Barnacle:
An assembly algorithm for Clone-based Sequences of Whole Genomes

Martin Farach-Colton

Joint work with: Vicky Choi

Outline

- Introduction to Sequencing
- Human Genome Project & the Sequence Assembly Problem
- The Barnacle Algorithm
  - Details of the input
  - The basic idea
- Comparison with NCBI's public assembly
- Conclusion

DNA Sequencing

- Sequencing is the process of determining the sequence of nucleotides of a region of DNA.
- How do we find the sequence of a piece of DNA?

Basic Operations for Sequencing

- Direct Sequencing
- Directed Reads
- End Sequencing
- Clone-Probe Incidence

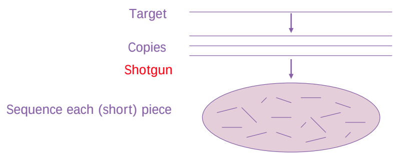
Direct Sequencing

- For short pieces (< 500bp)
  - We can determine complete sequence
    - Called *Direct Sequencing*
  - This is the workhorse of sequencing
  - Relatively fast & cheap
    - ~ 1% error rate

Greedy Assembly aka Shotgun Sequencing

- Make many copies of DNA
- Cut each piece in a different way
  - Now 500bp pieces have overlap
- Repeat until done:
  - Find sequences of maximal overlap
    - (must try reverse compliment)
  - Merge them, and add merged sequence to set
- Assembled pieces need not form one piece
  - So they have gaps once assembled into *contigs*

Shotgun Sequencing (Draft)



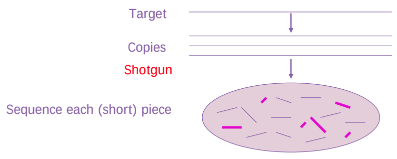
Target

Copies

Shotgun

Sequence each (short) piece

Shotgun Sequencing (Draft)



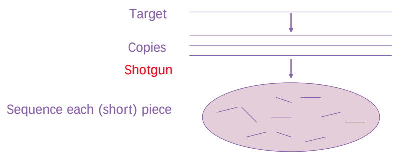
Target

Copies

Shotgun

Sequence each (short) piece

Shotgun Sequencing (Draft)



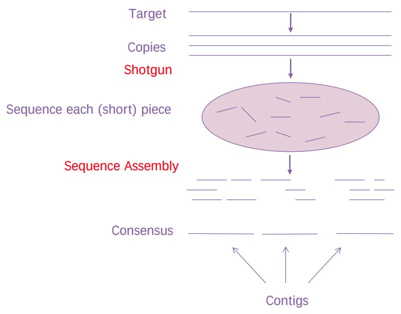
Target

Copies

Shotgun

Sequence each (short) piece

Shotgun Sequencing (Draft)



Target

Copies

Shotgun

Sequence each (short) piece

Sequence Assembly

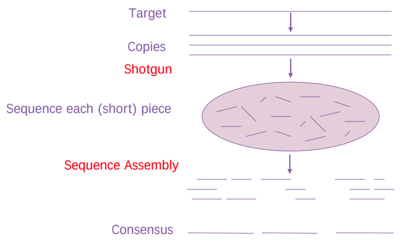
Consensus

Contigs

Directed Reads

- Given a long sequence that only occurs once in the genome…
  - It can be extended by *Directed Reads*
  - These are 500bp at a time.
  - You can iterate.
  - Each iteration is slow and expensive.
- You can connect contigs with directed reads

Shotgun Sequencing (Final)



Target

Copies

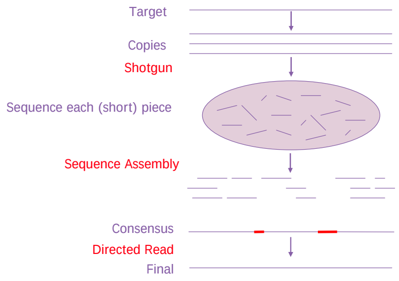
Shotgun

Sequence each (short) piece

Sequence Assembly

Consensus

Shotgun Sequencing (Final)



Why aren’t we done?

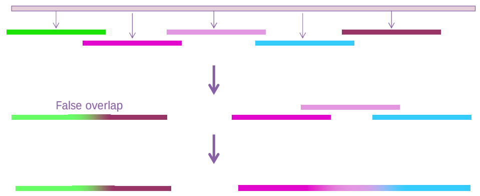
- Lab errors limit process.
  - Can get false matches or miss true matches
  - Can get more exotic errors (more later)
- Repeats
  - Human genome is repeat-rich
    - >50% repeats
    - 50-500kbp duplicated regions with >98% identity
  - 500bp fragments from different repeats can be merged.
    - How can we tell if we are merging from different repeats?
  - Repeats are the unsolved problem of sequencing!

Shotgun Sequencing History

- 1980s: 5 to 10 Kbp
- 1990: 40 Kbp
- 1995: 1.8Mbp (*H. Influenzae*)
- 2000: 120 Mbp (*Drosophila*)
  - » Except for repeated regions

Shotgun Sequencing Limitation

- We noted that you can have false merges.



Directed reads aren’t going to help merge false contigs!

Shotgun Sequencing Limitation

- We noted that you can have false merges.
- Once we’ve made a few bad choices, errors accumulate.
- This limits the length of DNA that can be reliably sequenced by this method.
- How can we shotgun longer sequences?

Medium Length DNA

- To scale methods up, we need operations to limit error propagation in longer pieces of DNA.
- The specific operations we care about depend on DNA length.
- Name of DNA pieces depend on how they are copied
  - Plasmid, Cosmids = a few Kbp
  - BACs, YACs = tens to a few hundred Kbp.

End-Sequencing

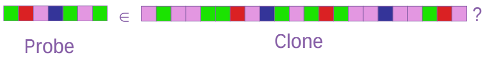
- You can sequence 500 bp at each end of DNA.
  - They can be used to:
    - Keep fragment merging on track, because if two fragments are known to be e.g. 2000 bp apart and your merging doesn't give that, you've got an error.
    - Tell the relative orientation of the pieces.
  - If it's too long, the information derived is too sparse.
  - Plasmids are the right length (~c x 10^3 bp)

Celera's Shotgun Sequence

- Get lots of plasmid information.
- This constrains which pairs can be merged in shotgun sequence.
  - You merge bogus pairs with lower probability.
  - So you can merge longer stretches more reliably.
    - Or at least, that's the idea.
- They claim to have complete human genome.
  - Once again, repeat regions are not yet sequenced.
  - Plasmids can easily fit within some repeats!

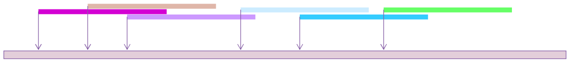
Probe-Clone Incidence

- You can tell if a piece of DNA (clone) has some particular substring (probe).
- If clone too short, unlikely to have the probe.
- If clone too long, too likely to have the probe.
- BACs are right length (~c x 10^4 or c x 10^5 bp)
- Used to tell if two BACs overlap.

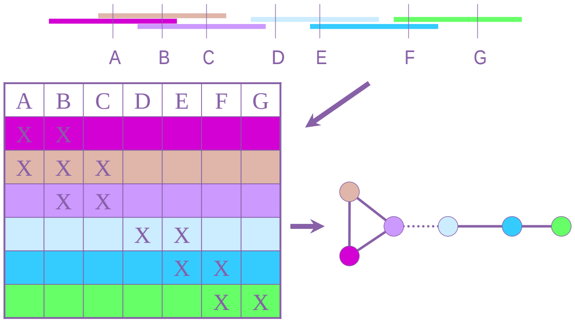


Clone-Probes & Physical Maps

- Given a set of BACs from a Chromosome
  - A *Physical Map* is the approximate location of each BAC
- Clone-Probe incidence matrices can be used to construct physical maps of BACs through
  - *Interval Graph* techniques

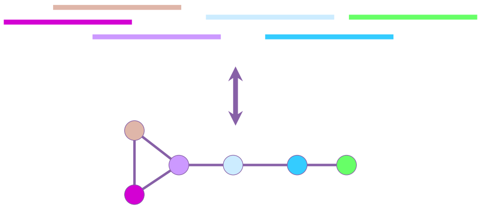


Physical Mapping by Probes



A	B	C	D	E	F	G
X	X					
X	X	X				
	X	X				
			X	X		
				X	X	
					X	X

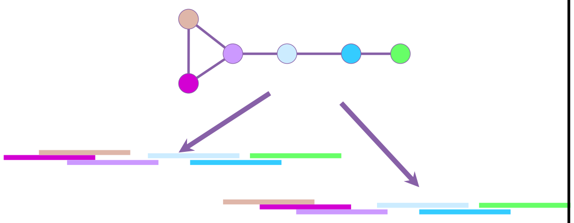
Interval Graph



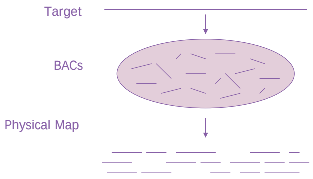
Interval Graphs

- Suppose you have intervals on a line
  - Make a graph with:
    - A node for each interval
    - An edge between overlapping intervals
- Suppose you have a graph so generated
  - Coming up with a set of matching intervals is called *Interval Realization*
  - A particular graph can have many different Interval Realizations

Interval Realizations



Hierarchical Shotgun Sequencing

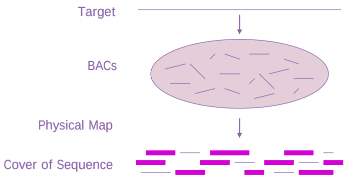


Target

BACs

Physical Map

Hierarchical Shotgun Sequencing



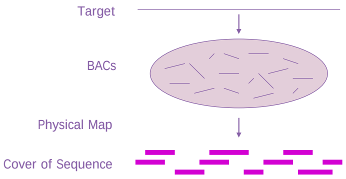
Target

BACs

Physical Map

Cover of Sequence

Hierarchical Shotgun Sequencing



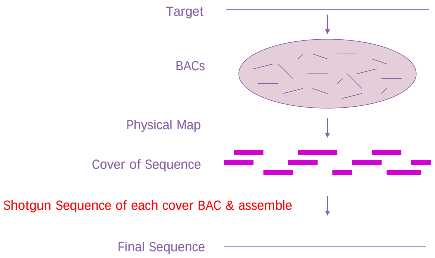
Target

BACs

Physical Map

Cover of Sequence

Hierarchical Shotgun Sequencing



Target

BACs

Physical Map

Cover of Sequence

Shotgun Sequence of each cover BAC & assemble

Final Sequence

Hierarchical Shotgun Sequencing

1. Copy target DNA
2. Make BAC library
3. Physically map all BACs
4. Find a subset of BACs that cover target DNA
5. Shotgun sequence only BACs in cover
6. Fill in gaps between BACs
7. Merge into consensus sequence

Hierarchical Shotgun Sequencing

- Sequencing each BAC lets you
  - Localize merging mistake to one BAC
- Physical map lets you get *covering* of genome by BACs, so you end up doing less sequencing.
  - If sequencing were expensive & physical mapping cheap, this would be a good idea.

Outline

- Biological Background
- Human Genome Project
- The Barnacle Algorithm
  - Details of the input
  - The basic idea
- Comparison with NCBI’s public assembly
- Conclusion

Human Genome Project (HGP)

- 1988: “Mapping and Sequencing the Human Genome”
- 1990: HGP started in US
- 2001: A “working draft” version
- 2003: Completed

Sequencing Approaches of HGP

- Hierarchical Shotgun Sequencing.
- The physical map was scheduled to take 5 years.
- Genome centers had two choices:
  - Start sequencing before physical map was done.
  - Twiddle thumbs.

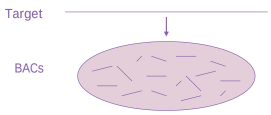
Clone-based Sequencing

or

Making a Virtue of Necessity

- Perhaps trading sequencing for physical mapping isn’t such a good idea.
- New Idea
  - Sequence every BAC, not just BAC in cover.
  - Release draft BAC as you have it.
    - Recall: getting BAC sequence in 1 piece is hard, so release sequences before directed reads.
  - Release finished BAC as you have it.
  - Release parts of physical map as you have them.

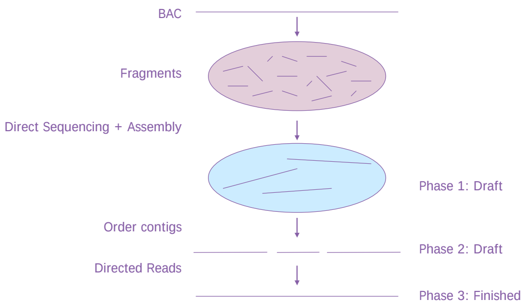
Clone-based Sequencing



Target

BACs

Sequencing of BACs



BAC

Fragments

Direct Sequencing + Assembly

Phase 1: Draft

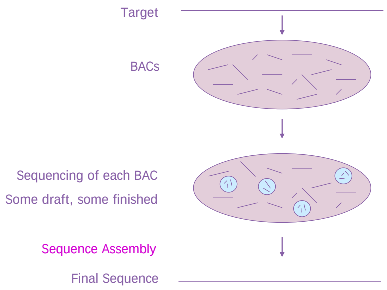
Order contigs

Phase 2: Draft

Directed Reads

Phase 3: Finished

Clone-based Sequencing



Target

BACs

Sequencing of each BAC

Some draft, some finished

Sequence Assembly

Final Sequence

Clone-based Sequencing The Input

- Clone-based sequencing wasn't so much planned as what's available
- Input is a mixed bag

Input: Sequence Information

- Recall: A BAC is a contiguous stretch of DNA from a chromosome. Each comes as a set of fragments.

Accession	Est. Length	Phase	Chrm	# frags
AC002092.1	95456	1	17	4

Frag acc.	length
AC002092.1~1	888
AC002092.1~2	45312
AC002092.1~3	38725
AC002092.1~4.1	10245

- Phase 1,2 = Draft
- Phase 3 = Finished

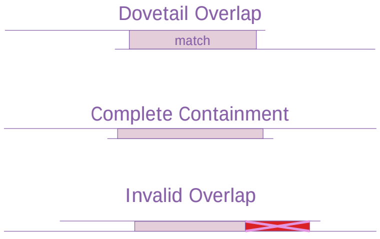
Input: Chromosome Assignment

- The chromosome of a BAC is assigned according to some additional info
  - – E.g. STS markers
- For some BACs, chromosome is marked **Unknown**
- Definition: Two BACs are *compatible* if they share a chromosome assignment or at least one is Unknown

Input: Pair-wise local alignments

- NCBI’s algorithm (and ours) need to know about shared sequences between fragments.
- NCBI preprocessing:
  - A local alignment between every pair of fragments with an compatible chromosome assignment.
  - This is slow.

Alignments are overlaps



Dovetail Overlap

match

Complete Containment

Invalid Overlap

We care about overlaps if…

- They are valid -- dovetail or containment.
- They have high sequence identity: 97%
- They have low end-allowed-error:
  - 350bp for phase 3
  - Min(15%,1500) for phase 1, 2.
- These thresholds give a *very conservative* measure of overlapping fragments
  - Lots of false negatives

Input: Plasmid Info

- Some labs produce plasmid (&mRNAs&…) with End Sequencing
- This lets us find orientation of some fragments.
  - This is different than Celera’s merging constraints (though they also use plasmids for orientation)

Input: A Mixed Bag

- BAC info.
  - Estimated Length
  - Chromosome Assignment
  - Fragment sequences
- Compatible Fragment Overlaps
- Plasmids
- Misc:
  - Genome Centers are also doing physical maps of BAC, so they report those.

Computed, not measured

NCBI Approach

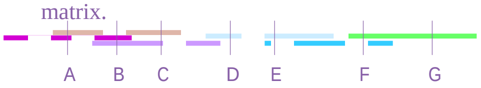
- Simulate physical map & Reduce to Hierarchical Shotgun Sequencing Assembly
- Find “probe” sequences that are shared by sequences in different BACs.
  - Use these as probes to create a Clone-Probe Incidence Matrix.
    - Give them weights: the longer the shared match, the less likely it’s due to chance, so give it a bigger weight.
  - Create Physical Map using known techniques.
  - Assemble using Physical Map.

Problems with NCBI Approach

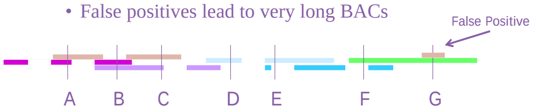
- Clone order is determined early on by physical map.
  - The Clone-Probe to Physical Map problem is noisy and error prone.
  - Any error is propagated badly in final answer.
- There are errors in the underlying data that confuse the physical map process badly.
- Clone-Probe incidence doesn't preserve information about *where* two clones overlap.

Physical Mapping Error Model

- Entire BAC sequence not available
  - Result: lots of false negatives in probe-clone matrix.



  - Physical map construction has to be insensitive to false negatives.
    - False positives lead to very long BACs



NCBI's Strength

- The Genome Centers report partial physical map information from time to time.
  - The good news: This can be incorporated directly into their algorithm by a very high scoring clone-probe pair.
  - The bad news: Their information is sometimes wrong!
    - And each wrong such piece of info causes lots of long BACs

NCBI Summary

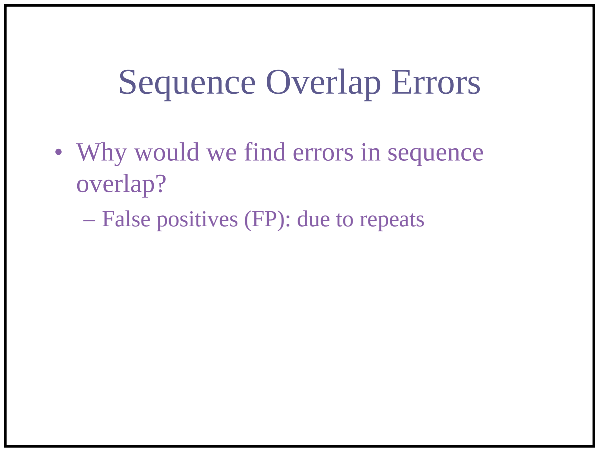
- Use clone-clone overlaps to compute clone-probe matrix
  - Reduces problem to known physical map problem
- Use physical map to align clones
- Produce Consensus Sequence
- Top-Down approach that first fixes BAC positions, then goes to sequence level

Outline

- Biological Background
- Human Genome Project & the Sequence Assembly Problem
- The Barnacle Algorithm
  - Details of the input
  - The basic idea
- Comparison with NCBI's public assembly
- Conclusion

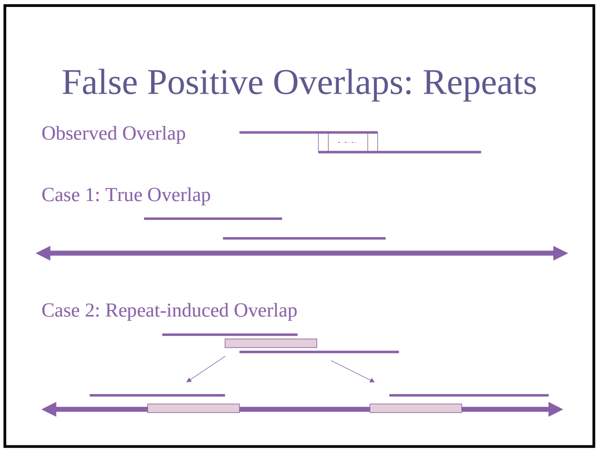
Our approach

- Bottom-Up
  - Sequence Data is most reliable
  - We can boost reliability by consistency conditions
- Use Sequence Overlap to Determine BAC Overlaps
  - Similar to NCBI, but only uses reliable overlaps
- Filter out inconsistencies in BAC overlap graph
  - Each error in BAC graph comes from some error in the underlying data
  - **We can detect and report the likely errors**



Sequence Overlap Errors

- Why would we find errors in sequence overlap?
  - False positives (FP): due to repeats

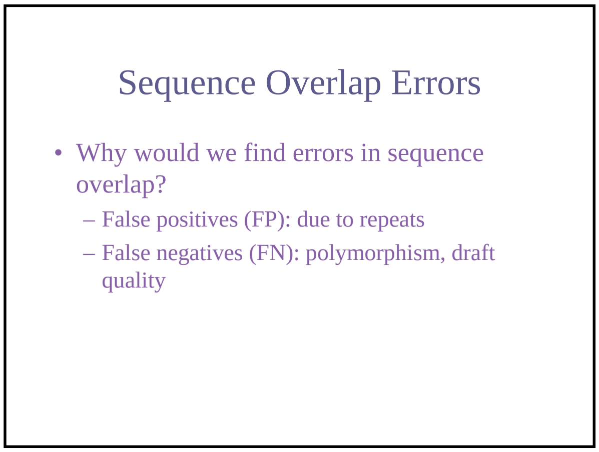


False Positive Overlaps: Repeats

Observed Overlap

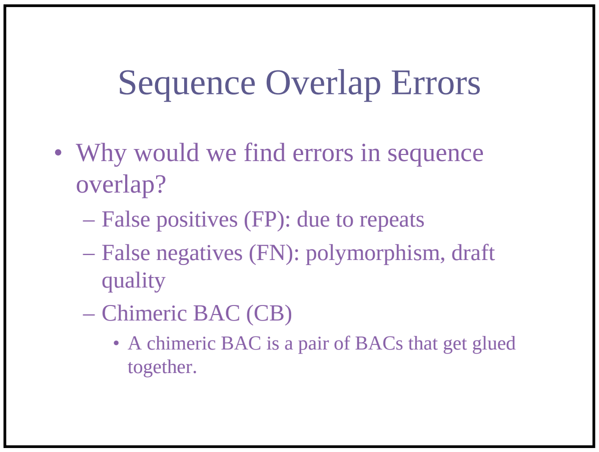
Case 1: True Overlap

Case 2: Repeat-induced Overlap



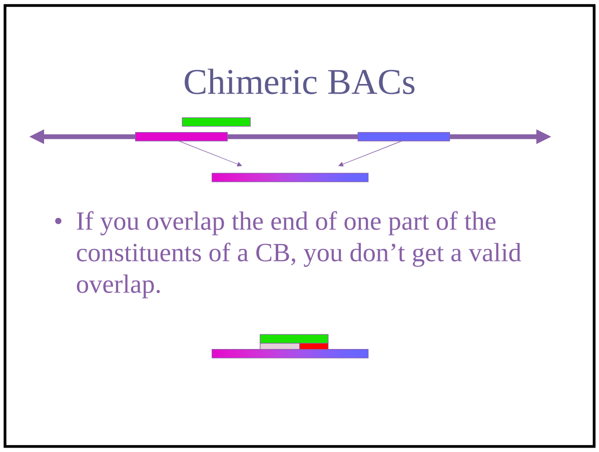
Sequence Overlap Errors

- Why would we find errors in sequence overlap?
  - False positives (FP): due to repeats
  - False negatives (FN): polymorphism, draft quality



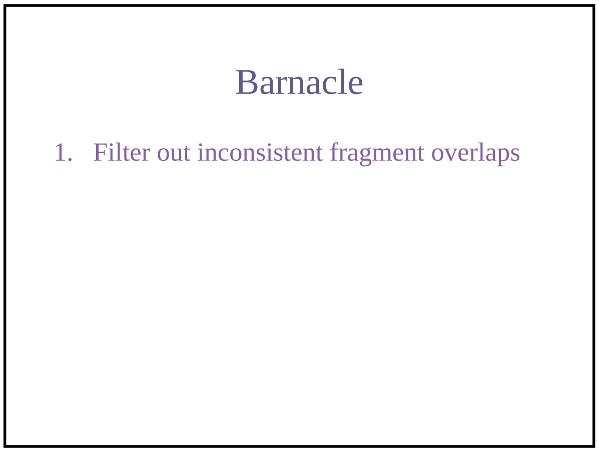
Sequence Overlap Errors

- Why would we find errors in sequence overlap?
  - False positives (FP): due to repeats
  - False negatives (FN): polymorphism, draft quality
  - Chimeric BAC (CB)
    - A chimeric BAC is a pair of BACs that get glued together.



Chimeric BACs

- If you overlap the end of one part of the constituents of a CB, you don’t get a valid overlap.

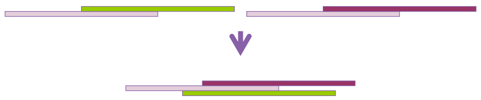


Barnacle

1. Filter out inconsistent fragment overlaps

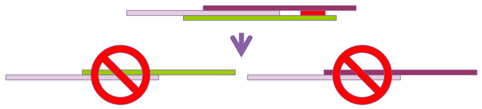
Remove Inconsistent Overlaps

- If two fragments overlap *the same end of another fragment*
  - They must overlap with each other!
  - Eliminate any overlaps that aren't consistent.
    - Using this and related considerations.



Remove Inconsistent Overlaps

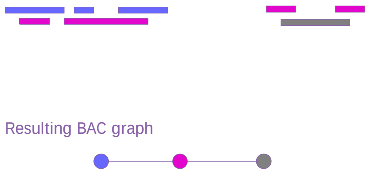
- If two fragments overlap *the same end of another fragment*
  - They must overlap with each other!
  - Eliminate any overlaps that aren't consistent.
    - Using this and related considerations.



Barnacle

1. Filter out inconsistent fragment overlaps
2. Form BAC overlap graph

BAC Graph from Overlaps



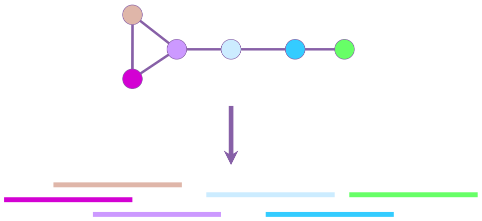
Consistent Overlaps of Fragments

Resulting BAC graph

Barnacle

1. Filter out inconsistent fragment overlaps
2. Form BAC overlap graph
3. Find *Interval Realization* of BAC Graph

Interval Graph



True BAC overlap graph is an interval graph!

We Still Have Errors

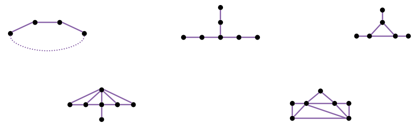
- The true BAC overlap graph is interval
- We only have the computed BAC overlap graph.
  - We’ve been very conservative in assembling it.
  - So we hope for not too many errors.
  - But the BAC graph we have might not be interval.
- We have to find places that keep BAC graph from being interval and decide what to do.

Interval Graph Recognition

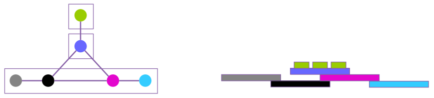
- There are lots of algorithms for recognizing an interval graph, e.g. using PQ-trees
- We use *5-sweep LBFS interval graph recognition algorithm* (Corneil, Olariu & Stewart 2000)
  - LBFS = Lexicographic Breadth First Search

Forbidden Subgraphs

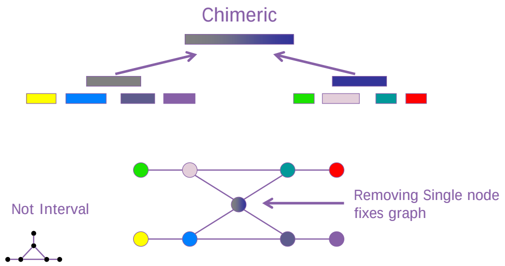
- Theorem: A graph is interval iff it does not contain one of the (induced) subgraphs below:



Forbidden Subgraph: Example



Errors Make BAC graph non-interval



Critical Nodes

- A node whose removal fixes non-interval subgraph is called *critical*
- The Interval Graph algorithm we use produces critical nodes
- Each one comes from some type of error
  - FP, FN, CB

Error Detection

- When we detect a critical node:
  - We find the most likely type of error that produced it.
  - We fix the graph by some local change in the edges or by removing the node.
  - If we remove a node, it’s because we have detected a *Reportable Error*
- Ours is the only algorithm available that does Error Detection on the genome center data.

Barnacle

1. Filter out inconsistent fragment overlaps
2. Form BAC overlap graph
3. Find *Interval Realization* of BAC Graph
   1. Resolve Critical Nodes if possible

Non-fixable Graphs

- Sometimes a graph can’t be fixed by critical node resolution.
  - We need a more global solution.
- We then break BAC graph into pieces at *articulation points* (graph theoretic mumbo-jumbo)
  - We fix each piece.
  - We glue them back together.

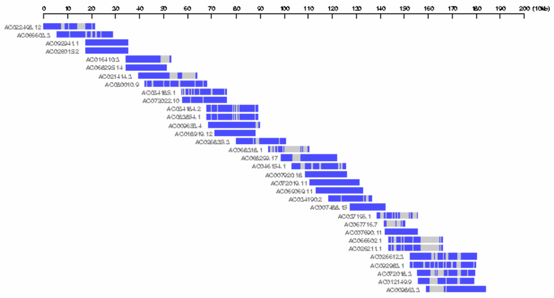
Barnacle

1. Filter out inconsistent fragment overlaps
2. Form BAC overlap graph
3. Find *Interval Realization* of BAC Graph
   1. Resolve Critical Nodes if possible
   2. Divide-n-Conquer at articulation nodes when needed

Once we have Interval BAC Graph

- We can produce interval realization of BAC graph.
  - But a single interval graph might have lots of interval realizations.
- Use plasmids to do final ordering and orientation
  - To disambiguate the BAC graph.

Our output: Contigs



Our Output: Errors in Data

- Suspected Chimeras & Repeats
- Chromosome Mis-assignment

Outline

- Biological Background
- Human Genome Project & the Sequence Assembly Problem
- The Barnacle Algorithm
  - – Details of the input
  - – The basic idea
- Experimental Results
- Conclusion

Statistics about Input (Dec 2001)

phase	BACs	frags	Total length in Gbp	Ave. number of frags
1	15298	246424	2.55	16.11
2	2154	8161	0.33	3.79
3	17624	17624	2.04	1
Total	35076	272209	4.992	7.76

Overlap Information: 403,466 fragment pairs
Orientation Information: 321,751 fragment pairs
Chromosome Info:
31543 by STS; 2450 by Genbank; 1083 unknown

Comparison Metrics

- BACs/Fragments deleted
  - – We delete BACs and Fragments when we put them in our error list.
  - – NCBI also deletes BACs/Fragments

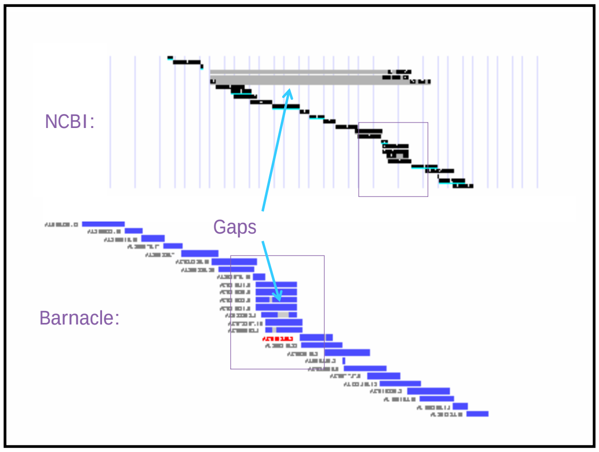
Fragments Used

Barnacle	BACs	Frags Used/ Frags	Contigs	Length (Gbp)
Singletons	1215	9967/9967	1215	0.142
Non-Singletons	33722	251041/259230	2443	2.708
	34937	261008/269197	3658	2.850

NCBI	BACs	Frags Used/ Frags	Contigs	Length (Gbp)
Singletons	836	9074/9074	836	0.112
Non-Singletons	32902	222391/251928	2292	2.745
	33738	231465/261002	3128	2.857

Comparison Metrics

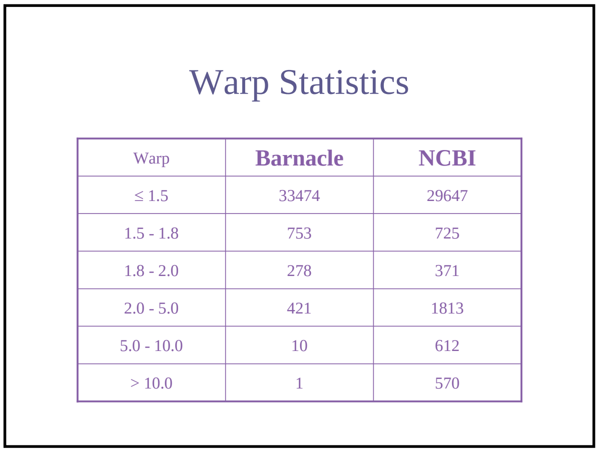
- BACs/Fragments deleted
  - – We delete BACs and Fragments when we put them in our error list.
  - – NCBI also deletes BACs/Fragments
- Warp
  - – Recall that each BAC has estimated length
  - – It has a length in the final answer alignment.
  - – Ratio of length/estimated length = *WARP*



NCBI:

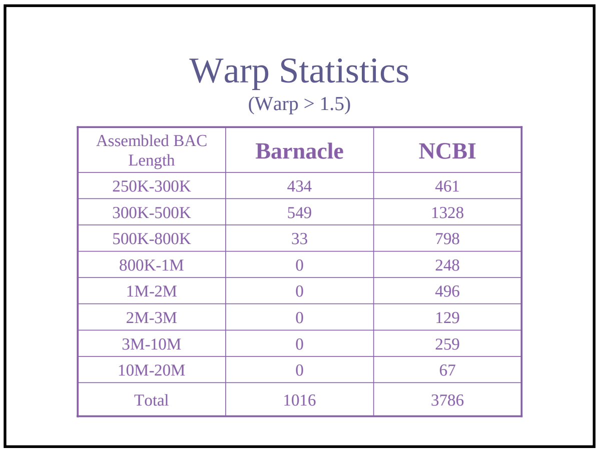
Gaps

Barnacle:



Warp Statistics

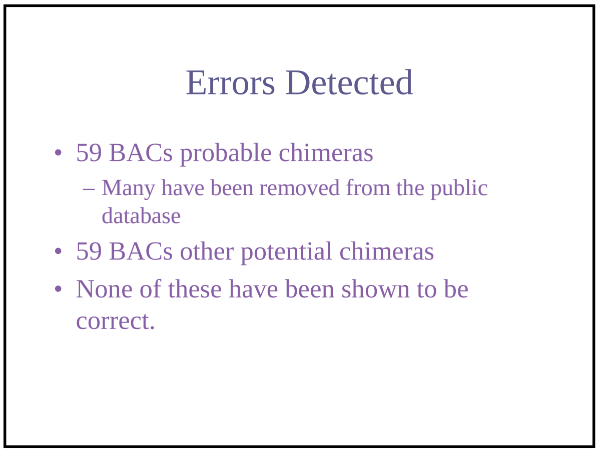
Warp	Barnacle	NCBI
≤ 1.5	33474	29647
1.5 - 1.8	753	725
1.8 - 2.0	278	371
2.0 - 5.0	421	1813
5.0 - 10.0	10	612
> 10.0	1	570



Warp Statistics

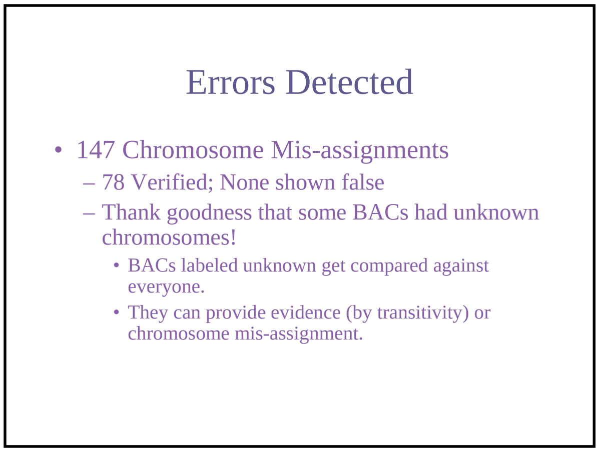
(Warp > 1.5)

Assembled BAC Length	Barnacle	NCBI
250K-300K	434	461
300K-500K	549	1328
500K-800K	33	798
800K-1M	0	248
1M-2M	0	496
2M-3M	0	129
3M-10M	0	259
10M-20M	0	67
Total	1016	3786



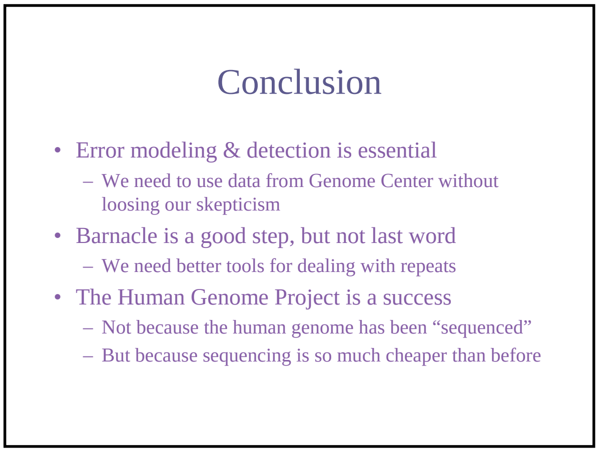
Errors Detected

- 59 BACs probable chimeras
  - Many have been removed from the public database
- 59 BACs other potential chimeras
- None of these have been shown to be correct.



Errors Detected

- 147 Chromosome Mis-assignments
  - 78 Verified; None shown false
  - Thank goodness that some BACs had unknown chromosomes!
    - BACs labeled unknown get compared against everyone.
    - They can provide evidence (by transitivity) or chromosome mis-assignment.



Conclusion

- Error modeling & detection is essential
  - We need to use data from Genome Center without loosing our skepticism
- Barnacle is a good step, but not last word
  - We need better tools for dealing with repeats
- The Human Genome Project is a success
  - Not because the human genome has been “sequenced”
  - But because sequencing is so much cheaper than before