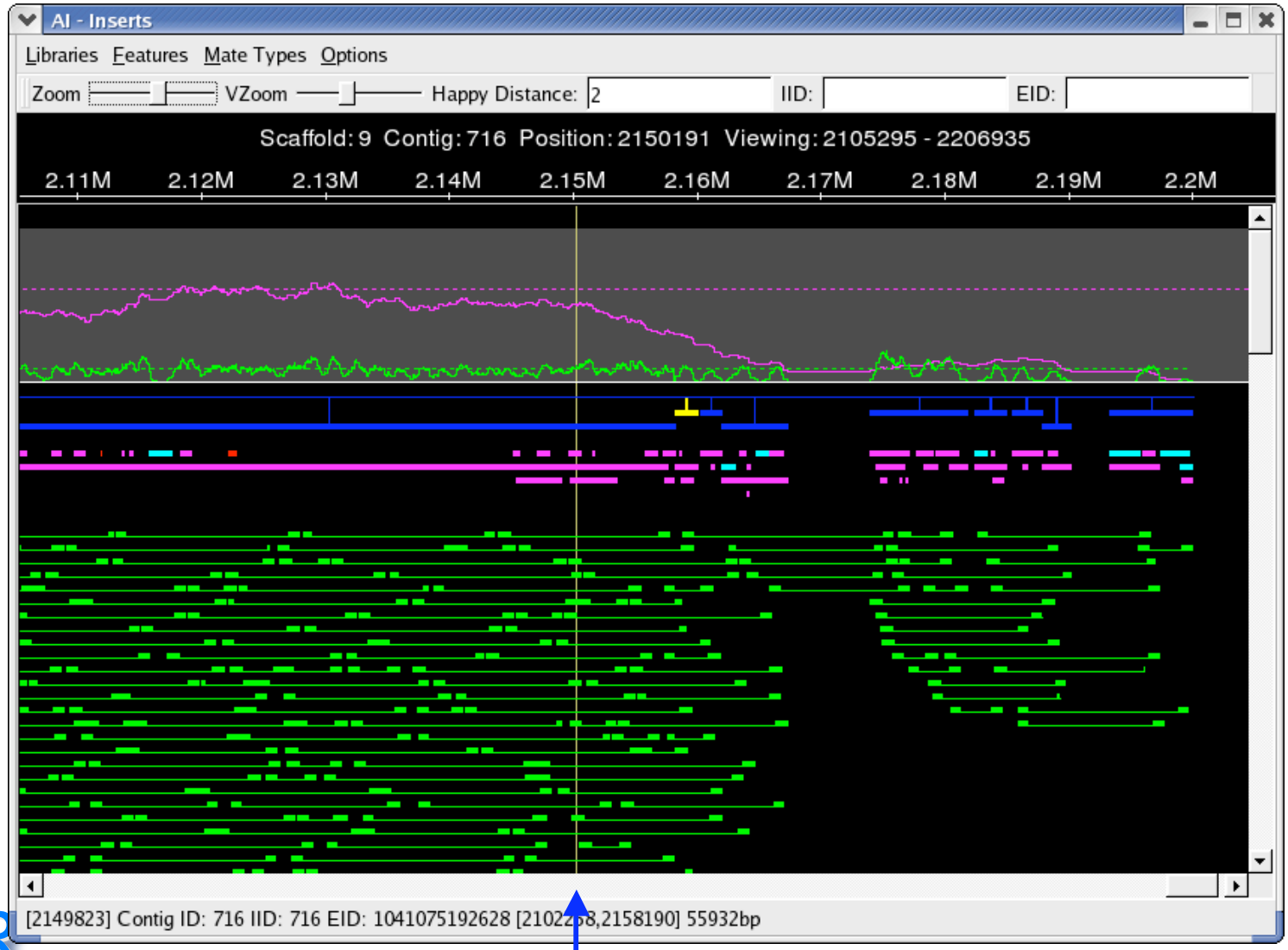
Resulting contig after repeat resolution



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

AMOS Assembly Investigator



Toolbar

Position

Insert and Read Coverage

Scaffold

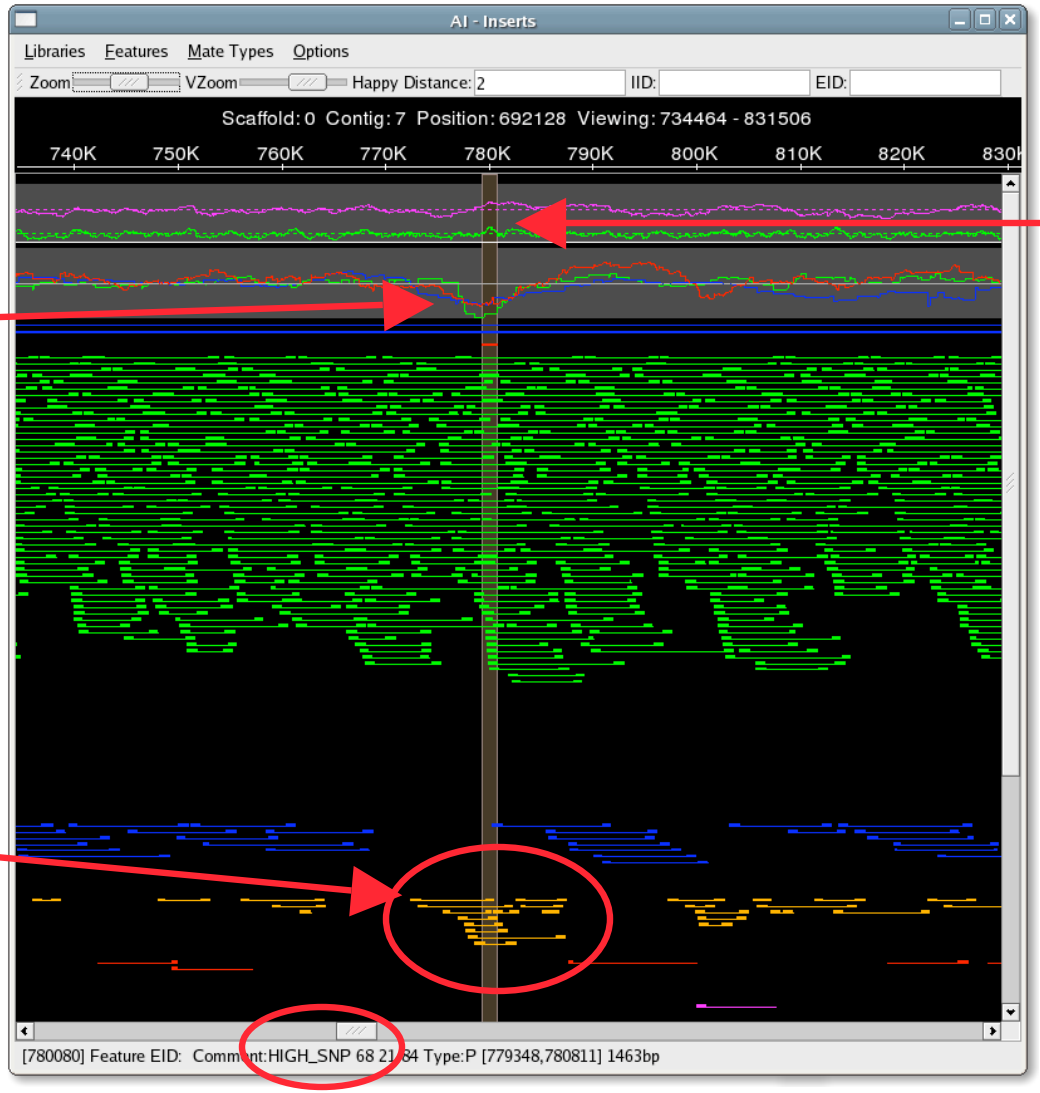
Features

Inserts

Current Contig Position

TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

Collapsed Repeat



Read Coverage Spike

-5.5 CE Dip

Individual Compressed Mates

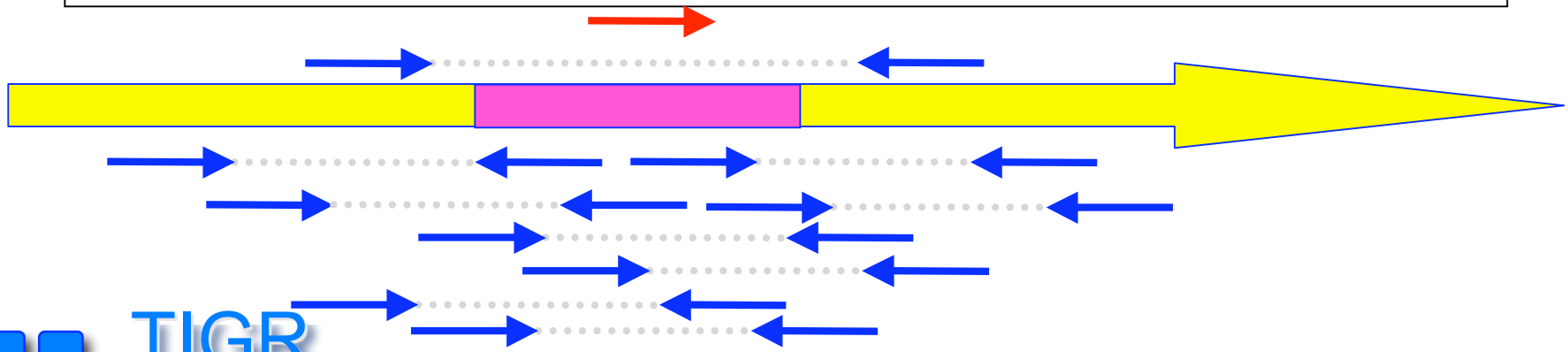
68 Correlated SNPs

TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

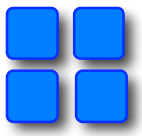
Resolved Repeat

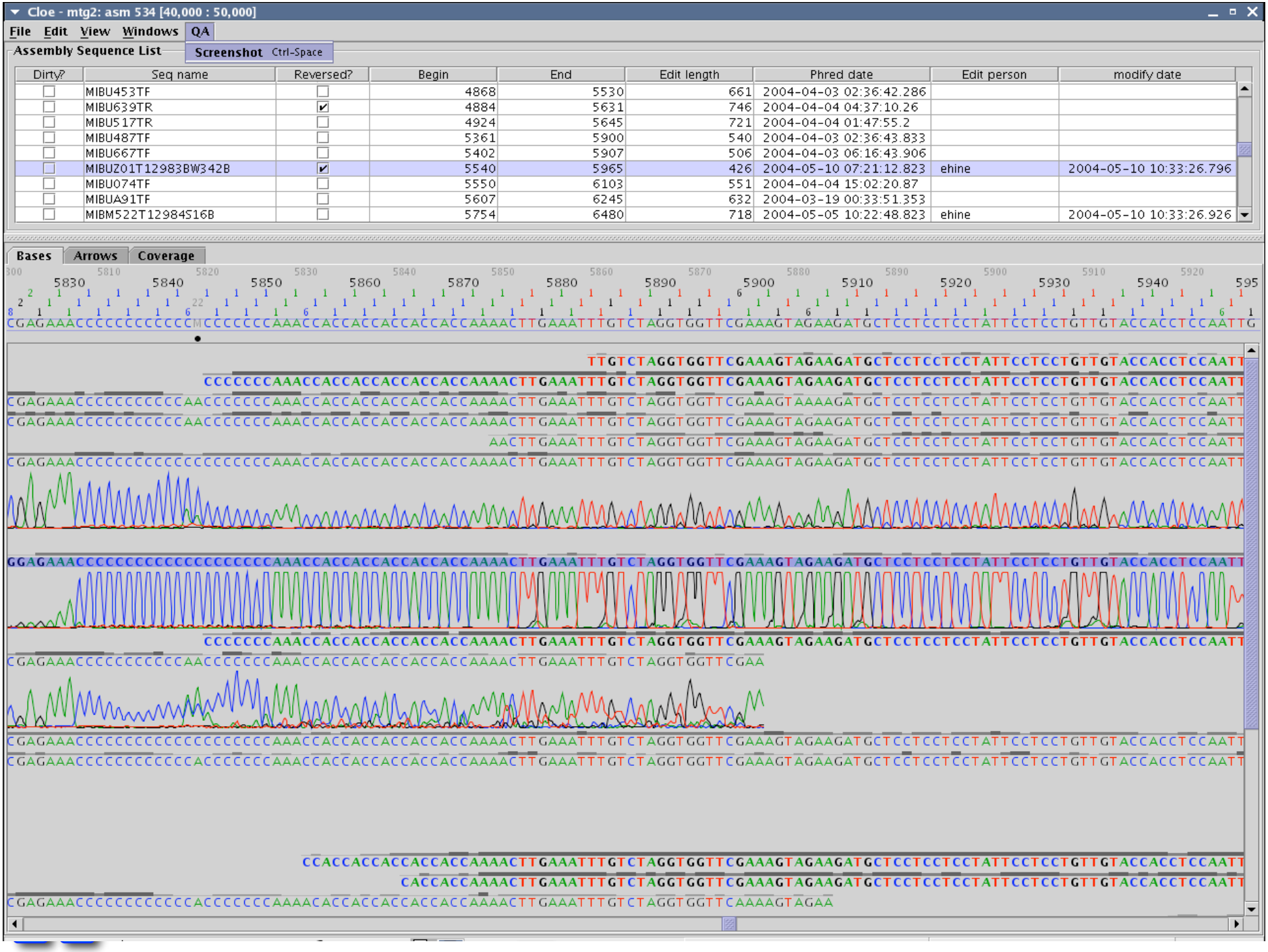
- Unique flank order is correct
  - Use linking information across the repeat (large insert clones or PCR)
- Consensus sequence is correct
  - Use linked clones that have one mate in the repeat and the other anchored in unique sequence
  - Transposon mediated libraries



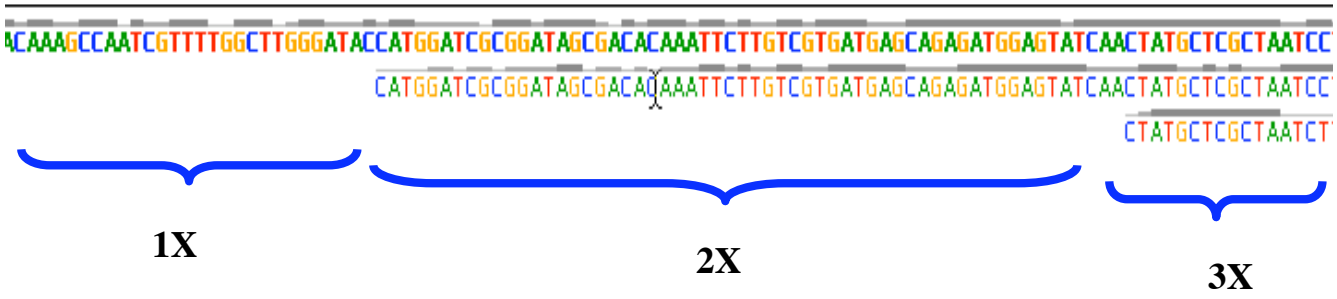
Sequence Validation: Editing

- Finishers look at all areas where the consensus sequence is below a certain quality threshold.
- Electropherograms are used to determine which base calls may need to be changed.
- Only supporting sequences are edited, not the consensus.
- Editing provides an integrity check, may aid in assembly, identifying repeat areas and other unique problems.




Cloe - mtg2: asm 534 [40,000 : 50,000]
File Edit View Windows QA
Screenshot Ctrl-Space
Assembly Sequence List
Dirty? | Seq name | Reversed? | Begin | End | Edit length | Phred date | Edit person | modify date
MIBU453TF | 4868 | 5530 | 661 | 2004-04-03 02:36:42.286
MIBU639TR | 4884 | 5631 | 746 | 2004-04-04 04:37:10.26
MIBU517TR | 4924 | 5645 | 721 | 2004-04-04 01:47:55.2
MIBU487TF | 5361 | 5900 | 540 | 2004-04-03 02:36:43.833
MIBU667TF | 5402 | 5907 | 506 | 2004-04-03 06:16:43.906
MIBUZ01T12983BW342B | 5540 | 5965 | 426 | 2004-05-10 07:21:12.823 | ehine | 2004-05-10 10:33:26.796
MIBU074TF | 5550 | 6103 | 551 | 2004-04-04 15:02:20.87
MIBUA91TF | 5607 | 6245 | 632 | 2004-03-19 00:33:51.353
MIBM522T12984S16B | 5754 | 6480 | 718 | 2004-05-05 10:22:48.823 | ehine | 2004-05-10 10:33:26.926
Bases Arrows Coverage

Sequence Validation:
Sequence coverage



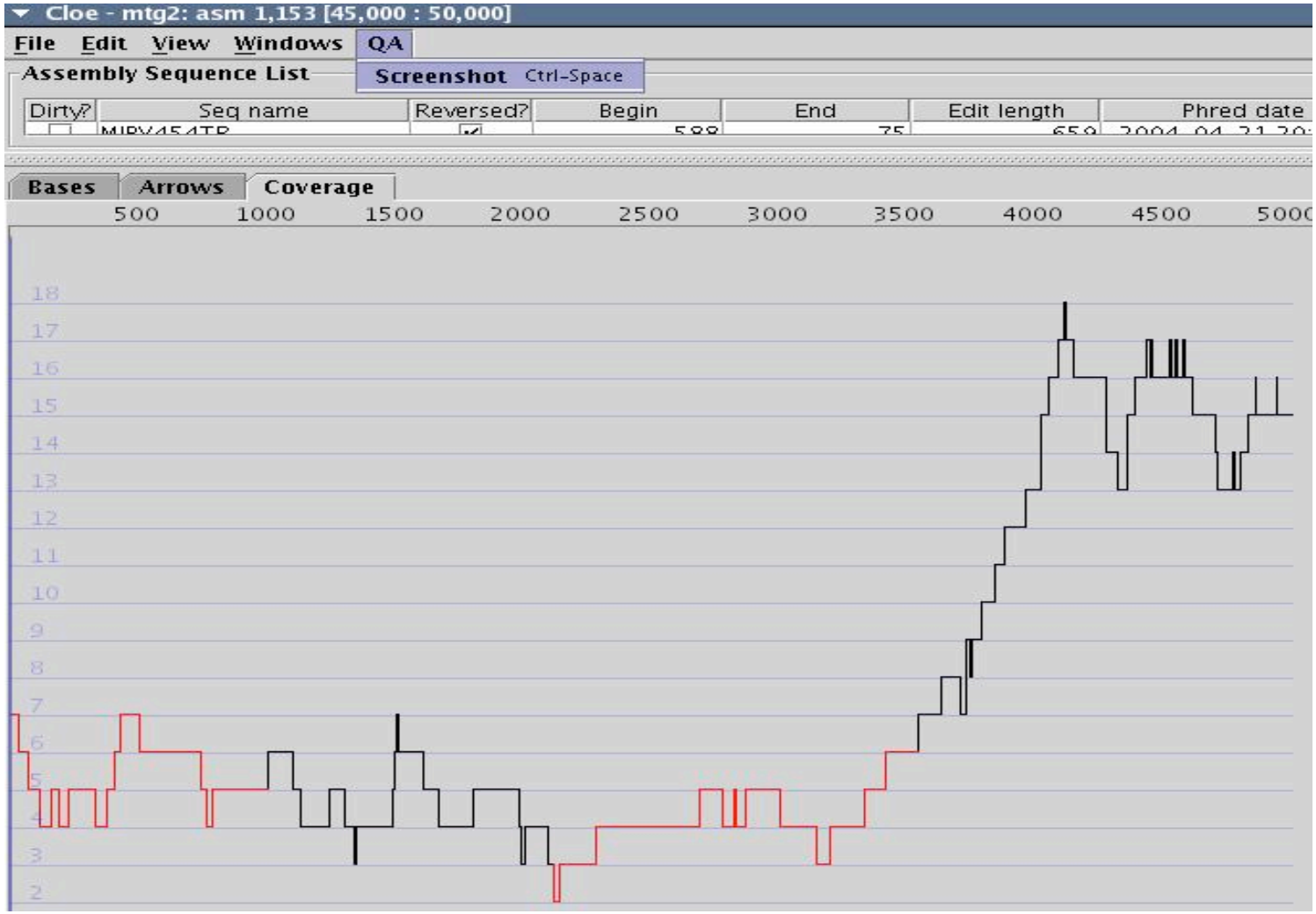
Sequence coverage rule:

Every base in an assembly must be covered by at least two sequences of high quality.

Why?

Validating sequence coverage provides a high degree of confidence in the consensus base calls.

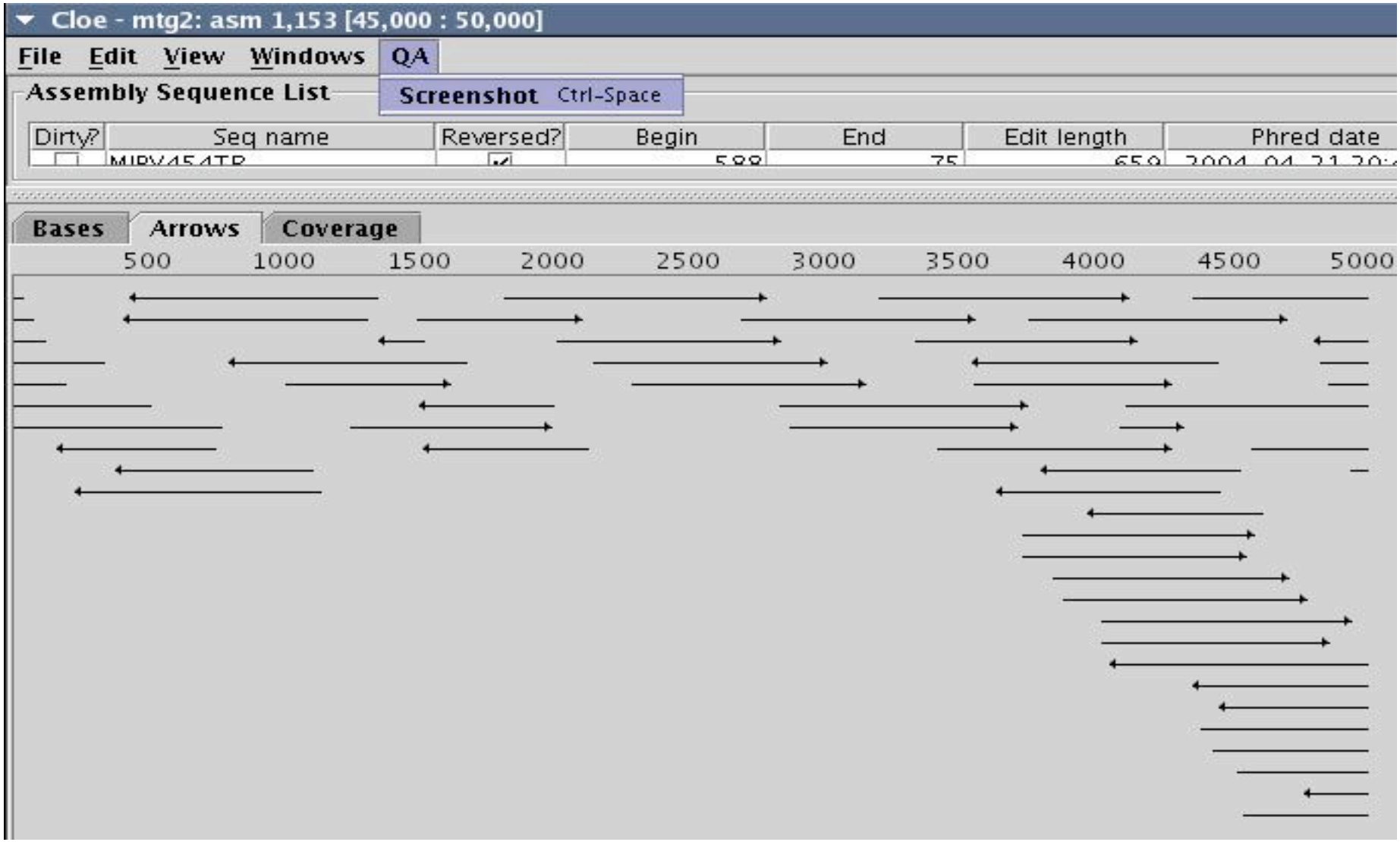
Sequence Coverage in Cloe



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Sequence Coverage in Cloe



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Assembly Verification: Clone Coverage

Clone Coverage Rule:

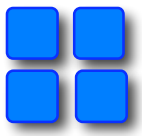
Every base must be spanned by two valid subclones.

A valid subclone is a pair of forward/reverse mates that are:

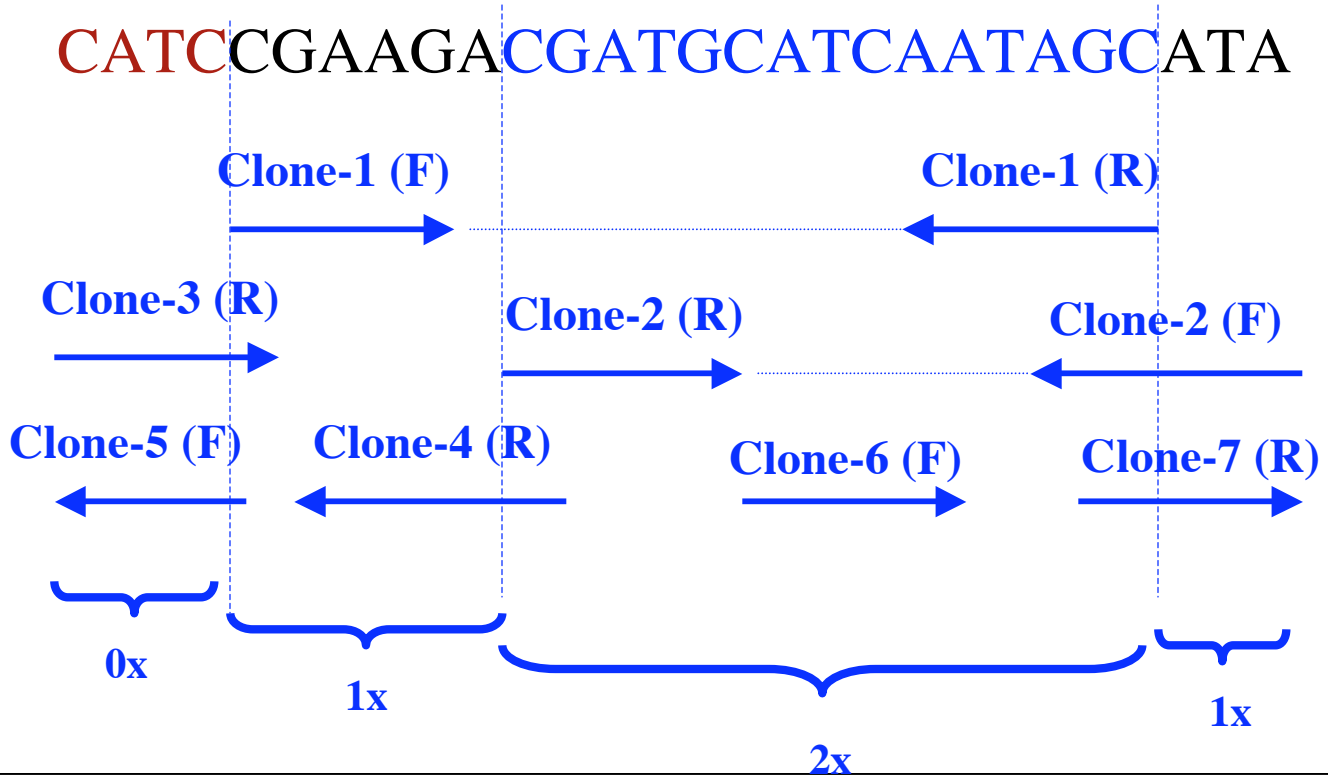
- Oriented correctly (pointing toward each other)
- Separated by an approximate length of sequence determined by the expected template size for the library.

Why?

Validating clone coverage provides a high degree of confidence that the contig was assembled correctly.

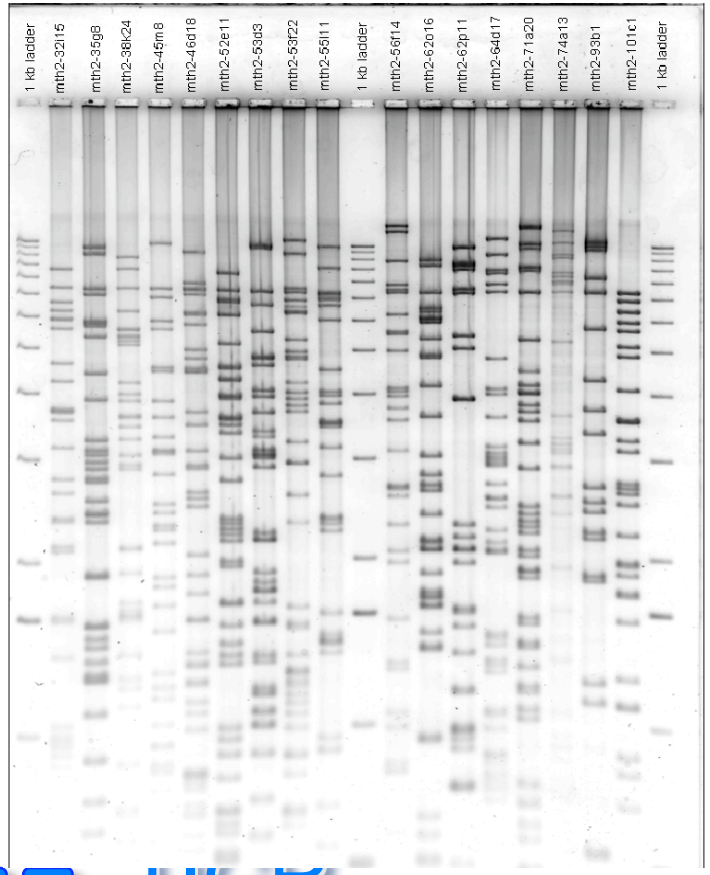


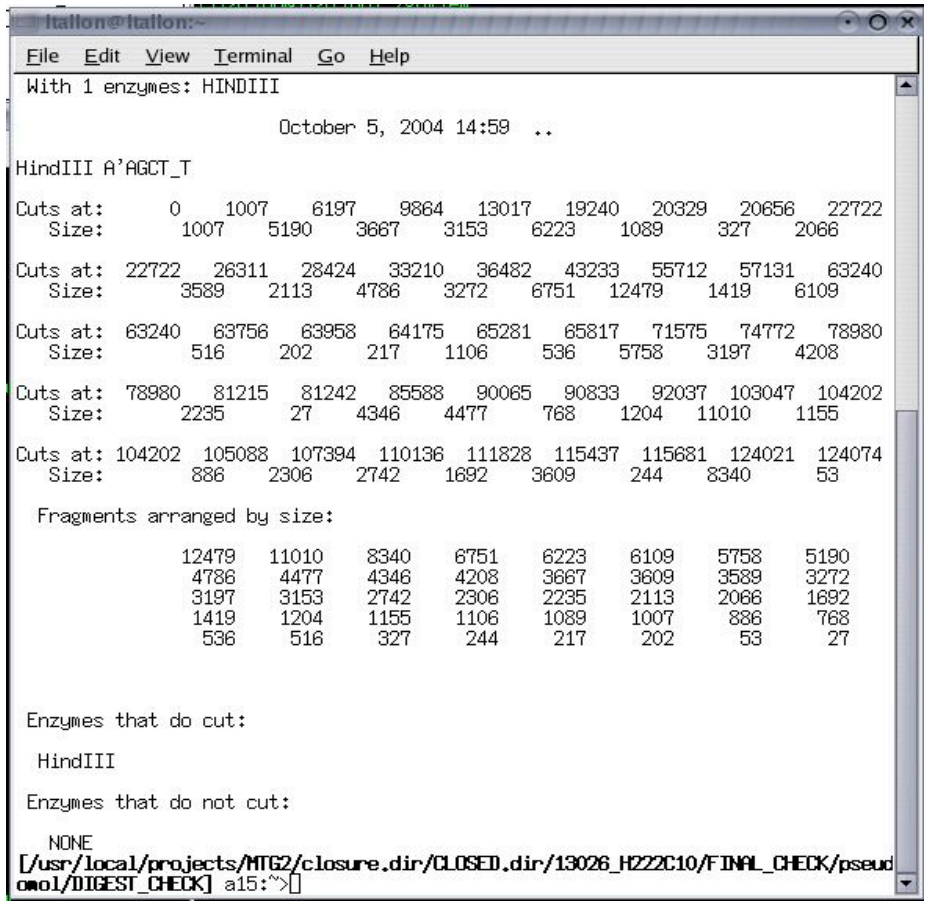
Clone Coverage Example



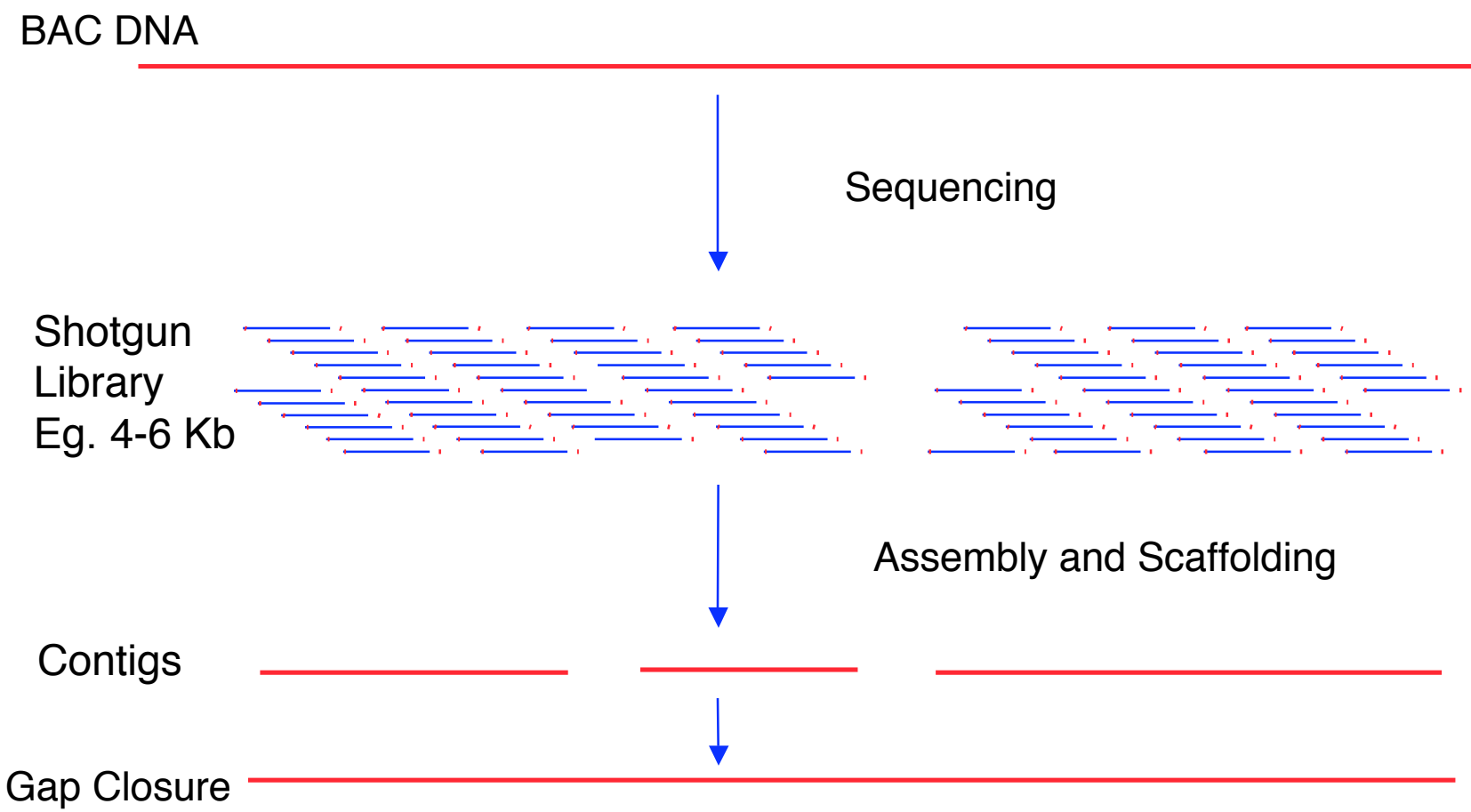
- Red: 0 x clone coverage
- Black: 1x clone coverage
- Blue: 2x clone coverage

Verify Closure by Fingerprint Comparison



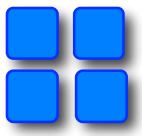


Finished BAC Ready for Annotation!



TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

EXTRA SLIDES START HERE…



TIGR

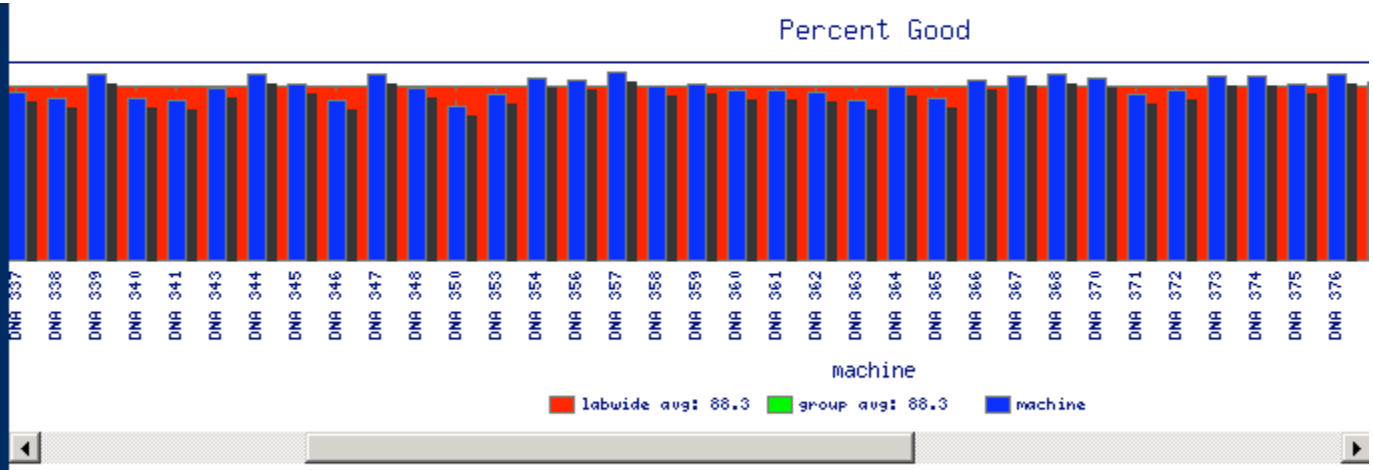
THE INSTITUTE FOR GENOMIC RESEARCH

Graph/top navigation

machine
catalyst
db
gel_run_pn
rxn_run_pn
rxn_lot
primer_lot

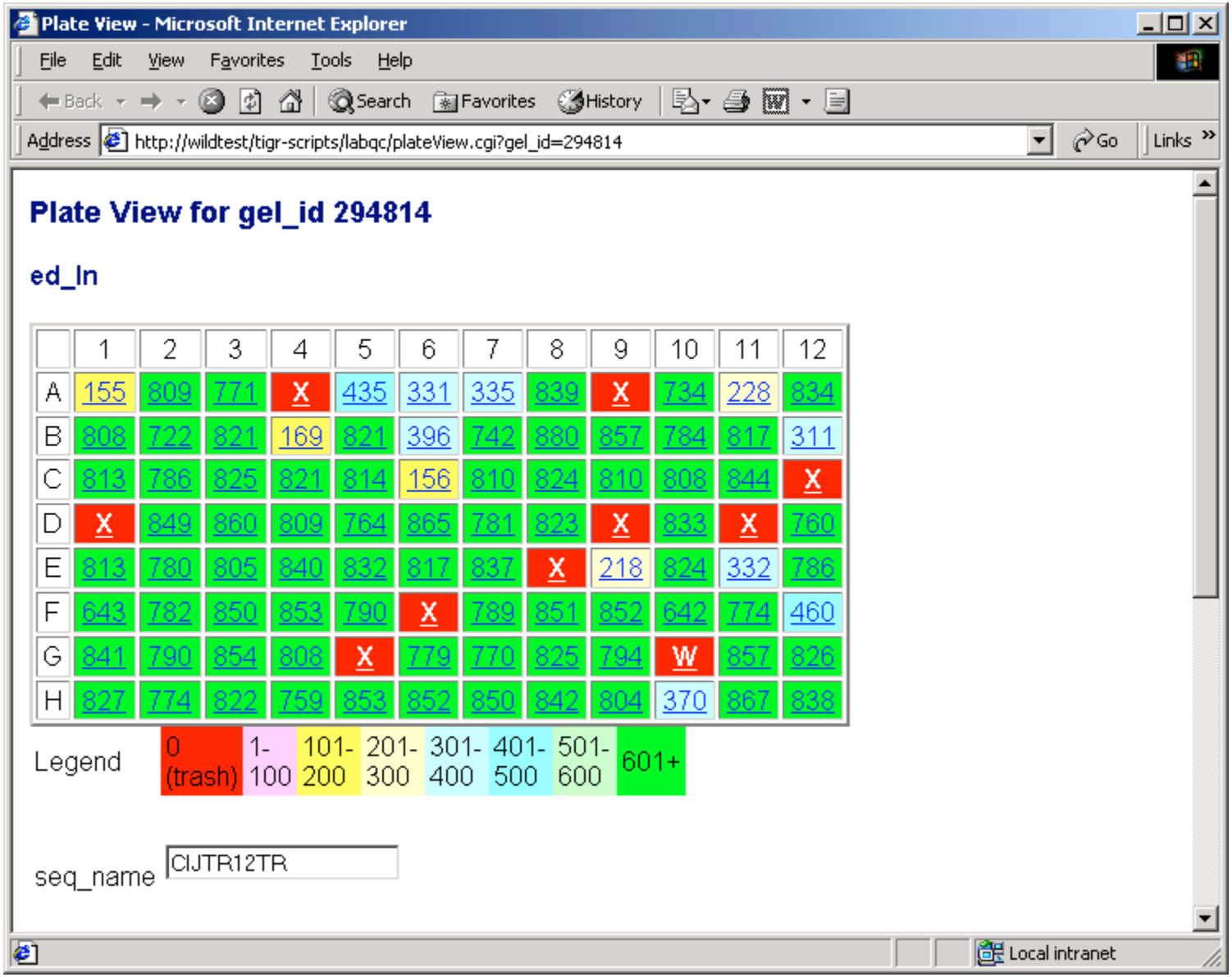
Data navigation

machine
catalyst
db
gel_run_pn
rxn_run_pn
rxn_lot
primer_lot
bad gels
group gels



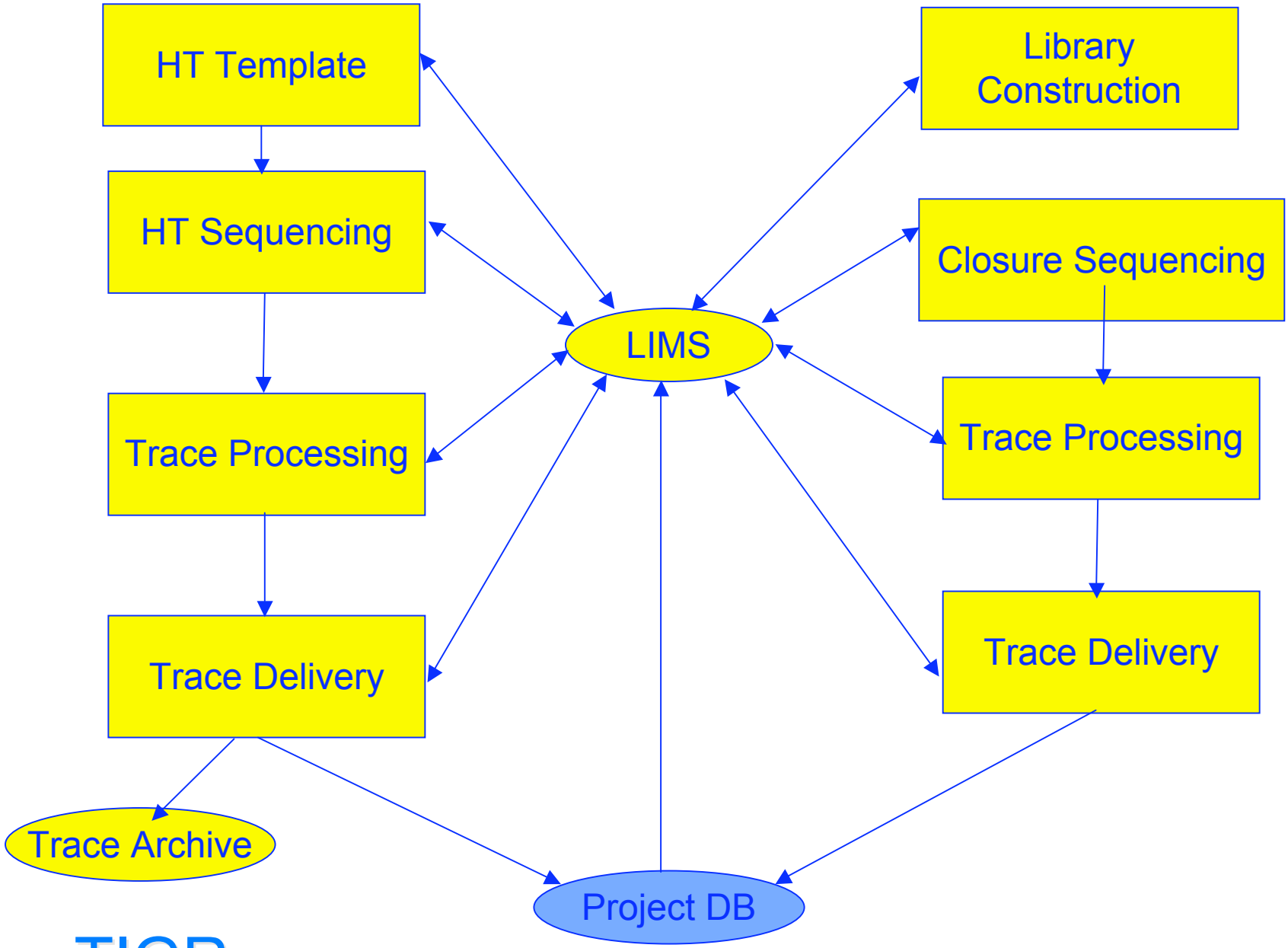
data for Jul 12 2003 to Jul 12 2003

db	Total Good	Total Lanes	% Good	VEIR	% VEIR	Average Edit Length	Average Quality
gssea	12375	14782	83.7 -4.5	698	4.7	827 -3	40.2 +0.6
mag	403	479	84.1 -4.1	30	6.3	853 +23	39.4 -0.2
pcg	572	672	85.1 -3.1	20	3.0	850 +20	38.7 -0.9
smag	2877	3157	91.1 +2.9	4	0.1	846 +16	40.1 +0.5
tbg	1112	1248	89.1 +0.9	68	5.4	874 +44	39.9 +0.3
tcrg	7021	7488	93.8 +5.6	28	0.4	829 -1	39.0 -0.6
tseest	2986	3552	84.1 -4.1	90	2.5	781 -49	35.7 -3.9
ttg	1905	2112	90.2 +2.0	2	0.1	839 +9	40.4 +0.8
tvg	2342	2592	90.4 +2.2	2	0.1	833 +3	40.5 +0.9
zmg	6602	7200	91.7 +3.5	203	2.8	829 -1	39.9 +0.3
Totals	53409 (avg: 2967)	60551 (avg: 3364)	88.2	1467	2.4	830	39.6



TIGR

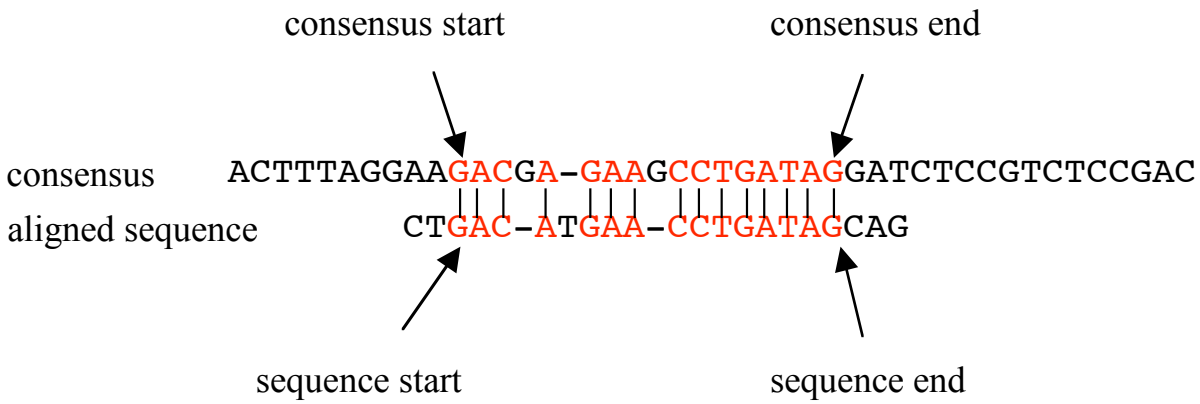
THE INSTITUTE FOR GENOMIC RESEARCH



TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Representing assemblies

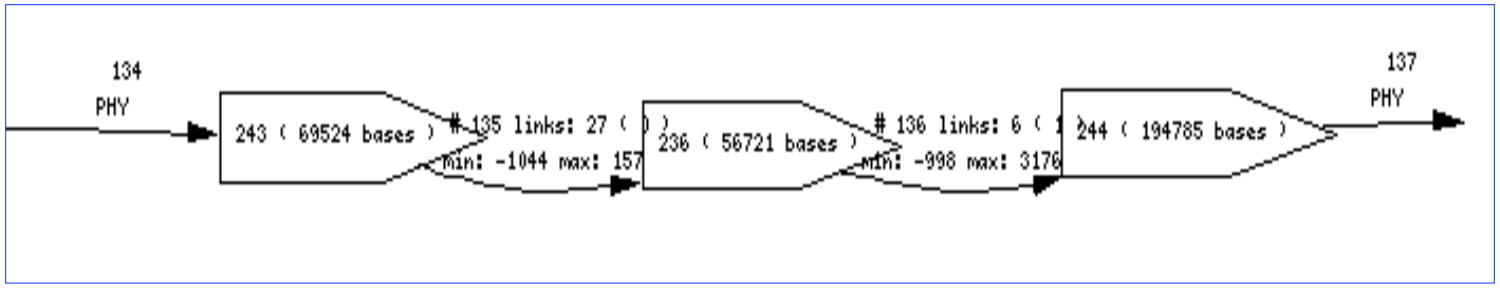


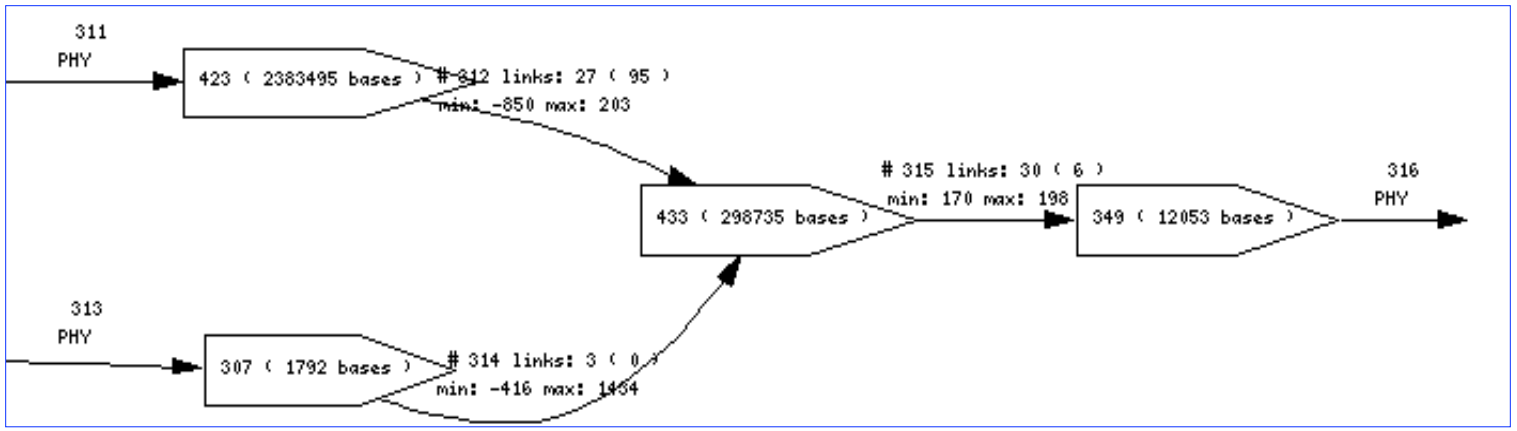
Assembly table:
consensus w/ and w/o gaps

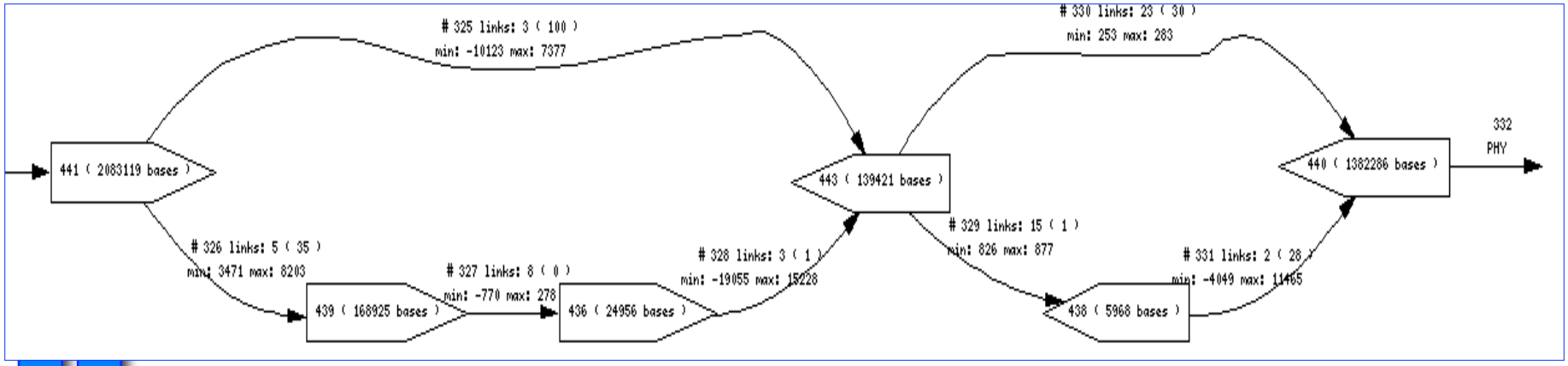
Asmbl_link table:
asmbl_id
seq_name
seq_lend, seq_rend
asm_lend, asm_rend
lsequence

Sequence table:
sequence
end5, end3

Schematic View of 3 Typical Scaffolds

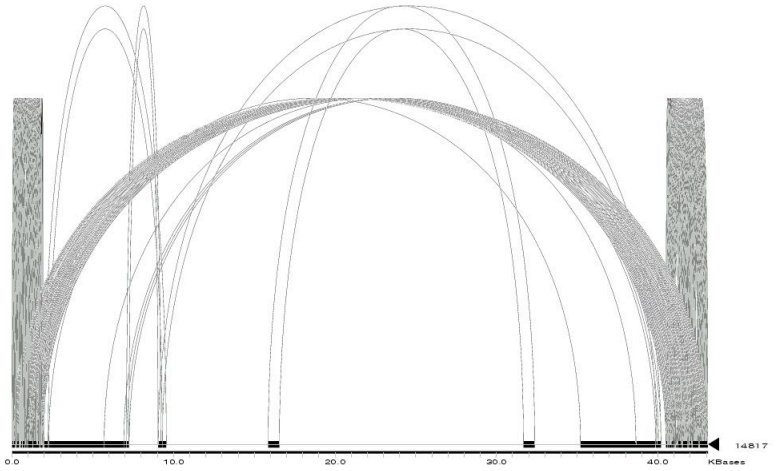






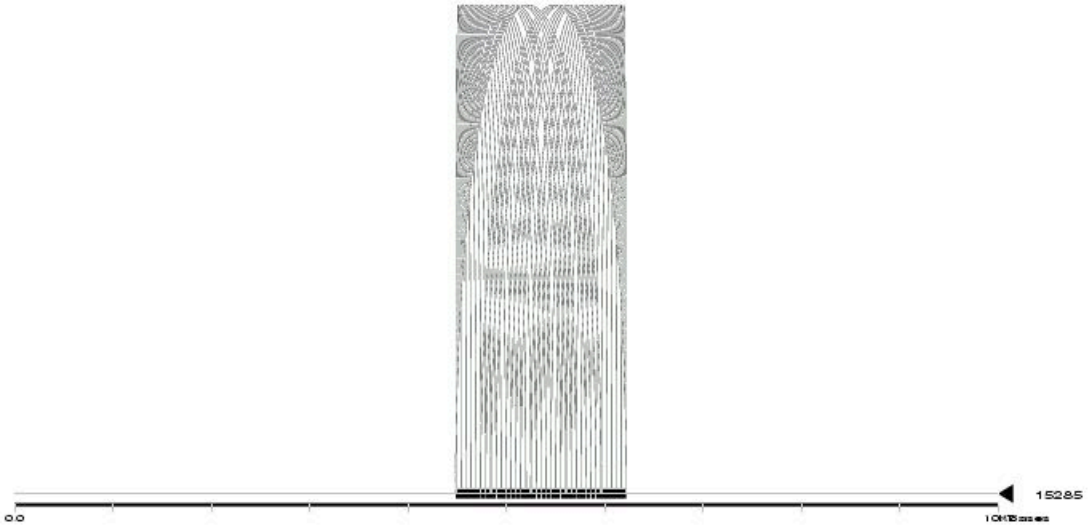
Types of repeats

- Short Tandem Repeats
- Large Multi Unit Tandem Repeats
- Multi Class Repeats
- Inverted Repeats



First file = asm.fasta Threshold Score = 30 Percent Match = 90

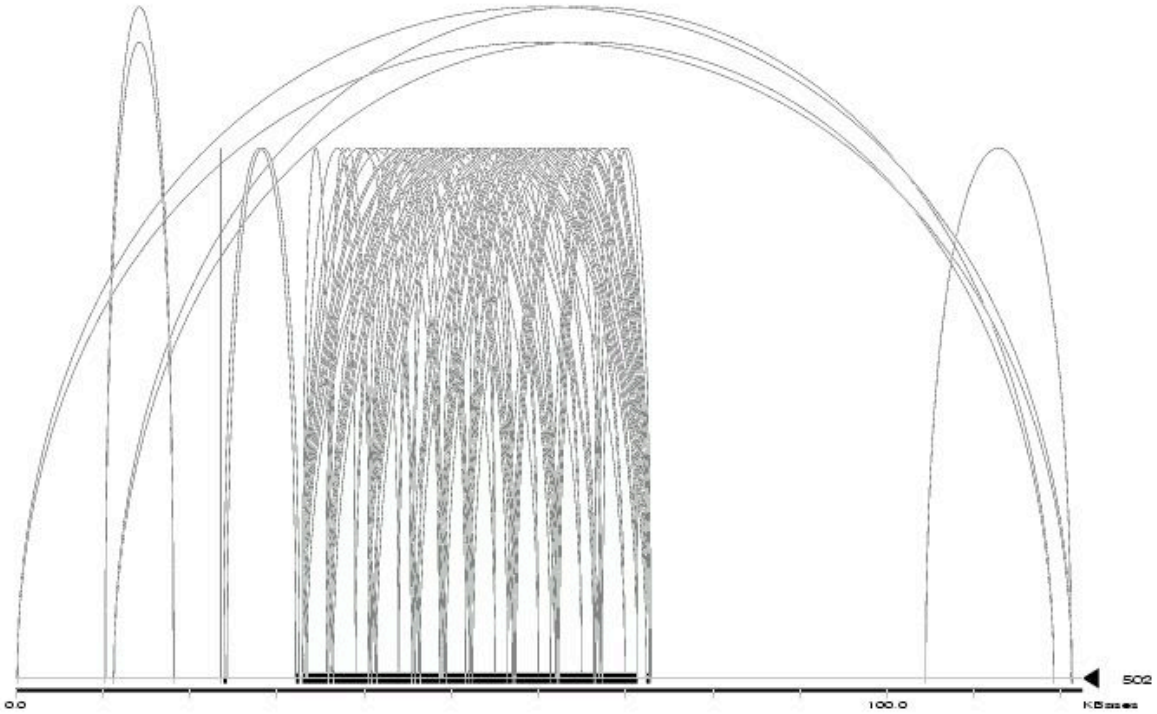
Short Tandem Repeats



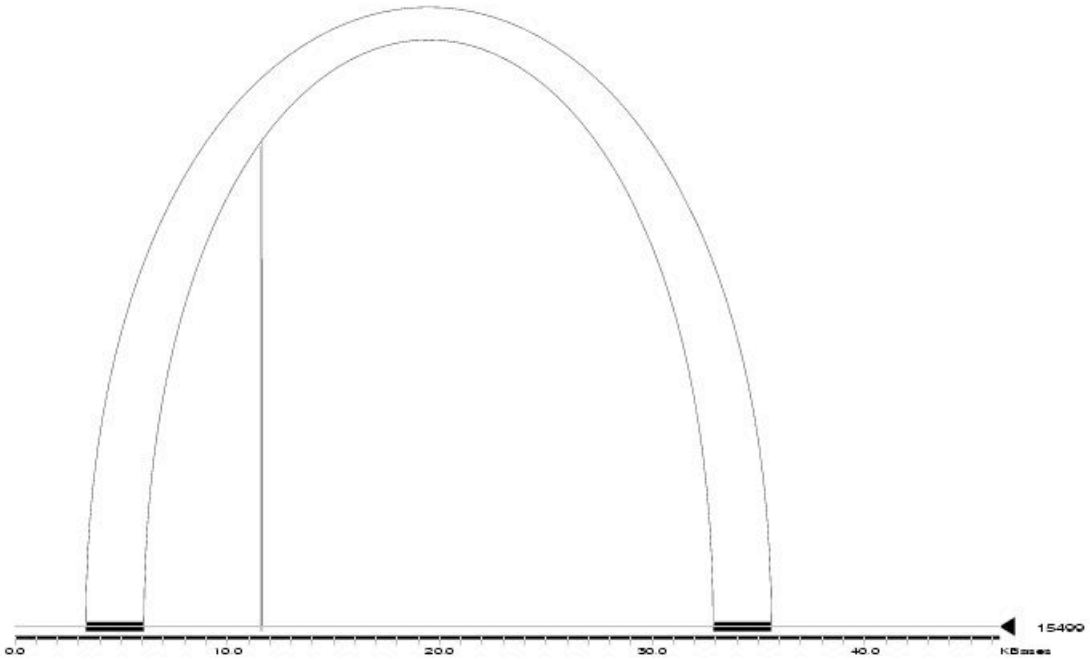
TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Large Multi Unit Tandem Repeats



Inverted Repeats

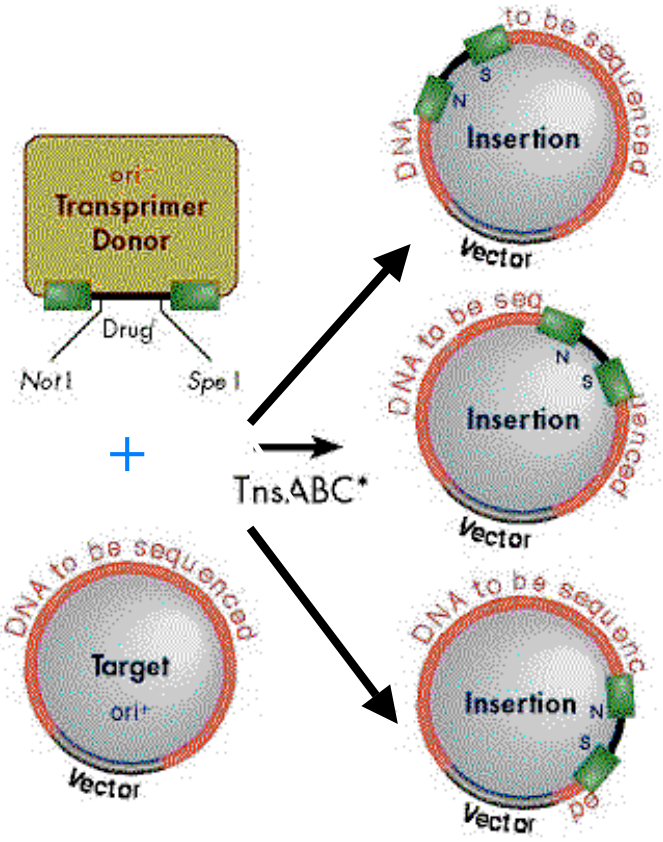


TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

Construction of Transposon-Mediated Libraries

- Transposons are genetic elements that provide mobile primer sites within the target DNA for sequencing
- The GPS-1 Genome Priming System (New England Biolabs E7100S) enables transposon insertions by using TnsABC* transposase
- Only one random insertion occurs per target DNA

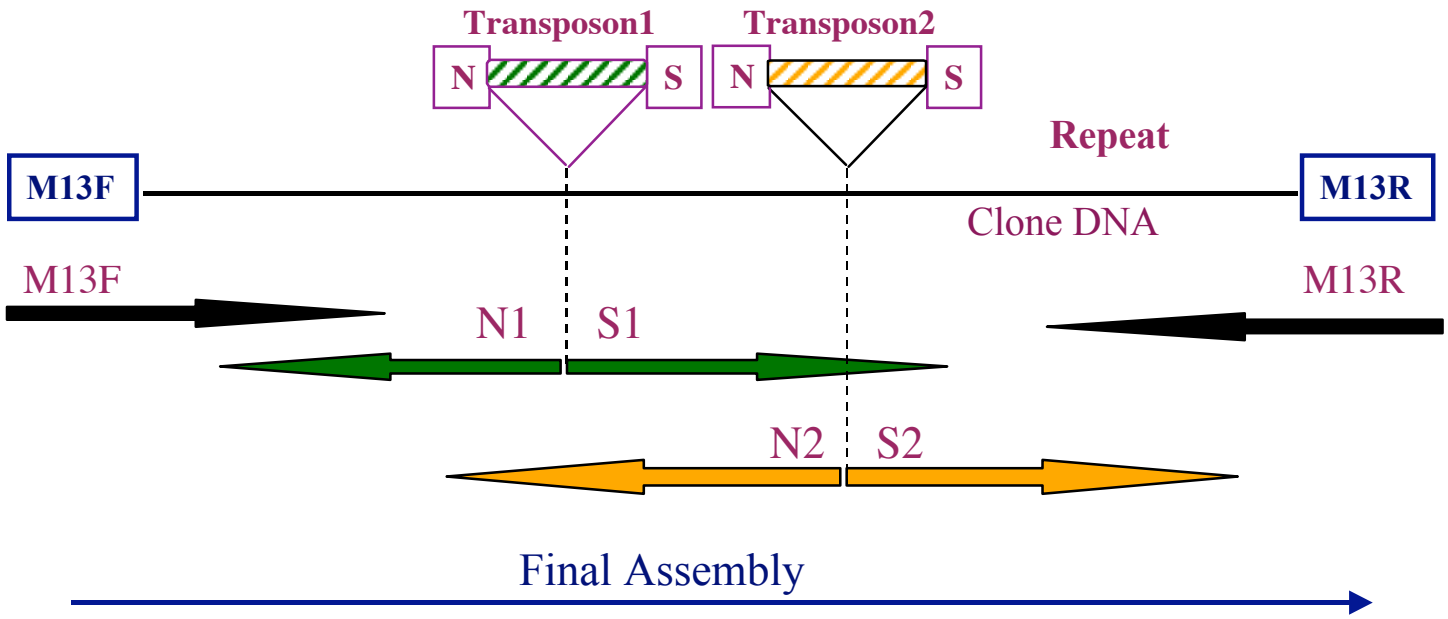


TIGR
THE INSTITUTE FOR GENOMIC RESEARCH

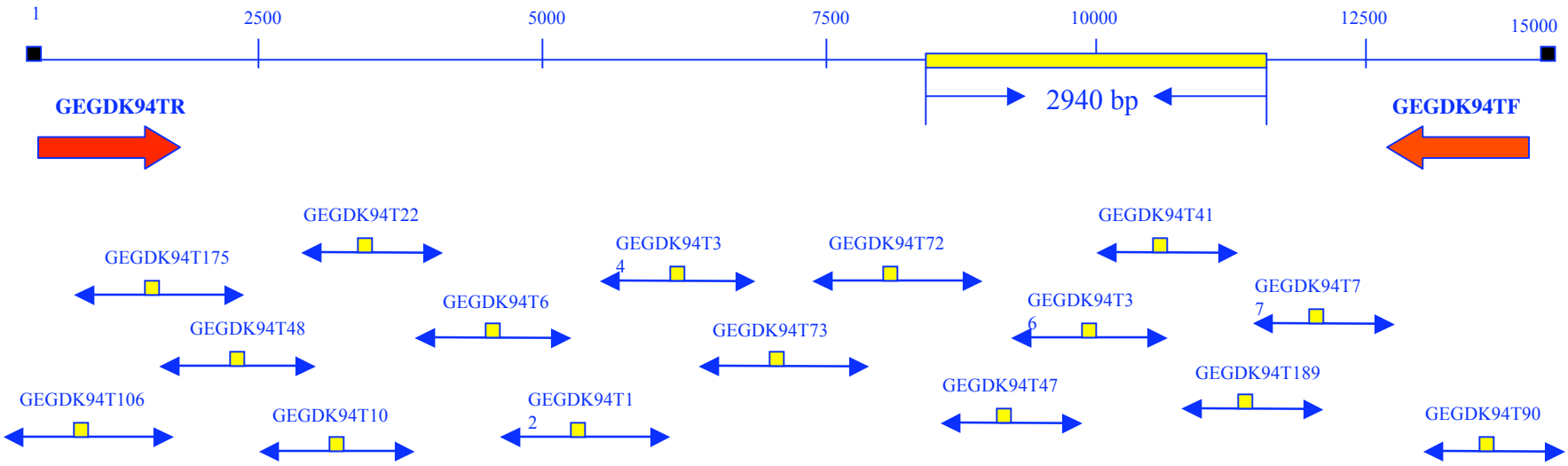
Use of Transposon-mediated libraries:

- Make a library using a transposon kit on a spanning clone.
- Sequence the clones using the transposon primers
- Assemble the sequences

Physical map of the transposon inserts:



Assembly of Transposon-Mediated Clone Walks



forward and reverse clone mates

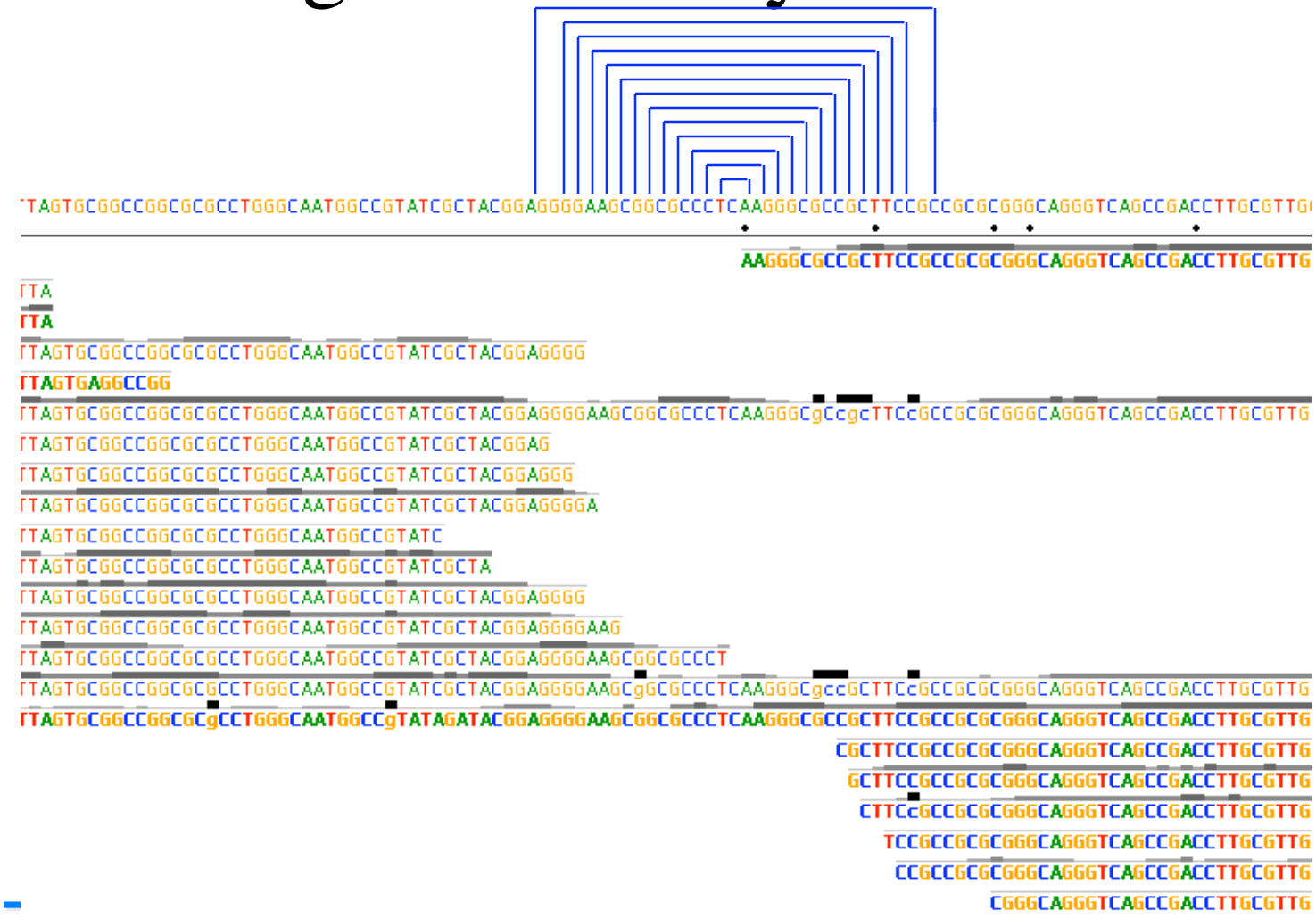
tandem repeat region

transposon insertion site

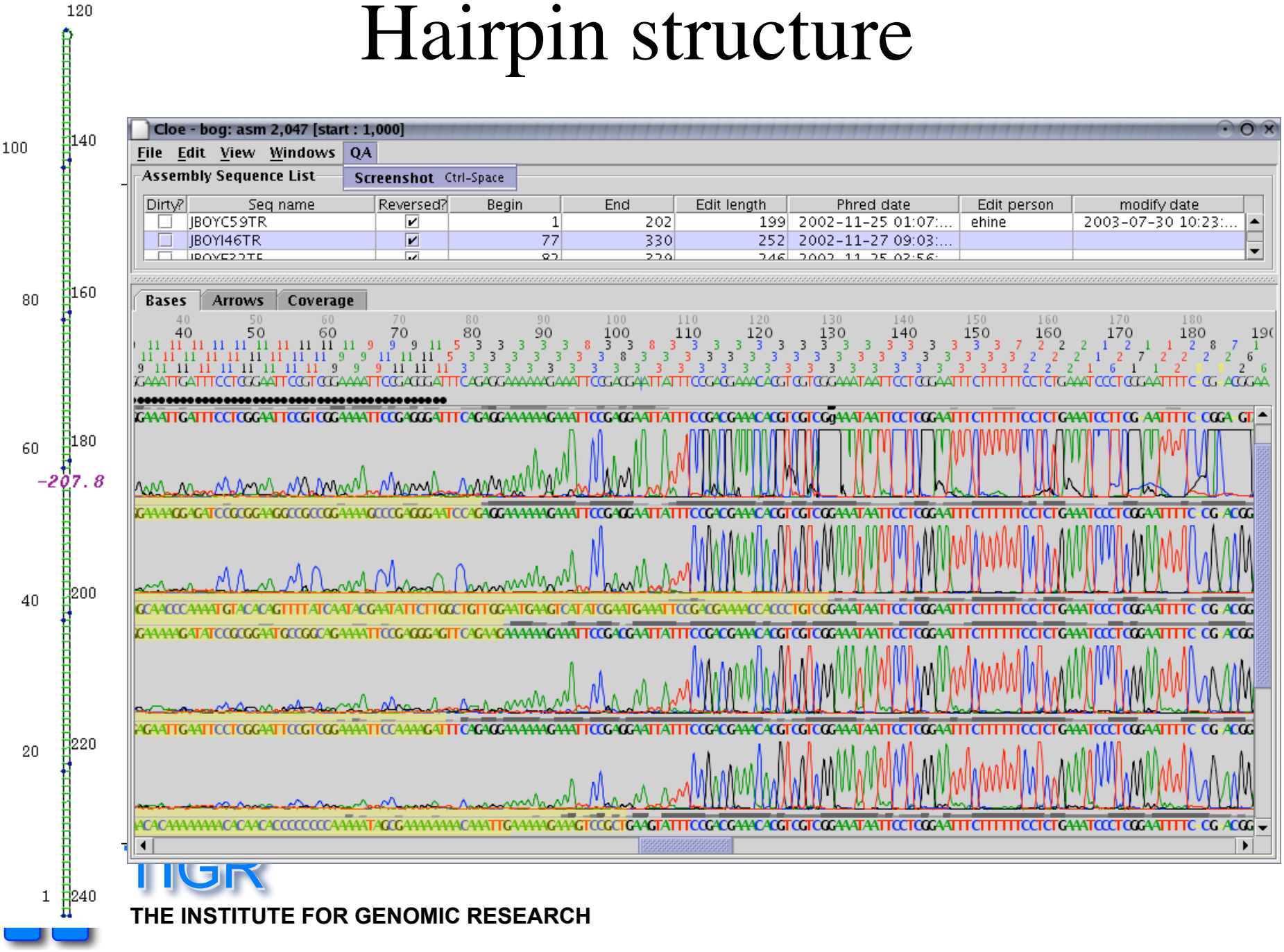
TIGR

THE INSTITUTE FOR GENOMIC RESEARCH

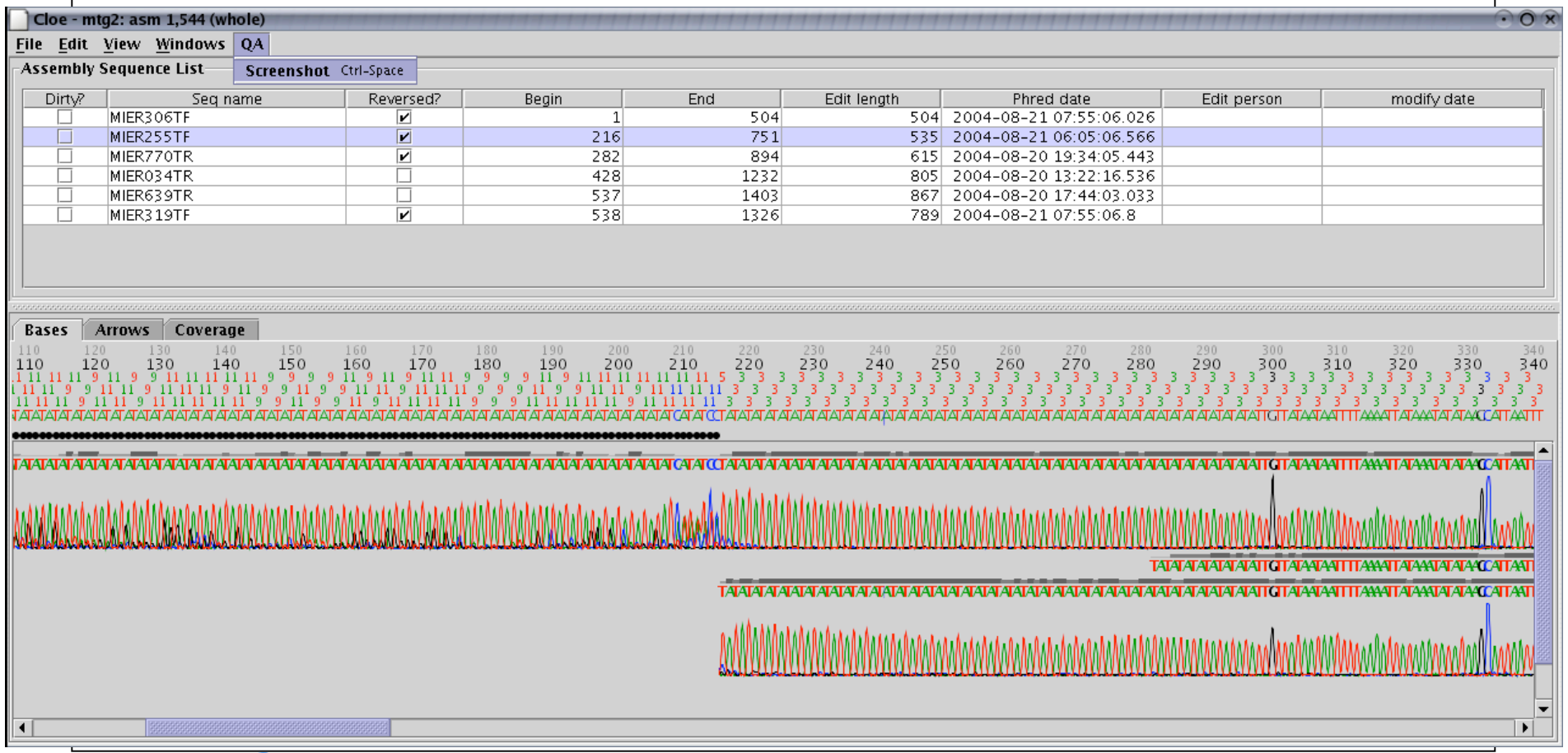
Closure Challenges: Sequencing Through Secondary Structures

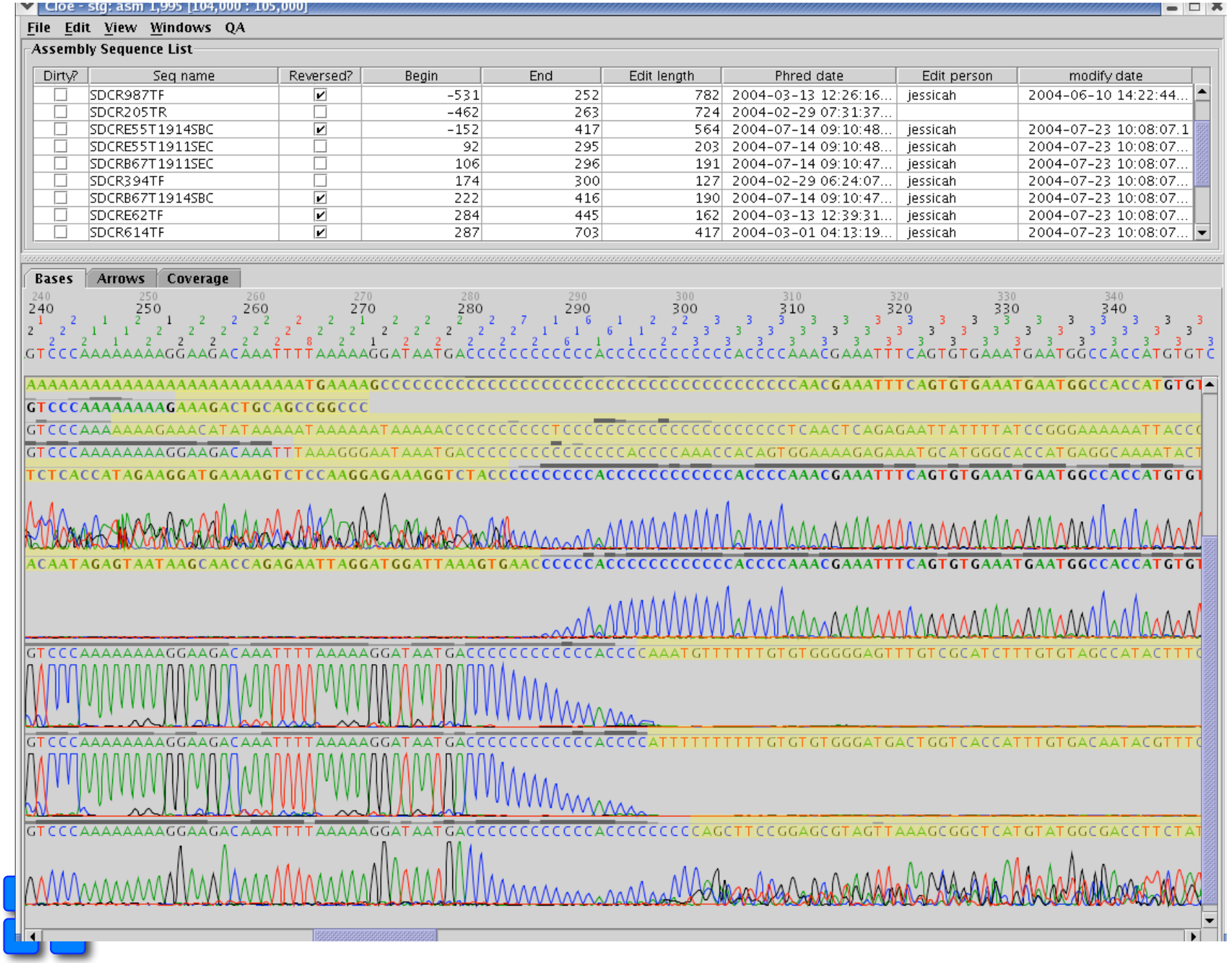


Hairpin structure



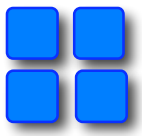
Homopolymeric tracts




Cloe - Stg: asm 1,995 [104,000 : 105,000]
File Edit View Windows QA
Assembly Sequence List
Dirty? Seq name Reversed? Begin End Edit length Phred date Edit person modify date
SDCR987TF -531 252 782 2004-03-13 12:26:16... jessicah 2004-06-10 14:22:44...
SDCR205TR -462 263 724 2004-02-29 07:31:37...
SDCRE55T1914SBC -152 417 564 2004-07-14 09:10:48... jessicah 2004-07-23 10:08:07.1
SDCRE55T1911SEC 92 295 203 2004-07-14 09:10:48... jessicah 2004-07-23 10:08:07...
SDCRB67T1911SEC 106 296 191 2004-07-14 09:10:47... jessicah 2004-07-23 10:08:07...
SDCR394TF 174 300 127 2004-02-29 06:24:07... jessicah 2004-07-23 10:08:07...
SDCRB67T1914SBC 222 416 190 2004-07-14 09:10:47... jessicah 2004-07-23 10:08:07...
SDCRE62TF 284 445 162 2004-03-13 12:39:31... jessicah 2004-07-23 10:08:07...
SDCR614TF 287 703 417 2004-03-01 04:13:19... jessicah 2004-07-23 10:08:07...
Bases Arrows Coverage

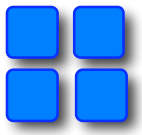
Solutions

- Apply different sequencing chemistries
  - Big-Dye terminator (default)
  - Dye-primer
  - dGTP mix (GC rich regions)
- Denature structures - Additives
  - Betaine
  - DMSO
- Break structure
  - Restriction digest
  - Transposon insertion
  - Micro-libraries



Automated Finishing at TIGR

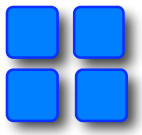
- Closure is a **feature driven** activity
- Feature Classification System
  - Identify/Classify Features
    - SEQ gaps
    - PHY gaps / scaffold ends
    - Low coverage
    - Repeat features
- Feature Reaction Design Strategy
  - Bin by Reaction Type and Feature Size
    - Primer walks
    - PCR
    - Transposon Bombing



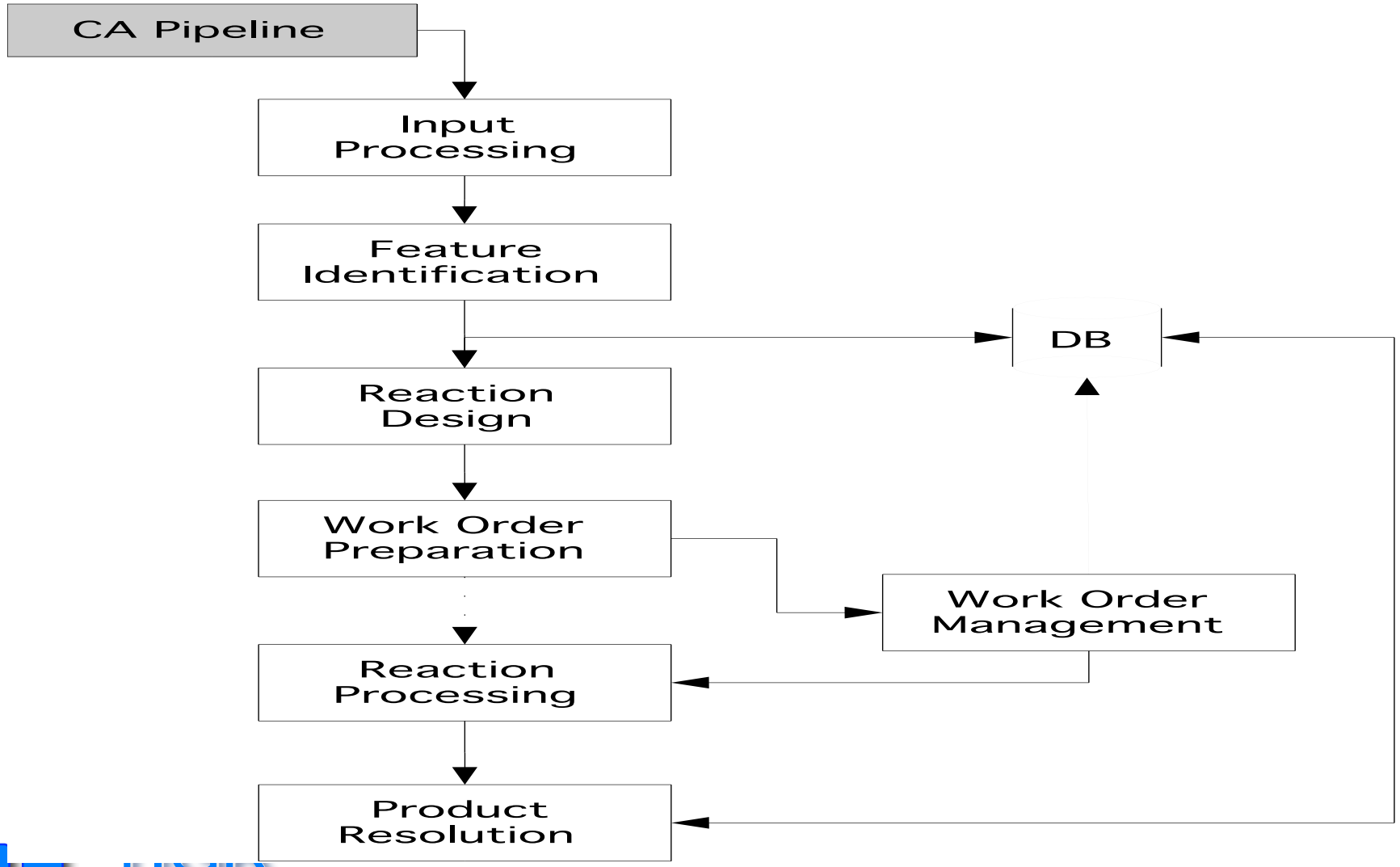
Automated Finishing (cont'd)

- Laboratory Process Integration
  - Order reactions from the lab
  - Group reactions by lab process
  - Integrate changing laboratory automation
  - Monitor results

- Reaction Product Resolution Strategy
  - Resolve reaction products back to features & contigs: conduct targeted assembly.
  - Evaluate resolution criteria for feature type.

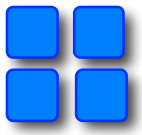


System Architecture



BAC Project Challenges

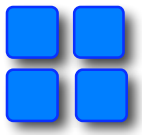
- Dead Ends
- Filling small gaps between adjacent supercontigs
- bp mismatches in BAC sequence overlaps
- Project management/organization



Dead Ends

A Dead End occurs when a BAC at the end of a sequence contig has no BAC-end hits or FPC BACs to extend the contig further

- Design primers and PCR small unique region near end of BAC
- Hybridize to BAC filters to identify candidate BACs
- End sequence and fingerprint candidate BACs to select minimal tiling BAC
- Sequence and Close new tiling BAC



Filling small gaps in a chromosome

- GAP < 15kb:
  - PCR using genomic DNA template
  - Sequence and close PCR product

- GAP > 15kb or unable to PCR:
  - Select gap-spanning BAC
  - Construct large insert library (10+kb inserts)
  - Sequence to low coverage (2-3x)
  - Scaffold mate pairs and select subclone tiling to close gap

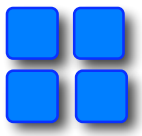
bp Mismatches In BAC Overlaps

Why?

- Basecalling error in sequencing
- Point mutation in BAC
- Misassembly of repeats near BAC end

What to do?

- Examine sequence quality in each BAC
- PCR region using genomic DNA template
- Sequence PCR to verify correct basecall



Project Organization

- Database to store various BAC data:
  - Tiling, overlap info
  - Mapping, FPC info
  - Sequencing and Closure data
  - Collaborator's BAC data
- Web-based BAC status tracking and closure task tracking tools