Genome Sequencing

Phil McClean
September, 2005

The concept of genome sequencing is quite simple. Break your genome up into many different small fragments, clone those fragments into a cloning vector, isolate many clones, and sequence each clone. All of the techniques used for sequencing are well established. These are the same techniques that scientists used for the past twenty years to characterize many different individual genes. Using these techniques, the most aggressive efforts to sequence a region around a gene might collect about 40,000 bases of sequence data.

What differs from previous sequencing efforts is the scale of the effort. For example, the dataset used by the public consortium to publish their first draft of human genome was based on 23 billon bases. That group used the hierarchial sequencing approach. By contrast, the whole genome shotgun approach of Celera Genomics used 27.2 million clones to read the 14.8 billion bases to construct the 2.91 billon basepair sequence of the euchromatic portion of the human genome. Clearly, tremendous advances have been made in sequencing.

The same basic technology initially used to sequence a small region around a gene is still in use today. That technology though has undergone significant improvements. The basic chain-termination sequencing procedure has been improved by the use of use of thermostable polymerase enzymes in the reaction to improve the quality of the sequencing products. We now routinely use fluorescently labeled nucleotides for the reaction. This coupled with laser-based detection systems have essentially eliminated the need for gel-based sequence detection systems. It is now very routine to rapidly collect 500-800 bases of high quality DNA sequence data in a single run. Coupled with the fact that 8, 16 or even 96 samples can be analyzed simultaneously, it is not surprising that, at peak output, the public human sequencing project was processing seven million samples per month and generating 1000 bases of data each second.

Robotics has also been a key addition to whole process. The human hand rarely touches the clone that is being sequenced. Robots pick subclones, distribute them into reaction plates, create the sequencing reaction, and load the load the plates onto the capillary detection system. Collectively, all of these innovations have increased the quality and quantity of the data while decreasing the cost. It is estimated that the cost of sequencing for these large centers has decreased over 100 fold since the advent of the Human Genome Project in 1990. These improvements have spilled over into the small research lab, where a sequence reads cost as little as $2.50 compared to the initial $15 in the early 1990s.

Hierarchical Shotgun Sequencing vs. Whole Genome Shotgun Sequencing

There are two basic approaches to sequencing large complex genome: hierarchical shotgun sequencing and whole genome sequencing. Historically, the general approach to hierarchical shotgun sequence came first. In the early days of sequencing, without the prospect of the degree of high throughput sequencing and sophisticated sequence assembly software, it was generally considered necessary to first order the genome into a series of overlapping

fragments. With such a clone-based contig (***contig***uous sequence) at hand, the assembly of the final genomic sequence seemed easier because each clone would provide a fixed sequence reference point from which a sequence assembly could be built. The advent of software that could assemble a large collection of unordered small, random sequence reads significantly changed perceptions regarding the best method of obtaining the sequence of a complex eukaryotic species.

Heirarchial Shotgun Sequencing. The first step in ***hierarchical shotgun sequencing*** is the construction of a large insert library from nuclear DNA of your species of interest. In the early days of sequencing, yeast artificial chromosomes (YACs) were used. YAC clones can contain up to a megabase (one million bases) or more of DNA. The concept was that only a few thousand overlapping clones would be necessary to develop a clone contig. But several technical difficulties with YAC clone (not the least of which was the fact that most researcher did not have the expertise to work with yeast), led researchers to look for an alternative cloning vector. The vector of choice was bacterial artificial clone (BAC) or the related P1 artificial clones (PAC). The primary advantage of these cloning vectors is that 1) they contained reasonable amounts of DNA [(about 100-200 kb (100,000 – 200,000 bases)], 2) unlike YACs, did not undergo rearrangements, and 3) that could be handled using standard bacterial procedures with which many scientists were familiar.

With a BAC or PAC library (or in the case of the public human project eight libraries) in hand, the next step is to create an ordered array of the clones. This is not a trivial endeavor. A key tool for this entire step is a ***molecular map*** of the species. A molecular map is a series of DNA markers that are aligned in the correct order along a chromosome. In genetic terminology, each chromosome is defined as a ***linkage group***. The map is used as the reference point to begin ordering the clones in a manner that represents sequence organization of the genome. Among the many clones in a library, the task is to find two clones that overlap either end of a seed clone. This is accomplished by "fingerprinting" each of the clones in the library. Each clone is cut with a restriction enzyme, such as the six-base cutter *Hin*dIII. The restriction fragment length pattern of each clone represents the fingerprint. The clones are seeded onto the map using markers from a dense STS (sequence tagged site) map. Overlapping clones are discovered based on shared fingerprint fragments. A series of overlapping clones were then defined for each chromosome. Collectively these overlapping clones define the ***physical map*** of the genome.

All totaled, the human project sequenced 29,298 large insert clones. This actually was more clones than necessary. The project began with the available technology that itself evolved over time. For example, sequencing actually began before the final physical map was completed. Therefore, the map itself was suboptimal. In contrast, the physical map of *Arabidopsis* was completed before the onset of sequencing. This genome is small (about 125 megabases). The physical map of this species consisted of 1,569 large insert clones that defined ten contigs. By contrast, the yeast genome was also sequenced using clones assigned to a physical map. That map consisted of 493 clones. Because that research began much earlier than the *Arabidopsis* project, it used the cosmid vector that accepts smaller-sized insert. Thus the number of clones is high relative to its genome size.

It is important to note that the construction of these maps represents only 1% of the costs of sequencing the genome. But at the same time, it is somewhat slow process. Yet at the same time, the physical map can be used to obtain preliminary sequence information (by BAC-end sequencing), is useful for researchers who wish to target a specific region of the genome for analysis prior to the release of the complete genome sequence, and provides a resource that can be used to develop more in depth genetic maps by comparing different genotypes at specific locations along the genome.

The physical map, the series of overlapping clones, is the raw material that is used for sequencing the genome. Collectively, these clones define the ***minimal tiling path*** for a particular region of the genome. The next step is to select clones for analysis. Ideally, the minimal tiling path consists of a redundant set of clones because not all clones are of equal quality. During the cloning process chimeric clones, which contain ligated fragments of DNA from two non-contiguous regions of the genome, can occur. For this reason, it is essential that each clone entering the sequencing mill be carefully analyzed to ensure that it contains sequences from a single genomic region. This is accomplished by a careful analysis of the fingerprint pattern. Ideally, the level of redundancy, the number of times a region is represented by a clone, is such that unique clones can be defined for the entire length of the genome.

Each of the large-insert clones is fractionated into small DNA fragments by physical means. The fragments are then modified by adding restriction-site cloning adapters. This permits the insertion of the fragments into standard plasmid cloning vectors. Currently plasmids are the vector of chose. Previously, the fragments were cloned into the M13 vector that generates high-quality, single stranded DNA that is ready for cloning. But sequence data could only obtained from one end. On the other hand, fragments placed into plasmid vectors can be sequenced from both ends. Sequence data collected from both ends of a clone are called ***read pairs*** or ***mate pairs***. The availability of read pairs (again of lengths of 500-800 bases) makes the subsequent assembly process simpler because sequences are known to reside in close proximity.

The fragments are cloned and subsequently sequenced. That sequence data is then assembled using computer algorithms. This process involves the alignment of sequences that overlap. Alignments are defined by the degree of accuracy. The human genome project adopted a level of 99.99% accuracy for what is called a ***full shotgun sequence***. As an example, it would take about 2000 sequence reads to generate an 8-10 fold coverage of a 100 kb BAC clone. This level of coverage is typically sufficient to achieve the standard level of accuracy. Midway to the full shotgun sequence is the ***working draft sequence***. This typically is achieved with a 3-5 fold coverage of the BAC. Such a draft could be useful as a first look at the genome, to identify potential SSR markers, and to get a general sense of what types of genes are located in a specific region of a genome. If an individual investigators wishes to study a particular BAC region in depth, the working draft provides a good starting point.

The final step is the development of the ***finished sequence***. Gaps can appear in the sequence of a BAC, or certain regions of the clone might not achieve the full shotgun standard. Efforts at finishing the sequencing of these specific regions of the genome are called ***directed sequencing.*** This process involves the sequencing of additional subclones of the clones. This, though, could be tedious and expensive. Alternatively, primers are developed to sequences

bordering the troublesome region. These primers are used to amplify BAC DNA that is then directly sequenced. Alternatively, this same approach could be employed using genomic DNA as the template for PCR amplification.

As you remember from above, the first step in the entire process is to develop a relatively high-density molecular map to seed the initial physical map. This molecular map again is useful to authenticate the finished sequence. Each marker should be placed in the correct order along the sequence. In addition, the fragment size distribution generated for each BAC during the fingerprinting phase should be readily observable from the sequencing data. Collectively all of these steps provide a high degree of reliability of the genome sequence.

So where does the major sequencing projects stand today? The immense amount of work on the human genome lead to the publication of a working draft sequence in February 2001. Two years later, in April 2003, the finished sequence was announced. Had you heard about it? Probably not. The fanfare associated with release of the draft sequence had much to do with the private effort that competed with the on-going public project. The major difference, and controversy centered around the whole genome shotgun approach of Celera Genomics.

Whole Genome Shotgun Sequencing. The hierarchical sequencing approach begins by first generating a physical map. The overlapping clones that define the map are then shotgun cloned and sequenced. The ***whole genome shotgun sequencing* (*WGS*)** approach bypasses this entire step. Instead, nuclear DNA itself is sheared, modified by the addition of restriction site adaptors, and cloned into plasmids. These plasmids are then directly sequenced. This process, even more so than determining the sequence of a BAC, requires pair reads; again sequence data from both ends of the clone. This is especially true because of the repetitive nature of complex genomes.

As an approach, WGS has proven very successful for smaller genomes. It is essentially the only approach that is used to sequence smaller genomes such as bacteria. The suggestion that it would be useful for complex genomes was consider bold statement. But to some degree, this approach has value. The *Drosophila* genome sequence currently in use is the product of shotgun sequencing. In addition, the two rice genomes that were sequenced were the product of shotgun sequencing. It is important to note that efforts are under way for both of these species to finish the genomes by using the hierarchical approach.

The major problem with shotgun sequencing is overcoming the alignment problems associated with the repetitive DNA. One approach to this problem is to generate a collection of fragment libraries. Celera Genomics used fragments that were two, ten and fifty kb in size. In this manner, data from fragments containing different types of sequences can be collected.

In many ways, the difference between the two sequences approaches is one of scale during the assembly process. The WGS approach involves the simultaneous analysis of gigabytes of sequence data, whereas the amount of data needed to assembly smaller datasets generated by shotgun sequencing a BAC clone is orders of magnitude less. Therefore, on-going research is focusing on developing new algorithms to handle and assembly the huge data sets generated by WGS.

The recently published mouse genome sequence represents a future direction. The genome was first assembled based on data generated by WGS sequencing. A total of 29.7 million pair reads were collected, a value representing a 7.7-fold coverage of the genome. This data was then assembled used software developed since the Celera Genomics project. Neither mapping data nor clone-based sequences were used. This generated 224,713 sequence contigs. Although many of these are short contigs, the average length was 24.8 kb. A subset of the contigs, called supercontigs have an average length of 16.9 MB, a significant improvement over the results obtained by earlier software. The 200 largest supercontigs represent 98% of the euchromatin regions of the genome. These were next anchored to the chromosomes using several sets of mapping data. This data set represents 187 Mb of the euchromatic DNA or 96% of the genome. These results demonstrate the potential of the WGS approach.

Combined Approach Used With Rat Genome. The rat genome project used a combination of the HSS and WGS approaches. First, data was collected using the traditional WGS approach. Subsequently, BACs were sequenced using the traditional shotgun approach but only to an average of 1.8x. The BAC information is called a BAC skim. The BAC skim data was used to seed the WGS data to a BAC. This results in enriched BACs (eBACs) which are then united to form long sequence structures (bactigs), and subsequently superbactigs and finally ultrabactigs. BAC end sequence, fingerprinting, and finally genetic data is to used to create the final assembled genome. The project developers claim this uses the advantages of the HSS and WGS approaches with few of the disadvantages.

Pyrosequencing of Genomes in Picolitre Reactors

The goal of genome sequencing is to capture large amounts of DNA sequence data in as short of time frame as possible. The dideoxy chain termination procedure has certainly proven to be a very successful approach to collecting a large amount of sequencing data. But as with everything, we want more data in a shorter period of time with less labor input. A recent publication by Margulies et al (Nature 2005) describes such an approach. The research group works for 454 Life Sciences (www.454.com), and they have developed a procedure system that utilizes pyrosequencing of DNA samples bound to beads that are located in a wells of fibreoptic plate. The company is selling a full system that will perform the sequencing. So how does the system work? First we need to discuss ***pyrosequencing***.

At its most basic, pyrosequencing generates visible light in a directly proportional to the number of a specific nucleotide that are incorporated into a growing DNA sequence. The reagents for pyrosequencing are listed in Table 1.

Table 1. Pyrosequence reagents

DNA template (DNA_n)	Adenosine 5'phosphosulfate (APS)
DNA polymerase	Luciferase
A deoxynucleotide (dNTP; deoxyadenosine thio triphosphate substitutes for dATP)	Luciferin
ATP sulfurlyase	Apyrase

The following series of reaction generates light.

$$(1)\quad DNA_n + \text{a dNTP} \xrightarrow{\text{DNA Polymerase}} DNA_{n+1} + PP_i$$

$$(2)\quad PP_i + APS \xrightarrow{\text{ATP Sulfurylase}} ATP$$

$$(3)\quad ATP + \text{Luciferin} \xrightarrow{\text{Luciferase}} \text{Oxyluciferin} + \text{Light}$$

$$(4)\quad \text{dNTP} \xrightarrow{\text{Apyrase}} \text{dNDP} + \text{dNMP} + \text{phosphate}$$

$$(5)\quad ATP \xrightarrow{\text{Apyrase}} \text{dADP} + \text{dAMP} + \text{phosphate}$$

The first important point is that only a single deoxynucleotide is used for each synthesis step. Therefore, if the dNTP in reaction is dTTP, and the next complementary deoxynucleotide in your template is adenine, then you will generate one unit of light from this reaction. If on the other hand, the next two complementary bases in the reaction are adenine, you will generate two units of light. And finally, if the next base is any of the other three deoxynucleotide (guanine, cytosine, or adenine) no light will be generated because pyrophosate (PP_i). The light is captured and quantified by a charge coupled device (CCD) camera.

Once this step is completed, it is important to clean the system of the residual dNTP and ATP before another dNTP is added and the process repeated. The enzyme apyrase [reactions (4) and (5)] degrades these reagents. Finally, a new nucleotide is added, and the entire process is repeated to determine the next base in the sequence.

Clearly it would be very labor intensive to perform this sequencing on individual templates. Therefore a useful system would be one that allows the simultaneous sequencing of many templates in parallel. 454 Life Sciences has also developed such a system. Their system also has the added feature of not requiring individual clones for sequencing, an absolute requirement for both hierarchical and whole genome shotgun sequencing. Here is how it works.

First, total genomic DNA is sheared, primer adaptors are added, and the DNA is made single-stranded. A single strand of DNA is then bound to a bead, and a single bead is encapsulated inside an emulsion. In this way, each emulsified bead can be considered to represent a single reaction vessel. All of the beads are then collectively treated enmass. Complementary primers are used, and the DNA is amplified using PCR. The result is that each bead has multiple copies of the same DNA fragment bound to the bead.

Next the beads are loaded into a specially manufactured fibreoptic plate (60 x 60 mm^2) that contains about 1.6 million wells. Each well contains a single bead and is the location where the pyrosequencing reaction will take place. The 454 machine then delivers all of the reagents

for a single pyrosequencing step simultaneously to all of the wells. The backside of the plate is in contact with a CCD camera. The camera collects data simultaneously from each well following each pyrosequencing step. Therefore, the system has the capability to collect sequencing data for up to1.6 million reactions simultaneously. The system is also designed so that it can deliver the sequencing reagents, wash the wells following a reaction, and deliver the next set of reagents.

In the Nature article, the researchers used their newly designed equipment to sequence the *Mycoplasma genitalium* genome. Table 2 lists some of the important summary statistics that demonstrates the utility of the 454 pyrosequencing system.

Table 2. Summary statistics for pyrosequencing of the *Mycoplasma genitalium* genome using the 454 system.

Activity/result	**Statistic**
Preparation of sequencing library	4 hr
Number of fibreoptic wells used	900,000
Pyrosequencing	6 hr (42 cylces of the 4 nucleotides)
Average read length	110 nt
Read length range	80 – 120 nt
Number of bases read	33.7 million bases
Number of bases with Phred 20 score	26.7 million bases
Number of contigs	25 (vs. 28 from Sanger approach)
Average contig length	22.4 kb

This paper demonstrates the feasibility of this sequencing approach for prokaryotic genomes. Two technical improvements are needed before this approach can be of wide use for the development of a draft genome sequence. The first issue is read length. These short read lengths are problematic for eukaryotic genomes with repetitive elements because it becomes difficult to assembly long contigs. One solution would be to improve the read length of each individual fragment. Secondly, the ability to perform pair-end reads would also improve contig development. The authors note these issues in the manuscript.

So where would it be useful to apply the 454 pyrosequencing technology today? One use would be for genetic diversity studies. Once a genome has a draft sequence available, the use of the 454 technology on other genotypes of the species would uncover genetic diversity at a significantly reduced cost relative to whole genome shotgun sequencing. The value of the draft sequence is that contig development using the 454 data would be easier and should lead to longer contigs.

Genome Sequencing

Concept of Genome Sequencing

- Fragment the genomic DNA
- Clone those fragments into a cloning vector
- Isolate many clones
- Sequence each clone

Sequencing Techniques Were Well Established

- Used for the past twenty years
- Helped characterize many different individual genes.
- Previously, the most aggressive efforts
    - Sequenced 40,000 bases around a gene of interest

How is Genomic Sequencing Different???

- 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
  - Basic technique is still the same
- Major changes
  - Thermostable polymerase enzymes
    - Improves quality of sequencing products
  - Fluorescently labeled nucleotides for the reaction
    - Allows for laser detection
  - Laser-based detection systems
    - 8, 16 or 96 samples analyzed simultaneously
  - Results for a single run
    - 500-700 bases of high quality DNA sequence data
  - Human Project peak output
    - 7 million samples per month
    - 1000 bases per second
- Robotics
  - 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
  - Sequence reads today
    - $2.50 vs. $15 in the early 1990s.

Hierarchical Shotgun Sequencing

- Two major sequencing approaches
  - Hierarchical shotgun sequencing
  - Whole genome shotgun sequencing
- Hierarchical shotgun sequencing
  - Historically
    - First approach
  - Why???
    - Techniques for high-throughput sequencing not developed
    - Sophisticated sequence assembly software not availability
- 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

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 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
- 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

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
    - 493 cosmid (smaller insert clones) clones
        - Relatively high number of clones for genome size

Other Uses Of A Physical Maps

- Rich source of new markers
- Powerful tool to study genetic diversity among species
- Prior to whole genome sequencing
    - Markers can locate a target gene to a specific clone
    - Gene can be sequenced and studied in depth

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

Sequencing Clones Of The Minimal Tiling Path

- Steps
    - Physically fractionate clone in small pieces
    - Add restriction-site adaptors and clone DNA
        - Allows insertion into cloning vectors
            - Plasmids current choice
    - 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**
      - Assembles sequence contigs only from high quality reads
  - Working draft sequence
    - 93-95% accuracy
    - 3-5 x coverage of 100 kb BAC clone

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

Confirming the Sequence

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

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

Whole Genome Shotgun Sequencing (WGS)

- Hierarchical sequencing approach
    - Begins with the physical map
    - Overlapping clones are shotgun cloned and sequenced
- WGS
    - Bypasses the mapping step
- Basic approach
    - Take nuclear DNA
    - Shear the 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

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 consider a bold suggestion
    - Large public effort dedicated to hierarchical approach
    - *Drosophila*
        - Sequenced using the WGS approach
    - Rice
        - Two different rice genomes sequenced using WGS approach
            - Only developed a working draft though
            - Public hierarchical sequence available; publication released in August 2005

WGS – Major Challenge 1

- Assembly of repetitive DNA is difficult
  - 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 is mostly useful to the developers
    - Examples: Celera Assembler, Arcane, Phusion, Atlas

WGS – Major Challenge 2

- For the two sequences approaches
- Assembly is a scale issue
    - WGS approach
        - Gigabytes 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 = 25 kilobases in length
- Super contig subset
    - Mean length =16.9 megabases
- 200 largest supercontigs
    - Anchored using mapping data
    - Represents 96% (9187 Mb) of the euchromatic region of genome

Rat Genome Project: A combined approach

- Nature (2004) 428:493
- Combination of hierarhcial shotgun and whole genome shotgun sequencing

- WGS sequence reads
  - 36 million quality reads (34 million used for assembly)
  - 7X coverage
  - 60%: Whole genome shotgun data
  - Insert size: <10 kb, 10 kb, 50 kb, >150 kb
  - 40% BAC data
  - Small insert clones from the BAC

- BAC Skim
  - A low density sequence analysis of a BAC
  - 21,000 clones analyzed
  - 1.6X coverage

- Enriched BACS
  - Sequences developed by combining WGS data and BAC skim data

- BAC Fingerprinting
  - 200,000 BACs fingerprinted
  - 12X coverage
  - 11,274 fingerprint contigs (FPC) developed
  - Clones selected from contigs for BAC skim

- Bactig
  - Overlapping BACs
  - 1MB in length

- Superbactigs
  - Bactigs joined by paired-end reads
  - Mean = 5MB in length
  - 783 total for the genome

- Ultrabactigs
  - Mean = 18 MB
  - 291 total for genome
  - Synteny data, marker data, and other data used to define the ultrabactig

Pyrosequencing in Picolitre Reactors

Pyrosequencing reagents

- DNA template (DNAn)
- DNA polymerase
- A dideoxynucleotide
  - dNTP
  - deoxyadenosine thio triphospate substitutes for dATP
- ATP sulfurlyase
- Adenosine 5' phosphosulfate (APS)
- Luciferase
- Luciferin
- Apyrase

Pyrosequencing reactions

(1) $\text{DNAn} + \text{a dNTP} \xrightarrow{\text{DNA Polymerase}} \text{DNAn+1} + \text{PPi}$

(2) $\text{PPi} + \text{APS} \xrightarrow{\text{ATP Sulfurylase}} \text{ATP}$

(3) $\text{ATP} + \text{Luciferin} \xrightarrow{\text{Luciferase}} \text{Oxyluciferin} + \text{Light}$

(4) $\text{dNTP} \xrightarrow{\text{Apyrase}} \text{dNDP} + \text{dNMP} + \text{phosphate}$

(5) $\text{ATP} \xrightarrow{\text{Apyrase}} \text{dADP} + \text{dAMP} + \text{phosphate}$

- Important points about pyrosequencing reaction
    - One nucleotide is introduced at a single time
    - Data is collected from a charge coupled devise (CCD)
        - Used to detect light emission
        - A single photon of light is detected for each nucleotide introduced
    - After reaction is complete
        - New set of reagents (different nucleotide) is introduced
        - Repeated many steps to collect sequence data

454 Life Sciences DNA Sequencing System

- Utilizes pyrosequencing to collect sequence data

Preparation of sequencing template

- Genomic DNA is
    - Sheared
    - Adapators added to the end
    - Made single-stranded
    - SS DNA is bound to a bead
    - Single bead/DNA combination encapsulated in emulsion
    - DNA is duplicated on the bead in the emulsion
    - Each emulsion DNA bead is a single DNA reaction vessel

Fibreoptic plates

- Plate contains 1.6 million wells
- CCD mounted to back of plate
- One DNA/bead emulsion is loaded per well

How it works

- Pyrosequencing reagents (one dNTP at a time) are added
- CCD collects sequence results for each well for that dNTP
- Residual sequencing reagents washed out
- New reagent for second dNTP added
- Process continues until finished

Results with Mycoplasma genitalium genome sequencing

Activity/result	**Statistic**
Preparation of sequencing library	4 hr
Number of fiberoptic wells used	900,000
Pyrosequencing	6 hr (42 cycles of the 4 dNTPs)
Average read length	110 nt
Read length range	80 – 120 nt
Number of bases read	33.7 million bases
Number of Phred 20 score bases	26.7 million bases
Number of contigs	25 (vs. 28 for Sanger approach)
Average contig length	22.4 kb

Major constraint

- Read length
  - 100-150 bases
- For large genomes
  - Reads to short construct contigs of significant length

Potential utility

- Rapidly collect sequence of additional genotypes of a species already sequenced
- Use original sequence as a reference
- Discover useful SNPs for the species

Sequencing the Gene Space: An alternative to whole gnome sequencing

Background

Major goal of genome sequencing

- Define the gene set
- Large genomes create a problem
    - High ratio of non-coding to coding DNA
        - ex. Human genome is 3% coding (or gene-based) DNA

Not all genomes are equally valued

- Many crop species have little support for complete genome sequencing
    - Why???
        - Not model species like Arabidopsis, rice or *Medicago truncatula* (legume model)
        - Genomes are complex with large amount of repetitive sequences

Features of the maize genome: an example

- 2500 Mega base pairs
- 80% repetitive DNA
- 60-70% retrotransposons
- 50 kb stretches of DNA without transposons are rare
- Whole genome shotgun assembly would fail
- Repetitive elements are too homologous to allow assembly

Alternative approach

- Sequence just the gene space
    - What current approaches are available???
        - EST sequencing
        - Methyl filtration
        - High Cot sequencing

Expressed **S**equenced **T**ag (EST) Sequencing

What nucleic acid fraction contains expressed genes???
mRNA

- Sequencing the mRNA fraction will discover genes
    - Drawback
        - Gene models (exon-intron) will not be discovered

Approach to discovering genes via ESTs

- Create multiple cDNA libraries from
    - Different developmental stages
    - Different tissues
    - Various environmental stresses
- Perform end sequencing = EST
- Assembly ESTs
    - **T**entatitive **C**onsensus (TC) sequences
        - TCs are full or nearly full genes
        - Complied from multiple ESTs
        - Unique sequence
    - Single EST
        - Don’t assembly into a contig
        - Unique sequence from those in TC
    - Lack
        - Introns
        - 5’ and 3’ controlling regions

Current Status of Plant EST Projects

Species	# of ESTs in assembly	# of total unique	# of TCs
Arabidopsis thaliana	353,003	62,010	28,900
Potato (*Solanum tuberosum*)	189,864	38,239	21,063
Soybean (*Glycine max*)	330,436	63,676	31,928
Barley (*Hordeum vulgare*)	370,546	50,453	23,176
Tomato (*Lycopersion esculentum*)	162,621	31,838	16,268
Medicago truncatula	226,923	36,878	18,612
Rice (*Oryza sativa*)	274,018	89,147	36,381
Bean (*Phaseolus vulgaris*)	21,290	9,484	2,906
Pine (*Pinus taeda*)	327,484	45,557	23,531
Sorghum (*Sorghum bicolor*)	187,282	38,148	20.029
Wheat (*Triticum aestivum*)	580,155	122,282	44,954
Corn (*Zea mays*)	407,423	58,582	31,375

from: The Institute for Genomic Research (http://www.tigr.org/tdb/tgi/plant.shtml)

Methyl Filtration Sequencing

Methylation status of the maize genome

- 25% of cytosines (C) in maize are methylated
- 50% of CG and CXG sites are methylated
- Important feature of exons
    - 95% of maize exons are under methylated
- Conclusion
    - Sequencing the under methylated region = sequencing the gene space

Principles of methyl filtration

- Use specific bacterial restriction modification systems
    - 5' methyl cytosine system
        - mcrBC+ strain
            - Cloned DNA with methylated cytosine are not propagated
            - DNA containing expressed genes (exons) greatly enriched in the library
- Sequencing
    - Sequence each clone as a read pair

Maize methylation filtration results; a preliminary analysis

	***E. coli* strain**	
Sequence type	**mcrBC-**	**mcrBC+**
Total reads	439	303
Total repeats	53.7%	13.2%
Annotated repeats	48.7%	7.6%
Unannotated repeats	5.0%	5.6%
Total exons	2.3%	10.2%
Known exons	1.4%	8.2%
Hypothetical exons	0.9%	2.0%

from: Rabinowicz et al. 1999. Nature Genetics 23:305.

Large scale 5’ methyl cytosine analysis of maize genome

- Palmer et al Science 302:2115

	Strain	
Sequence type	**mcrBC-**	**mcrBC+**
Total reads	5,679	96,576
Total repeats	57.0%	24.0%
Total exons	1.4%	8.6%

- 93% of repeats removed by methyl filtration

Comparison to rice sequences

- 81% of reads matched a rice gene
- More genes discovered than using expressed sequence tags (ESTs)

High Cot Sequence Analysis

High Cot analysis

- Another approach to obtain DNA from gene rich region
- Based on well defined solution hybridization procedure
- Time course during hybridization
    - Early (Low Cot DNA): highly repetitive sequences
    - Mid (Middle Cot DNA): middle repetitive sequences
    - Late (High Cot DNA): low copy or single copy DNA; enriched for genes
- Approach
    - Collect high Cot DNA
    - Clone and sequence the DNA
    - Annotate

Current Status of Maize Genome Sequence Survey (GSS)

Project of Plant Genome Database at Iowa State University

- Data set available:
    - 2.7 million reads from:
        - Methyl filtration clones
        - High Cot clones
        - BAC end sequences
        - Random genomic sequencing
- Assembly
    - 294,425 contigs
        - based on 1,661,712 reads
    - Total contig size
        - 503 million bp
- Full length genes discovery
    - Align contigs and compare to known gene structures
        - 4,062 maize genes
            - 32 super families
            - 252 gene families