Genome Sequencing--Strategies

Bio 4342
Spring '04

What is a genome?

A genome can be defined as the entire DNA content of each nucleated cell in an organism.

Each organism has one or more chromosomes that contain all of its genetic information.

Humans, for example, have a genome that is encoded on 46 chromosomes, organized into 23 pairs. One chromosome of each pair is inherited from the mother and one from the father. One pair of chromosomes determines gender (“sex chromosomes”) and the other 22 pairs are called autosomes.

The object of the Human Genome Project was to determine the entire DNA sequence of each of these long DNA molecules (chromosomes), and to locate and identify each of the human genes.

Factors that determine sequencing strategy

- Genome size
- Chromosomal structure
- Repeat content and character
- Polymorphism/inbreeding
- Desired product (draft, finished)
- Supporting information (physical or genetic maps, closely related sequenced genomes, EST sequences)

Genome sequencing process

- Library creation
- Production sequencing
- Assembly
- Prefinishing
- Finishing
- x2
Genome size determines...
- Number and types of subclones
- Number of sequence reads
- Compute power/algorith for sequence read assembly
- In general, genome size and complexity scales with the size of the organism

Chromosomal structure
- Centromeres and telomeres
- Pairing at meiosis—even numbers??
- Numbers and types/extent of sequence duplications

Repeat content and character
- Length and degree of degeneracy of repeats is important to know
- Cot-based methods can help to characterize the repeat classes in a genome
- Some newer strategies eliminate repeat content from libraries via “Cot subtraction”

Polymorphism/inbreeding
- The extent to which an organism is inbred determines the degree of polymorphism in the genome
- A high degree of polymorphism can complicate assembly, depending upon the assembly algorithm used
- A preliminary assessment of polymorphism can determine the sequencing strategy
### The desired end product determines strategy

- “Finished” means completed to contiguity and high quality with no/few gaps in the sequence
- “Comparative draft” means ordered and oriented contigs with gaps that are sized
- “Draft” means assembled into contigs (and supercontigs if possible), without order and orientation

### Supporting information

- Genetic maps typically provide context in terms of simple sequence repeats that occur “near” genes
- Physical maps provide a context for the sequence assembly to scaffold onto, or provide an ordered series of clones for clone-based sequencing
- The availability of genome sequence for a closely related organism can provide some support for assembly validation
- ESTs are partial gene sequences obtained by cloning and sequencing mRNA populations

### General considerations

- Timeline
- Desired end product
- Library making capacity/ability
- $$
- Algorithms available for sequence assembly

### Common sequencing strategies

- Whole genome shotgun (WGS)
- Clone-by-clone (physical map required)
- Hybrid approach (WGS and clone-by-clone)
- Map-assisted WGS (physical map required)
Whole genome sequencing strategy

- Sequencing by whole genome shotgun using several subclone insert sizes/types
- Large insert clone end-sequences used for increased contig alignment and scaffolding
- Anchoring of assembled contigs to the genome accomplished through computational identification of known markers

Whole genome shotgun

- Advantages include few libraries and rapid accumulation of sequence information
- Disadvantages include lack of genome-scale order/organization confirmation including repeats assembly
- WGS is best suited for low complexity, small genomes (e.g. bacterial)

WGS Strategy

- Arachne
- PCAP
- Apollo
- Celera assembler
- Etc.
- All WGS algorithms deal poorly with repeated regions

WGS Assembly algorithms
Clone-by-clone strategy

- A clone-by-clone strategy involves the initial creation of a physical map of large-insert clones (BACs or fosmids).
- Physical maps are built by generating restriction digests of individual clones in excess (typically 15x coverage), then using computer-based algorithms to assemble contigs of related clones.
- A “minimal tile path” is then generated, to provide minimally overlapping large insert clones that span the genome.
- These clones then form the basis of sequencing efforts, including the creation of shotgun sequencing libraries from the individual tile path clones.
- Finishing is then done on a large insert clone basis, with computer assembly of finished clones to provide whole chromosome sequence.

Bacterial Artificial Chromosomes (BACs)

- BACs are pieces of DNA called “vectors” that contain specific sequences:
  - a BAC can be put into a bacterial cell, such as E. coli
  - the BAC DNA is replicated and copies go to daughter cells during cell division
  - we try to put ~100,000 base long pieces of DNA into BACs
  - BAC clones containing genomic DNA pieces represent a manageable size of DNA for mapping studies—they can be amplified, isolated and characterized by restriction digestion to determine what sequences they share.

BAC Fingerprints

- Digested BAC DNA is electrophoresed on 1.2% agarose gels for 8 hours at 3 V/cm. Each gel contains 96 sample and 25 marker lanes.

How are BAC fingerprints used??

- We use them to create a fingerprint map...
- A fingerprint map is a set of fingerprints (restriction digests of BACs) that are assembled into “contigs”.
  - Contigs are clusters of related BAC clones (they share component pieces, judging by their fingerprints)
  - Together, the BAC clones in a contig represent the genomic region from which the clones were derived
- FACT: The human and mouse genomes each required over 300,000 BAC fingerprints!!

How did we put all those together?
Computer-aided BAC fingerprint assembly

Individual gel lanes are identified

Peaks are used to represent the sizes of the fragments

Fingerprint contig assembly

A computer program called FPC is used to look at each BAC clone fingerprint and determine relationships.

Using FPC, overlapping clones with common restriction fragments are identified.

The process is repeated to assemble the contig until no further clones can be incorporated.

Manual editing of contigs is usually required.

A contig from the human BAC fingerprint map

What is the end result of fingerprint mapping?

1. Most BAC clones have been assembled into contigs.
2. The length (in base pairs) of the contigs is roughly equal to the anticipated size of the genome.
3. The BAC fingerprint map can be used to select a set of BAC clones that overlap one another to a minimal extent. These so-called “minimal tiling path” BACs will be used for the next phase of genomic mapping...
Clone-by-clone strategy

- Advantage is provided by the physical map—an independent means of confirming sequence assembly in advance of sequencing. A well developed algorithm for clone-based assembly is freely available (phrap)
- Disadvantages include large numbers & types of libraries required, need to wait for physical map to be ~complete before selecting minimal tile path for sequencing
- Mainly appropriate for large complex genomes with high degree of polymorphism

Mouse/Rat genome sequencing

- No need to follow the same path as the human genome
- Two main components:
  - Whole genome shotgun
    - high genome coverage
    - rapid genome survey (aid human annotation)
  - BAC-by-BAC
    - BACs selected from fingerprint database
    - low sequence coverage
  - BAC end-sequences of good quality

Hybrid approach

- Assemble WGS reads and add to BAC shotgun reads

Hybrid approach strategy

- Advantages include lots of early data from WGS, time to build physical map, pre-determined genome assembly
- Disadvantages include need to generate many different libraries/types, few/poor assembly algorithms available, expensive
- Mainly applied to large, more complex genomes
Map-assisted WGS strategy

- Physical map constructed of large insert clones concurrent with WGS sequencing of variable insert size subclones
- End sequences of mapped clones enable linking of map contigs to sequence assembly contigs, supercontigs result
- Linkage enables the organization of finishing into discrete units (supercontigs) of known order

Maplink Viewer

Map-assisted WGS strategy

- Ultimately, this strategy seeks to organize the sequence assembly using the map, and to complete the map using the sequence assembly
- Gap closure of the map and sequence can be accomplished by identifying gap-spanning clones, shotgun sequencing them, and assembling in the completed clone sequence to the assembly
- This is a newer strategy, and algorithms are currently being developed or modified to enable it

Conclusions

- Many factors contribute to selecting a genome sequencing strategy
- Strategies and the algorithms to enable them are constantly being developed and refined
- Generating the production-style data for a genome project is rarely the rate-limiting step—finishing and gap closure are
- Having a physical map is key to being able to verify a genome assembly, but newer strategies entwine the sequence and map building processes