Computational Molecular Biology Symposium

March 12th, 2003

Carnegie Mellon University

Organizer: Dannie Durand

Sponsored by the Department of Biological Sciences and the Howard Hughes Medical Institute

Human Genetics and Genomics

12:30 Introduction
Dannie Durand, Carnegie Mellon

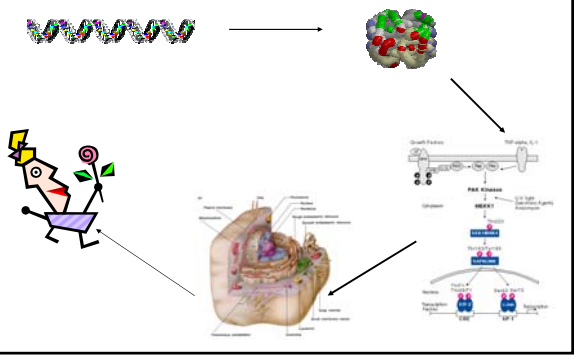
12:45 Genome Assemblies and Interval Graphs
Martin Farach-Colton, Rutgers

2:15 Patterns of Human Genetic Diversity:
Implications for Human Evolution and Disease
Sarah Tishkoff, Maryland

3:15 Algorithms for extracting information from Human Genetic Variation
Russell Schwartz , Carnegie Mellon

4:30 Meet the speakers, 3301 NSH

From Genes to Organism



Human Genetics and Genomics

- Acquisition
- Interpretation

Human genome sequence

February, 2001

Human Genome

Complete set of genetic information in each cell
"Human blueprint"
Consensus sequence:
Genetic material common to all humans

- Cellular function
- Tissue type differentiation
- Development
- Species specific traits

Human Genome

Complete set of genetic information in each cell
"Human blueprint"
Consensus sequence:
Genetic material common to all humans
Genomic variation:
Differences between individuals

- Genetic basis for disease
- Human history
- Human evolution

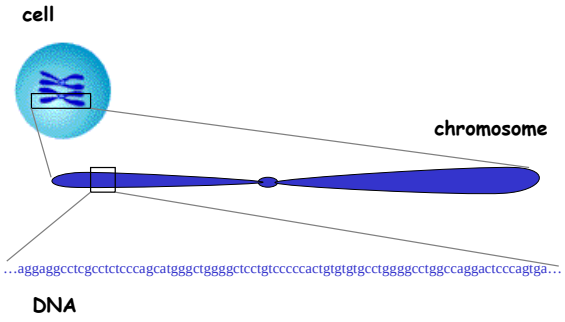
Human Genetics and Genomics

- **Acquisition**
- Interpretation

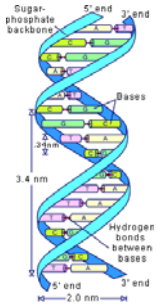
Human genome sequence

February, 2001

DNA Sequencing:
A whirlwind introduction

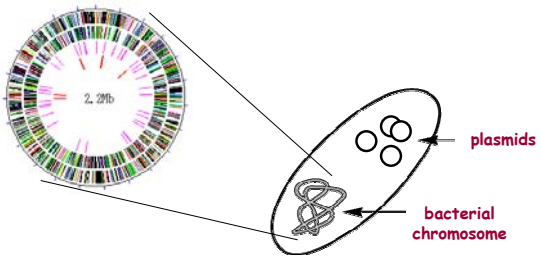


DNA

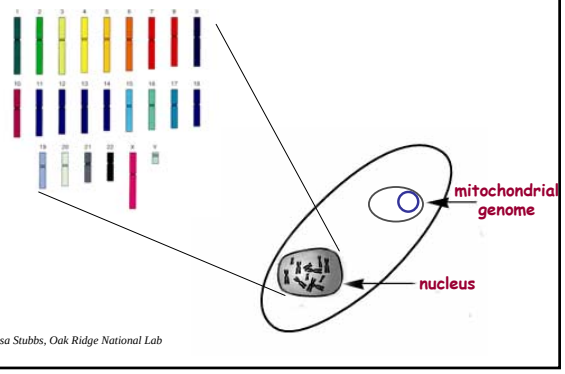


Double-stranded polymer
Four letter alphabet: **A, C, G, T**
Base pairing: **A:T, G:C**
Orientation

E. Coli Genome



Human Genome



Lisa Stubbs, Oak Ridge National Lab

Genome Complexity

Organism:	***Mbases***	***# of genes:***
E. Coli	4.6	~4,000
Baker's yeast	12.1	~6,000
Fruit fly	180.0	~13,000
Worm	97.0	~18,000
Human	3200.0	~30,000
Human mitochondrion	17kb	37

Genome Features

- Genes

…aggaggcctcgcctctcccagcatgggctggggctcctgtcccccactgtgtatgcctggggcctggccaggactcccagtga…

protein coding sequence

A gene is a location on a chromosome that encodes a protein

Genome Features

- Genes
- Regulatory regions

…aggaggcctcgcctctcccagcatgggctggggctcctgtcccccactgtgtatgcctggggcctggccaggactcccagtga…

promotor

Regulatory regions: non-coding sequences that determine transcription

Genome Features

- Genes
- Regulatory regions
- Repeated regions
  - Signature pattern
  - Length of pattern
  - Copy number
  - Distribution within the genome

…aggcgagagagagagagagcctggggcctggccaggctggggctcctgtccagagagagagaagtga…

Genome Features

- Genes
- Regulatory regions
- Repeated regions
  - Make up >50% of the human genome
  - Complicate sequence assembly
  - Vary from one individual to the next

    Useful markers in studying diversity

…aggcgagagagagagagagcctggggcctggccaggctggggctcctgtccagagagagaagtga…

DNA Sequencing

➢ Tools for manipulating DNA

Enzymes that cut, paste and modify

- Methods for copying DNA fragments
- Methods for determining the sequence of fragments
- Assembling fragments into finished sequence

DNA Sequencing

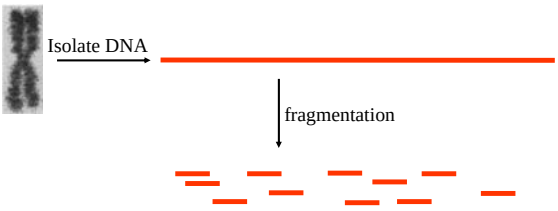
- Tools for manipulating DNA

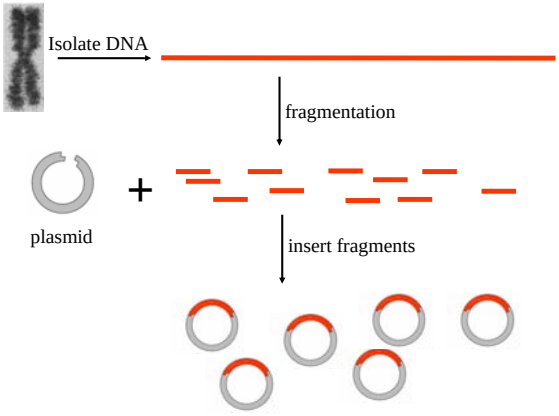
Enzymes that cut, paste and modify

➢ Methods for copying DNA fragments

- Methods for determining the sequence of fragments
- Assembling fragments into finished sequence

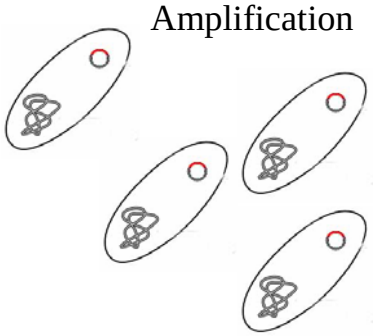
Cloning – copying fragments



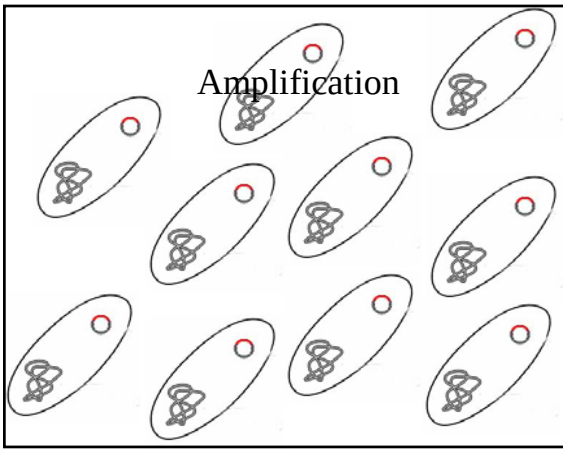


Amplification

Amplification



Amplification



Cloning vectors

Properties
- Size of insert
- Host
- Stability of insert

Examples

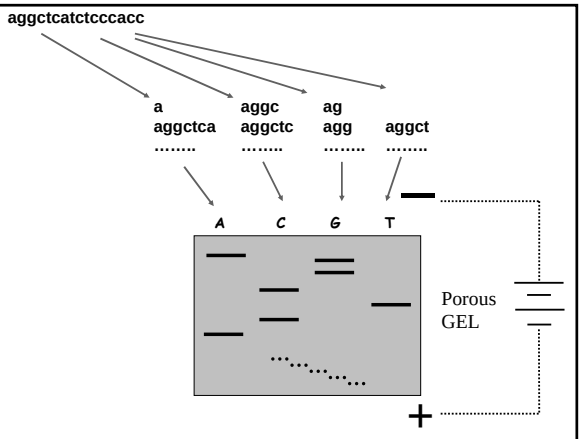
– *Plasmids*	5-10 Kb
– Lambda Phage	20 Kb
– *BAC (Bacterial Artificial Chrom.)*	100-200 Kb
– YAC (Yeast Artificial Chrom.)	1000 Kb

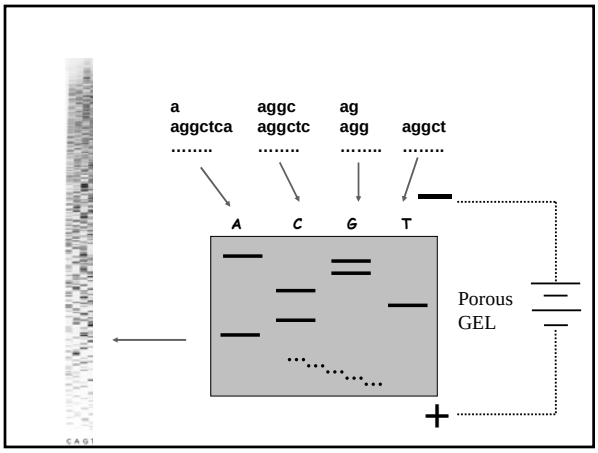
Fragment Sequencing

- Tools for manipulating DNA
  Enzymes that cut, paste and modify
- Methods for copying DNA fragments
- ➢Methods for determining the sequence of fragments
- Assembling fragments into finished sequence

Fragment Sequencing

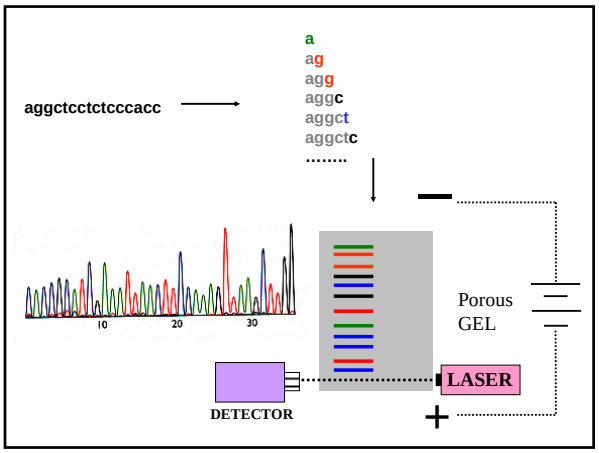
Given a pool of a particular DNA fragment
- Generate all prefixes (those enzymes again)
- Sort them by size (gel electrophoresis)
- Read base composition of fragment





Improvements in Sequencing Technology

- Fluorescent bases



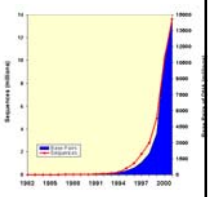
Improvements in Sequencing Technology

- Fluorescence bases
- Automation
- Polymerase Chain Reaction (PCR)
- Capillary-based sequencing machines

ABI 3700 sequencer

History of Sequencing

1971 Nobel prize for restriction enzymes
1973 First recombinant DNA
1980 Nobel prize for DNA sequencing
1988 Congress establishes Genbank
1995 First genomic sequence
1998 First multicellular organism
2000 Fly genome
2000 First plant genome
2001 Human genome
2003 Mouse genome



22 million sequences
28 billion base pairs

DNA Sequencing

- Tools for manipulating DNA
  Enzymes that cut, paste and modify
- Methods for copying DNA fragments
- Methods for determining the sequence of fragments

➢ Assembling fragments into finished sequence

Sequence Assembly

Limits of gel electrophoresis: ~ 500bp in one "read"

To sequence more than 500 bp:

Sequence 500bp fragments separately

Combine *computationally* using sequence comparison

Human Genetics and Genomics

- Acquisition
- **Interpretation**

Human genome sequence

February, 2001

Variation within the Human Genome

Polymorphism –

- Occurrence of more than one type of genetic feature within a population.
- A common variation in the sequence of DNA among individuals.

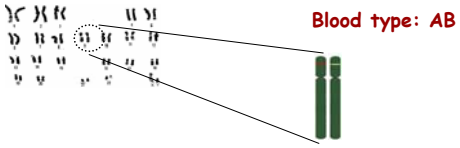
Variation within Human Populations

Polymorphisms:

- **Alleles – variant types of the same gene**
- Single Nucleotide Polymorphisms (SNPs)
- Haplotypes
- Tandem Repeats
- Indels

Example: Blood Groups

Variation within an individual:



Variation within a population:

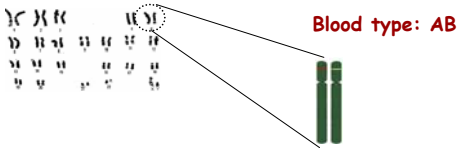
Blood type is *polymorphic* in the human population.

Blood groups: A,B,O

allele – One of the variant forms of a gene at a particular locus

Example: Blood Groups

Variation within an individual:



Variation within a population:

Blood type is *polymorphic* in the human population.

Blood groups: A,B,O

Variation within Human Populations

Polymorphisms:

- Alleles – variant types of the same gene
- **Single Nucleotide Polymorphisms (SNPs)**
- Haplotypes
- Tandem Repeats
- Indels

Single Nucleotide Polymorphisms

GCTTTAATTATCACGATATAATATAACGATT
GCTTTAATTTTCACTATATAATATAACGATT
GCTTTAATTTTCACGATATACTATAACGATT
GCTTTAATTATCACGATATAATATAACGATT
GCTTTAATTTTCACTATATAATATAACGATT

SNP:

Variation at a single nucleotide position.

Roughly one every 1,000 bases in the human genome.

Variation within Human Populations

Polymorphisms:

- Alleles – variant types of the same gene
- Single Nucleotide Polymorphisms (SNPs)
- **Haplotypes**
- Tandem Repeats
- Indels

Haplotypes

GCTTTAATTATCACGATATAATATAACGATT
GCTTTAATTTTCACTATATAATATAACGATT
GCTTTAATTTTCACGATATACTATAACGATT
GCTTTAATTATCACGATATAATATAACGATT
GCTTTAATTTTCACTATATAATATAACGATT

A genomic region that is inherited as a unit

Variation within Human Populations

Polymorphisms:

- Alleles – variant types of the same gene
- Single Nucleotide Polymorphisms (SNPs)
- Haplotypes
- **Tandem Repeats (copy number)**
- Indels

…aggcgagagagagagagagagcctggggcctggccaggctggggctcctgtccagagagagagaagtga…

Variation within Human Populations

Polymorphisms:

- Alleles – variant types of the same gene
- Single Nucleotide Polymorphisms (SNPs)
- Haplotypes
- Tandem Repeats
- **Insertions and deletions**

gccaggctgagagcctgctcctgtccaga

…aggcgcctggacttgtgcctggccaggctgagagcctgctcctgtccagagtgag…

Interpretation of genomic variation

- Genetic basis for disease
- History
  - Patterns of migration, epidemics, population subdivision
- Anthropology
  - Linguistics, religion
- Evolution

Interpretation of genomic variation

Use polymorphism to address questions like:

- Are all jews biologically related?
- Where did the Basques come from?
- Did humans originate only in Africa?
- When did humans come to the new world?
- Why is sickle cell disease more prevalent among African Americans?
- Did humans interbreed with Neanderthals?

Acknowledgements

- Annette McCleod
- Jennifer Sciullo
- Nicole Reading
- Narayanan Ragupathy, Nan Song
- Ken Pesanka
- Lauren Ward
- Catherine Copetas
- Beth Jones