

# Why we must finish the rice genome and how we can do it.

Dick McCombie

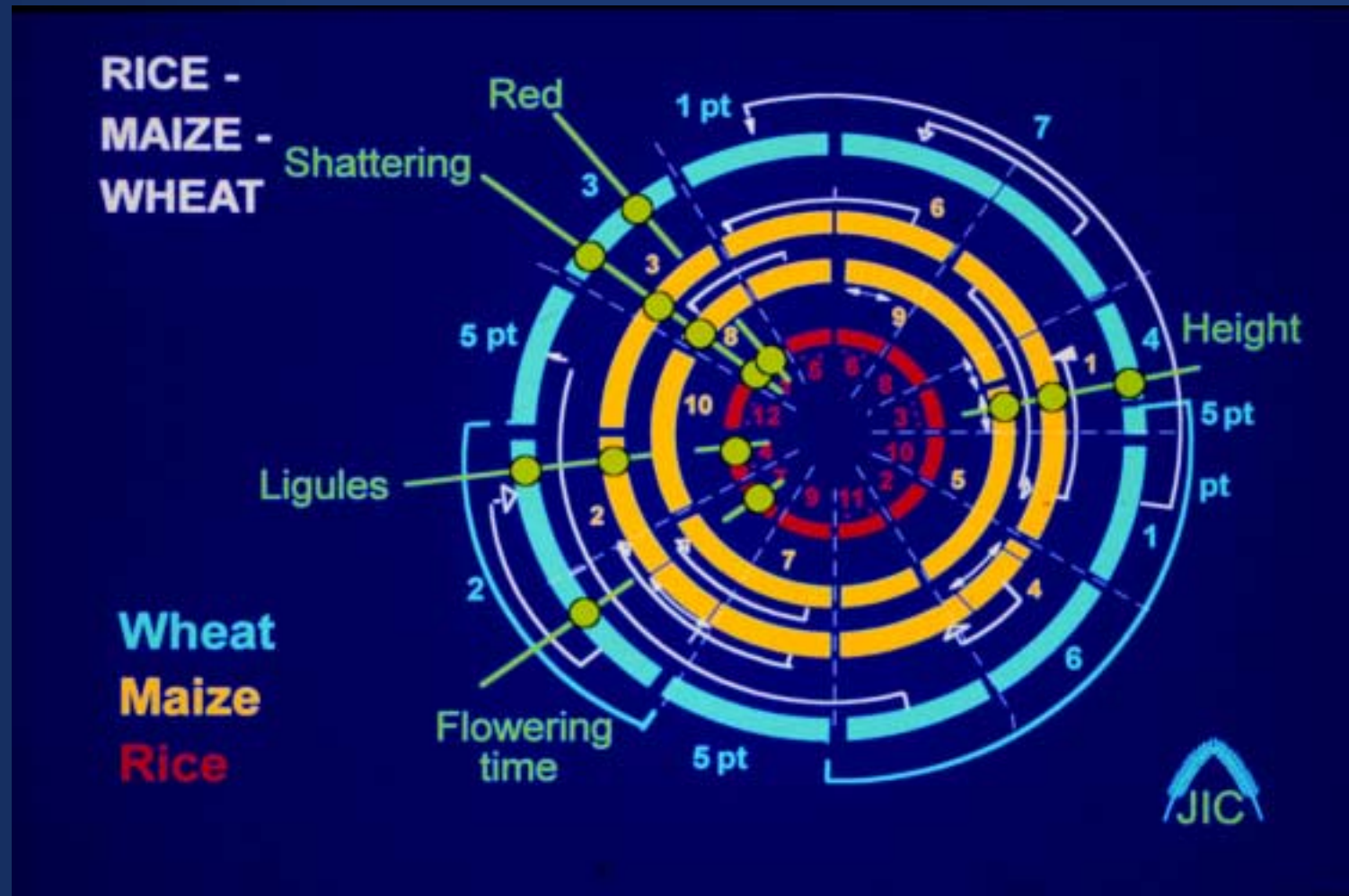
CCW

2/6/02

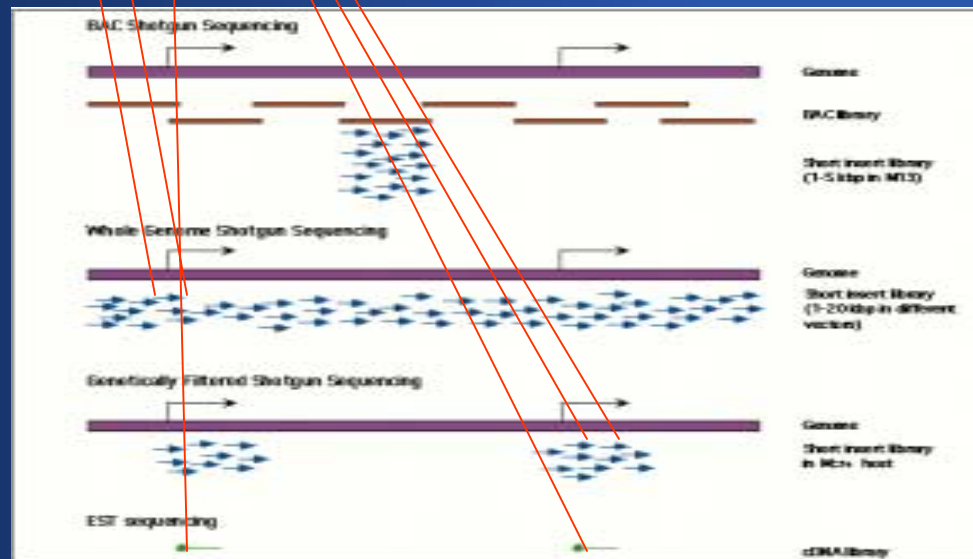
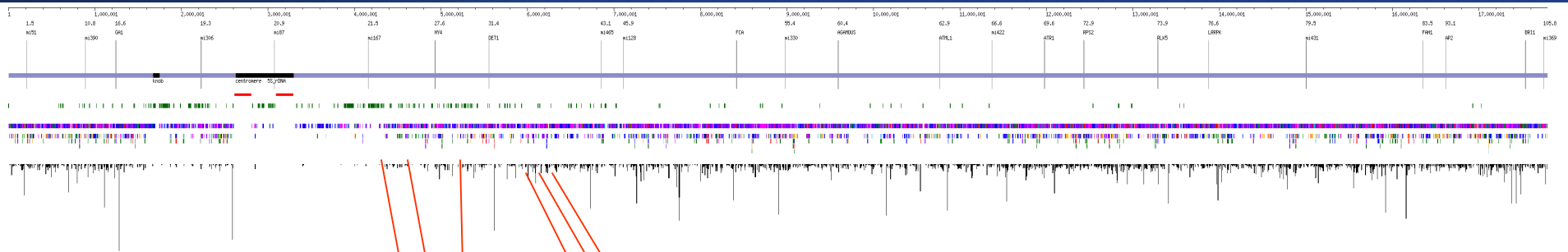
# Goals of Presentation

- Need for finished sequence
- Strategy we are implementing
- Software that has been developed
- Initial test results of the software and finishing strategy

# Co-linearity among cereal genomes



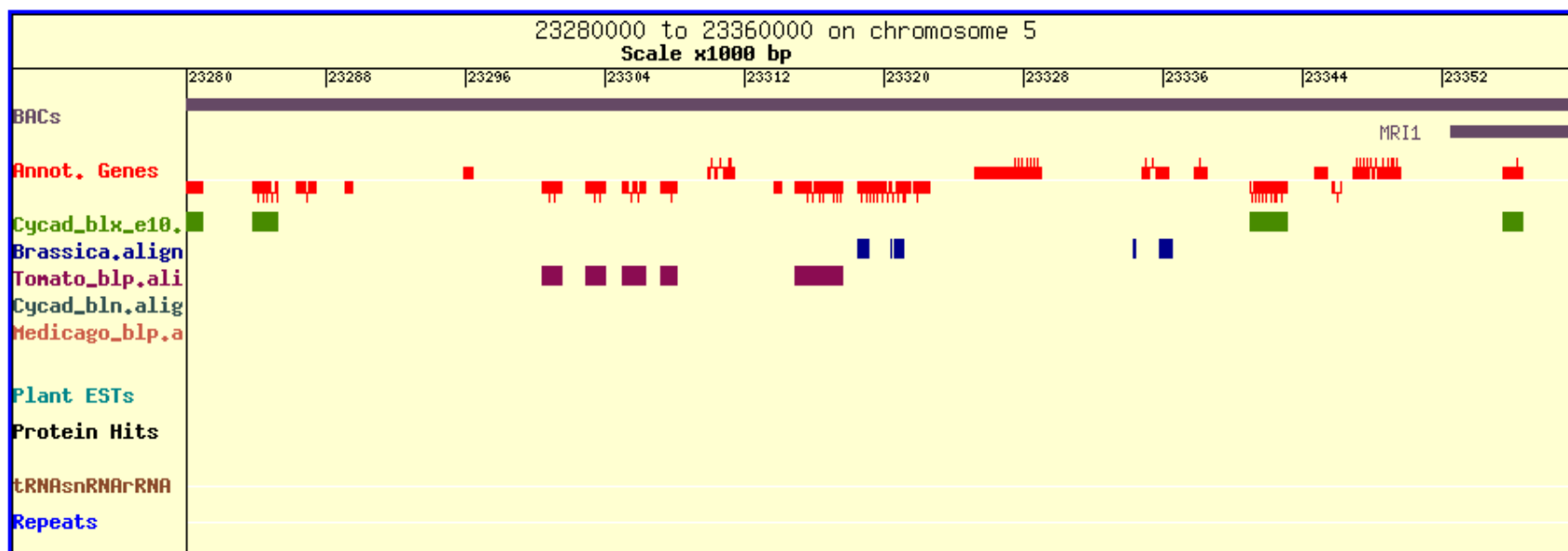
# The Arabidopsis and rice sequences can provide knowledge infrastructure for plant genomics data



## Brassica oleracea and Arabidopsis thaliana Alignment

[HOME](#) | [FIGURE](#) | [GENES](#)

### From 23280000 on Chromosome 5



<<< << < > >> >>>

**COORDINATE SEARCH**

Select Chromosome :  1  2  3  4  5

Start Coordinate

Choose scale (bp/pixel)

Window Size (pixels)

Detail

**NAME SEARCH**

gene  reads

Protein Code/Read Name

Choose scale (bp/pixel)

Window Size (pixels)

# Finishing Strategy

Run software to identify misassembled regions

Use PHRAP/CONSED to sort misassembled regions  
-add tags and re-phrap  
-manually sort read pairs

Run software to pick

Plasmids for transposition reactions

Custom primers for BAC sequencing

Assemble new sequences into database  
-re-assemble all data with PHRAP  
-use Add New Reads function of CONSED  
for highly repetitive sequences

# Detailed Strategy - Overview

- Large insert clones (1-3 plates per BAC)
- Worklist generated by computer
- Biomek FX to rearray clones based on worklist and barcoded plates
- Rearranged plates subjected to transposon mutagenesis on Biomek FX
- Transposants put through production queue
- Data reassembled to complete project
- Remaining areas finished by direct sequencing on BACs

# Finishing Software

Identifies Problem sequence areas:

- Misassembled reads/regions
- Gaps
- Low Quality Regions

Picks

- double stranded templates
- primers

Interfaces with mySQL database to

- track progress in finishing projects

# Sequencing using Transposons

**Transpose individual/pooled plasmids using the GPS-1 system from NEB**

- Tn7 transposon
- random insertions
- bi-directional priming sites attached

**Transform plasmids containing Tn7  
Plate on double selection media**

**Pick and prep 12-24 transposed clones per plasmid with Amersham Temphlphi kit**  
-number of clones depends on gap size and number of spanners available  
-- Temphlphi kit provides simple prep and clean sequencing templates

**Sequence transposed clones with both bi-directional primers  
Use Big Dye terminator reactions**

# Individual and Pooled Transposition Reactions Give Similar Performance

Clones	Contig no.	Individual (Clones assembled)	Pooled (Clones assembled)
A1-A12	13	9	9
	14	9	9
	15	6	6
	16	6	6
	17	3	2
B4-B12	14	7	4
	15	36	35
	16	32	31
	17	0	4

# Gap Closure with Transposons

Clone	# of gaps	# of plasmids transposed	# of transpositions per plasmid	# of gaps closed
OSJNBa0048F08*	7	10	24	4
OSJNBa0004A10#	3	7	12	2

\*Two gaps in OSJNBa0048F08 had no plasmid spanners available. One gap had a small insert plasmid which extended into the gap but did not close it. Spanners were chosen manually. Plasmids were in pOT vector.

#OSJNBa0004A10 had mis-sorted read pairs across the last gap which prevented it's closure in the first round. Spanners were identified by the finishing software. Plasmids were in Pzero vector.

# Sequencing from BACs

- Suggested by data from RGP at last years meeting
- 1 DNA preparation per BAC
- Computer generated list of oligos
- Oligos ordered in 96 well plates and reacted in production mode (highly automated)
- Data assembled and analysed
- This process requires no intervention with personnel at the finisher level until completed

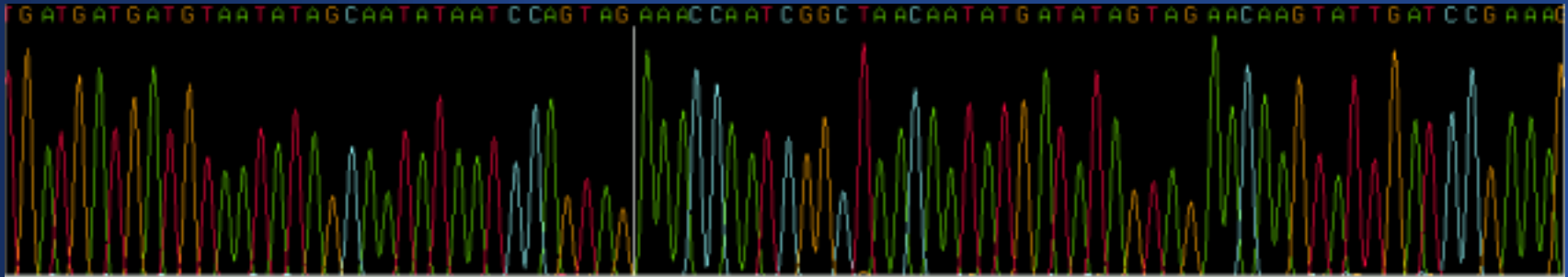
# Sequencing From BAC

**Isolate BAC DNA from 250 ml culture  
~150 µg DNA obtained**

**Nebulize the isolated BAC DNA  
To get -.5-1 KB fragments**

**Sequence transposed clones with custom primers  
Use 1 µg DNA per reaction  
Use Big Dye terminator reactions**

## Direct Sequencing from BAC DNA



# Direct BAC Sequencing Statistics

Clone	# of primers	# of reactions	Success rate (%)	Mean Read Length (bp)	# of reads into assembly
OSJNBa0019A16	2	2	100	580	2
OSJNBa0071I20	9	9	95	585	8
RP23-37A14	48	48	58*	400*	28*
OSJNBa58P118	26	96	80	-	77

\* Failures attributed to instrument problems

# Summary of 58P18 Direct BAC Sequencing First Pass

- 4 of 7 gaps filled
- 5 of 6 low quality areas resolved

# Future Plans

- **Plans to further automate the process and reduce finishing time include**
  - automatic tagging and re-phraping misassembled reads
  - automatic production of worklist for finishing
  - interfacing with Biomek for assembling reaction by robot

# Advantages of this approach

- Computer generates worklist and does much of the finishing
- Simple workflow, highly automated
- Makes extensive use of high throughput production pipelines
- Minimizes need for experienced finishers (they only do the very difficult parts)

# Summary

- The finishing software substantially reduces human intervention and the requirement for highly trained finishers.
- Preliminary data indicates that a combination of transposons for large gaps and regions with difficult secondary structures and direct BAC sequencing for other areas will significantly reduce cost.
- Per person finishing of 750,000/person/month has been achieved
- This rate should be at least doubled within 6 months
- We estimate the cost of finishing in this manner is less than \$0.05/base

# Acknowledgements

Transposition and BAC sequencing

Lisa King

Vivekanand Balija

Finishers

Melissa De La Bastide

Lori Spiegel

Finishing Software

Sujit Dike

Genome Research Center  
Cold Spring Harbor Lab

RGP

USDA

NSF

USDOE

NHGRI