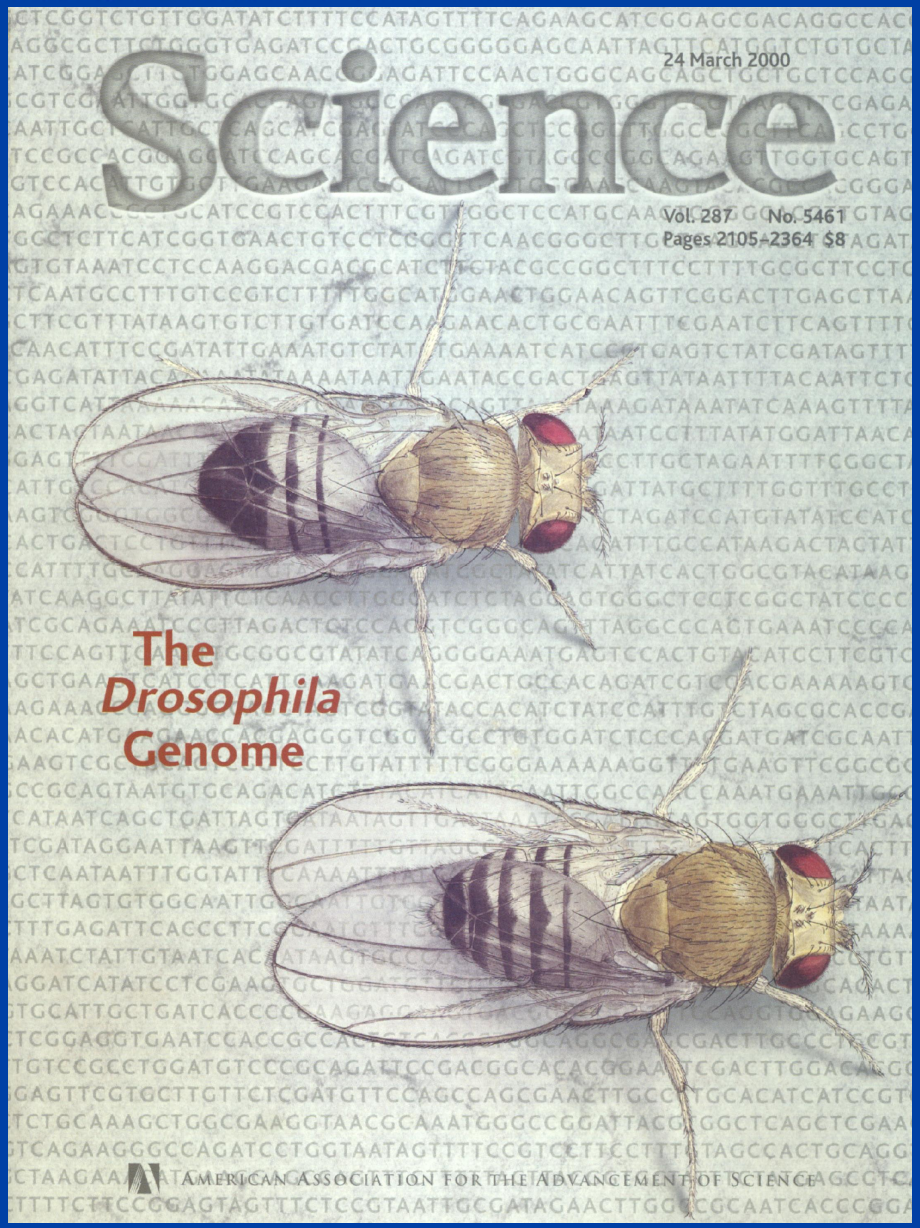
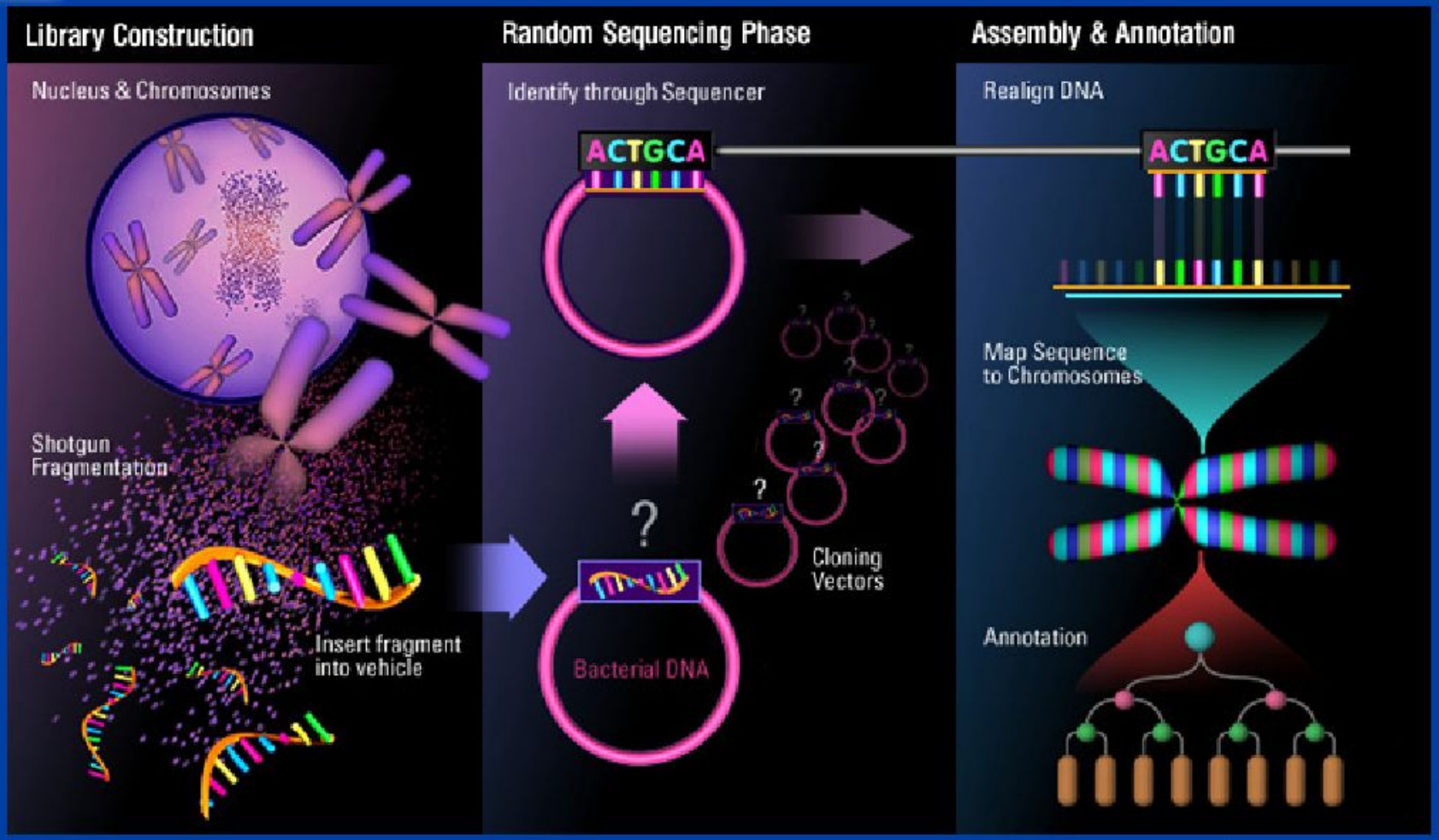


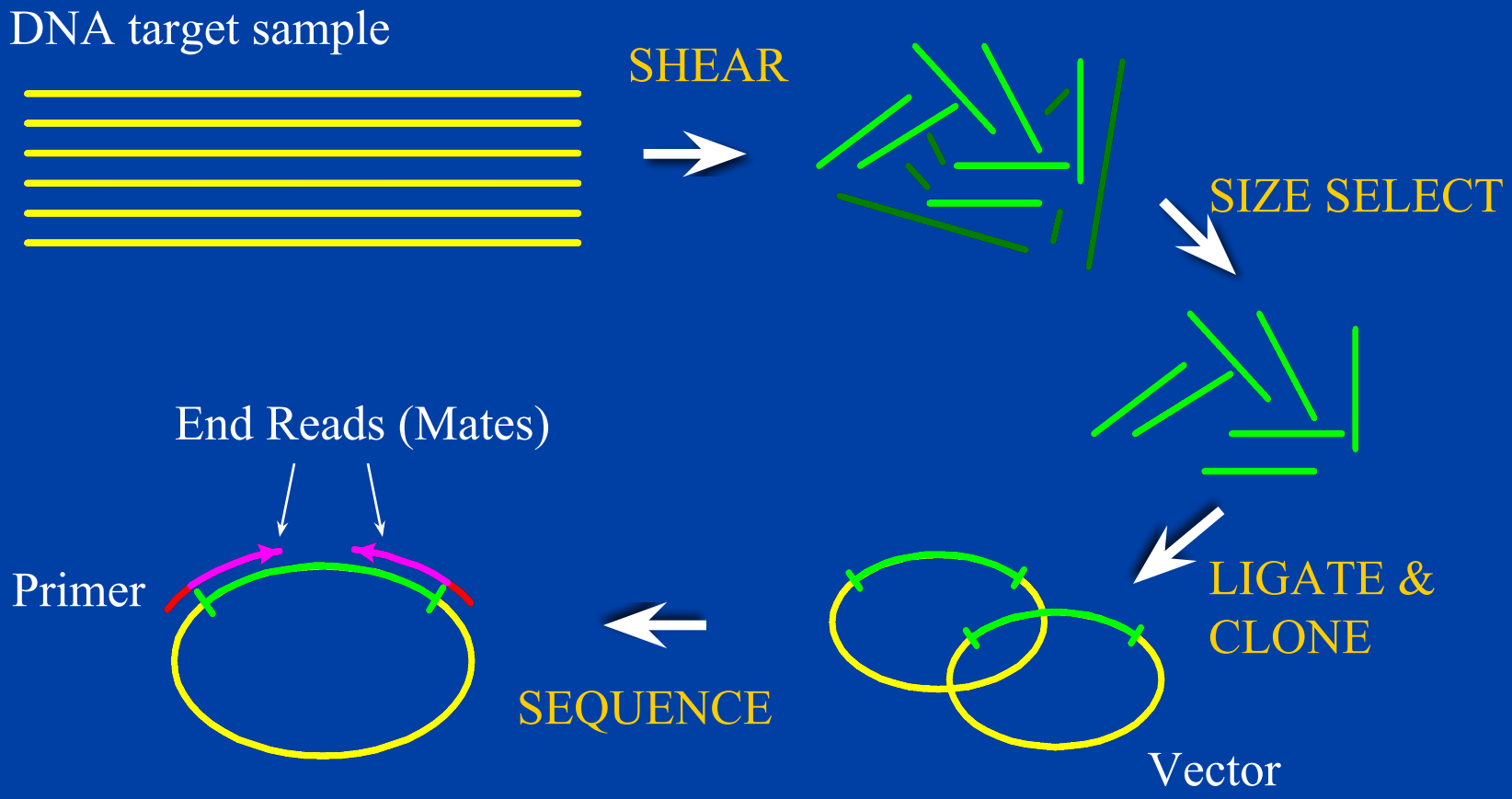
CELERA
A Unit of PE Corporation

Whole Genome Shotgun Sequencing



CELERA
A PE Corporation Business

Shotgun DNA Sequencing (Technology)

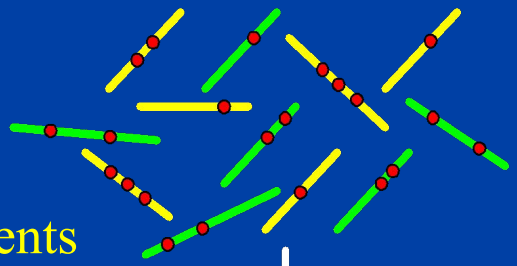


CELERA
A Unit of PE Corporation

Shotgun DNA Sequencing (Computation)

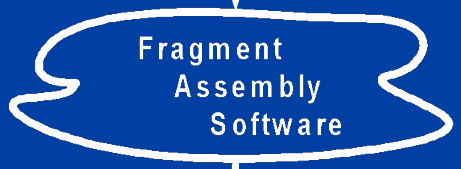
Unknown "Target" DNA Sequence

Randomly Sample
("Shotgun") Fragments



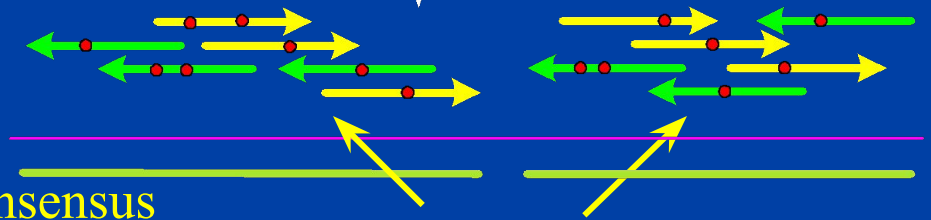
Fragments

G = 100Kbp Target Length
(e.g., BAC, P1, PAC)
F = 1600 # of Fragments
L = 500 Avg. Fragment Length
N = FL = 800Kbp Total Bases Sequenced
c = N/G = 8 Avg. Coverage



- UNKNOWN ORIENTATION
- SEQUENCING ERRORS
- INCOMPLETE COVERAGE
- CONSTRAINTS (MATES)
- REPEATS

Layout



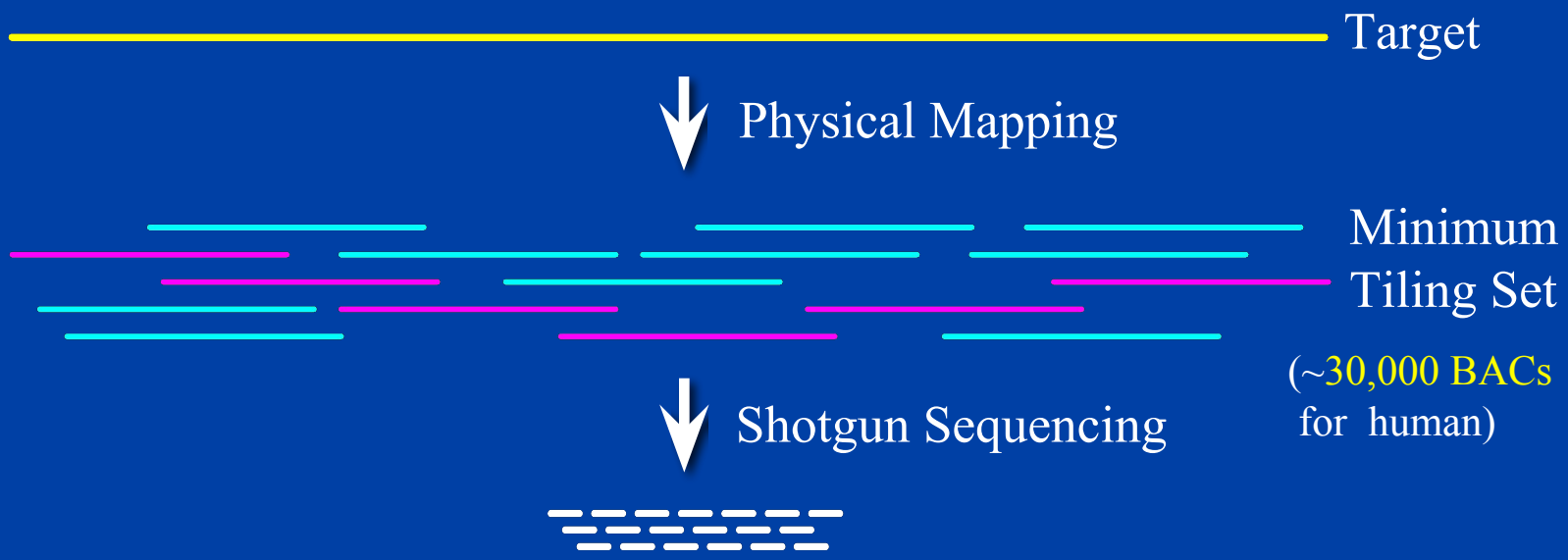
Consensus

Contigs

CELERA
A Unit of PE Corporation

Whole Genome Sequencing Approaches

- Hierarchical HGP Approach:

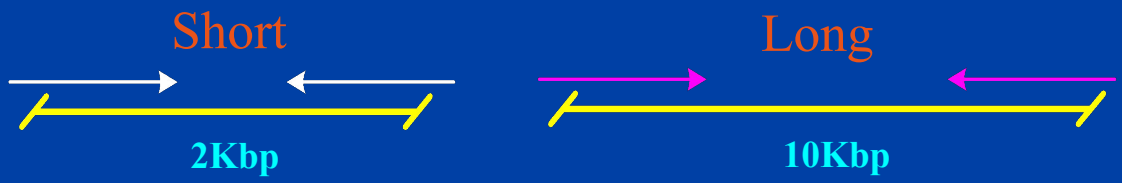


– 2 separate processes
– maps very hard to complete, libraries unstable
– must make shotgun library of each BAC
+ infrastructure is already developed
+ quality of outcome is known

CELERA
A Unit of PE Corporation

Whole Genome Sequencing Approaches

- Whole Genome Shotgun Sequencing:
  - Collect 10x sequence in a 1 -1 ratio of two types of read pairs: ~ 70million reads for Human.



  - Collect 10-15x BAC inserts and end sequence: ~ 300K pairs for Human.
  - Early simulations showed that if repeats were considered black boxes, one could still cover 99.7% of the genome unambiguously.

+ single process, three library constructions
– assembly is much more difficult

CELERA
A Unit of PE Corporation

Sequencing Factory

- 300 ABI 3700 DNA Sequencers installed
- 50 Production Staff
- 40 Support Staff (R&D, QC/QA, Service)
- 20,000 sq. ft. of wet lab
- 20,000 sq. ft. of sequencing space
- 800 tons of A/C (160,000 cfm)
- 4,000 amps electrical service

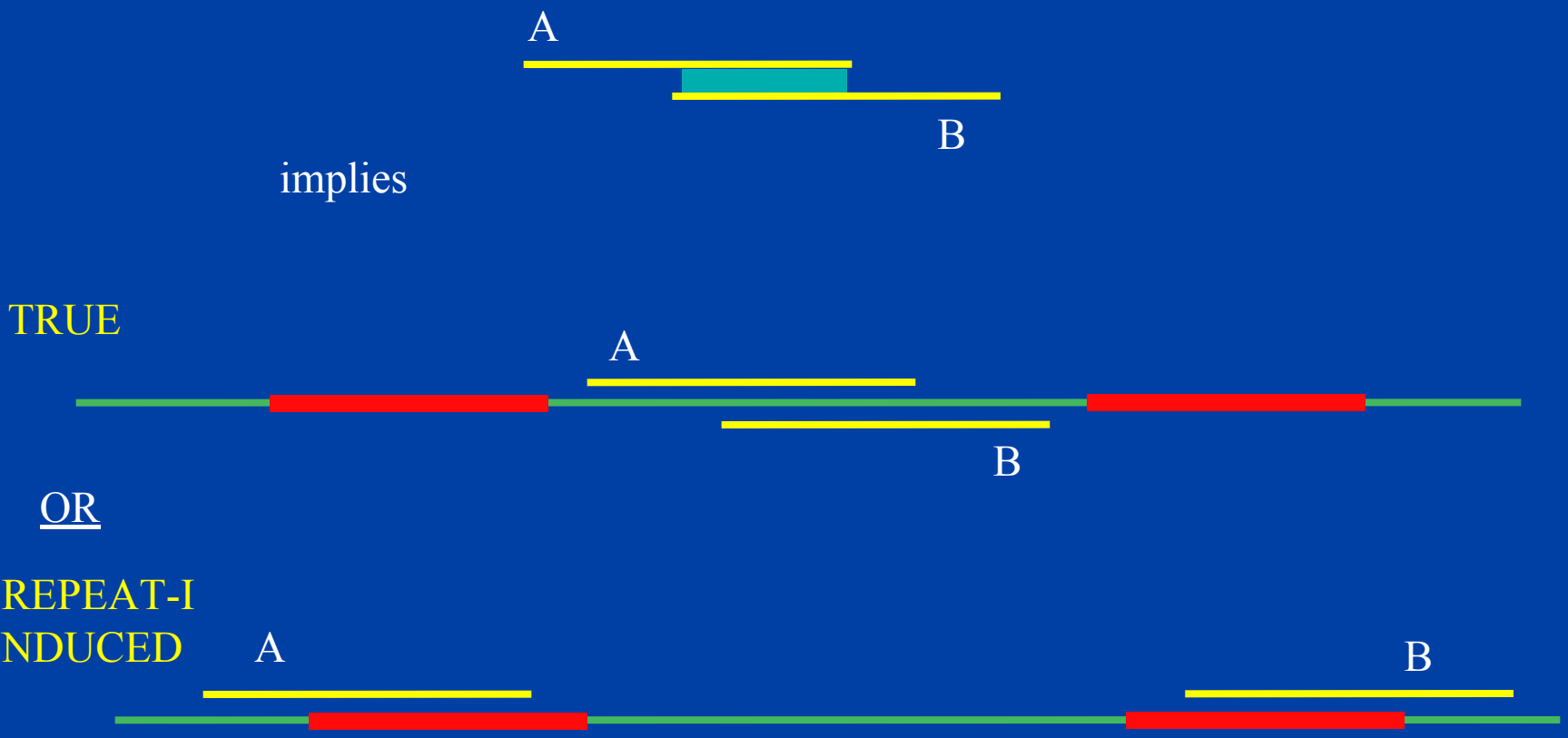
CELERA
A Unit of PE Corporation

The DNA is loaded into automated sequencers. Celera's automated sequencers run 24-7 and have the ability to decipher more than 100 million letters of genetic code per day - the equivalent of 3 percent of the entire human genetic code every day.

The sequencers create am image of the DNA samples being decoded. The four letters of the genetic code -- A, C, T, G -- each are assigned a color.

CELERA
A Unit of PE Corporation

True vs. Repeat-Induced Overlaps



CELERA
A Unit of PE Corporation

Assembly Pipeline

167:41 cpu hrs. for Dros

Screener

8:37

Overlapper

86:25

Unitiger

38:29

Scaffolder

4:12

Repeat Rez I, II, III

5:44+4:21+19:53

Consensus

(~25)

Mask heterochromatin and ribo-DNA,
Tag known interspersed repeats.

Find all overlaps ≥ 40bp allowing 6% mismatch.
(1000X Blast)

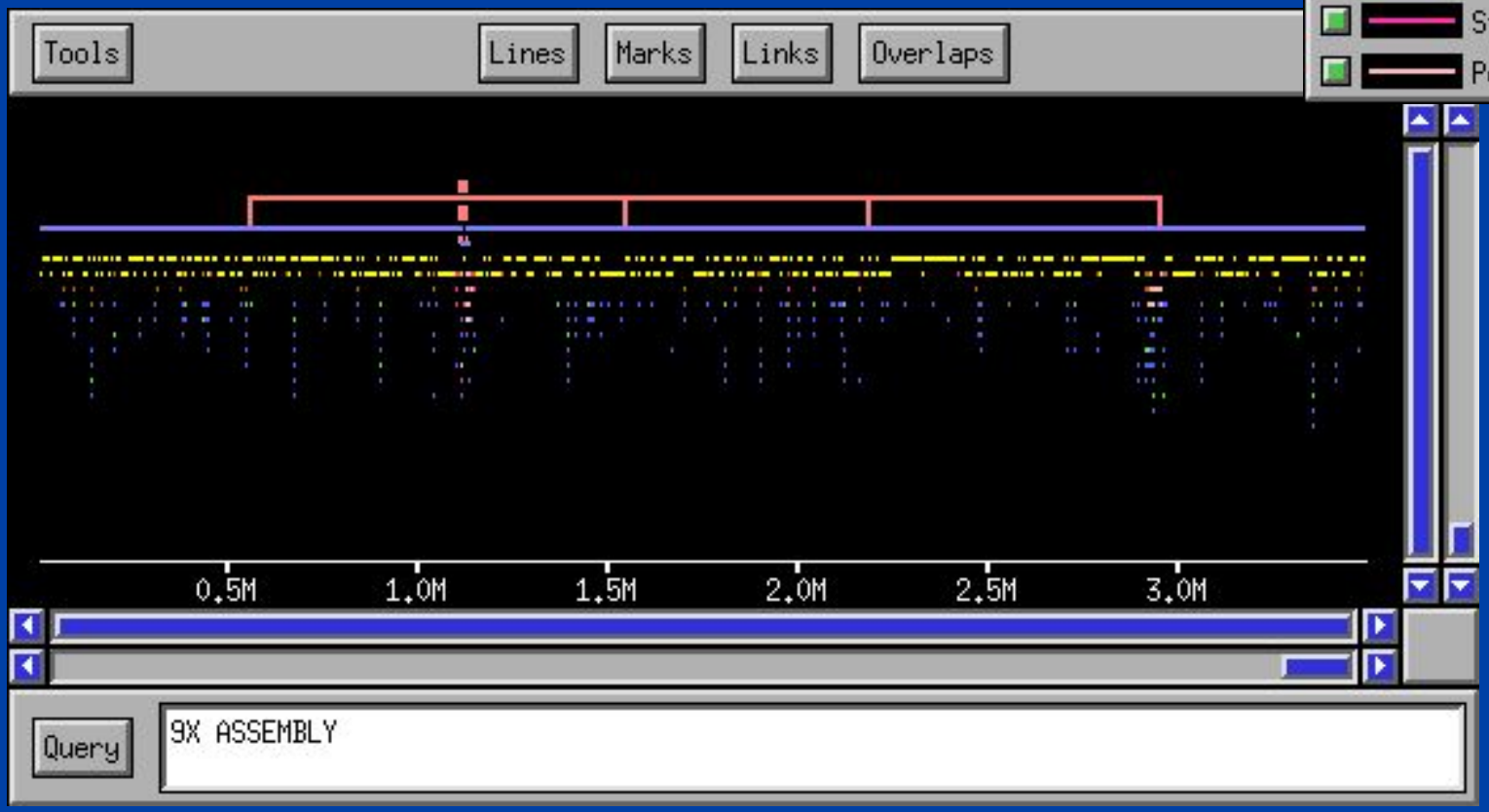
ASSEMBLER CORE:

- Compute all consistent sub-assemblies = unitigs
- Identify those that cover unique DNA = U-unitigs
- Scaffold U-unitigs with confirmed shorts & longs
- Then with BAC ends
- Fill repeat gaps with:
  - I. Doubly anchored mates
  - II. O-path confirmed singly-anchored mates
  - III. Greedy path completion using QVs

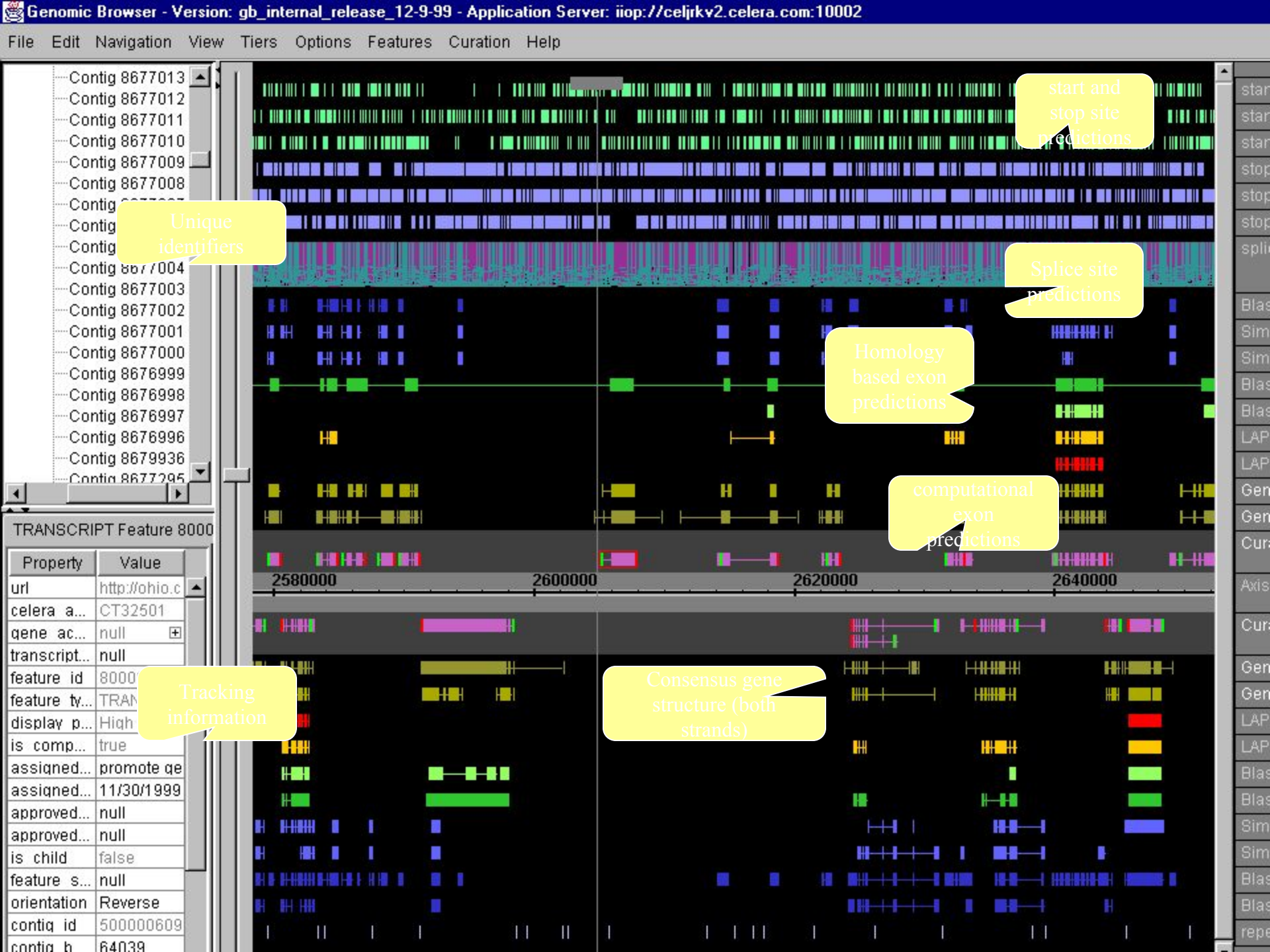
Bayesian "SNP" consensus using quality values.
Occurs throughout assembler core.

CELERA
A Unit of PE Corporation

Assembly Progression (Macro View)



CELERA
A Unit of PE Corporation


Genomic Browser - Version: gb_internal_release_12-9-99 - Application Server: iiop://celjrkv2.celera.com:10002
File Edit Navigation View Tiers Options Features Curation Help
Contig 8677013
Contig 8677012
Contig 8677011
Contig 8677010
Contig 8677009
Contig 8677008
Contig 8677004
Contig 8677003
Contig 8677002
Contig 8677001
Contig 8677000
Contig 8676999
Contig 8676998
Contig 8676997
Contig 8676996
Contig 8679936
Unique identifiers
start and stop site predictions
Splice site predictions
Homology based exon predictions
computational exon predictions
2580000
2600000
2620000
2640000
Consensus gene structure (both strands)
Tracking information
TRANSCRIPT Feature 8000
Property
Value
url
http://ohio.c
celera a...
CT32501
gene ac...
null
transcript...
null
feature id
8000
feature ty...
TRAN
display p...
High
is comp...
true
assigned...
promote ge
assigned...
11/30/1999
approved...
null
approved...
null
is child
false
feature s...
null
orientation
Reverse
contig id
500000609
contig b
64039