Suquencing the Whole Genome

Problem: we only know how to sequence about 500 bps at a time in the lab.

1. Linear sequencing
2. The shotgun method
3. Hierarchical shotgun method
4. Whole genome and double-barreled shotgun methods

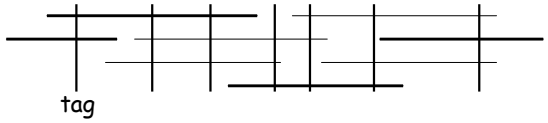
Shotgun on whole genome?

Problems:

- Computationally very expensive
- 50% of genome consist of repeats. Causes major problems.
- Hard to partition work among multiple labs.

Hierarchical Shotgun

1. Generate clone Libraries (100K – 1M per clone)
2. Order the clones by finding "tags" that overlap multiple clones. Use these for ordering.
3. Identify a set of clones that cover the whole length (minimum tiling path)
4. Use shotgun technique on each identified clone
5. Put the results together.



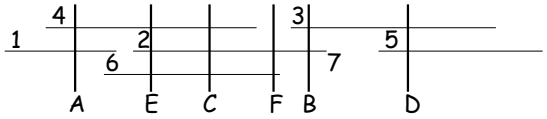
1. Clone Libraries

A "BAC" library will contain sequences of about 200K bps each. These can be cloned using "BAC Vectors" (Bacterial Artificial Chromosome)

A "YAC" library will contain sequences of about 1M bps each. These can be cloned using "YAC Vectors" (Yeast Artificial Chromosome)

These are typically stored at a common site and can be ordered. Many can be purchased from companies.

2. Ordering Clones



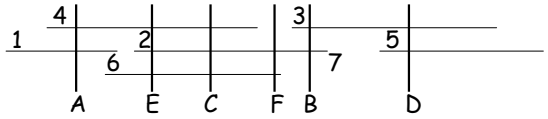
We have the clones, but we don't know their order or how they overlap.

Pick random small sequences that only appear once in one location covered by the library.

These are called STS (Sequence Tagged Sites)

Figure out which clones contain which STSs using PCR (use tag site to start copy…will only copy of the sequence contains the site).

2. Ordering Clones (cont.)



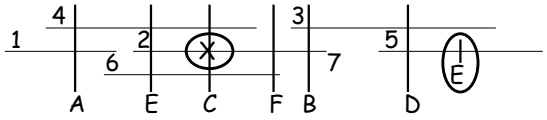
	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	1	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	0	0
6	0	0	1	0	1	1

Goal: Reorder the columns so that all the 1s in each row are contiguous.

Can be done in O(n) time, where n is the number of entries in the array.

But!!!, what about **errrors**?

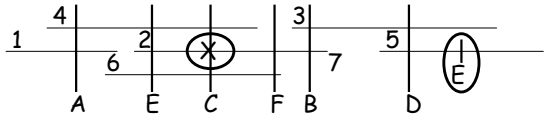
2. Ordering Clones (cont.)



	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	0	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	1	0
6	0	0	1	0	1	1

	A	E	C	F	B	D
1	1	0	0	0	0	0
2	0	1	0	1	1	0
3	0	0	0	0	1	1
4	1	1	1	0	0	0
5	0	1	0	0	0	1
6	0	1	1	1	0	0

2. Ordering Clones (cont.)



Find ordering that minimizes the number of zero-one and one-zero transitions (i.e. errors).

This is NP-hard, but can be posed as a Traveling Salesman Problem (TSP).

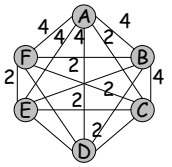
Any ideas?

2. Ordering Clones (cont.)

Create graph with one vertex per STS.

Edge weights = hamming distance (number of bits that differ).

	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	**0**	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	**1**	0
6	0	0	1	0	1	1

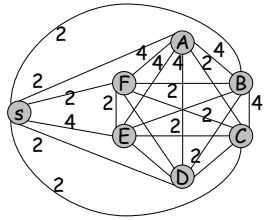


2. Ordering Clones (cont.)

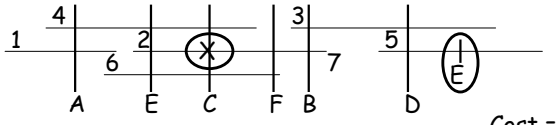
Add in source (s) node with weights equal to number of 1s in each row.

Solve TSP. Answer gives min number of transitions.

	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	**0**	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	**1**	0
6	0	0	1	0	1	1

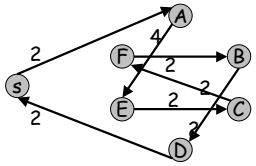


2. Ordering Clones (cont.)

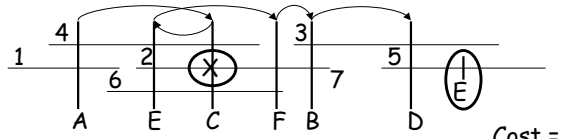


Cost = 16

	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	**0**	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	**1**	0
6	0	0	1	0	1	1

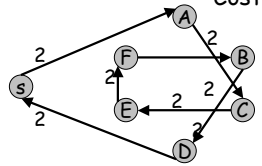


2. Ordering Clones (cont.)



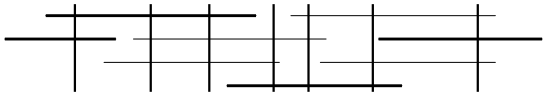
Cost = 14

	A	B	C	D	E	F
1	1	0	0	0	0	0
2	0	1	**0**	0	1	1
3	0	1	0	1	0	0
4	1	0	1	0	1	0
5	0	0	0	1	**1**	0
6	0	0	1	0	1	1



The "wrong" answer has smaller cost

3. Find "Minimum Tiling Path"



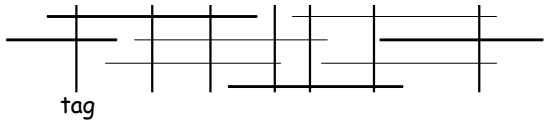
Minimum Tiling Path: Find a set of clones that cover the whole length and for which the total number of bps is minimized.

Can be posed as a shortest path problem.

Any ideas?

Hierarchical Shotgun (revisited)

1. Generate clone Libraries (100K – 1M per clone)
2. Order the clones by finding "tags" that overlap multiple clones. Use these for ordering.
3. Identify a set of clones that cover the whole length (minimum tiling path)
4. Use shotgun technique on each identified clone
5. Put the results together.

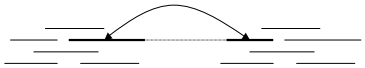


Celera's Method

Whole genome shotgun:

Use shotgun method on whole genome.

Use **double-barreled** approach: some sequences of known length (e.g. 2-5K) are sequenced at both ends. These can be used to bridge across repeats.



In practice they used some mapping (hierarchical) data from the NIST effort, which was freely available. This was needed to deal with long repeats.