

# *Genome Sequence Analysis*

*Ping Xu*

Philips Institute, School of Dentistry  
Virginia Commonwealth University  
Richmond, Virginia

# Analyze complexity genome

1. Mapping
  - a) Genetic map
  - b) Physical map
2. Expressed gene analysis
  - a) EST and full length cDNA
  - b) SAGE (serial analysis of gene expression)
  - c) Transcriptome or Microarray
3. Genome sequence analysis
  - a) Filtering
  - b) Skimming
  - c) Rough draft
  - d) Complete sequence

# Genome Filtering

1. The Way to remove a large percentage of the repetitive regions of the genome.
2. Methyl filtration is one approach
3. Physical methods may also be useful  
(hybridization methods)

# Genome skimming

1. Carry out 1-3 fold coverage of the region
2. Can be whole genome or clone based
3. Clone based can therefore be targeted
4. Covers ~66 – 97% of the target sequence
5. 99% or greater accuracy on average

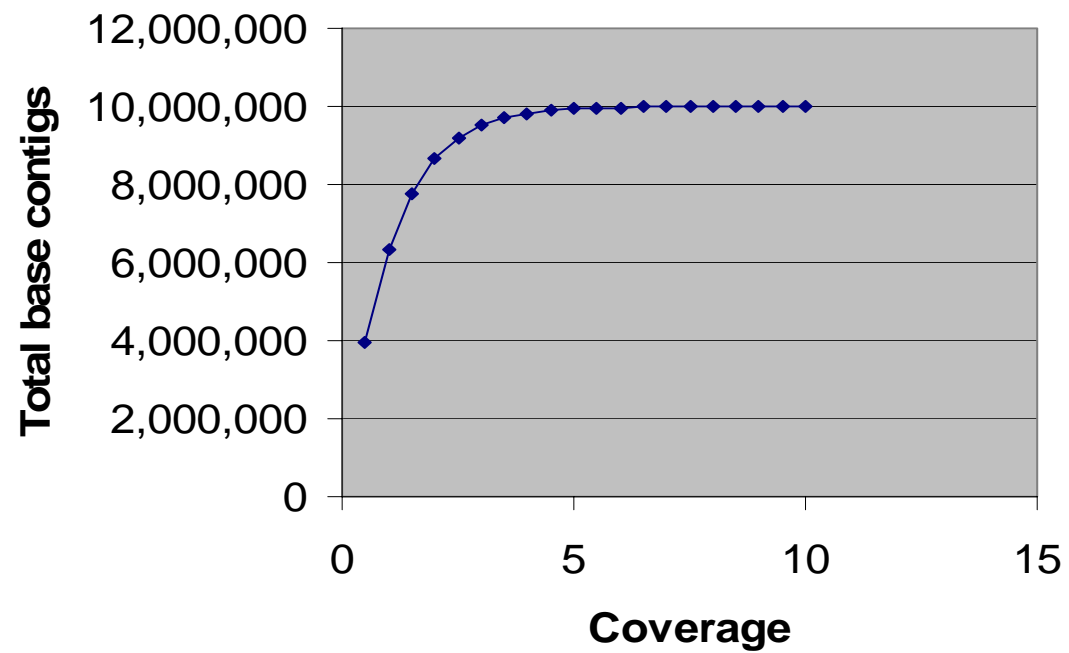
# Genome rough draft

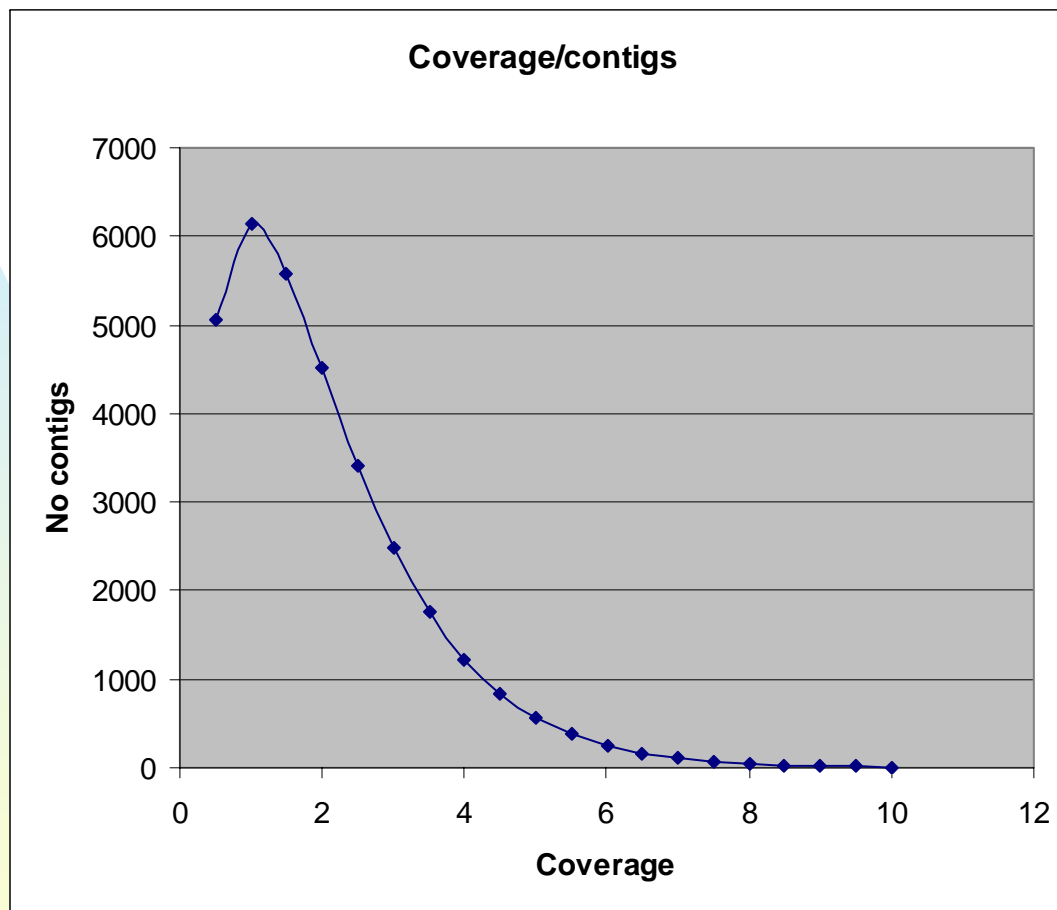
1. Can be thought of as:
  - High coverage skimming
  - Low coverage complete sequencing
2. Advantages and disadvantages are intermediate between skimming and complete sequencing (dependent on the coverage)
3. Typically 5X coverage
4. Some are proposing 10X rough draft as “finished”

# Genome complete Sequence

1. High coverage sequence, usually more than 10 X coverage
2. No gaps
3. All bases have high accuracy to “Gold Standard”
4. All bases have been double confirmed

## Length of contigs/coverage for 10 Mb genome





# Steps in large-scale Sequencing

## Subclone

Shotgun genomic library  
BAC or YAC library

## Production sequencing

Template isolation  
Sequencing reactions  
Fragment separation  
Data acquisition  
Base calling

## Finishing

Assembly  
Gap filling  
Conflict resolution  
Verification

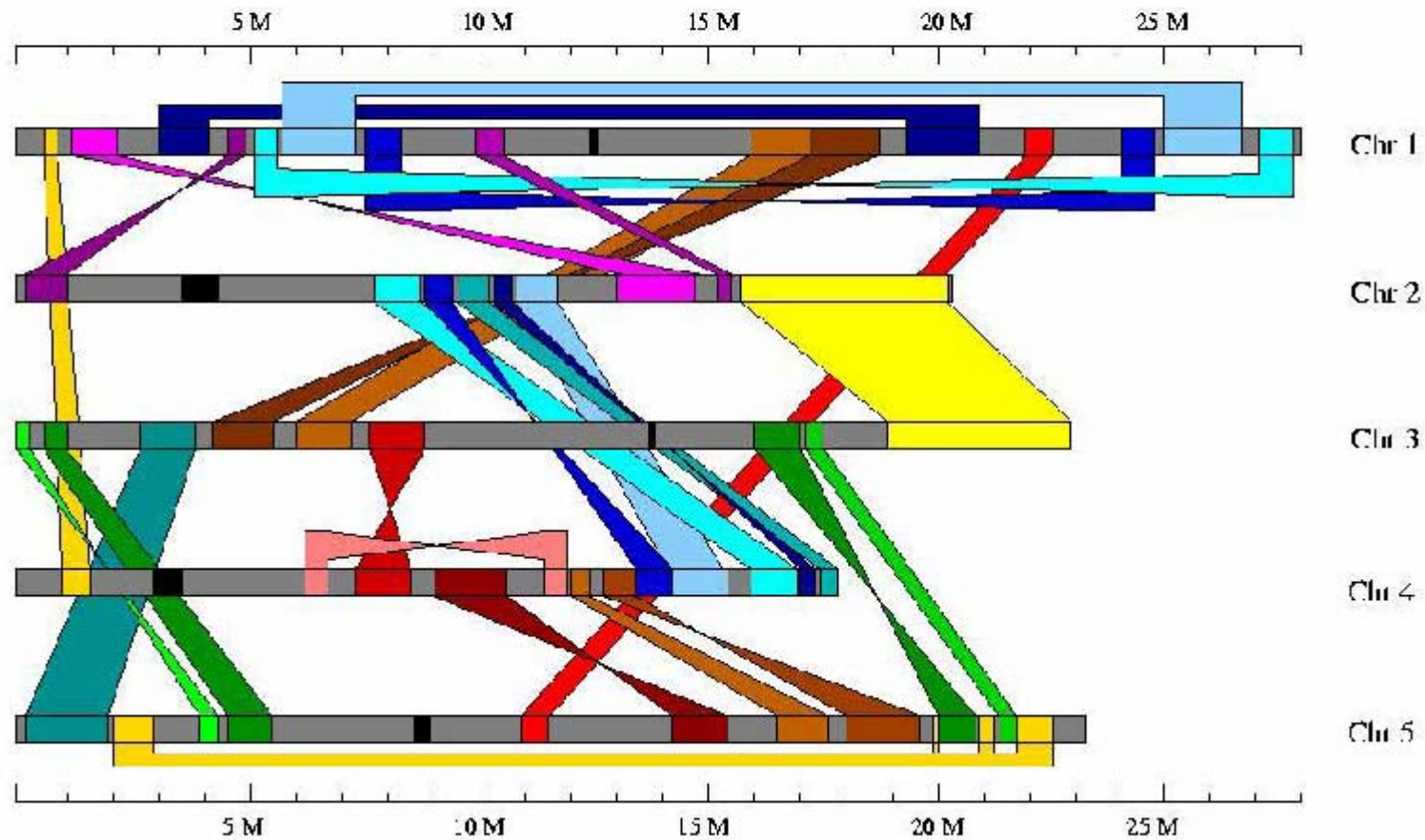
## Analysis

Gene predictions  
Homology searches  
Annotation

# Shotgun based genome sequencing

1. No up front large insert clone preparation
2. No up front mapping
3. Generates data immediately
4. Very successful on many small genomes (bacteria)
5. Difficulties on complicated genomes (e.g. repeats)

# Segmental duplications

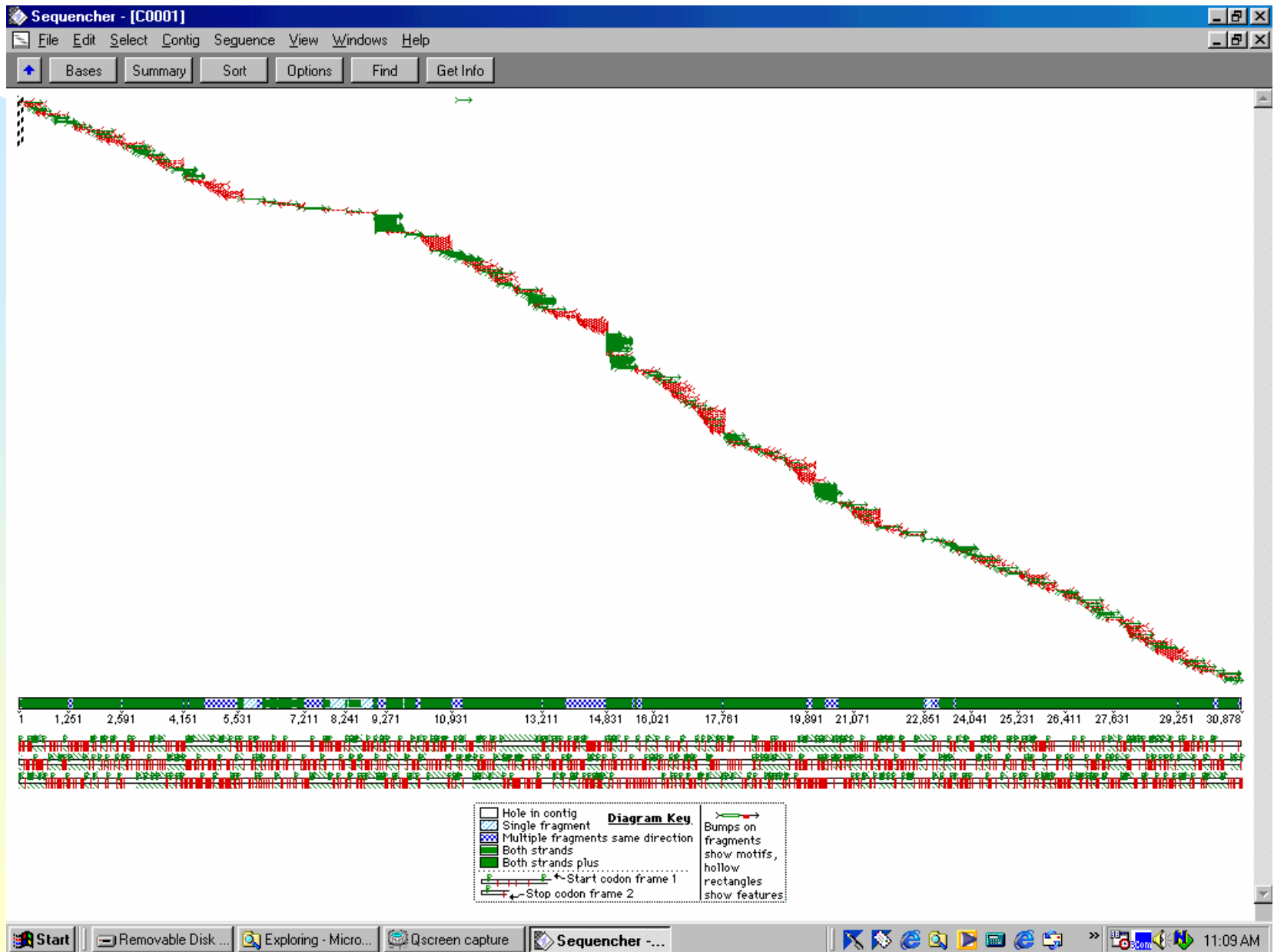


From Dr Rob Martienssen, Cold Spring Harbor Lab

Ping Xu, Philips Institute, VCU

2/16/2005

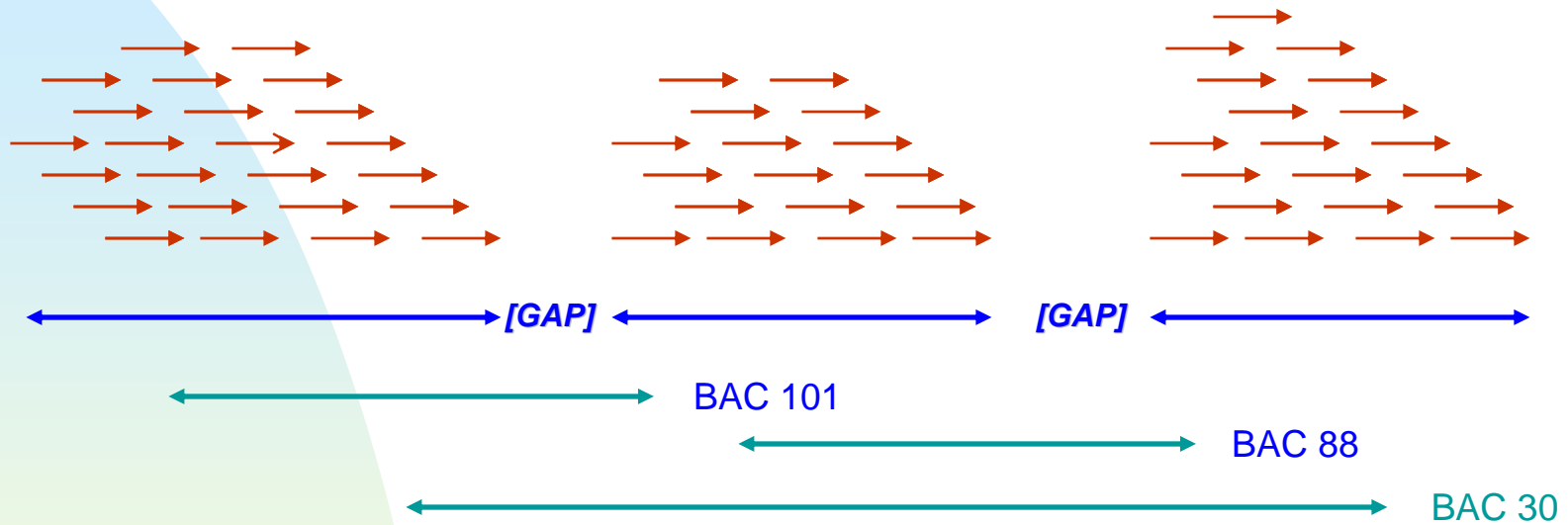
11



# Map based genome sequencing

- Large insert clone library (YAC, BAC)
- Take effort on mapping preparation
- Very successful on many large genomes with repeat (plants)
- Easy to finish whole sequence based on map

## Gap Closure Strategies



**Shotgun contigs aligned by large insert clones (BAC101 and BAC30 selected)**

# Different Maps

1. Hybridization derived map
2. BAC end or sequence tagged map
3. Whole genome restriction enzyme map  
(Fingerprint or Optical)

# Hybrid strategy for complete genome sequence

1. Combine shotgun sequencing with genome mapping
2. Shotgun sequencing to 10 X coverage
3. BAC map clones
4. Combine some percentage of sequencing of mapped clones with shotgun sequences
5. Overlay whole genome reads on reads from mapped clones when completed



# Sequence processing Phred

Phred quality score:

10 means 1 error in 10 bases

20 means 1 error in 100 bases

30 means 1 error in 1000 bases

40 means 1 error in 10000 bases

# Requiring high Sequence accuracy

1. Diagnostics, forensics, etc.
2. Protein coding predictions
3. Repeat sequencing looking for SNPs, etc.
4. Many of the high accuracy applications are actually re-sequencing, rather than de novo sequencing

# Genetics: the inherited contribution to phenotypic variation



Genetic identity



Genetic diversity

From Davis Altshuler lecture in Cold Spring Harbor Laboratory

# Correlating biological variation and variation in DNA sequence

```
aattggaagc aatgacate acagcaggtc agagaaaaag ggtt  
gagtagtagg tctttggcat taggagcttg agcccagaag  
gcccagagaga ccatgcagag gtcgctcttg gaaaagge  
tceagctgga ccagaccaat tttgaggaaa ggata  
atataccaaa tccctctgtg tgattctgct gac  
tgggatagag agctggcttc aaagaaaaat c  
ttttctgga gatttatggt ctatggaaat tttt  
gtacagcctc tcttactggg aagaatc gctt  
cgetctatcg cgaattatct aggcaggc ttat  
ctcctacacc cagccatttt tgccttcat caca  
tttagtttga tttataagaa actttaag ctg  
attggacaac ttgttagt cctttccaac aaectg  
ttggcacatt togtgtgat cgetcctttg caagtggc  
gagttgttac aggtctctgc etctctgtga cttggttcc  
caggctgggc tgggagaat gatgatgaag tacagagatc agaga  
gaaagacttg gattacctc aaaaatc cctctgtaa ggcatactgc  
tgggaagaag caatggaaaa aatgattgaa aacttaagac aaacagaact gaaactgact  
cggaaggcag cctatgtgag ataactcaat agctcagcct tctctctctc agggttcttt  
gtggtgtttt tatctgtgct tccctatgca ctaatcaaaag gaatcctctc ccggaaaaata  
tteaccacaa tctcattctg cattgtcttg cgcctggggg teacteggea atttccctgg  
gctgtacaaa catggtatga ctctcttggg gcaataaaca aaatacagga tttcttacia  
aagcaagaat ataagacatt ggaataaac ttaacgacta cagaagtgtg gatggagaat  
gtaacagcct tctgggagga gggatttggg gaattatttg agaaaacaaa acaaaacaat  
aaccaatgaa aaacttctaa tggtagtgac agcctctctc tcaqtaattt ctcactctt  
ggtaactctg tctgaaaaga tattaatttc aagatagaaa gaggacagtt gttggcgggt  
gctggatcca ctggagcagg caagacttca cttctaatga tgattatggg agaactggag  
ccttcagagg gtaaaaattaa gcacagtgga agaatttcat tctgttctca gtttctctgg  
attatgctg gcaccattaa agaaaaatc atCTTgggtg ttctctatga tgaatatag  
tacagaagcg tctcaaaagc atgccaacta gaagaggaca tctccaagtt tgcagagaaa  
gacaatatag tttctggaga aggtggaatc acactgagtg gaggtcaagc agcaagaatt
```

agaatttcat  
at T/C gtg  
gaagaggaca



3.2 billion letters of human DNA

The screenshot displays a DNA sequencing software interface with three panels, each showing sequence data and chromatograms. The panels are labeled as follows:

- Panel 1 (Top):** 10920\_a211\_g11\_08. The sequence is:
 

```

      con rd 330 335 340 345 350 355
            359 364 369 374 379 384
      con A T A T A A A A T T A A T A G T C A T T A A T A A T T T C
      edt t a a t t a a t t t a a a a a t a a a a a t a a a
      phd t a a t t a a t t t a a a a a t a a a a a t a a a
      
```
- Panel 2 (Middle):** 10910\_a205\_f09\_07. The sequence is:
 

```

      con rd 330 335 340 345 350 355
            176 181 186 191 196 201
      con A T A T A A A A T T A A T A G T C A T T A A T A A T T T C
      edt A T A T A A A A T T A A t a g t c a g t a g t a a t t t c
      phd A T A T A A A A T T A A t a g t c a g t a g t a a t t t c
      
```
- Panel 3 (Bottom):** 10910\_a172\_b01\_00. The sequence is:
 

```

      con rd 330 335 340 345 350 355
            90 95 100 105 110 115
      con A T A T A A A A T T A A T A G T C A T T A A T A A T T T C
      edt A T A T A A A A T T A A T A G T C A T T A A T A A T T T C
      phd A T A T A A A A T T A A T A G T C A T T A A T A A T T T C
      
```

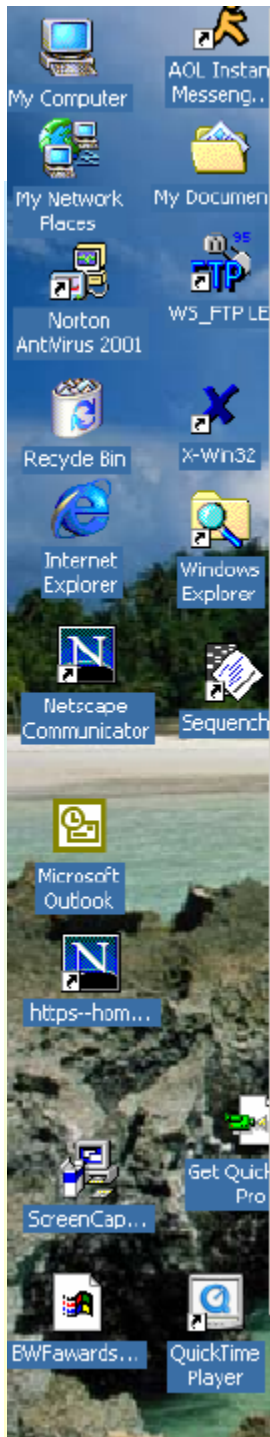
Each panel includes a chromatogram below the sequence. The software interface also features a 'Dismiss' button at the top, a 'Compare Cont' button, and a 'Find Main Min' button. A Microsoft Word window is visible in the background, showing a document with a 'Phred score' of 9 and a table of values:

→	→	18
→	→	40



# Sequence Assembly Phrap

**Contig formation + Vector masking with  
PHRAP/Crossmatch**



**Consed Main Window** [ \_ ] [ □ ] [ × ]

File    Navigate    Info    Options    Help

allsequence.fasta.screen.ace.5

Search for String    Undo Edit...    Quit Consed

Assembly View    Add New Reads

Single Click and then Click OK or  
Double Click to Select Contig

**Contig List**

Contig3027	(120 reads, 7363 bps)
Contig3028	(122 reads, 10468 bps)
Contig3029	(131 reads, 6887 bps)
Contig3030	(151 reads, 7481 bps)
Contig3031	(199 reads, 7046 bps)
Contig3032	(236 reads, 5541 bps)
Contig3033	(564 reads, 3087 bps)
<b>Contig3034</b>	<b>(770 reads, 6481 bps)</b>

**Read List**

cp010918_a212_h06_058.r	Contig3018 (~925 bps from 4629 to 55)
cp010918_a212_h07_062.f	Contig3018 (~1029 bps from 3196 to 4)
cp010918_a212_h08_074.f	Contig3007 (~919 bps from 551 to 145)
cp010918_a212_h08_074.r	Contig3007 (~935 bps from 2542 to 34)
cp010918_a212_h09_078.f	Contig3015 (~940 bps from 2232 to 31)
cp010918_a212_h09_078.r	Contig3015 (~909 bps from 1151 to 20)
cp010918_a212_h10_090.f	Contig592 (~865 bps from -77 to 786)
cp010918_a212_h10_090.r	Contig1389 (~897 bps from -50 to 846)

Find read (\*'s allowed):

Find read (old way):



# Finishing

Finishing is the process of assembling and refining raw sequence data into a highly accurate final genomic sequence. There are five finishing goals:

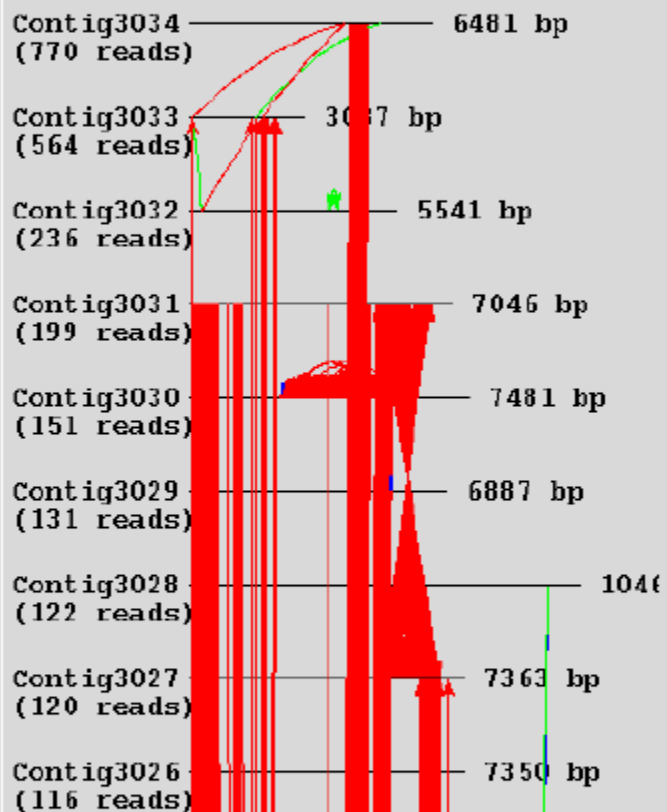
1. Filling gaps.
2. Address regions of low sequence quality that may contain sequence errors.
3. Add coverage to single-clone regions
4. Examine high quality discrepancies.
5. Confirm the sequence by comparing the restriction map *in silicon* to a real restriction fingerprint.

34104 reads total:  
 23038 in 3034 contigs;  
 0 exact duplicates;  
 11066 singletons.  
 445 chimeras.

7217 contig matches:  
 6941 problems, 276 grayzone

NOTHING SELECTED

Color code: red = problem;  
 black = ok; blue = grayzone



Show Depths	Show Contig Matches	Show Fwd-Rev Links	Show Quality
Show Reduced Depths	Show Chimera Matches	Show Same-Strand Links	Clear Display
Horiz mag: <input type="text" value="100"/>	min LLR: <input type="text" value="0"/>	min fwd-rev: <input type="text" value="0"/>	<input type="button" value="Quit"/>
Spacing mag: <input type="text" value="100"/>	max unalign: <input type="text" value="50"/>	max fwd-rev: <input type="text" value="5000"/>	
Depth mag: <input type="text" value="100"/>	qual cutoff: <input type="text" value="25"/>	min ss: <input type="text" value="0"/>	
Qual mag: <input type="text" value="100"/>		max ss: <input type="text" value="1000"/>	

# Finishing Step

1. Blast search for orientation
2. Add more reads from finishing reactions (dye terminator, dye primer, PCR and dGTP kits)
3. Edit to check quality of ends of contigs (trim or extend)
4. Create and edit a consensus of contigs and check repeat structures
5. Check forward-reverse pairs to bridge a gap
6. Along with BAC clones and compare digestion

# Gene Prediction

Based on:

1. Codon usage
2. Base composition
3. Splicing site
4. Poly A signal
5. Di, tri, hexa-nucleotide
6. Transcription signals
7. Translation signals
8. Whole genome with accuracy

# Genome Annotation

1. Repeat identification (RepeatMasker)
2. Homology Searches (BLAST or FASTA)
3. Gene prediction (GenScan, Grail or Glimmer)
4. Characterization of genes (GCG)

# Analyze sequence

1. Computational
  - Homology searches
  - Compositional analysis
  - Comparative genomics
  - Map comparisons
2. “Wet lab”
  - Individual gene analysis
  - Chip analysis
  - Knock out
  - Insertion tagging

# Why Comparative Genomics?

1. Gene knock out is impossible.
2. Controlled breeding is impossible.
3. Manipulation on genes is impossible.
4. Many specific materials are difficult to obtain
5. There is no a comprehensive collection of mutants

.

# Gene Orders on Chromosomes

Rearrangements

Clusters of genes on chromosomes

Composite genes with a multiple domains

# Proteome analysis

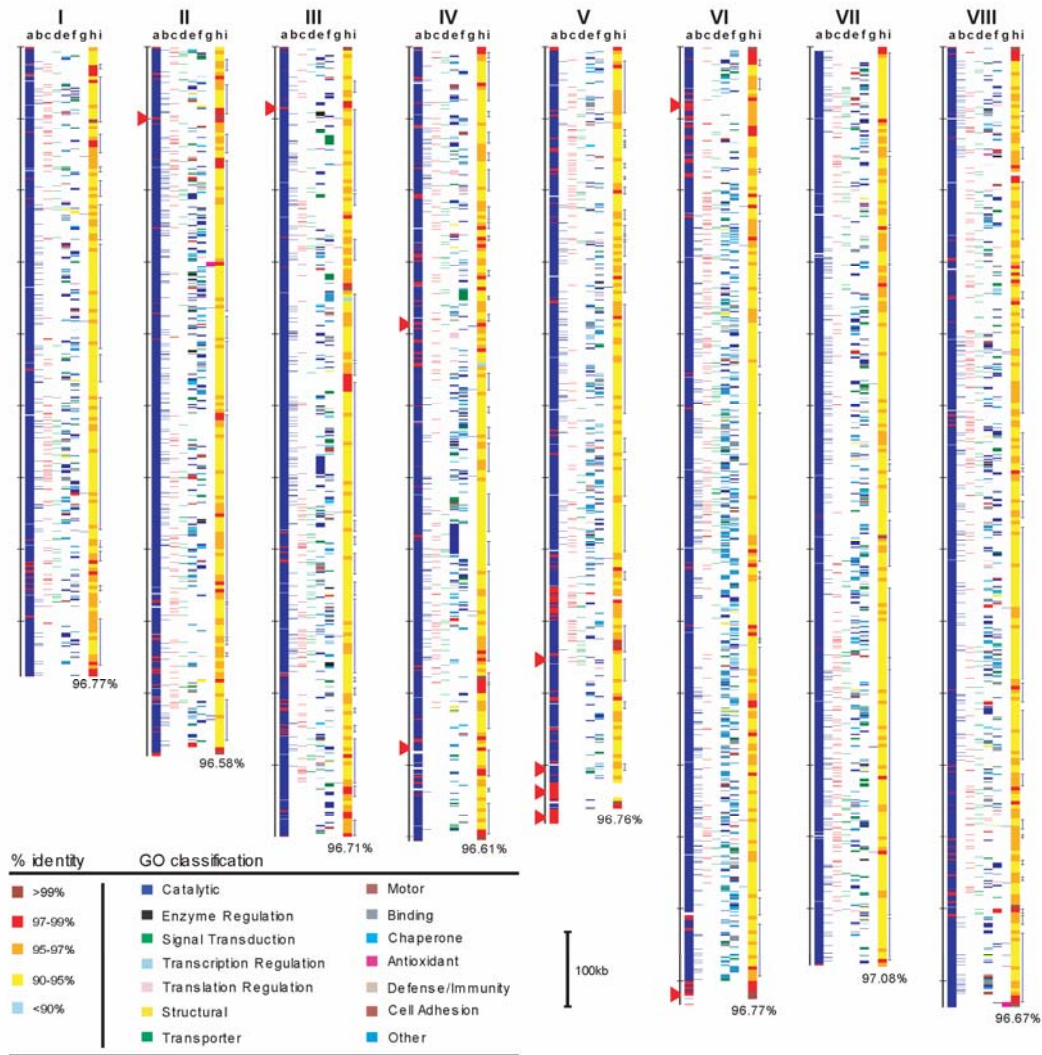
**1. All against all for gene family and duplications.**

**Find orthologs, gene family and domains**

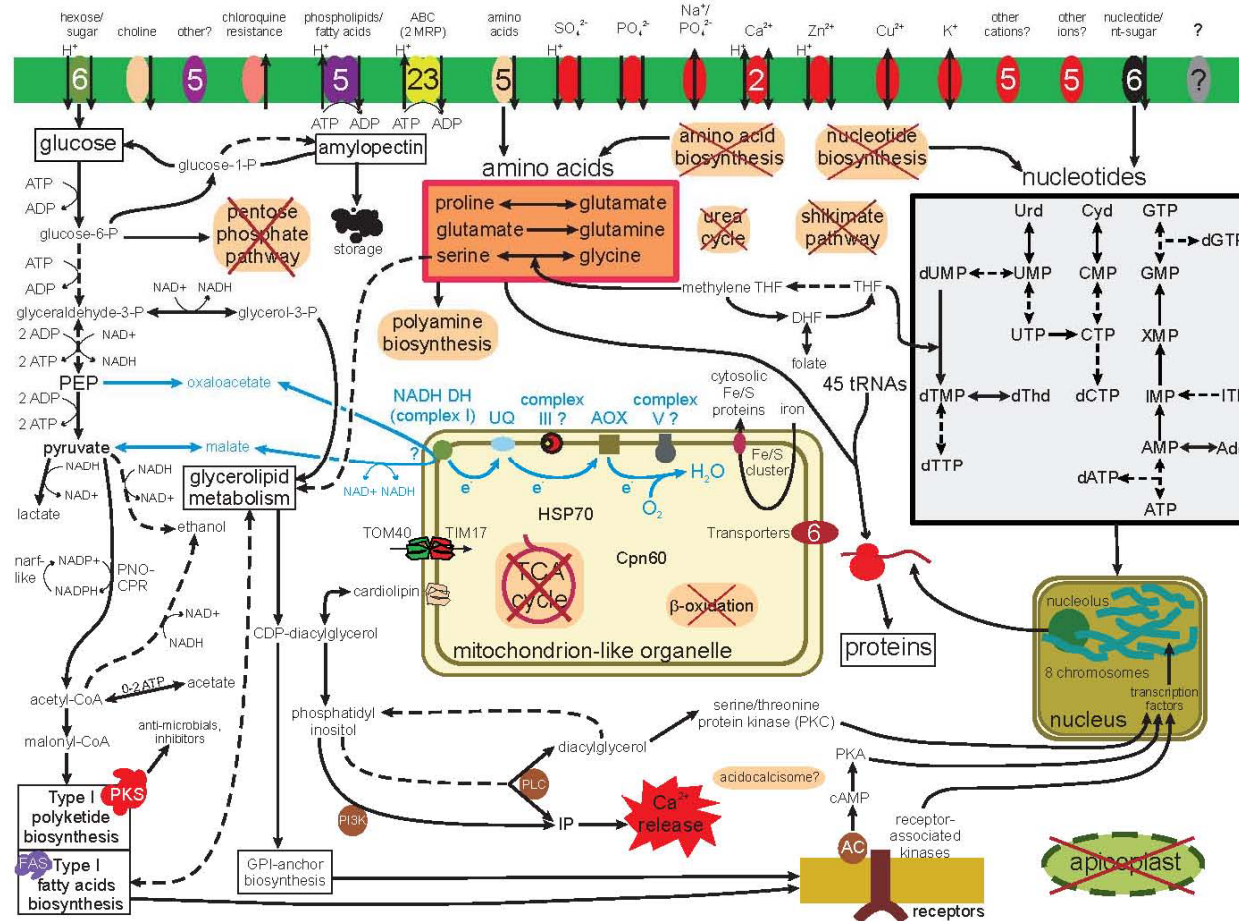
**2. Ancient conserved region analysis**

**3. Horizontal gene transfer**

**4. Phylogenetic analysis**



Xu et al., Nature. 2004 Oct



Xu et al., Nature. 2004 Oct