Romancing (all the) Genomes

Ernie Retzel, Ph.D.
Program Leader, NCGR
efr@ncgr.org

Café Scientifique, October 2008

National Center for Genome Resources

"Take-homes" from my training

- You don't have to be from a scientist-family to make a science-life.
- Public schools and public colleges will get you where you want to go.
- It's not how *smart* you are, but how *stubborn* you are.
- **Don't give up. Ever.**

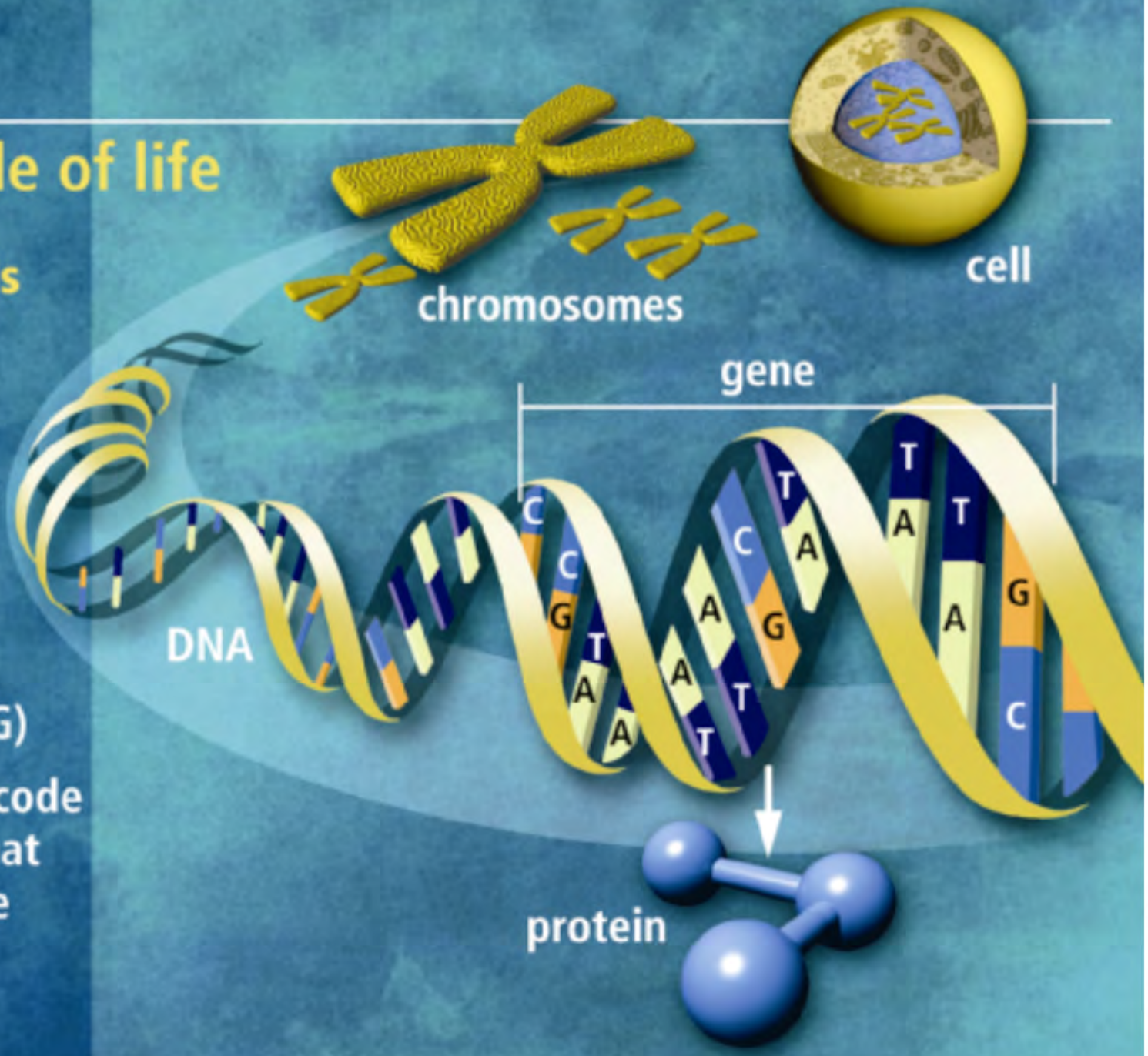
DNA

the molecule of life

Trillions of cells

Each cell:

- 46 human chromosomes
- 2 m of DNA
- 3 billion DNA subunits (the bases: A, T, C, G)
- 80,000 genes code for proteins that perform all life functions

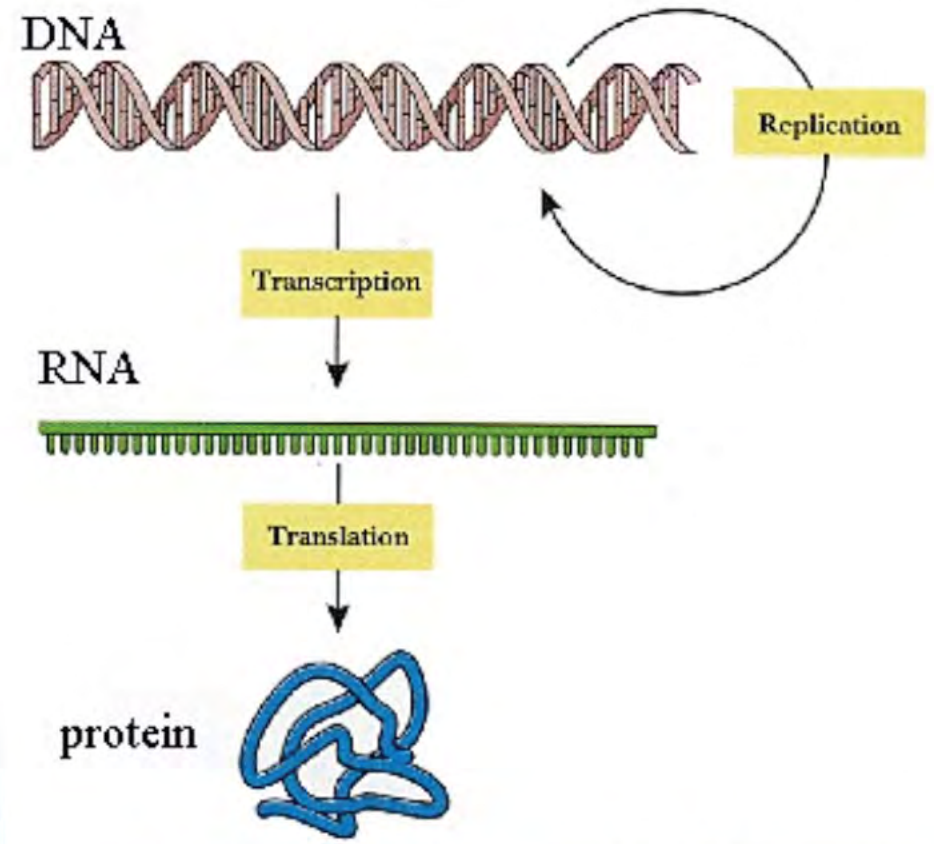


Y-GA 98-090R

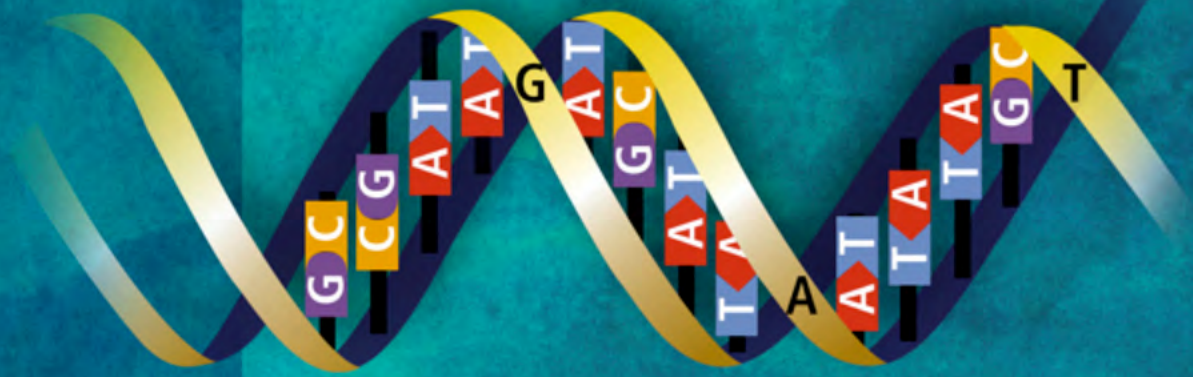
How Does A Genome Work?

- 2% genome is genes
- Active genes in cells are copied into RNA

- 1 RNA → 1 protein
- Proteins direct all bodily activities
- 98% of genome regulates the gene – protein steps



DNA Genetic Code Dictates Amino Acid Identity and Order



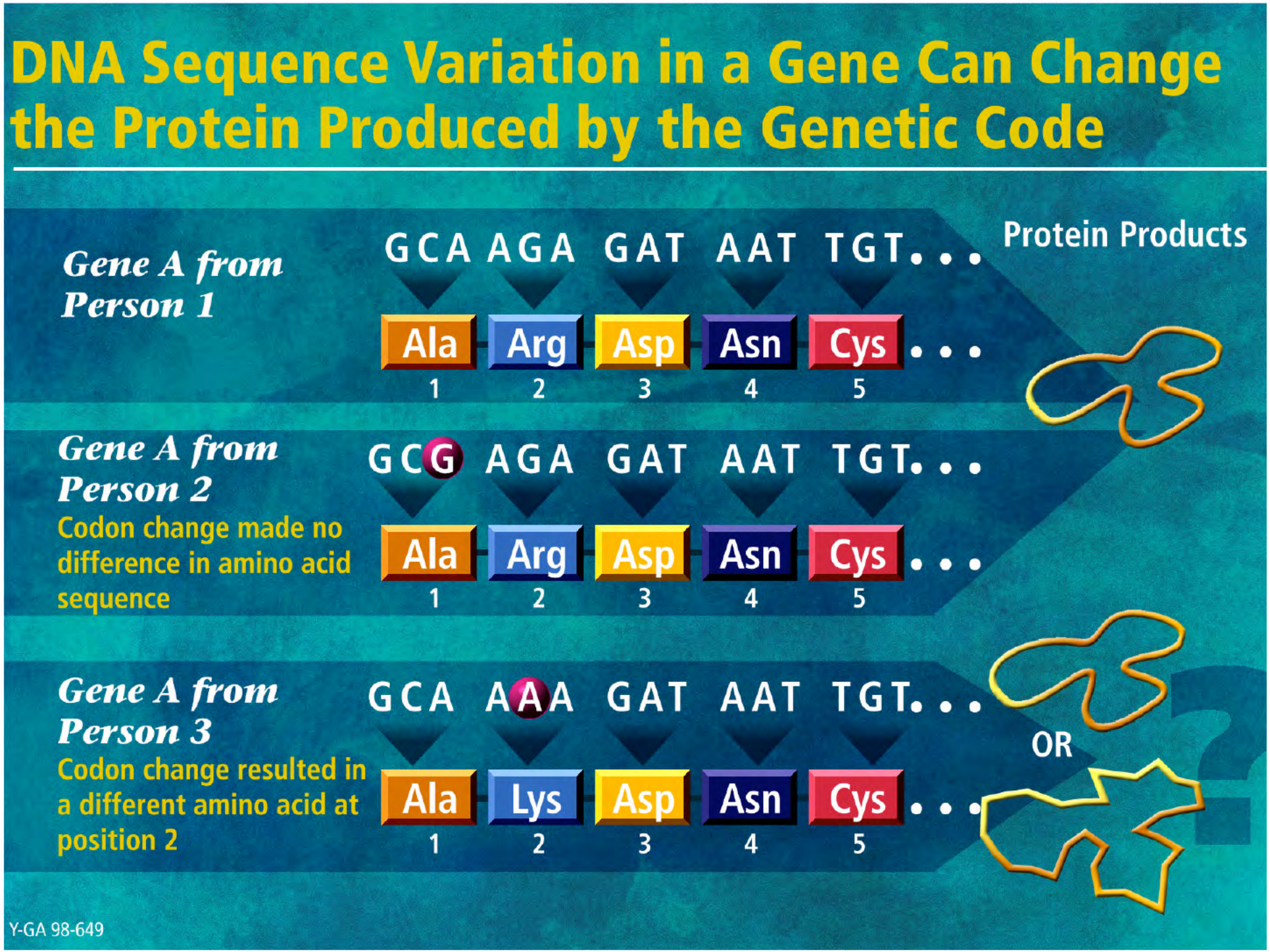
DNA Sequence Is the Genetic Code.

GCA AGA GAT AAT TGT . . .

. . . *Growing Protein Chain*

Y-GA 98-648

DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code



Y-GA 98-649

Health or Disease?

DNA Sequence

Person 1 A A A T T T

Normal protein

Some DNA variations have no negative effects

Person 2 A A T T T T

Low or nonfunctioning protein

Person 3 A A C T T T

Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

YGG-00-0480

Genomes - how big? (OK, *about* how big)

Viruses - 1,400-100,000 bases

Bacteria - 3-12 million bases

Yeast - 12 million bases

Mustard weed - 115 million bases

Soybean - 1 billion bases

Human - 3 billion bases

Pine tree - 22 billion bases

Plant, animal (& human) genomes

Historically (five years ago), considered static and comprised mostly of "junk" DNA. Coding regions [those coding for genes] are a small fraction of the total (1-2%).

Contemporary reality is: there is virtually **NO** "junk" DNA, it is almost all transcribed, and may either be genes (that 1-2%) or many levels of control elements (~93%).

Sequencing technology progress

1977 - 1-200 bases per multiple months

1985 - 1000 bases per day

2008 - >1 billion bases (>1,000,000,000) per day per machine

(Human genome is 3 billion bases)

National Center for Genome Resources

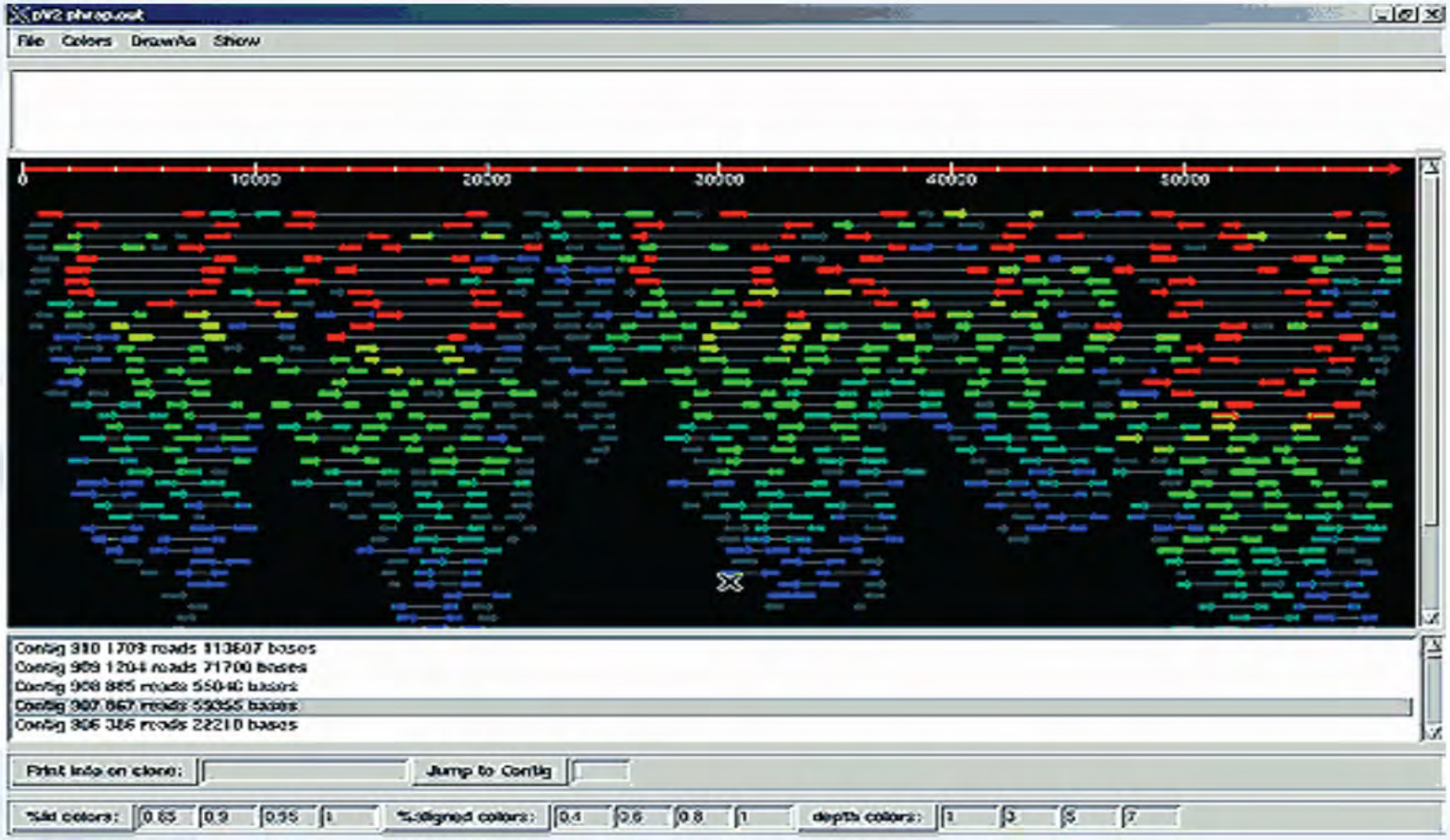
Sequencing - how-to

Fragment the genome into random, overlapping "bite-size" pieces, about 200 bases each.

With enzymes or chemistry, determine the sequence of bases in each bite-size piece.

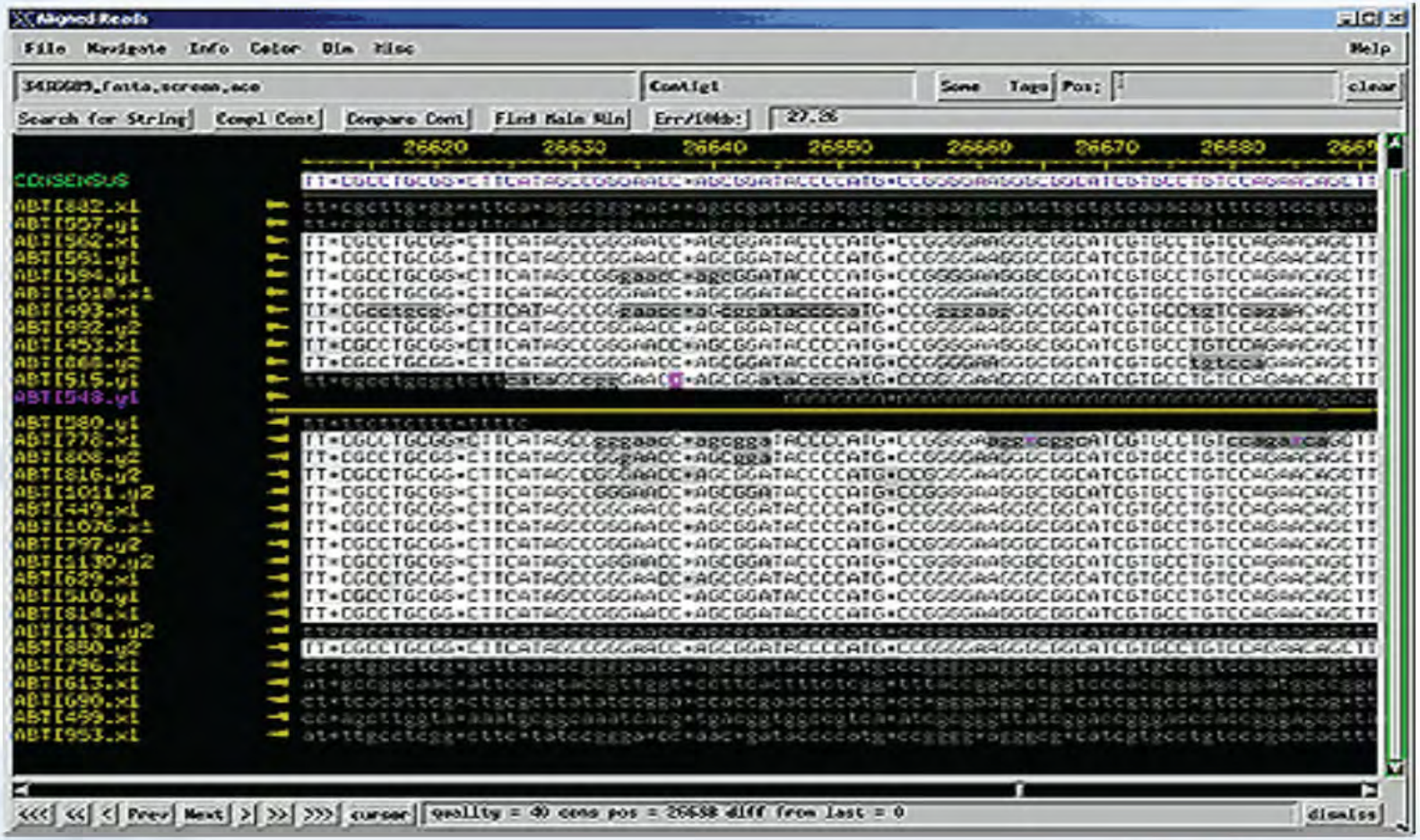
Arrange overlapping pieces by comparing their sequences.

Fragment assembly



National Center for Genome Resources

The detailed version …



National Center for Genome Resources

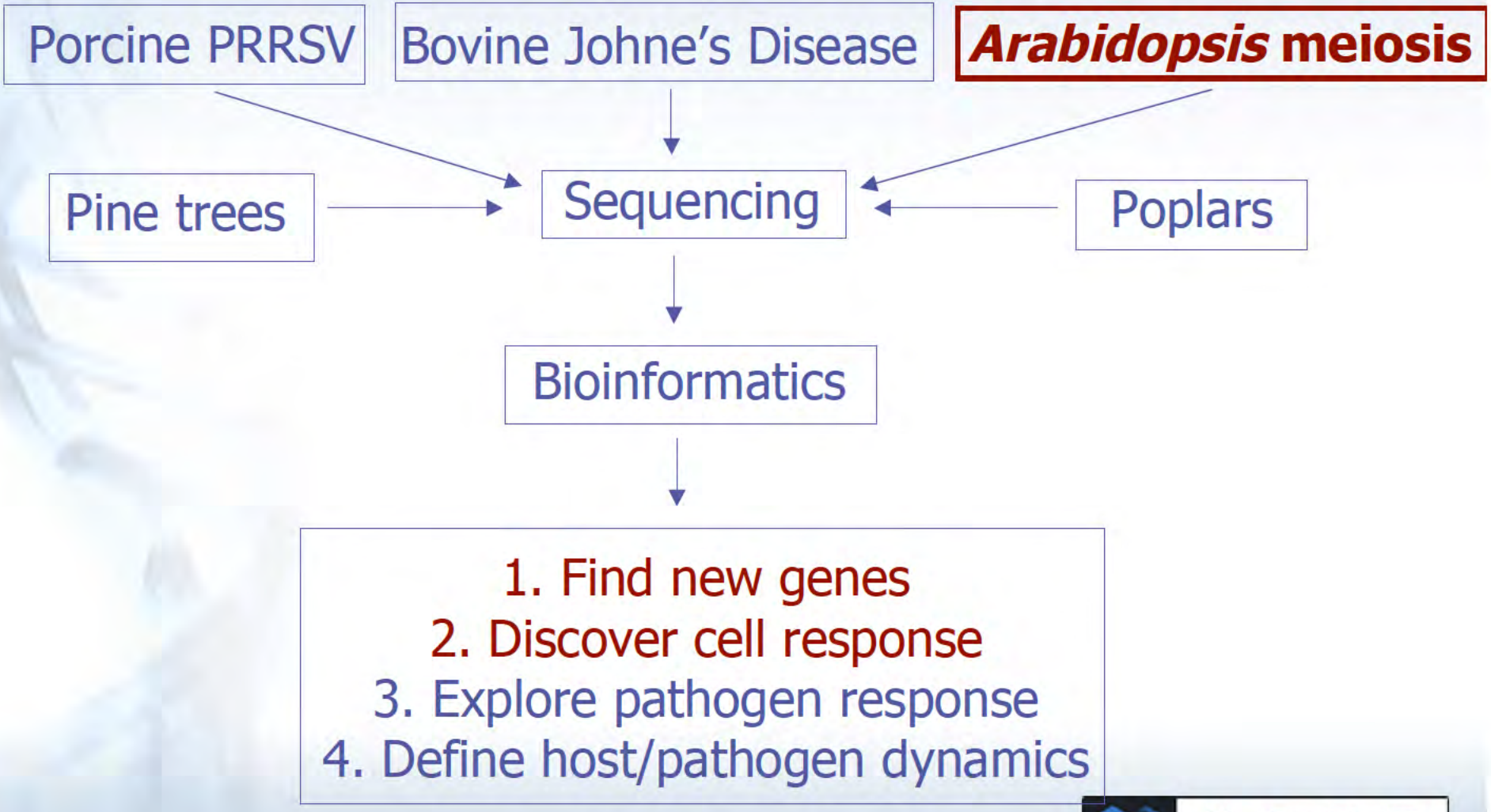
A tiny example from GenBank

Pig circovirus 2 strain YJK 0703 capsid protein mRNA, partial

```
001 cacggatatt gtagtcctgg tcgtatatac tgttttcgaa cgcagtgccg aggcctacgt
061 ggtctacatt tccagcagtt tgtagtctca gccacagctg atttcttttg ttgtttggtt
121 ggaagtaatc aatagtggaa tctaggacag gtttgggggt aaagtagcgg gagtggtagg
181 agaagggctg ggttatggta tggcgggagg agtagtttac ataggggtca taggtgaggg
241 ctgtggcctt tgttacaaag ttatcatcta gaataacagc actggagccc actcccctgt
301 caccctgggt gatcggggag cagggccaga attcaacctt aacctttctt attctgtagt
361 attcaaaggg cacagagcgg gggtttgagc cccctcctgg gggaagaaag tcattaatat
421 tgaatctcat catgtccacc gcccaggagg gcgttttgac tgtggttcgc ttgacagtat
481 atccgaaggt gcgg
```

National Center for Genome Resources

Systems I work in



NCGR

National Center for Genome Resources

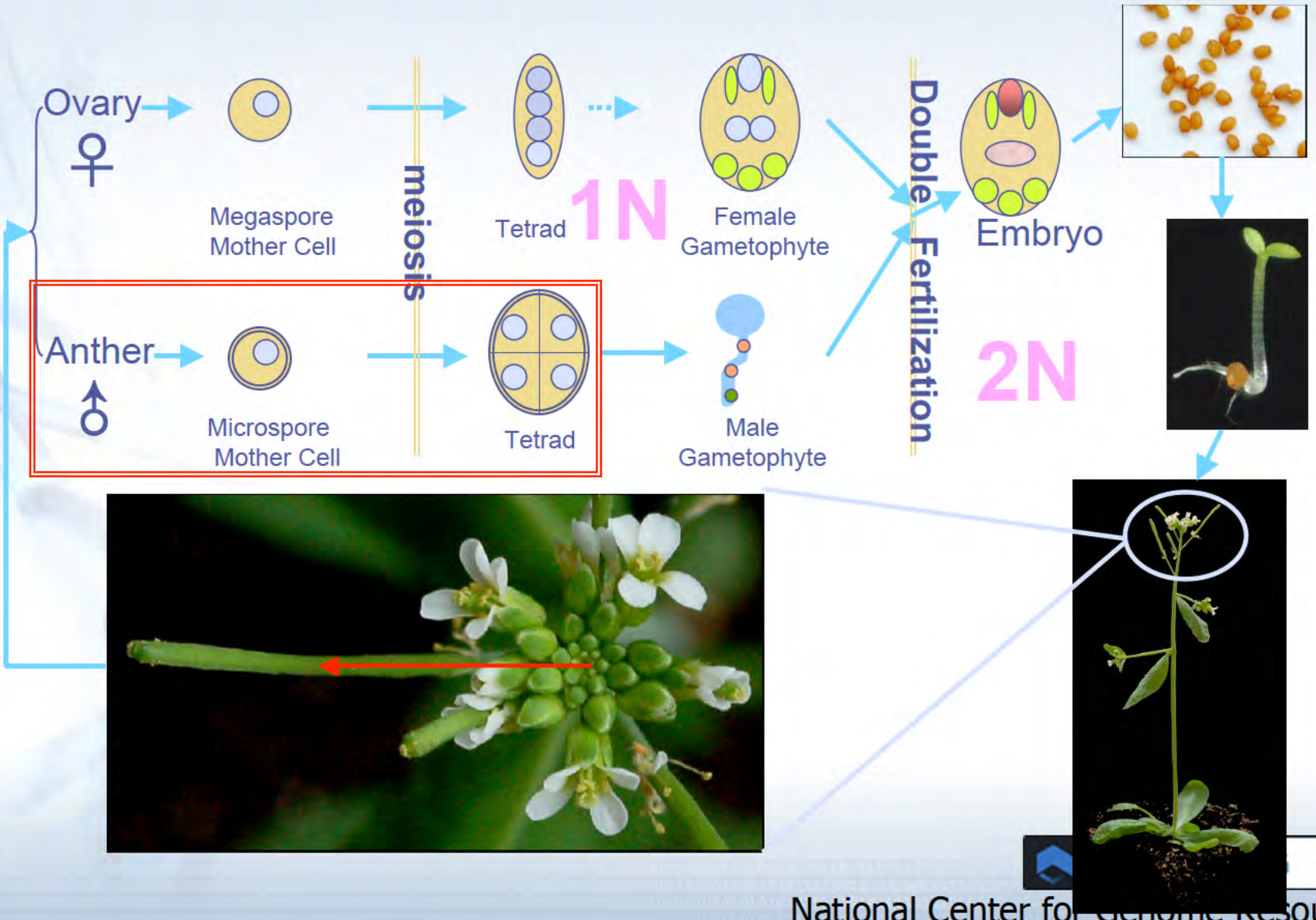
Example: Cell division in *Arabidopsis*

(Also Known as Mustard Weed)

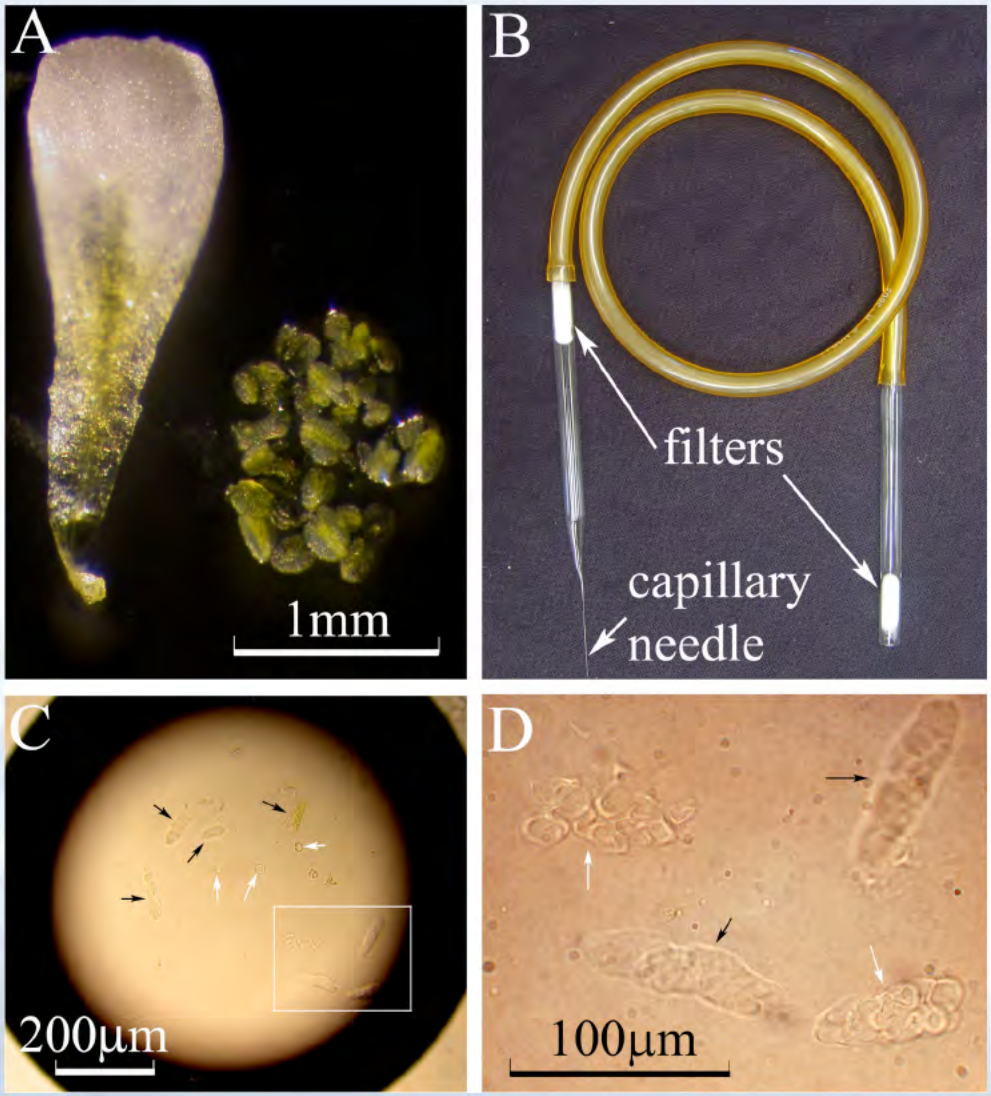
- A key feature in all sexually-reproducing eukaryotes (plants and animals).

- **Critical in re-sorting genetic information and maintaining the inheritance of traits from parents.**

The *Arabidopsis* Life Cycle



Capillary Collection of Meiocytes (CCM)



Using this method, more than 1000 meiocytes are collected per hour per person.

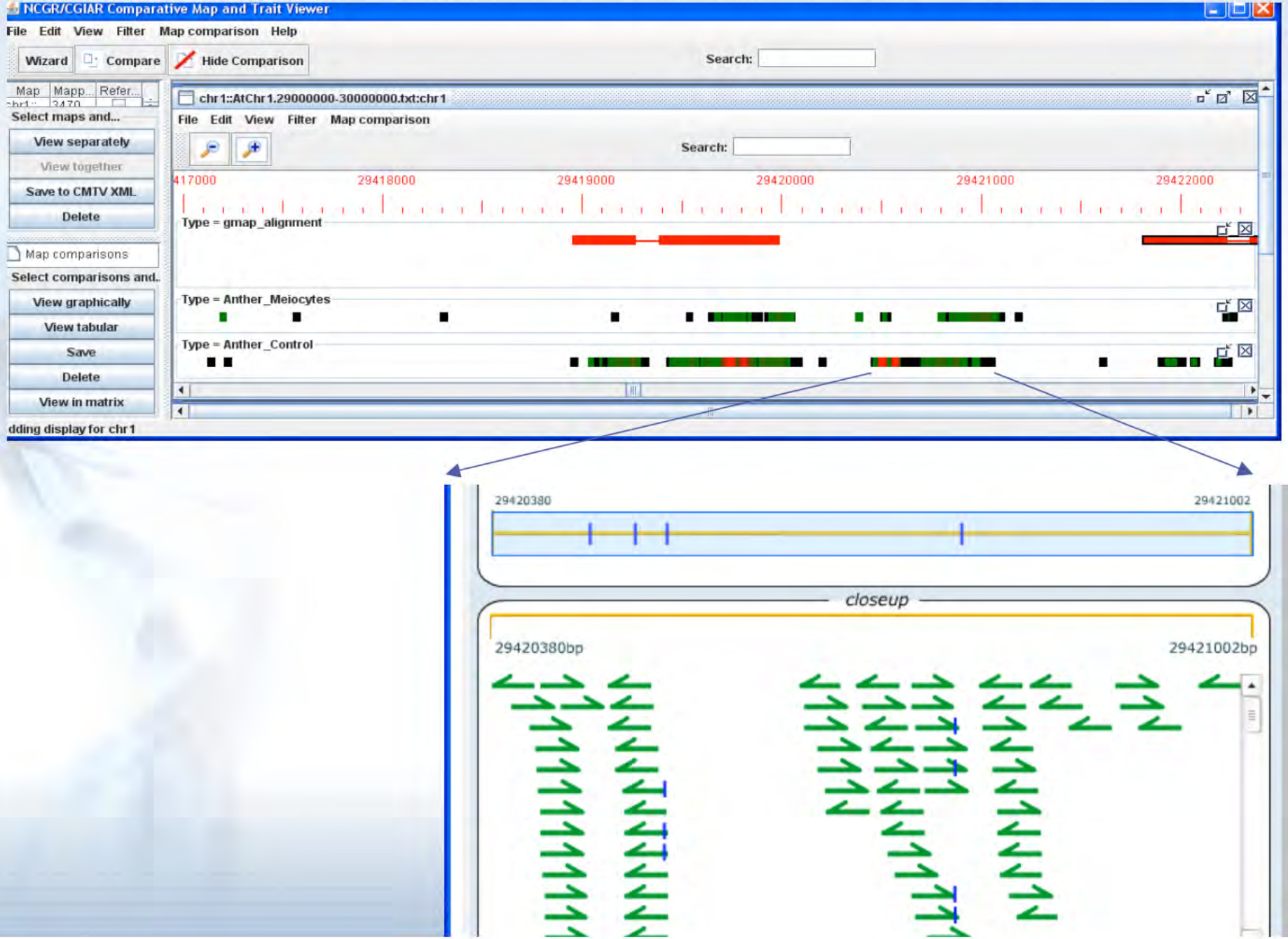
Three trails of 25 hours collection:

~15,000 (96%),
~28,000 (100%*),
~30,000 (98%).

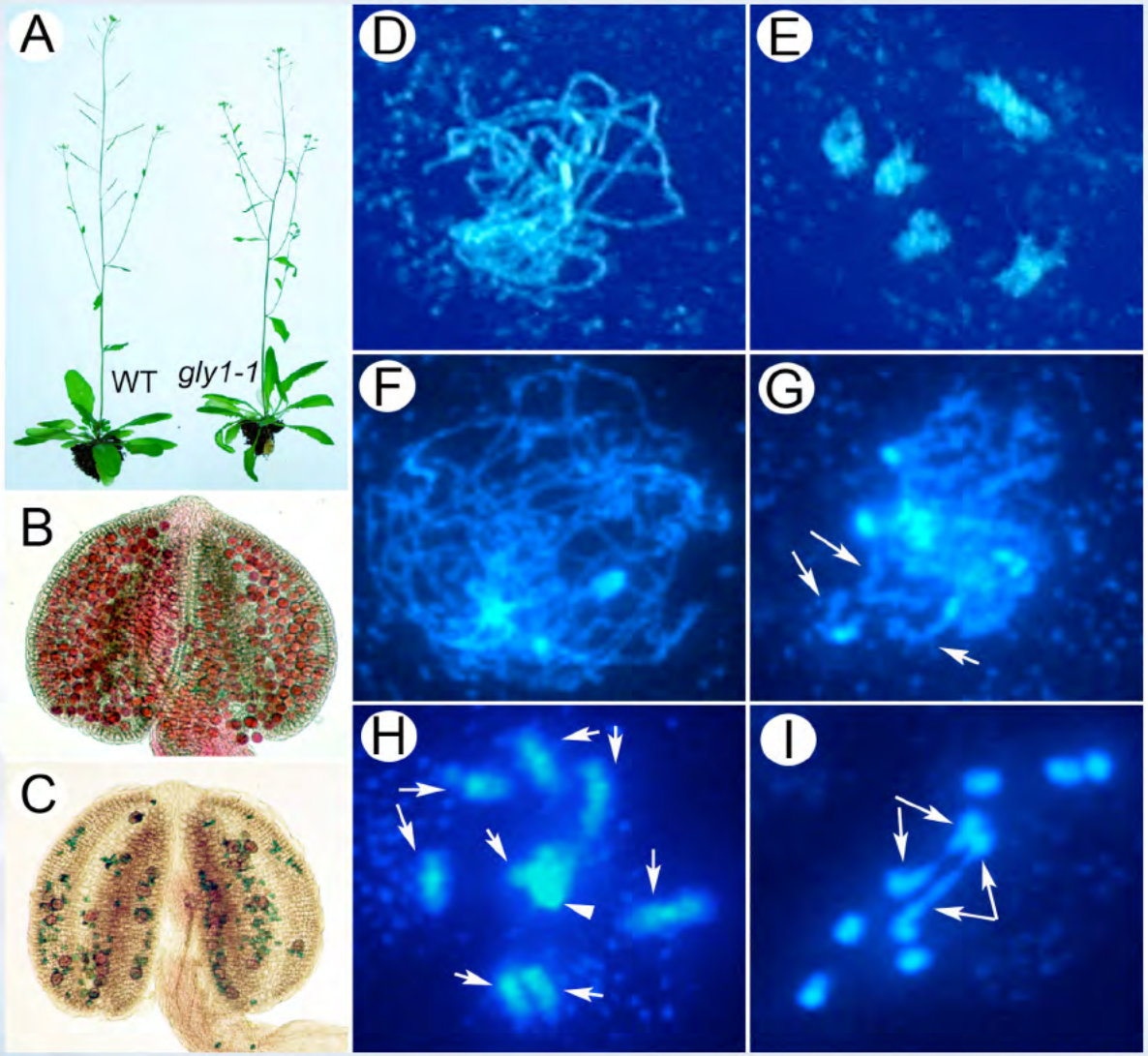
* A sample examination of this collection observed no somatic cells or microscopes.

National Center for Genome Resources

Un-described, but transcribed region



Example: Effects of a mutation in one gene on recombination



National Center for Genome Resources

Mutant Screening: Insertions in the Identified Genes

NCGR

National Center for Genome Resources

Open questions ....

- If we can sequence anything, what would you like to see sequenced?

- We are getting to the point where we can both breed and engineer plants and animals to have properties we are interested in. What do you think about that?

- We are nearly at the point where we can sequence your genome for the cost of a diagnostic test. There are positives and negatives with this possibility. What are they? And how should we deal with them?

National Center for Genome Resources

The harder you work, the harder it is to surrender.

-Vince Lombardi

National Center for Genome Resources