**Primer walking:** Custom primers designed to be complementary to the known sequence can be used to extend the DNA sequence iteratively from the distal ends of known regions into unknown regions of the same molecule.
The Human Genome Project adopted a map-based strategy for all of the genomes that it initially tackled. In this approach, one starts with a well-defined physical map, usually consisting of overlapping BAC or PAC clones. Among the overlapping clones, those that form what's called the shortest tiling path are chosen. The shortest tiling path is simply the set of clones that have the shortest overlaps between them. The reason for using the shortest tiling path is that fewer clones then need to be sequenced. Each clone is then randomly sheared, and the sequence read and assembled. From the sequence of each BAC or PAC, the entire sequence of each chromosome is assembled based on the physical map. This procedure requires finding the sequence overlaps of each large-insert clone (in the shortest tiling path) and removing the overlapping portion of one of the two sequences in order to form a single contiguous sequence.

An analogy for the map-based sequencing strategy is to think of the genome as a set of books. Each volume contains one BAC's worth of sequence. In the map-based strategy, the order of the volumes is first determined. Then the information in each of the individual volumes is determined. Finally, the knowledge of the order of the volumes and the material in each volume are combined to form a single assembly.
As an alternative to map-based sequencing, the biotechnology company Celera applied shotgun sequencing to large genomes. Although this method had been applied to small genomes such as viruses, it was thought to be inappropriate for large genomes containing repetitive sequences. (Celera was formed as a subsidiary of Applied Biosystems, the maker of automated sequencing instruments.) In the whole-genome shotgun (wgs) approach, there is no physical map made. Instead, the whole genome is treated like a single large-insert clone. It is randomly sheared into fragments. (Celera used 2-kb, 10-kb, and 50-kb subclones.) These fragments are then randomly sequenced until there is sufficient coverage.

To understand whole-genome shotgun sequencing, let us use the book analogy again: The information on random fragments of pages from each of the volumes is determined. This procedure is done multiple times so that overlapping sentences can be found. Based on these overlaps, the whole work is then assembled.

The major challenge of whole-genome shotgun sequencing is the assembly of the huge number of individual sequence reads into one contiguous sequence. A computer program was written by a team headed by Gene Myers to assemble the DNA sequence. To run the program required some of the fastest computers available. One of the keys to assembling the sequence was the use of paired reads: During the production sequencing process, the clones were bar coded so that one could identify when the same clone was read from both ends.
One way to resolve the controversy has been to combine the two approaches. This hybrid approach starts with a physical map as well as sequencing of some of the clones in the shortest tiling path. At the same time, whole-genome sequencing is performed. Then the information is combined, using the sequence from the mapped clones as a scaffold to help assemble the whole-genome shotgun reads. This approach was used by the publicly funded effort to sequence the mouse genome.

To use the book analogy again, in the hybrid approach, the order of the volumes is determined and a small amount of information from each volume is acquired. At the same time, random fragments from all the volumes are read. These random fragments are then aligned with the help of the information gained by ordering the volumes and of the small amount of information known about each volume. In essence, the random fragments are used to fill in the gaps that remain when the information from individual volumes is assembled. The graph in the top right corner of this slide depicts the open question of what mix of whole-genome shotgun and map-based sequencing is optimal.
The majority of the genome sequencing projects is by WGS. Some projects have not reported the methodology and are listed as undefined.
Physical mapping, which involves determining the physical distance in base pairs between two points on a chromosome, has many uses, including isolating genes by chromosomal walking and providing the basis for large-scale sequencing. For the former, the goal is to find a gene, starting from a linked physical marker. This task requires finding a set of overlapping fragments of DNA that spans the distance between the marker and the gene. To make this process more efficient, scientists have developed vectors that can accommodate large pieces of DNA. Examples of large-insert cloning vectors are cosmids, YACs, and BACs. These vectors are described on the next set of slides.

<table>
<thead>
<tr>
<th>Steps for Map-based sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Development of high-density map (David section)</td>
</tr>
<tr>
<td>• Construction of large insert libraries</td>
</tr>
<tr>
<td>• Construction of physical map</td>
</tr>
<tr>
<td>• Connection of the genetic and physical map</td>
</tr>
<tr>
<td>• Minimum tilling path</td>
</tr>
<tr>
<td>• Sequencing of the BACs from the contig</td>
</tr>
<tr>
<td>• Assembling using Phred Phrap Consed</td>
</tr>
<tr>
<td>• Finishing</td>
</tr>
</tbody>
</table>
Among the first large-insert vectors were lambda phage vectors, which can take a maximum insert size of between 20 and 30 kilobases. Cosmids combine phage vectors' ability to be packaged with the DNA propagation and purification properties of plasmids. Their maximum insert size is between 35 and 45 kilobases. Bacterial and P1 artificial chromosomes (BACs and PACs) can be used for inserts in the 100–300 kilobase range, while yeast artificial chromosomes (YACs) can hold up to a megabase of DNA.

The insert sizes of lambda phage and cosmid vectors (shown in the schematic on the slide) are too small for large-scale sequencing projects, even though their inserts are generally very stable during propagation in bacteria. Although one would imagine that YACs, with their very large insert size, would be the vector of choice for genome sequencing, in fact they have proven to be fairly difficult to work with. Among the problems are difficulties in isolating the artificial chromosome away from the normal yeast chromosomes and stability issues with the larger inserts.

Thus, although BACs and PACs do not hold the largest inserts, they have come to be the most commonly used vectors for large-scale sequencing. They represent a good compromise between insert size and ease of use. In particular, the growth and isolation of these cloning vectors is similar to that for bacterial plasmids. The diagram in this slide is of a typical BAC cloning vector, showing the various restriction sites for cloning and the origin of replication (ori).
Wheat Large Insert libraries

http://agronomy.ucdavis.edu/Dubcovsky

**BAC library of *T. monococcum***
- **276,480** BAC clones of 115-kb
- Library coverage: 5.6 genome equivalents.
- Probability of having any gene represented in this library: > 99.6%.

**BAC library of *T. dicoccoides***
- **516,096** BAC clones of 130-kb
- Library coverage: 5.0 X each genome
- Probability of having any gene represented in this library: > 99%

GENOME (1999) 42:1176-1182
Picking to plates and gridding into HD-filters

Transformation

Plating

Picking in 384 plates

Copying

O-bot

Serial Number

Filter Label

Duplicate Pattern

Field 2

Field 6

Field 3

Field 4

Field 1

Field 5

Field 2, position 4, plate 20, BAC 20J4

Field 3, position 6, plate 33, BAC 33M20

6 x 384 x 8
18,432 clones
Per HD-filter
There are several ways of generating contigs. One is to cut all of the available inserts (from cosmids, BACs, PACs, or YACs) with various restriction enzymes and compare the pattern of bands generated. When several bands appear to be of the same size in two different inserts, this indicates that the two fragments overlap. This process is known as “fingerprinting.” The steps in this process can then be repeated until a contig of the desired length is generated.

When restriction enzymes are used to build contigs from large-insert clones (e.g., BACs, PACs, or YACs), a large number of fragments are usually generated, due to the size of the insert. The potential problem is that two fragments of slightly different size will be counted as being the same size. To alleviate this problem, high-resolution gel electrophoresis (frequently employing gels similar to those used for sequencing) can be used to separate the fragments so that very small size differences can be detected.
Global contig construction with FPC

Production of BAC libraries
\[\downarrow\]
High-throughput Fingerprinting 20X BACs
\[\downarrow\]
Assembly of contigs FPC
\[\downarrow\]
Map genetic markers into the contigs
(GAPCs or sequence-based markers)
\[\downarrow\]
Contigs Integrated Into genetic and physical maps
\[\downarrow\]
Construction of physical maps and minimum tiling path
\[\downarrow\]
Parallel sequencing of the contigs using 454-ILLUMINA combinations

DNA isolation
\[\rightarrow\]
Fingerprinting
\[\rightarrow\]
Contig assembly
SELECT CLONES FOR SEQUENCING

Section of BAC fingerprint

12
FPC takes as input a set of clones and their restriction fragments (called Bands) and assembles the clones into contigs; for more information, see summary. FPC contributions are by the GSC, BCGSC, Sanger Centre, Clemson University, and University of Arizona.

Latest release: FPC V9.3 (3 Dec 2008) details; Published in Nucleic Acids Research Download

it is increasingly common for a species to have both a fingerprint map and draft sequence, along with BESs to link them together; indeed, this is the case for soybean, sorghum, medicago, poplar, xenopus and is soon to be the case for several other species such as Brachypodium and cotton.
The overlapping pieces of DNA that span a region of the chromosome are known as “contigs,” a term that derives from “contiguous,” as in, producing a contiguous series of DNA fragments. For efficient sequencing, the goal is to start from the smallest number of DNA fragments that form a contig covering the region of the chromosome that you want to sequence. This set of fragments is known as the “minimum tiling path.”
Genomic sequencing can be broken into several steps. First, libraries of large fragments of genomic DNA are made. The second stage of production sequencing can itself be broken into smaller steps. These steps include generating the small DNA fragments to be sequenced, performing the actual sequencing reactions, and running the reactions through the automated sequencer. The third step is called finishing, and it involves assembling the raw sequence reads into a continuous sequence and then filling any remaining gaps in the sequence.
The difficulty and cost of large-scale sequencing depend on two parameters: accuracy and coverage. Accuracy means how many errors can be tolerated in the final sequence. Coverage is a measure of the number of times that the same region is sequenced. Of course, the two parameters are related. With increasing coverage, one usually gets higher accuracy. However, accuracy is dependent on more than just coverage; the quality of the finishing effort also plays a critical role in determining the level of accuracy.

### Sequencing parameters

Cost of sequencing projects depend on the following parameters:

**Accuracy**
- How many errors are tolerated: protein prediction, forensic application require high-quality
- Increasing accuracy from one error in 100 to one error in 10,000 increases costs three to fivefold.
- Depending on finishing effort

**Coverage**
- How many times the same region is sequenced
- Accuracy increases with coverage, but not much is gained above 10x coverage:
  - **Finishing**: 8x to 10X coverage
  - **Rough-draft**: 5x coverage
  - **Skimming**: 1x to 3x coverage (67% 97% sequence, 99% accurate)
Gap4 is a Genome Assembly Program. The program contains all the tools that would be expected from an assembly program plus many unique features and a very easily used interface. The original version was described in Bonfield, J.K., Smith, K.F. and Staden, R. A new DNA sequence assembly program. Nucleic Acids Res. 24, 4992-4999 (1995).

The random sequences generated from the sheared fragments are first assembled into overlapping sequences. Depending on the desired coverage, this step can require up to 10 reads for each portion of the DNA. From these overlapping reads, one contiguous sequence is identified. This task is usually accomplished using an automated editing program known as PHRAP, developed by Phil Green. The program analyzes each position in the sequence and determines the quality of each read. It then generates a consensus sequence based on the different sequences acquired from the same region and their internal consistency (i.e., how frequently the same base was found at any given position).
- Phred reads traces and produces:
  - **bases** \( (T) \), **quality** \( (Q) \), **position** \( (173) \)
  - The quality value is a log-transformed error probability:
    \[ Q = -10 \times \log_{10}(Pe) \]
  - \( Pe \) is the probability of error,
  - \( Q_{20} \approx 0.01 \) probability of error
  - \( Q_{30} \approx 0.001 \) probability of error
  - Phred quality values have been tested for accuracy and power to discriminate between correct and incorrect base calls.
  - Phred can use the quality values to perform sequence trimming.
Phrap creates an index containing the locations of all occurrences of every possible word of length 10 \((4^{10})\). Then each read as well as its complement is compared at each of its complete index words with the sequences at every other occurrence of that word. Each match is extended to its maximum exactly matching length. Each of the word matches corresponds to a particular diagonal in the alignment matrix for two reads. Rather than fill the entire matrix with scores, phrap creates a “band” centered on the diagonal indicated by the match and with width equal to “bandwidth” default 14. Since these are supposed to be identical sequences no large gaps are expected, and the correct alignment is expected to be contained entirely within the band. After the assembly phrap calculates quality scores for each base position in each contig. Phrap also classifies some of the reads as quimeric (If the distance between the paired read exceeds the maximum size established by the user). Phrap uses its adjusted quality scores to evaluate the likelihood for the hypothesis that the reads actually overlap against the hypothesis that the reads come from separate copies of a repeat that is 95% identical.
Consed provides a flexible visualization tool
Although automated editing programs like PHRAP have greatly increased the efficiency of sequencing, there remains a need for human judgment and intervention. This occurs during the finishing step, which is defined as the process of assembling the raw sequence reads into an accurate contiguous genomic sequence. For genomic sequencing with an accuracy of one error in 10,000 bases, a manual finishing step is essential. The finisher looks at positions where the automated editing program can’t tell which base is the correct one. By examining the various raw sequence reads, the finisher then makes a judgment call as to the correct base or sends the region back for additional sequencing. Similarly, when there are gaps in the sequence or insufficient coverage, the finisher will flag the region and send it back to the production sequencing team for more work.
BAC sequences exist at different levels of finishing.

**Phase 1**: Draft sequence; *non-ordered, non-oriented* contigs, contains gaps.

**Phase 2**: Draft sequence; *ordered and oriented* contigs, contains gaps

**Phase 3**: Finished sequence; high quality sequence with no gaps
The final step of finishing is to verify the sequence. All regions are checked for the extent of coverage (i.e., how many times the same region has been sequenced, and in what direction), for sequence quality (i.e., whether ambiguity has been removed for all positions in the sequence), and for contiguity (i.e., whether the sequence forms one uninterrupted stretch of DNA). A good test of sequence quality that is frequently used in the finishing stage is to determine the sites where restriction enzymes would cut in the newly acquired sequence. A restriction map is generated from the sequence and then compared with the known fingerprint generated from the clone during the mapping step. If both show the same pattern, then it is considered to be an indication that the sequence is of high quality and relatively error free.
### Positional cloning of *Vrn2*

1. **Genetic map:** *T. monococcum* genetic map (3,700 gametes)

   Distance based on recombination events

2. **Physical maps**

   *T. monococcum*

3. **BAC sequencing**

4. **Candidate gene validation**

<table>
<thead>
<tr>
<th>Sequence comparison</th>
<th><em>AY485644.3</em></th>
<th><em>ZCCT1</em></th>
<th><em>ZCCT2</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted protein</td>
<td>winter = spring</td>
<td>winter ≠ spring</td>
<td>winter = spring</td>
</tr>
<tr>
<td>5' promoter region</td>
<td>winter = spring</td>
<td>winter = spring</td>
<td>winter = spring</td>
</tr>
<tr>
<td>3' region</td>
<td>winter = spring</td>
<td>winter = spring</td>
<td>winter = spring</td>
</tr>
</tbody>
</table>

---

**ThdZCCT1**

DP92: PRR7KVEFGQKDYPVTYIVGQKPR

**ThdZCCT1-O33116**

DP92: PRR7KVEFGQKDYPVTYIVGQKPR

---

Science 2004, 303:1640
A successful genetic screen results in the identification of interesting mutants. Once these mutants have been characterized, the next step is usually to identify the gene in which the mutation has occurred. If the mutation was produced by a chemical mutagen or by ionizing radiation, gene identification normally requires genetic mapping and chromosomal walking. These methods are described in detail in the chapter entitled, “Fundamentals of Mapping and Sequencing.”
Candidate gene validation by TILLING

TILLING (Target induced local lesions in genomes) generates mutations using the chemical EMS and then detect them using gene specific primers and pools of DNAs from mutagenized plants.
EMS produces random mutations in genetic material by nucleotide substitution; specifically by guanine alkylation. This typically produces only point mutations. It can induce mutations at a rate of 5x10^-4 to 5x10^-2 per gene without substantial killing. The ethyl group of EMS reacts with guanine in DNA, forming the abnormal base O-6-ethylguanine. During DNA replication, DNA polymerases that catalyze the process frequently place thymine, instead of cytosine, opposite O-6-ethylguanine. Following subsequent rounds of replication, the original G:C base pair can become an A:T pair.
Detecting mutations

- Denaturing step
EMS treatment of seeds results in alkylatation of guanine bases and leads to mispairing such as G pairs with T instead of C.
Detecting mutations

*Cell* cut heteroduplexes

**Cell** from celery juice extract

Targets hetero-duplexes and cuts

### No cut

- T
- A

### Cut

- T
- G

3% Polycrylamide gel
Detecting mutations

- Screen 96-well plate
  4x pools
- Identify positive individual DNA
  AB CD BC AD
RPS-BLAST: Reversed Position Specific Blast RPS-BLAST is one of the BLAST series of searching programs from NCBI. It reverses the popular PSI-BLAST search by comparing a protein query with a database of PSI-BLAST checkpoint files made from multiple alignments of protein families. Here the checkpoint files are made from the Blocks and Prints alignments.
PARSE-SNP (Project Aligned Related Sequences and Evaluate SNPs) is a web-based tool for the analysis of polymorphisms in genes. It determines the translated amino acid sequence from a reference DNA sequence (genomic or cDNA) and a gene model, and the effects of the supplied polymorphisms on the expressed gene product. If a homology model is provided, predictions can be made as to the severity of missense changes.

### PARSE-SNP: Select best mutation

For help interpreting these results, view the PARSE-SNP Instruction page.

<table>
<thead>
<tr>
<th>#</th>
<th>View Or Sequence</th>
<th>Nucleotide Change</th>
<th>Effect</th>
<th>Restriction Enzyme Differences from RefBASE</th>
<th>PNNM Difference</th>
<th>SIFT Score</th>
<th>Description</th>
<th>Zygosity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GLC C209A</td>
<td>W24*</td>
<td>Mut</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>2</td>
<td>GLC C374T</td>
<td>N42=</td>
<td>XbaI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>3</td>
<td>GLC G444A</td>
<td>V561I</td>
<td>BseI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>4</td>
<td>GLC L291A</td>
<td>M23*</td>
<td>XbaI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>5</td>
<td>GLC G511A</td>
<td>C268Y</td>
<td>BseI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>6</td>
<td>GLC C508T</td>
<td>A309Y</td>
<td>BseI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>7</td>
<td>GLC G109A</td>
<td>G110R</td>
<td>BseI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>8</td>
<td>GLC G658A</td>
<td>E133A</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>9</td>
<td>GLC G671A</td>
<td>D443N</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>10</td>
<td>GLC G933A</td>
<td>E229K</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>11</td>
<td>GLC C1006T</td>
<td>S255F</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>12</td>
<td>GLC G1006A</td>
<td>E272V</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>13</td>
<td>GLC G1101A</td>
<td>E258K</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>14</td>
<td>GLC G1131A</td>
<td>V294I</td>
<td>Not Available</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
<tr>
<td>15</td>
<td>GLC Y1366T</td>
<td>A101</td>
<td>CuTSI</td>
<td>Not Available</td>
<td>0.00</td>
<td>0.00</td>
<td>Homo</td>
<td>Homo</td>
</tr>
</tbody>
</table>

http://www.proweb.org/parsesnp/
Re-sequencing: aligning short-sequences to a reference

**Bowtie** - an ultrafast, memory-efficient short read aligner that aligns short DNA sequences (reads) to the human genome at a rate of 25-35 million reads per hour on a typical workstation with 2 GB of memory.

Bowtie indexes the genome with a Burrows-Wheeler index. Used in data compression since it generates more runs of the same letter.

BWT-based indexing allows large texts to be searched efficiently in a small memory footprint. The Burrows-Wheeler transformation of a text T, BWT(T), is constructed as follows.

1) The character $ is appended to T, where $ is not in T and is lexicographically less than all characters in T.
2) The Burrows-Wheeler matrix of T is constructed as the matrix whose rows comprise all cyclic rotations of T$.
3) The rows are then sorted lexicographically.
4) BWT(T) is the sequence of characters in the rightmost column of the Burrows-Wheeler matrix (Figure 1a).
5) BWT(T) has the same length as the original text T.
6) This matrix has a property called 'last first (LF) mapping'.

**BWT in bioinformatics**

The advent of high-throughput sequencing (HTS) techniques at the end of the 2000 decade has led to another application of the Burrows–Wheeler transformation. In HTS, DNA is fragmented into small pieces, of which the first few bases are sequenced, yielding several millions of "reads", each 30 to 500 base pairs ("DNA characters") long. In many experiments, e.g., in ChIP-Seq, the task is now to align these reads to a reference genome. A number of alignment programs were published, which initially relied on hashing (e.g., Eland, SOAP[1][3], or Maq[4]). Here, the enormous memory requirements for the hash table was a problem, and so, the next generation of alignment programs, notably Bowtie[5], BWA[6], and SOAP2[7], uses the Burrows–Wheeler transform of the reference genome as the basis for very efficiently aligning the reads.