# Details of Whole Genome Sequencing Details

**Principals of a genome project for a species**

- *Identify the reference line*
  - Typically the line has been used historically in genetic research
  - Or, a resource is available for the line
    - ~~Clone library~~
    - Mutant lines
    - Mapping parent

**Obsolete. Reason.**
**Clones not used now!!!**

- *Isolate high quality DNA*
- *Apply some sequencing approach that fits the budget*
- *Collect DNA reads*
- *Assembly the reads into*
  - Contigs
  - Scaffolds

**Or Hi-C reads!!!**

- *Order scaffolds using a genetic map into pseudochromosomes*
- *Annotate the genes*
- *Use the reference genome in research*
  - Discover candidate gene(s) that control a phenotype
  - Develop markers to tract important genes/regions of the genome

REFERENCE
SEQUENCE

**Approach to the Actual Genome Sequencing**

- Fragment the genomic DNA
- *Old school*
  - ***Clone those fragments into a cloning vector***
  - ***Isolate many clones***
  - ***Sequence each clone***
- ***Today***
  - ***Massively parallel sequencing***

**How is Whole Genomic Sequencing Different than Gene Sequencing?**

- ***The scale of the effort***
  - Example
    - Public draft of human genome
      - Hierarchical sequencing
      - Based on 23 billon bases of data
    - Private project (Celera Genomics) draft of human genome
      - Whole genome shotgun sequencing approach
      - Based on 27.2 billon clones
      - 14.8 billion bases

**Result:**

- Human Genome = 2.91 billion bases

## Changes That Facilitated Genomic Sequencing

- **Sequencing**
  - Massively parallel sequencing is universally adopted
- **Major changes over time**
  - Thermostable polymerase enzymes
    - Improves quality of sequencing products
  - Fluorescently labeled nucleotides for the reaction
    - Allows for laser detection
  - Massively parallel sequencers
    - Millions to billions of simultaneous reads
- **Robotics (early on a major innovation)**
  - Key addition to genomic sequencing
  - Human hand rarely touches the clone that is being sequenced
  - Robots
    - Pick subclones
    - Distribute clones into reaction plates
    - Create the sequencing reaction
    - Load the plates onto the capillary detection system
  - Result
    - Increased the quality and quantity of the data
    - Decreasing the cost
      - Dropped over 100 fold since 1990
    - Improvements felt in small research lab
  - Not as important for massively parallel sequencing

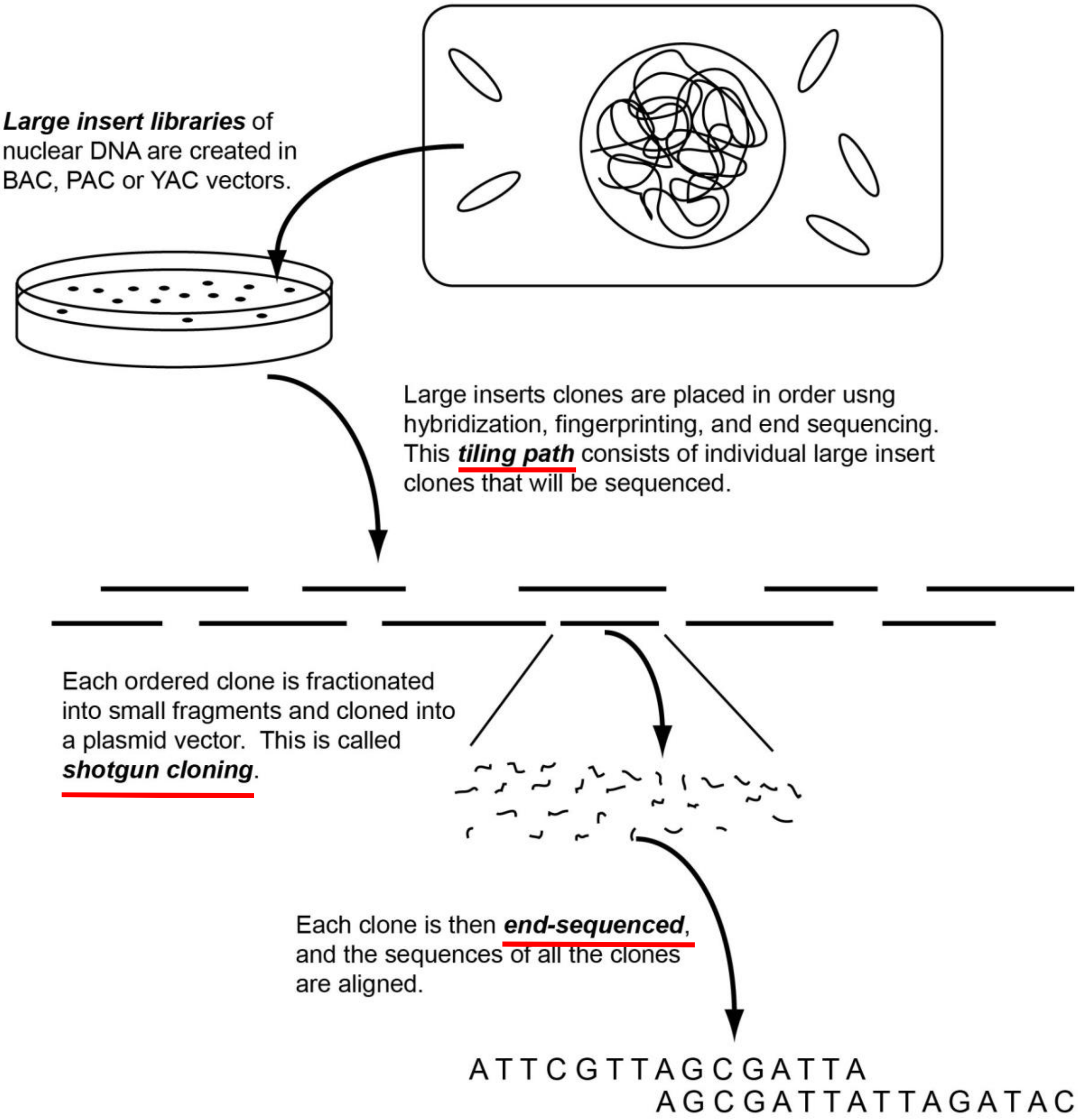
**Sequencing Methods**

- **Two major sequencing approaches**
  - Hierarchical shotgun sequencing
  - Whole genome shotgun sequencing
- **Hierarchical shotgun sequencing**
  - Historically
    - First approach
  - Why???
    - *Techniques for massively parallel sequencing not developed*
    - *Sophisticated sequence assembly software not availabile*
- **Concept of the approach**
  - **Necessary to carefully develop physical map of overlapping clones**
    - Clone-based contig (***contig***uous sequence)
  - Assembly of final genomic sequence easier !!!!!
  - Contig provides fixed sequence reference point
- But
  - *Advent of sophisticated software permitted*
    - Assembly of a large collection of unordered small, random sequence reads might be possible
  - Lead to **Whole Genome Shotgun** approach

# Hieracrchical Shotgun Sequencing of Genomes

## A. The Concept

Hierarchical shotgun sequencing requires that large insert libraries be constructed. A series of these clones are ordered by several techniques. Once these clones are ordered, each clone is separately fractionated into small fragments and cloned into plasmid vectors. The plasmid clones are sequenced, and the sequence is assembled. This is the procedure used to sequence the ***Arabidoposis*** genome, and by the public project to sequence the human genome.



***Large insert libraries*** of nuclear DNA are created in BAC, PAC or YAC vectors.

Large inserts clones are placed in order usng hybridization, fingerprinting, and end sequencing. This ***tiling path*** consists of individual large insert clones that will be sequenced.

Each ordered clone is fractionated into small fragments and cloned into a plasmid vector. This is called ***shotgun cloning***.

Each clone is then ***end-sequenced***, and the sequences of all the clones are aligned.

**Steps Of Hierarchical Shotgun Sequencing**

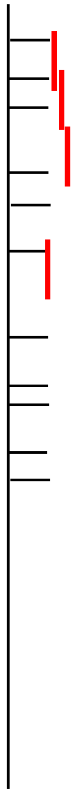
- Requires large insert library
    - **Clone types**
        - YAC (yeast artificial chromosomes)
            - Megabases of DNA
            - Few (several thousand) overlapping clones necessary for contig assembly
            - But
                - YACs are difficult to manipulate
                - Most research skilled with bacteria but not yeast culture
            - **Rarely, if ever, used today**
        - **BAC or P1 (bacterial artificial chromosomes)**
            - Primary advantages
                - Contained reasonable amounts of DNA
                    - about 75-150 kb (100,000 – 200,000) bases
                - Do not undergo rearrangements (like YACs)
                - Could be handled using standard bacterial procedures

## Developing The Ordered Array of Clones

- **Using a *Molecular Map***
  - DNA markers
  - Aligned in the correct order along a chromosome
  - *Genetic terminology*
    - *Each chromosome is defined as a **linkage group***
  - Map:
    - Is reference point to begin ordering the clones
    - Provides first look at sequence organization of the genome

### Tomato High Density Marker Collection

| Chromosome | # Markers |
|---|---|
| 1 | 363 |
| 2 | 310 |
| 3 | 242 |
| 4 | 238 |
| 5 | 158 |
| 6 | 202 |
| 7 | 191 |
| 8 | 173 |
| 9 | 184 |
| 10 | 160 |
| 11 | 149 |
| 12 | 136 |



BAC clones

Marker Map

**Genetic approach**
**Minimum Tiling** path $$$$$

**Marker Type** Early map; few SNPs

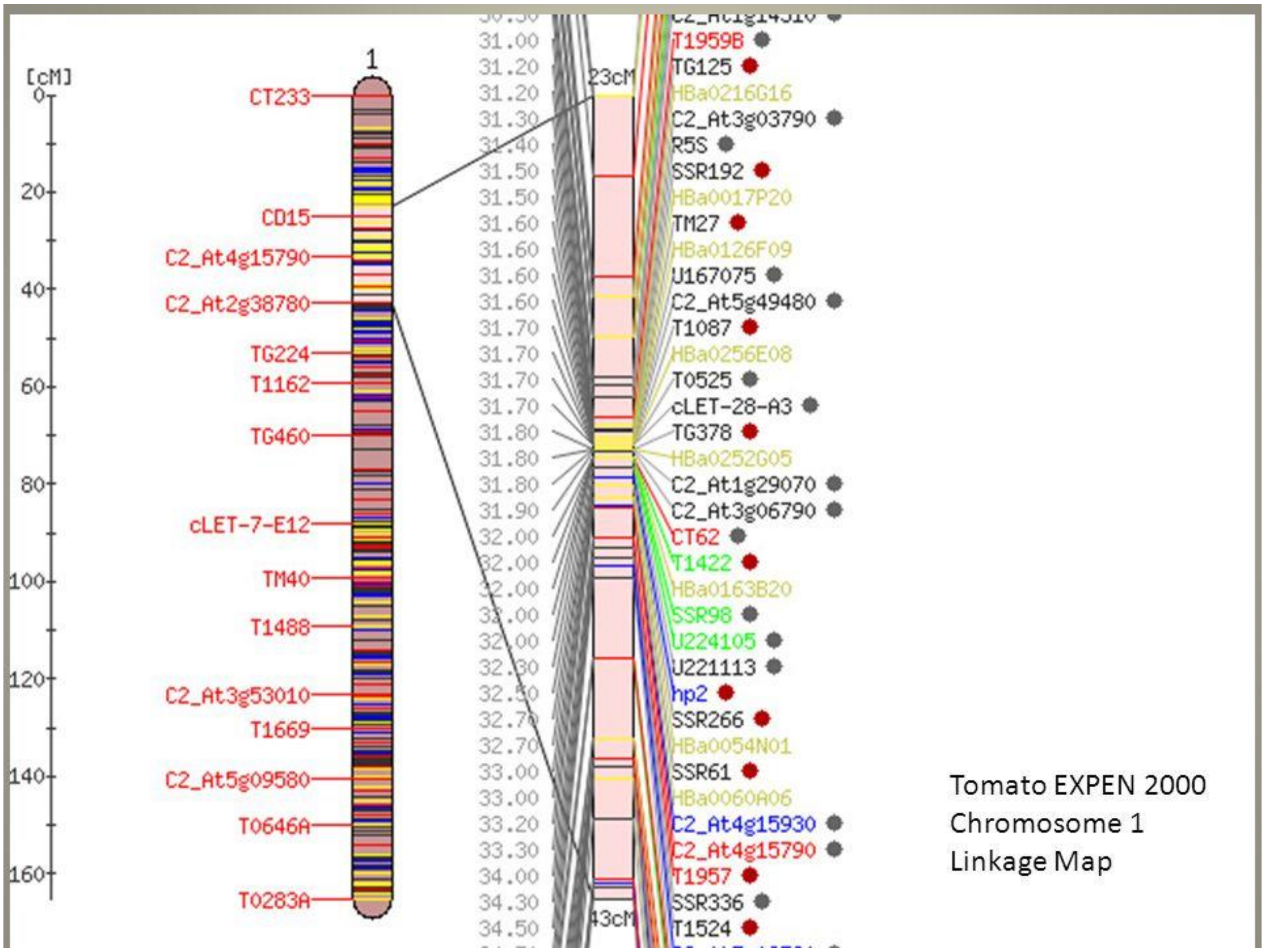
- CAPS (1088), RFLP (1342), SNP (19), SSR (155)

**Mapping population**

- 88 F2 individuals

**Today**

- Exclusively SNP markers (n>6,000 SNPs)
- Populations larger (n>200 individuals)



Tomato EXPEN 2000
Chromosome 1
Linkage Map

**Developing a Minimal Tiling Path**

- **Definition**
    - *Fewest clones necessary to obtain complete sequence*
- **Caution is needed**
    - *Clones must be authentic*
    - *Cannot contain chimeric fragments*
        - Fragments ligated together from different (non-contiguous) regions of the genome
    - How to avoid chimeras and select the minimum path
        - *Careful fingerprinting*

Physical Minimum Tiling Path

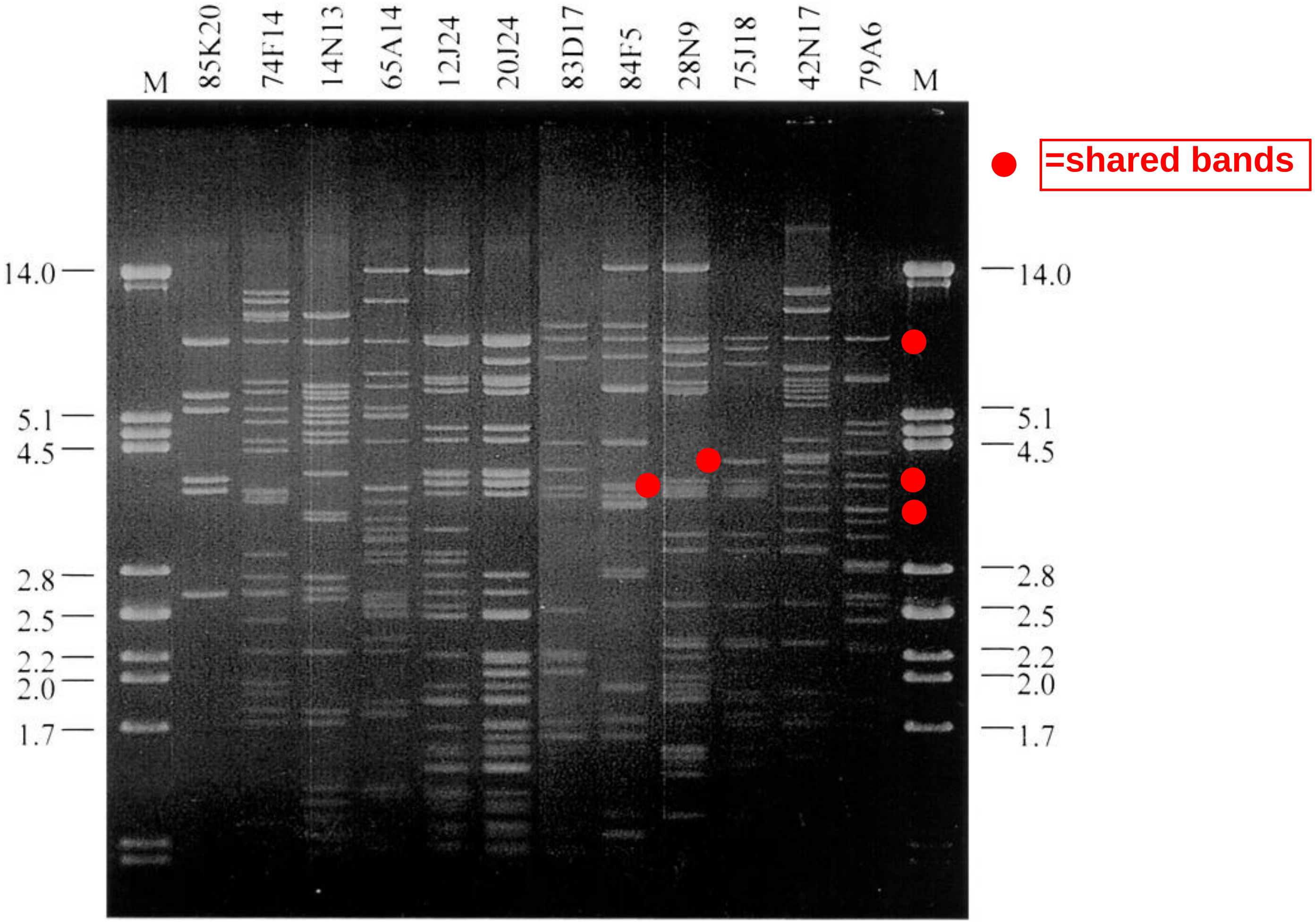
- **Overlapping the clones**
    - Maps not dense enough to provide overlap
    - ***Fingerprinting clones***
        - Cut each with a restriction enzyme (*Hin*dIII)
        - Pattern is generally unique for each clone
        - Overlapping clones defined by
            - Partially share fingerprint fragments
    - Overlapping define the ***physical map*** of the genome

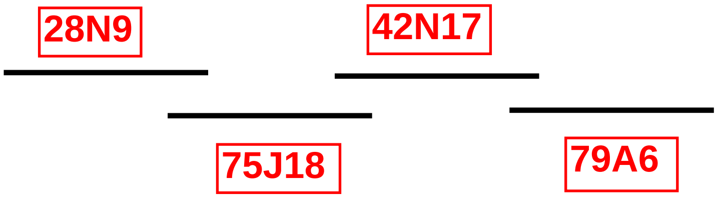
**BAC clone fingerprinting**

- Restriction enzyme digestion
- Digital, imaging, scoring and aligning

## Gel Photograph of digested BAC clones

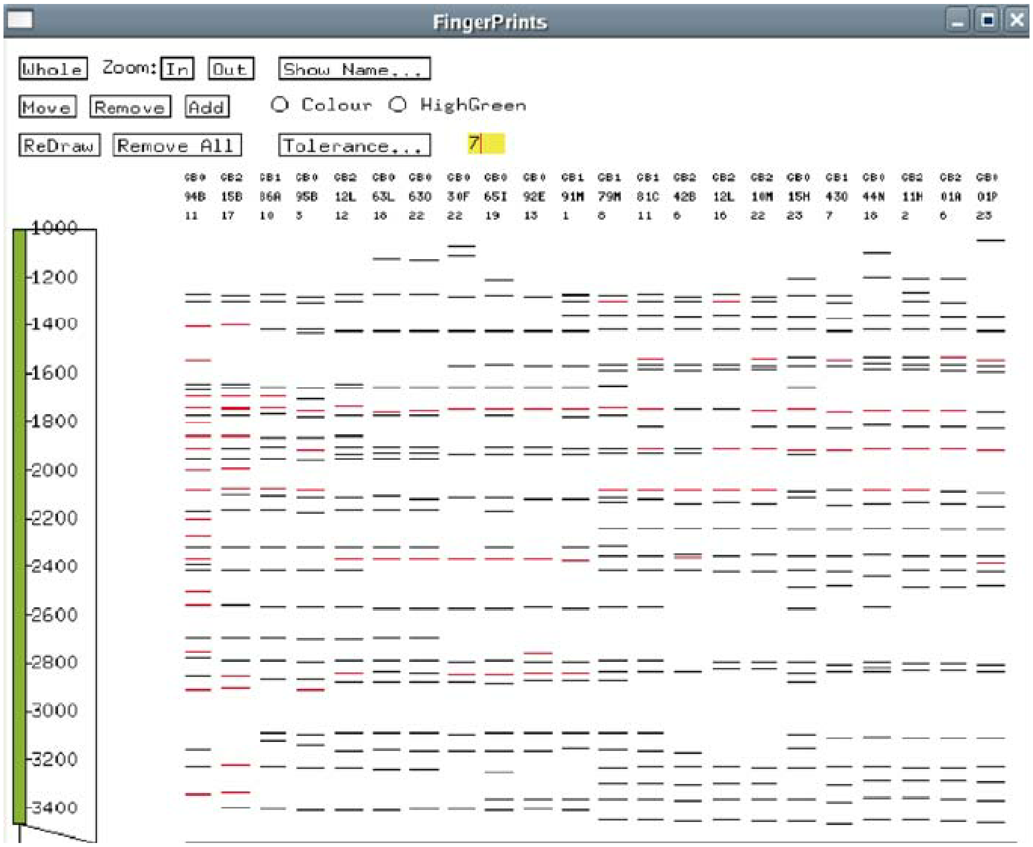
(https://www.researchgate.net/profile/T_Mirkov/publication/11896126/figure/fig3/AS:394567875612700@1471083720903/Fig-3-HindIII-fingerprinting-gel-of-the-major-BAC-clones-in-the-12-Mb-contig-that.ppm)





# Digital Image of Clones

(https://www.researchgate.net/figure/6535067_fig5_Fig-5-Example-of-the-clone-order-fingerprints-of-a-BAC-contig-of-the-apple-physical)



Tolerance... 7

## Genomic Physical Maps

- **Human**
  - *29,298 large insert clones sequenced*
    - More than necessary
    - Why???
      - Genomic sequencing began before physical map developed
      - Physical map was suboptimal
- ***Arabidopsis***
  - *1,569 large insert clones defined ten contigs*
    - Map completed before the onset of sequencing
    - Smaller genome
      - about 125 megabases
- **Yeast**

  Cosmid clones = 50 kb
  - *493 cosmid (smaller insert clones) clones*
    - Relatively high number of clones for genome size

## Sequencing Clones Of The Minimal Tiling Path

- **Steps**
  - *Physically fractionate clone in small pieces and clone into plasmid*
  - *Sequence data can be collected from both ends of insert*
    - ***Read pairs or mate pairs***
      - Sequence data from both ends of insert DNA
      - Simplifies assembly
      - Sequences are known to reside near each other

## Assembly of Hierarchical Shotgun Sequence Data

- **Process**
  - Data collected
  - Analyzed using computer algorithms
  - Overlaps in data looked for
- **Accuracy levels**
  - Analyzing full shotgun sequence data for a BAC clone
    - Goal: 99.9% accuracy
    - 100 kb BAC clone
      - 2000 sequence reads
      - Equals 8-10x coverage of clone
      - Typical level of accuracy that is sought
    - Primary software used is Phrap
      - **Phrap** = f(**ph**)ragment **a**ssembly **p**rogram
      - Efficient for a “small” number of clones
        - Small relative to number from a whole genome shotgun approach
      - Each sequence read is assesses for quality by the companion software **Phred**

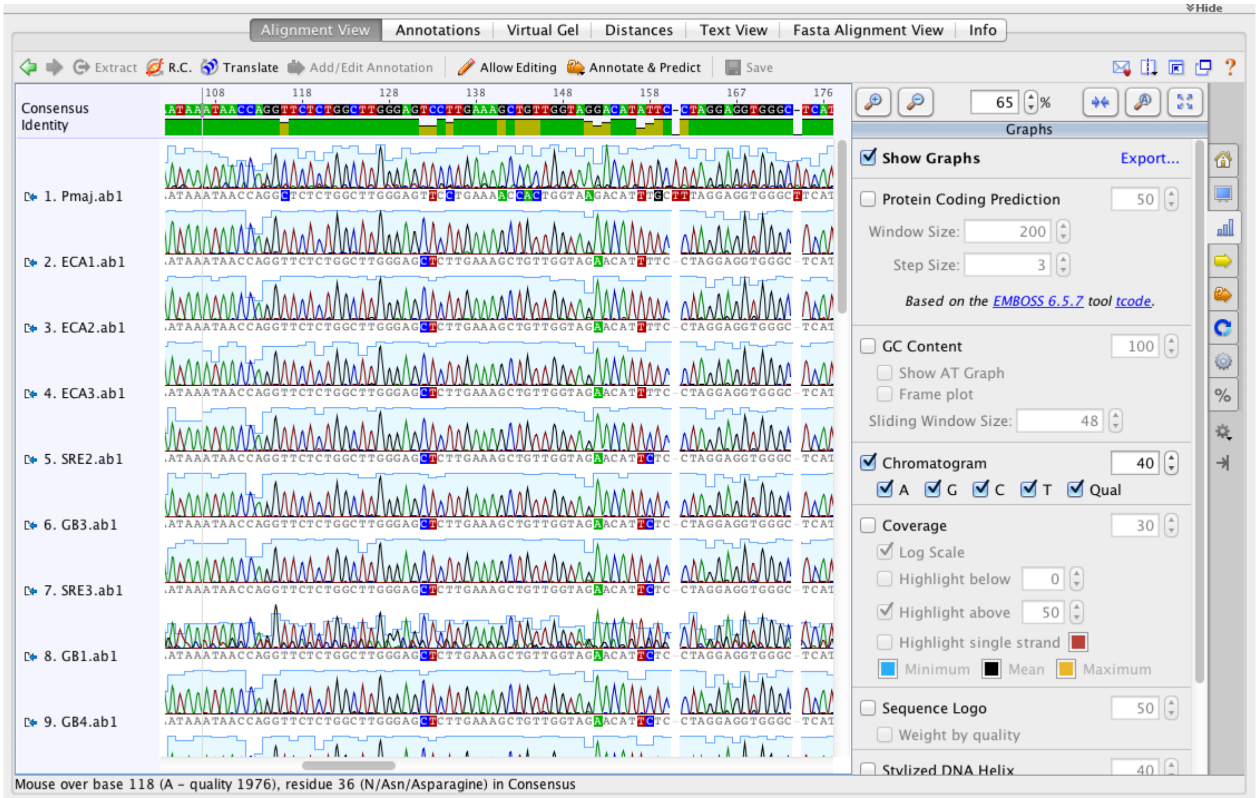
        Pfred = q>30; gold standard
      - Assembles sequence contigs only from high quality reads
  - Working draft sequence
    - 93-95% accuracy
    - 3-5 x coverage of 100 kb BAC clone

# Viewing the Quality Score Data

**Here the Phred scores are overlaid on the chromatogram of a Sanger sequencing output.**

- This is just one format the data can be visualized.
- The visualization comes from a quality score data file generated by base-call machine.

From:
http://assets.geneious.com/manual/8.1/GeneiousManualse29.html



65 %

☑ Show Graphs

☐ Protein Coding Prediction 50

Window Size: 200

Step Size: 3

*Based on the EMBOSS 6.5.7 tool tcode.*

☐ GC Content 100

☐ Show AT Graph

☐ Frame plot

Sliding Window Size: 48

☑ Chromatogram 40

☑ A ☑ G ☑ C ☑ T ☑ Qual

☐ Coverage 30

☑ Log Scale

☐ Highlight below 0

☑ Highlight above 50

☐ Highlight single strand 🟥

🟦 Minimum ⬛ Mean 🟨 Maximum

☐ Sequence Logo 50

☐ Weight by quality

☐ Stylized DNA Helix 40

Mouse over base 118 (A – quality 1976), residue 36 (N/Asn/Asparagine) in Consensus

**Finishing The Sequence**

- **Gaps need to be filled**
  - Study more clones
    - Additional subclones sequenced
    - Tedious and expensive
  - Directed sequencing of clones or genomic DNA used
    - Create primers near gaps
      - Amplify BAC clone DNA
        - Sequence and analyze
      - Amplify genomic DNA
        - Sequence and analyze

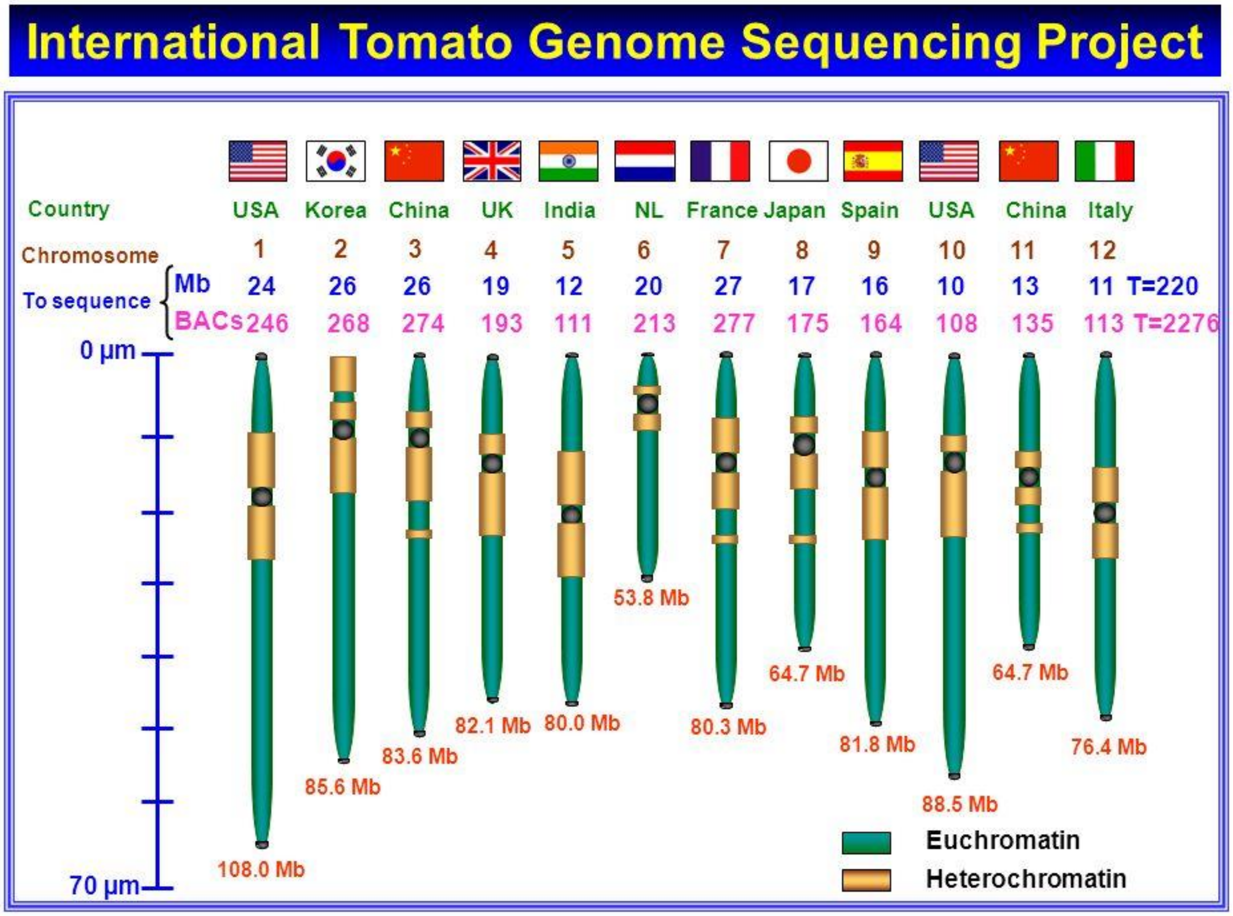
Genetic maps are the proper order of the genome

**Confirming the Sequence**

- **Molecular map data**
  - Molecular markers should be in proper location
- **Fingerprint data**
  - Fragment sizes should readily recognized in sequence data

## Tomato Genome Project

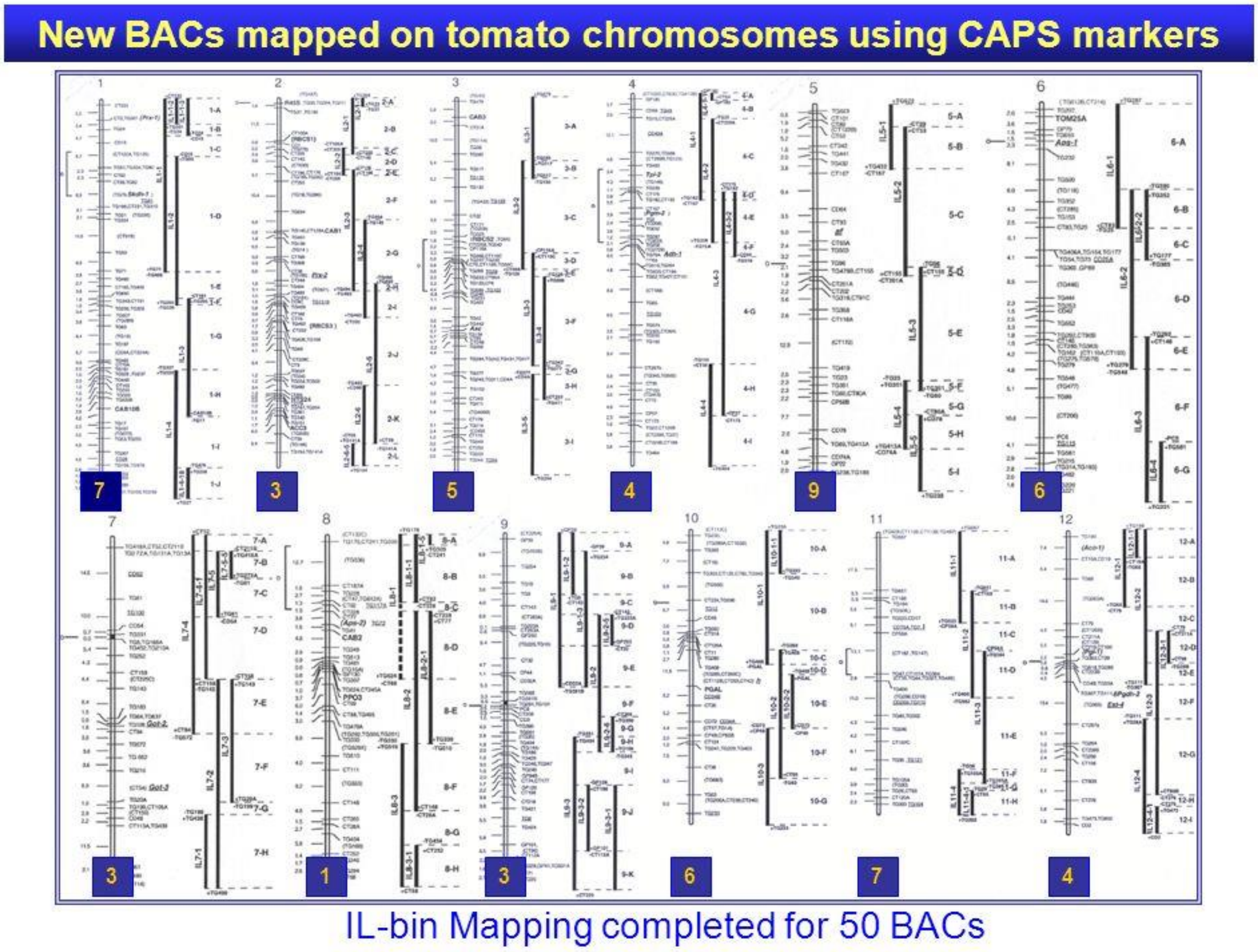
- Distribution of  BAC sequencing effort across the genome



International Tomato Genome Sequencing Project

| Country | USA | Korea | China | UK | India | NL | France | Japan | Spain | USA | China | Italy | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chromosome | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | |
| To sequence Mb | 24 | 26 | 26 | 19 | 12 | 20 | 27 | 17 | 16 | 10 | 13 | 11 | T=220 |
| To sequence BACs | 246 | 268 | 274 | 193 | 111 | 213 | 277 | 175 | 164 | 108 | 135 | 113 | T=2276 |

- **BAC selection for the tomato genome project**

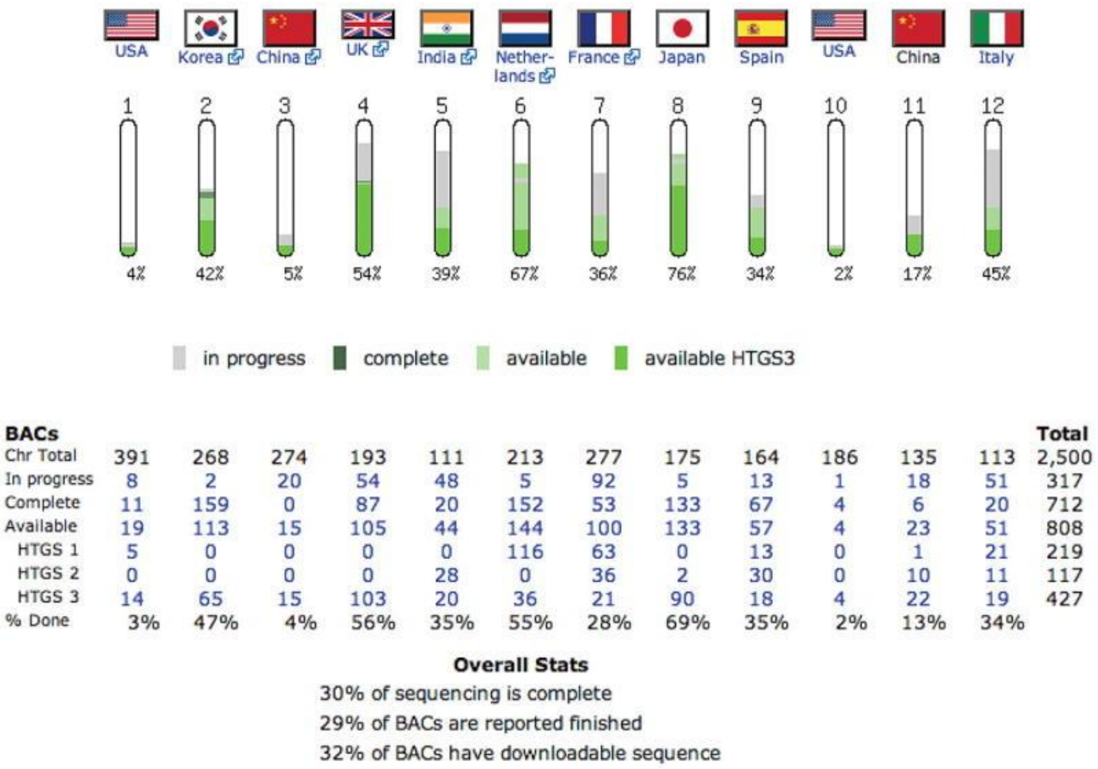


New BACs mapped on tomato chromosomes using CAPS markers

IL-bin Mapping completed for 50 BACs

-

## Visualizing the progress in the Tomato Genome Project



| BACs | | | | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chr Total | 391 | 268 | 274 | 193 | 111 | 213 | 277 | 175 | 164 | 186 | 135 | 113 | 2,500 |
| In progress | 8 | 2 | 20 | 54 | 48 | 5 | 92 | 5 | 13 | 1 | 18 | 51 | 317 |
| Complete | 11 | 159 | 0 | 87 | 20 | 152 | 53 | 133 | 67 | 4 | 6 | 20 | 712 |
| Available | 19 | 113 | 15 | 105 | 44 | 144 | 100 | 133 | 57 | 4 | 23 | 51 | 808 |
| HTGS 1 | 5 | 0 | 0 | 0 | 0 | 116 | 63 | 0 | 13 | 0 | 1 | 21 | 219 |
| HTGS 2 | 0 | 0 | 0 | 0 | 28 | 0 | 36 | 2 | 30 | 0 | 10 | 11 | 117 |
| HTGS 3 | 14 | 65 | 15 | 103 | 20 | 36 | 21 | 90 | 18 | 4 | 22 | 19 | 427 |
| % Done | 3% | 47% | 4% | 56% | 35% | 55% | 28% | 69% | 35% | 2% | 13% | 34% | |

**Overall Stats**

30% of sequencing is complete

29% of BACs are reported finished

32% of BACs have downloadable sequence

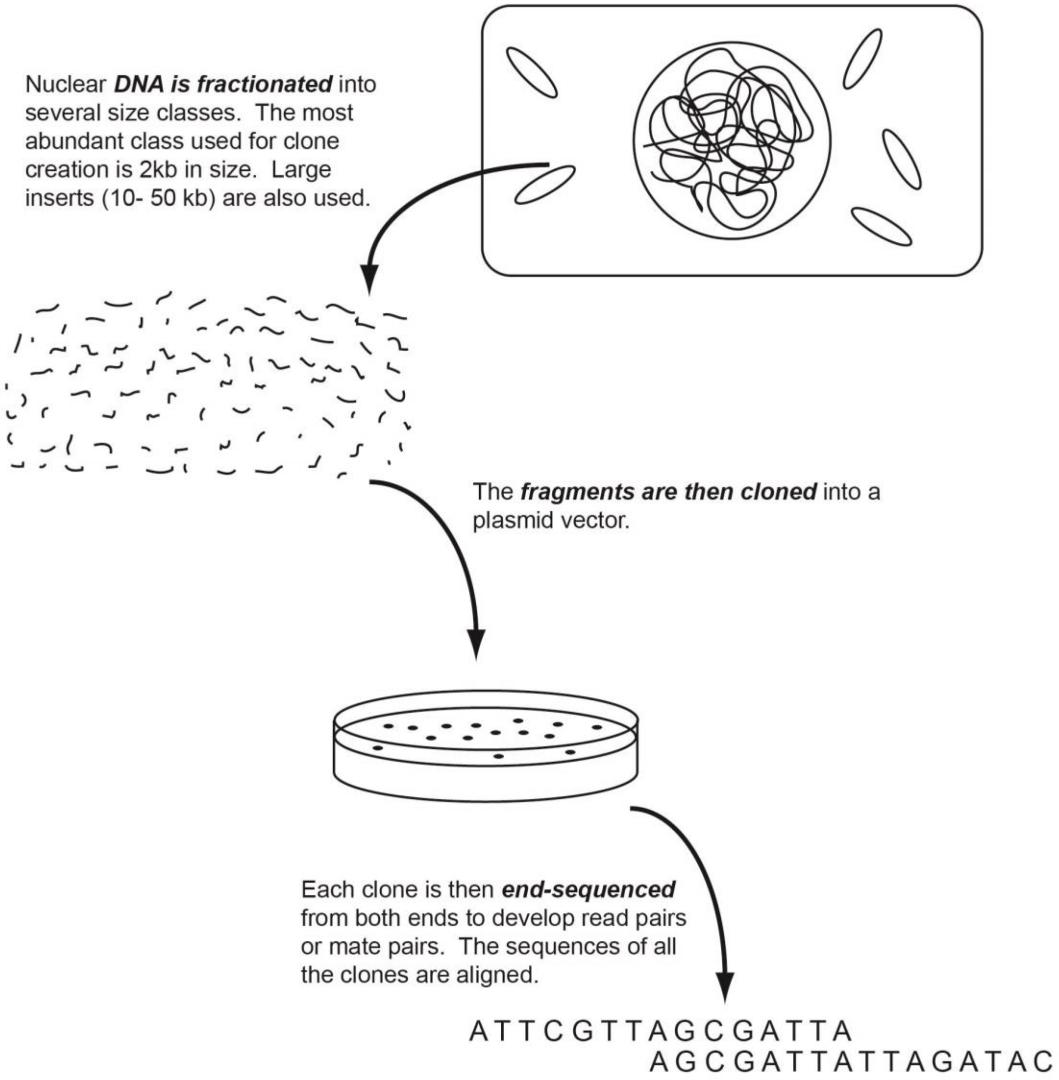
## Headlines of Human Genome Sequencing Project

- **February 2001**
  - **Working draft announced**
  - Major worldwide news event
- **April 2003**
  - **Finished draft announced**
    - *Little fanfare*
    - *Data more useful*
- **Many more $$s will be spent on projects to improve and understand the genome after the reference genome is released to the public**

# Whole Genome Shotgun Sequencing

## A. The Concept

Shotgun sequencing requires that random, small insert libraries are created from the total nuclear DNA of the species of interest. A plasmid cloning vector is used for this step. These clones are then sequenced. This step is analogous to the shotgun cloning and sequencing step used for each large-insert clone used in hierarchial shotgun. The sequences of the clones are then aligned. This is the procedure used to sequence the ***Drosophila*** genome, and by Celera to sequence the human genome.



Nuclear ***DNA is fractionated*** into several size classes. The most abundant class used for clone creation is 2kb in size. Large inserts (10- 50 kb) are also used.

The ***fragments are then cloned*** into a plasmid vector.

Each clone is then ***end-sequenced*** from both ends to develop read pairs or mate pairs. The sequences of all the clones are aligned.

```
ATTCGTTAGCGATTA
         AGCGATTATTAGATAC
```

## Whole Genome Shotgun Sequencing (WGS)

- **Original WGS approach**
    - *Isolate and shear nuclear DNA*
        - Modify DNA by adding restriction site adaptors
    - *Clone into plasmids*
    - *Plasmids are then directly sequenced*
        - Approach requires **read-pairs**
            - Especially true because of the repetitive nature of complex genomes
- **Modern WGS approach**
    - *Isolate and shear nuclear DNA*
    - *Create massively parallel sequencing library*
        - Sequencing billions of bases
            - NO plasmid cloning

## WGS

- **Proven very successful for smaller genomes**
    - *Essentially the only approach used to sequence smaller genomes like bacteria*
- **Is WGS useful for large, complex genomes?**
    - *Initially considered a bold suggestion*
        - Large public effort dedicated to hierarchical approach
    - ***Drosophila (2002)***
        - **Sequenced using the WGS approach**

## WGS – Major Challenge 1

700 bp repeat;
1000x in genome
4 chromosomes

Read is 400bp

To which chromosome should the repeat be assigned?

- **Assembly of repetitive DNA is difficult**
  - *Difficult to place in the correct genomic position*
  - Many types
    - Retrotransposons (RNA mobile elements)
    - DNA transposons
    - Alu repeats (human)
    - Long and Short Interspersed Repeat (LINE and SINE) elements
    - Microsatellites
- **Solution**
  - *Use sequence data from 2, 10 and 50 kb clones*
    - Data from fragments containing different types of sequences can be collected
    - Paired-end reads collected
  - *Assembly Process*
    - Repeat sequences are initially masked
    - Overlaps of non-repeat sequences detected
    - Contigs overlapped to create supercontigs
  - *Software available but wass mostly useful to the developers*
    - Examples: Celera Assembler, Arcane, Phusion, Atlas

**Recent solution**
**PacBio Reads
**Span the full repeat

## WGS – Major Challenge 2

- For the two sequences approaches, assembly is a scale issue
  - **WGS approach**
    - *Giga~~bytes~~* bases *of sequence data*
  - **Hierarchical approach**
    - *Magnitudes less*
  - On-going research focuses on developing new algorithms to handle and assembly the huge data sets generated by WGS

**Mouse WGS Data**

- 29.7 million reads
- 7.4x coverage
- Newer software
- **Assembled without mapping or clone data**
    - Human WGS had access to this data from the public project
    - **225,000 contigs**
        - ***Mean length*** (Not Contig L50) = 25 kilobases in length
    - **Super contig subset**
        - ***Mean length*** (Not Scaffold L50) = 16.9 megabases
    - **200 largest supercontigs**
        - Anchored using mapping data
        - Represents 96% (9187 Mb) of the euchromatic region of genome

# Illumina Short-read Sequencing Projects

**The Panda Project (genome = 2.4 Gb)**

- **First genome assembled fully from short reads**
- **2010**

Human genome released in 2001

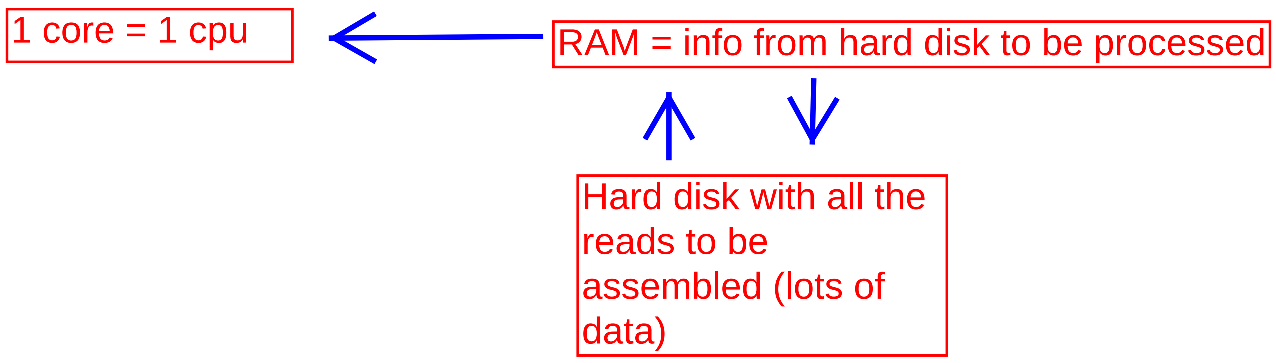
**Sequencing**

- 37 paired-end libraries
  - 150 bp, 500 bp, 2 kb, 5 kb, 10 kb in size
- 176 Gb usable sequence data
  - 73x coverage

100-150 bp reads???

**Assembly**

- **SOAPdenovo used for assembly**
  - Part of SOAP software package
    - Short Oligonucleotide Analysis Package
      - "SOAPdenovo uses the **de Bruijn graph algorithm** and applies a stepwise strategy to make it feasible to assemble the panda genome using a supercomputer (32 cores and 512 Gb random access memory (RAM)."
- Poor library and low-quality reads excluded
  - 134 Gb sequence data used
    - **56x coverage**



1 core = 1 cpu

RAM = info from hard disk to be processed

Hard disk with all the reads to be assembled (lots of data)

**Step 1. Contig building**

- Data from 500 bp or smaller libraries used first
- Assembly halted when repeat region encountered
    - 39X coverage achieved
    - **N50 = 1.5 kb**
    - Length = 2.0 Gb

**Step 2. Scaffold building**

- Paired-end data from all libraries used
    - **N50 = 1.3 Mb**
    - Total length = 2.3 Gb

**Step 3. Closing the gaps**

- Local assembly (within a specific gap) using paired end read with one end in a contig and the other in a gap
    - 223.7 Mb gaps closed
    - 54.2 remained unclosed

**Step 4. Compare with other carnivores**

- **Determined that gaps most likely repetitive elements**

# Genome Assembly

## Goal of assembly

- Create contigs based on similarity of sequence reads

## Assembly problem

- Finding the shortest supersting (*T*) from a set of strings ($s_1$,$s_2$,…$s_n$)
    - **Strings** = **Reads**
    - **Superstring** = **Contig**

## Issues that make assembly difficult

- **Sequencing errors**
    - ***Hard to ascertain, so ignored during assembly***
- **Repetitive sequences**
    - Some found 100,000 times
        - Repeats will lead to incorrect assembly
        - Hard to know which sequences overlap
    - Brings two regions together that are not in fact together
    - Resolving some repeats
        - ***If repeat is shorter than read length, assembly is possible***

**Features of original assembly algorithms**

- Greedy Algorithm Approach
    - Search for the best solution at each step
    - Does not go back and correct previous steps
        - Compute all possible overlaps between strings and assign a quality score
        - Merge strings with highest score
        - Continue until no other strings can be merged
- **Fastest method to a solution**
    - **Doesn’t guarantee optimum solution**
        - Approach doesn’t work for large genomes
        - Large RAM memory requirements

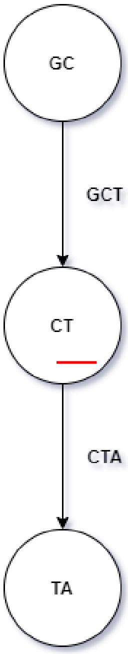
**Next generation assembly algorithms**

- **Graph theory approach**
    - Graph definition
        - A mathematical structure that models pairs of objects from a collection of objects
            - For sequencing, the objects are sequence reads
- **Overlap-Layout Consensus approach**
    - Early algorithm for used small genomes
        - Set a *sequence* as a *node*
        - *Overlaps* are *edges* that connect nodes
            - **Contig is a path of nodes and edges**
        - **Process**
            1. Find all possible alignments
            2. Remove overlap duplications
            3. Construct consensus node/edge combination to create contig
- **DeBruijn Graph**
- For perfect sequencing
    1. Break reads in k-mer length (often k=24)
    2. Form left and right k-1-mers for each k-mer that are **nodes**
    3. Draw **directed edge** from left k-1mer to right k-1 mer
    4. Add new k-1mer as node if not yet included
    5. Add directed edge to new node
    6. Repeat until k-mers for all reads are mapped as **edges**

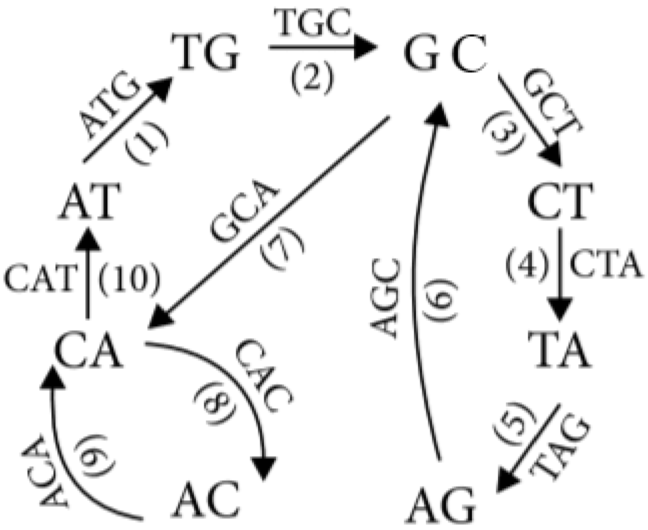
- Challenges for De Bruijn assemblers
    1. Sequencing errors
    2. Repetitive sequences

k-mers: GCT, CTA

**This is the multigraph for the two k-mers.**



**Multigraph for the sequence; ATGCTAGCAC**



## Arachne Whole Genome Assembler

Genome Research 12:177 (2002)

1. Breaks 600 nt read into 24 nt sequences and note read origin of the sequence
2. Create database with each sequence as main entry
   - Each sequence entry contains frequency and read identifier data
     - Read 1 = ATGCATTATTCGAGATT
     - k-mer = 4 nucleotides

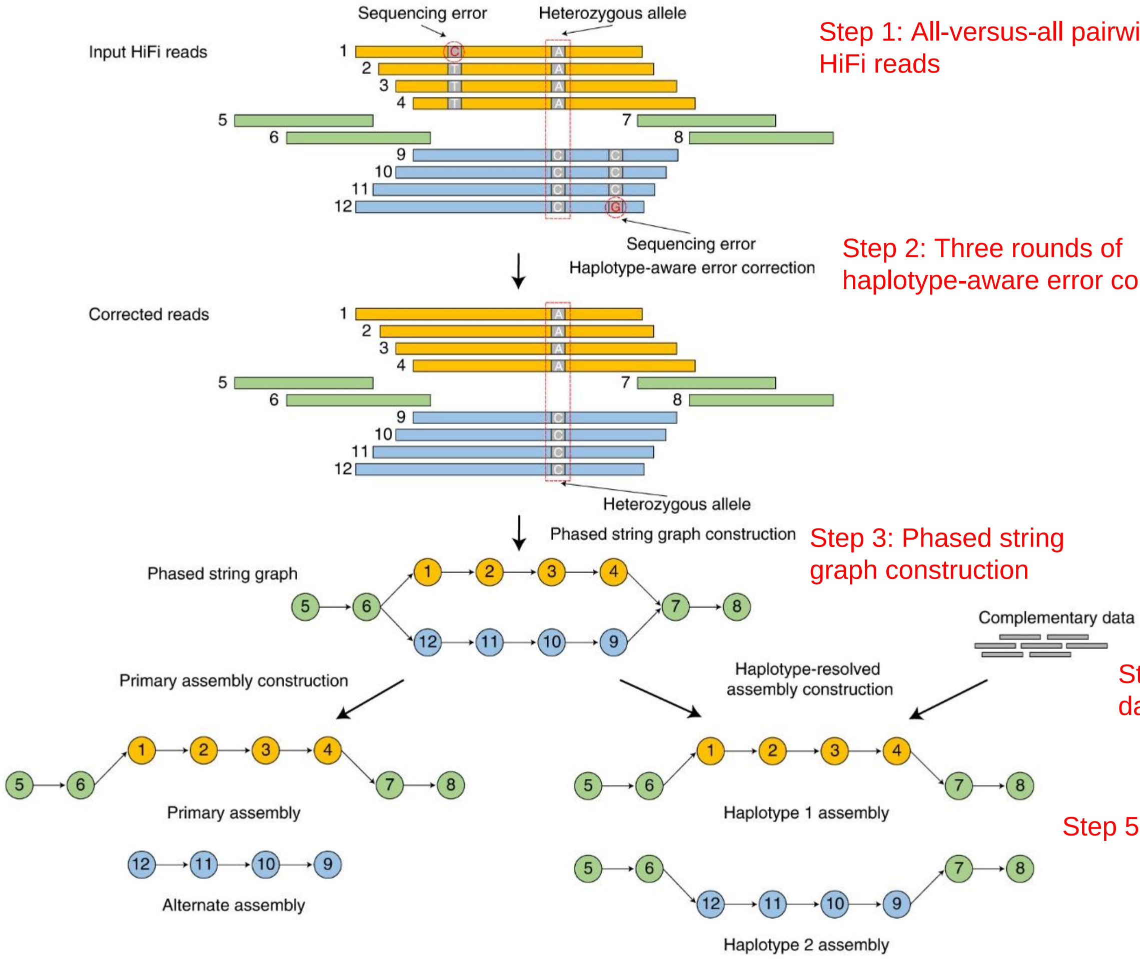
| K-MER | Read | Position |
|---|---|---|
| ATGC | 1 | 1 |
| TGCA | 1 | 2 |
| GCAT | 1 | 3 |
| CATT | 1 | 4 |
| ATTA | 1 | 5 |
| TTAT | 1 | 6 |
| TATT | 1 | 7 |
| ATTC | 1 | 8 |
| TTCG | 1 | 9 |
| TCGA | 1 | 10 |
| CGAG | 1 | 11 |
| AGAT | 1 | 12 |
| GATT | 1 | 13 |

- **File with k-mer data**

| k-mer | read | start position | direction |
|---|---|---|---|
| ATGC | 1,2,21,24 | 1,18,30,10 | f,f,f,r |
| TGCA | 1,8,31 | 2,4,51 | r,r,f |
| GCAT | 1,2 | 3,81 | f,r |
| CATT | 1,41,51,52,53 | 4,6,10,15,67 | f,f,f,r,r |
| Etc… | | | |

4. Discard high copy reads (these are repeats)
5. Align reads from low frequency sequences
6. Discover mate pairs represented in two plasmids of same length
    - These are paired pairs
7. Find a mate pair that matches only one end of the paired pair
    - Sequences are considered to be a single large read
8. Process continues until a repeat is encountered
9. Assembly stops and a unique **contig is declared**

**Assembly of PacBio reads using hifiasm. From: Cheng et al. 2021.** Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18:170-175.



Step 1: All-versus-all pairwise alignment of HiFi reads

Step 2: Three rounds of haplotype-aware error correction

Step 3: Phased string graph construction

Step 4: Complementary data (HiC) added

Step 5: Haplotype assembly

**Fig. 1 |** Outline of the hifiasm algorithm. Orange and blue bars represent the reads with heterozygous alleles carrying local phasing information, while green bars come from the homozygous regions without any heterozygous alleles. In the phased string graph, a vertex corresponds to the HiFi read with the same ID, and an edge between two vertices indicates that their corresponding reads are overlapped with each other. Hifiasm first performs haplotype-aware error correction to correct sequence errors but keep heterozygous alleles, and then builds a phased assembly graph with local phasing information from the corrected reads. Only the reads coming from the same haplotype are connected in the phased assembly graph. With complementary data providing global phasing information, hifiasm generates a completely phased assembly for each haplotype from the graph. Hifiasm also can generate an unphased primary assembly only with HiFi reads. This unphased primary assembly represents phased blocks (regions) that are resolvable with HiFi reads, but does not preserve phasing information between two phased blocks.

Notes on hifiasm:
**Minimum coverage for hifiasm: >13x per haplotype
**hifiasm also incorporates Hi-C data in the assembly

# Common Bean: 454-based Project

## Sequencing Libraries

Majority of sequencing data used for assembly

| Library | Sequencing Platform | Average Read/Insert Size | Read Number | Assembled Sequence Coverage |
|---|---|---|---|---|
| Linear | 454 XLR & FLX+ | 362 | 38,107,155 | 18.64x |
| GPNB | 454 XLR paired | 2,798 ± 1,047 | 589,346 | 0.11x |
| GGAS | 454 XLR paired | 3,922 ± 643 | 1,940,576 | 0.41x |
| GXSF | 454 XLR paired | 3,991 ± 337 | 467,414 | 0.07x |
| HYFA | 454 XLR paired | 4,729 ± 497 | 1,648,022 | 0.25x |
| HYFC | 454 XLR paired | 4,736 ± 504 | 1,491,648 | 0.24x |
| HYFB | 454 XLR paired | 4,759 ± 528 | 1,196,104 | 0.17x |
| HXTI | 454 XLR paired | 8,022 ± 1,016 | 1,364,808 | 0.22x |
| GXNX | 454 XLR paired | 9,192 ± 1,058 | 878,832 | 0.16x |
| HXWF | 454 XLR paired | 11,903 ± 1,928 | 724,196 | 0.13x |
| HXWH | 454 XLR paired | 12,231 ± 1,902 | 413,396 | 0.08x |
| VUK (Fosmid-end) | Sanger | 34,956 ± 4,536 | 240,384 | 0.20x |
| VUL (Fosmid-end) | Sanger | 36,001 ± 4,632 | 88,320 | 0.08x |
| PVC (BAC-end) | Sanger | 121,960 ± 16,572 | 81,408 | 0.08x |
| PVA (BAC-end) | Sanger | 126,959 ± 25,658 | 89,017 | 0.09x |
| PVB (BAC-end) | Sanger | 135,292 ± 21,487 | 92,160 | 0.09x |
| **Total** |  | N/A | 49,412,786 | 21.02x |

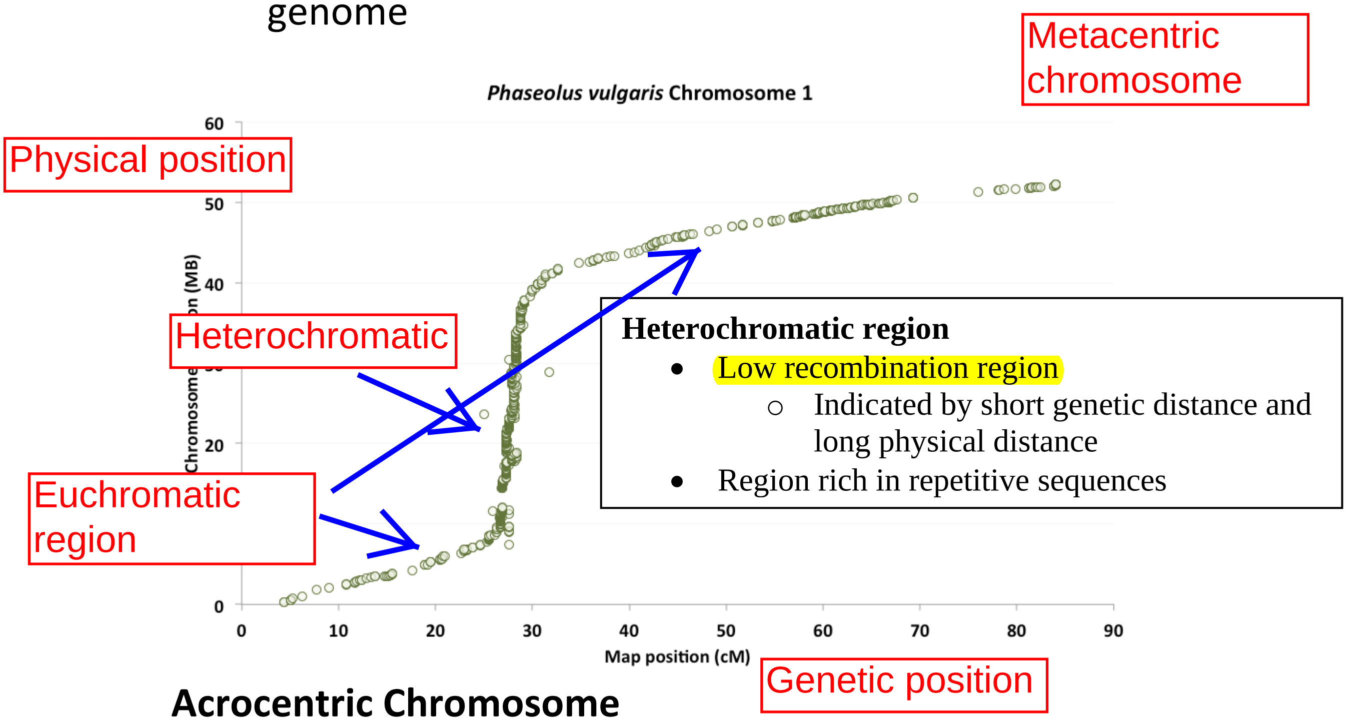
# Genome Assembly Statistics

## Comparison of Two Legume Species and Sanger and Short Read Sequence Data Collection

| **Statistic** | **Soybean** | **Common Bean** |
|---|---|---|
| Sequencing method | WGS, Sanger | WGS, 454 & Illumina |
| Genome size (contig) | 955 Mb (1.9% gap) | 473 Mb (9.3% gap) |
| Genome size (scaffold) | 973 Mb | 521 Mb |
| Contig number | 16,311 | 41,391 |
| Contig N50 | 1,492 | 3.273 |
| Contig L50 | 189 kb | 39.5 kb |
| Scaffold number | 1,168 | 708 |
| Scaffold N50 | 10 | 5 |
| Scaffold L50 | 47.8 Mb | 50.4 Mb |
| Genetic map loci | 1.536 SNPs | 7,015 SNPs |

## Typical Chromosome

- High recombination on ends of chromosome
- Low recombination in the center (heterochromatic) region of the genome

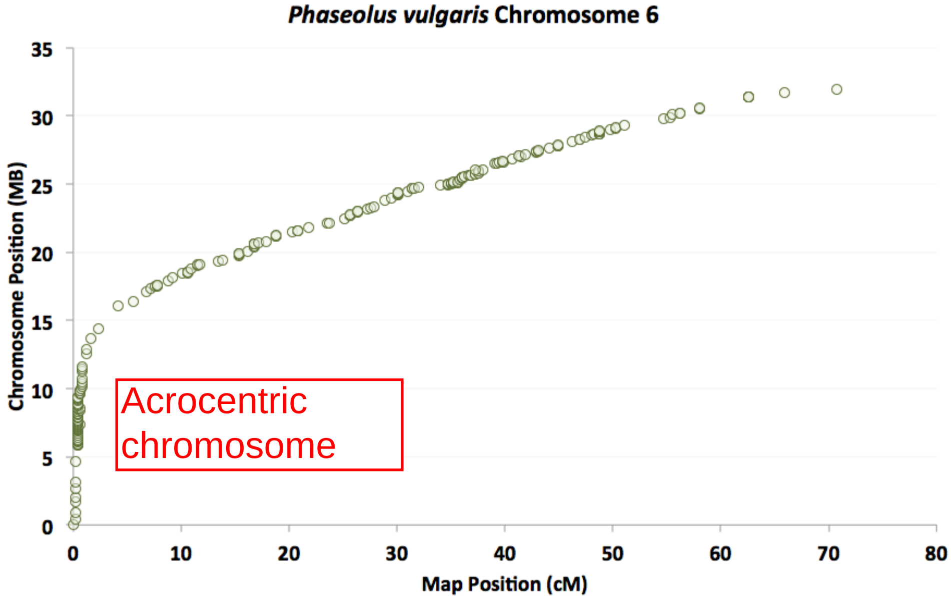


**Heterochromatic region**

- Low recombination region
  - Indicated by short genetic distance and long physical distance
- Region rich in repetitive sequences

## Acrocentric Chromosome

- Heterochromatic repeat rich region at end of chromosome


*Phaseolus vulgaris* Chromosome 6
Chromosome Position (MB)
Acrocentric chromosome
Map Position (cM)