BF528 - Introduction to Genetics and Genomics

NHGRI Lecture Series

These materials were developed in part from this excellent lecture series at the NIH:

http://www.genome.gov/12514288

Origin of “Genomics”: 1987

“For the newly developing discipline of [genome] mapping/sequencing (including the analysis of the information), we have adopted the term GENOMICS… The new discipline is born from a marriage of molecular and cell biology with classical genetics and is fostered by computational science.”

- McKusick and Ruddle, A new discipline, a new name, a new journal, Genomics, Vol. 1, No. 1. (September 1987), pp. 1-2

What is genomics?

"**Genomics** is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes (the complete set of DNA within a single cell of an organism)."

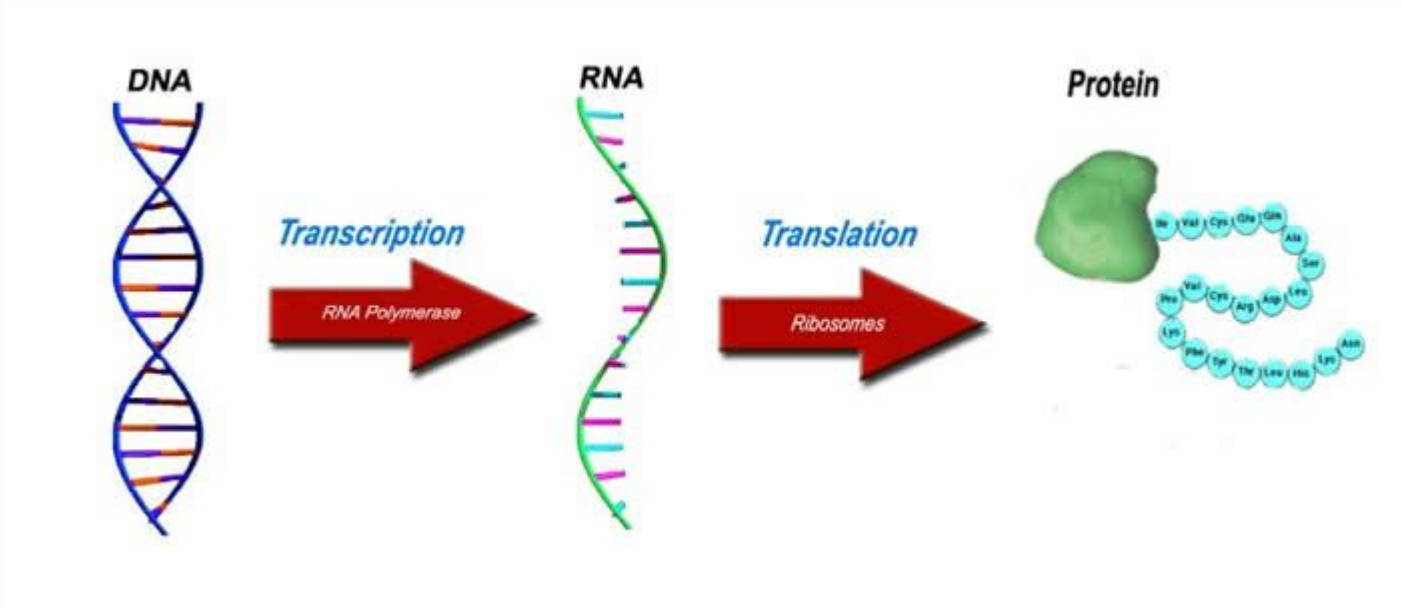
- Wikipedia

What is genomics?

“Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks.”

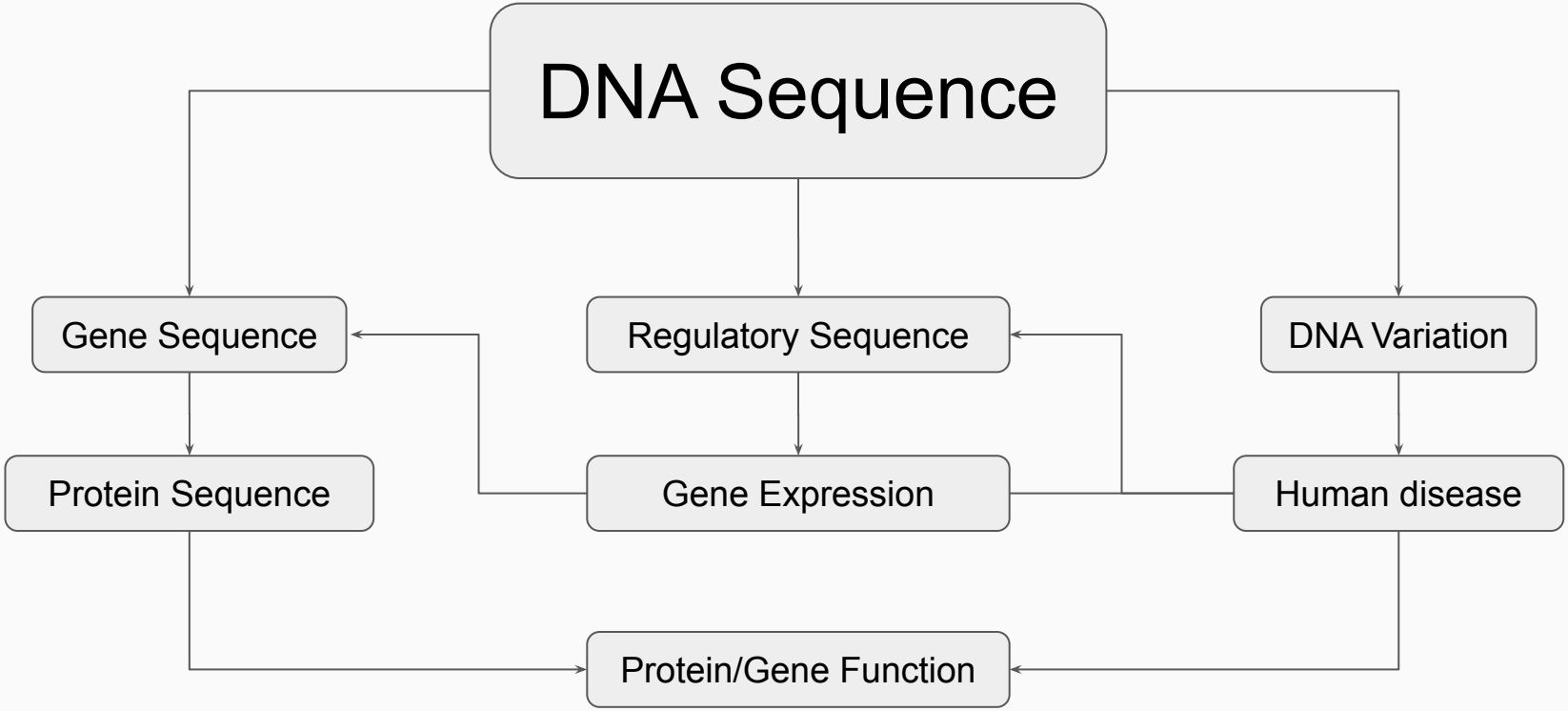
- *Wikipedia*

Central Dogma of Biology



http://www.lhsc.on.ca/Patients_Families_Visitors/Genetics/Inherited_Metabolic/Mitochondria/DiseasesattheMolecularLevel.htm

What can genomics tell us?

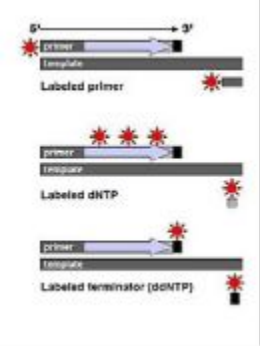


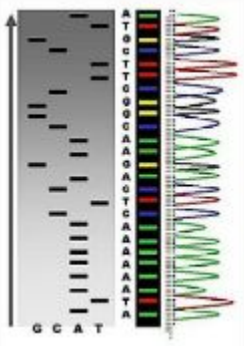
How do we study genomics?

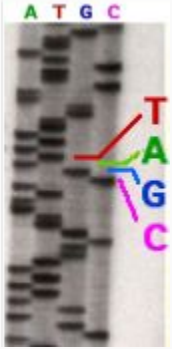
1. Isolate nucleic acid molecules from biological samples
2. Determine nucleotide sequence using biochemical techniques
3. Digitize nucleotide sequence
4. Examine digital sequences to identify patterns with algorithms and statistics
5. Relate patterns to biological observations by:
6. Comparing patterns detected across many samples
7. Manipulating a system to see how patterns change

Sequencing Techniques

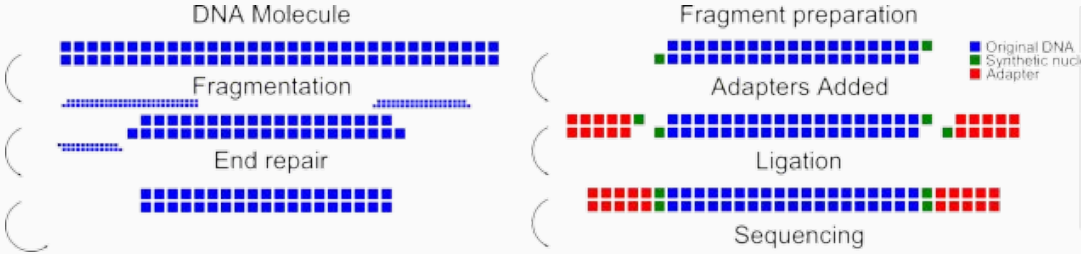
Sanger sequencing – fluorescent-labeled DNA fragments







Sequencing by synthesis, pyrosequencing



Adapted from http://web.uri.edu/gsc/next-generation-sequencing/

Sequence Analysis

- **Assembly** – putting short sequences together to reconstruct a longer, source sequence
- **Mapping** – locating where one short sequence is found in a longer sequence
- **Pattern recognition** – looking for specific patterns within sequences that have special meaning

In each of these cases, sequences are **aligned** to one another

Sequence Alignment

- Provides a *measure of relatedness*
- Alignment quantified by *similarity* (% identity)
- Useful for any sequential data type:
  - DNA/RNA
  - Amino acids
  - Protein secondary structure
- High sequence similarity *might* imply:
  - Common evolutionary history
  - Similar biological function

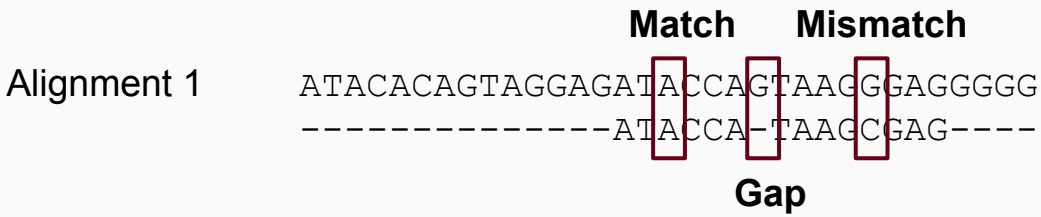
What Alignments Can Tell Us

- Homology - Orthologs, Paralogs
- Genomic identity/origin of a sequence/individual
- Genome/gene structure
  - Genic structure (exons, introns, etc)
  - RNA 2D structure
  - Chromosome rearrangements/3D structure

DNA Sequence Alignment Example

Sequence 1 ATACACAGTAGGAGATACCAGTAAGGGAGGGGG

Sequence 2 ATACCATAAGCGAG



Scoring/Substitution Matrices

- Given alignment, how “good” is it?
- Higher score = better alignment
- Implicitly represent evolutionary patterns

	A	C	G	T	-
A	2	-3	-1	-3	-3
C	-3	2	-3	-1	-3
G	-1	-3	2	-3	-3
T	-3	-1	-3	2	-3
-	-3	-3	-3	-3	NA

```
ATACCAGTAAGGGAG
ATACCA-TAAGAGAG
```
Score = 22

```
ATACCAGTAAGG-GAG
ATACCA-TAAG-AGAG
```
Score = 19

```
ATACCA-GTAAGGGAG
A-TACCATAAGAGAG-
```
Score = -20

Sequence Alignment Algorithms

- **Global** alignments - beginning and end of both sequences must align
- **Local** alignments - one sequence may align anywhere within the other
- Multiplicity:
  - Pairwise alignments (2 sequences)
  - Multiple sequence alignment (3+ sequences)

Global Alignment

Both sequences are aligned from end to end

```
AAANTAIYYDPNPDMP
A--NTAI-YDPN--M-
```

Interior sequences are aligned as well as possible

```
AERAKDNLCRLEHTTLRKVTAAANTAIYYDPNPDMPVVAEDQEWVNVYYEM
A-----N------T-----------AI-YD--P------------N----M
```

However, sequences of vastly different length can produce meaningless alignments

Local Alignment

Alignment may begin and end at any position

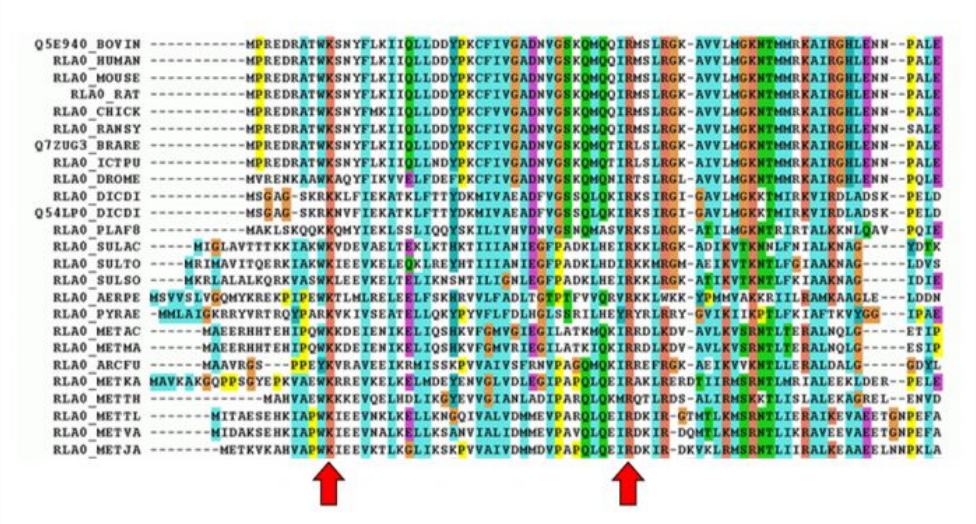
```
AAANTAIYYDPNPDMP
-AANTAI-YDPN--M-
```

```
AERAKDNLCRLEHTTLRKVTAAANTAIYYDPNPDMPVVAEDQEWVNVYYEM
--------------------AANTAI-YDPN--M---------------
```

Local alignment may produce better alignments when sequence lengths differ greatly

Multiple Sequence Alignment

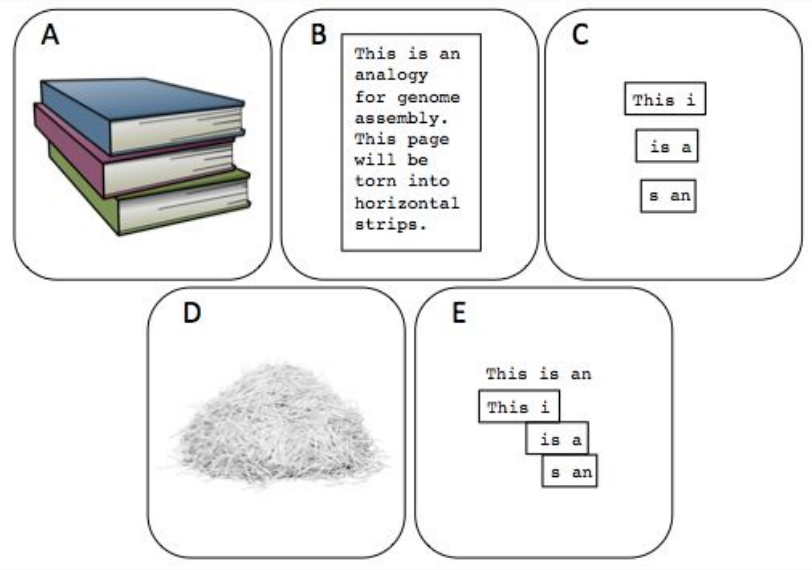
Like pairwise alignment, but with N sequences



Sequence consensus among many species suggests evolutionary pressure

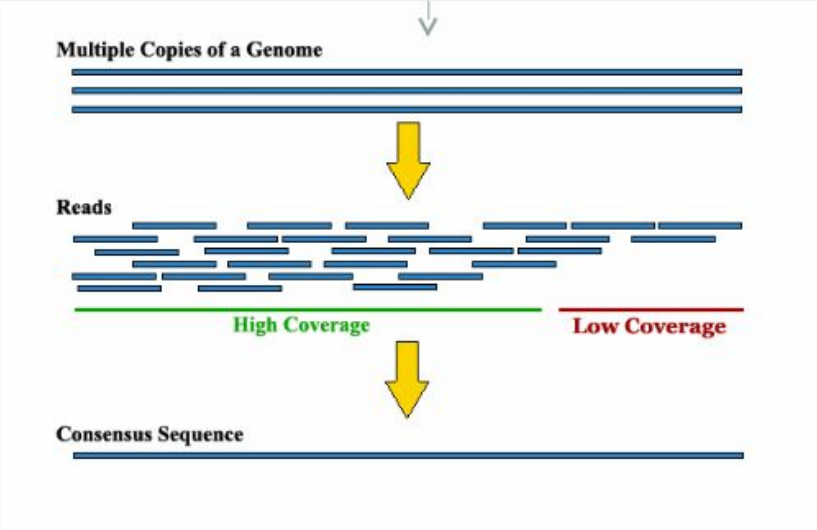
Alignment Examples

Example: Genome Assembly



If your genome was a book that had its sentences chopped into fragments, assembly is analogous to reconstructing all the sentences.

We need multiple copies of each book (genome) to arrive at a *consensus* text (DNA sequence) of the original



Great explanation of DNA sequence assembly: http://gcat.davidson.edu/phast/

Example: Genome Assembly

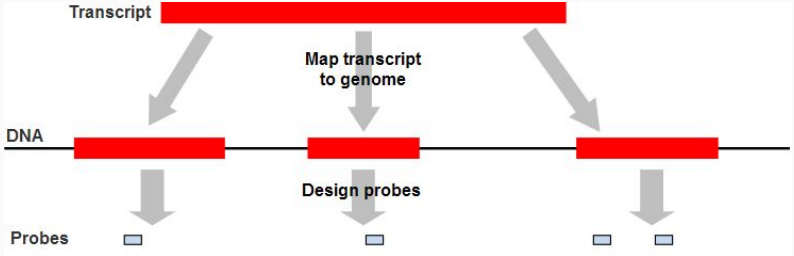
An error?
A polymorphism?
A different allele?
Incorrect alignment?

Greedy approach: take most frequent nucleotide at each aligned position

Great explanation of DNA sequence assembly: http://gcat.davidson.edu/phast/

Example: Exon Microarray Probes

- Microarray *probes* are short single-stranded DNA sequences from a reference genome
- Exon Microarrays have probes only from exons

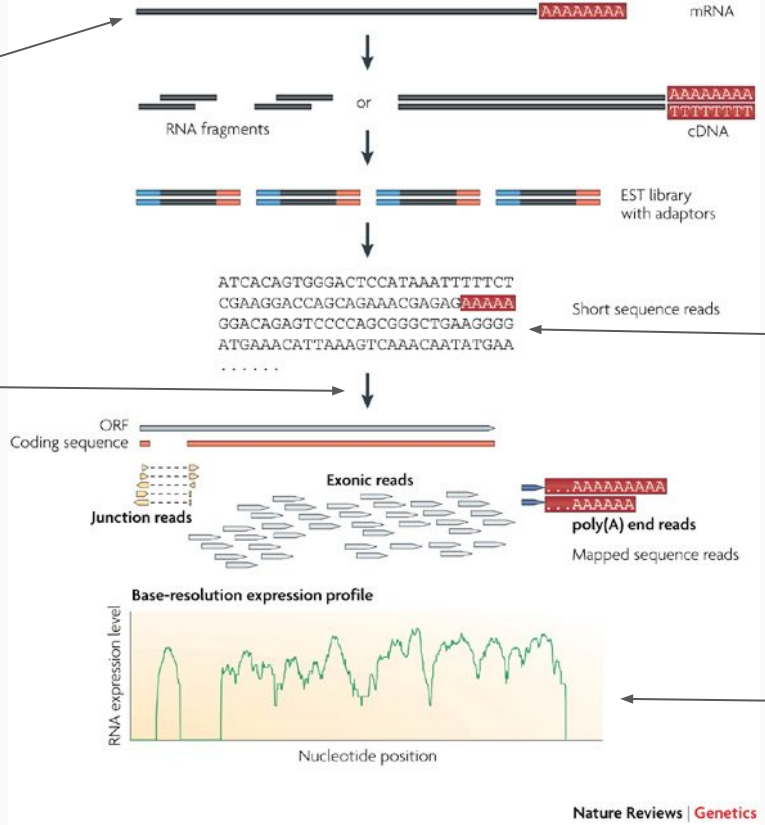


- Exon probes must map to the correct exon, BUT
- Probes must NOT map anywhere else, they must be *unique in the genome*

Example: mRNA-Seq Analysis

Start with a pool of mRNA molecules

Find all locations where sequences map in genome

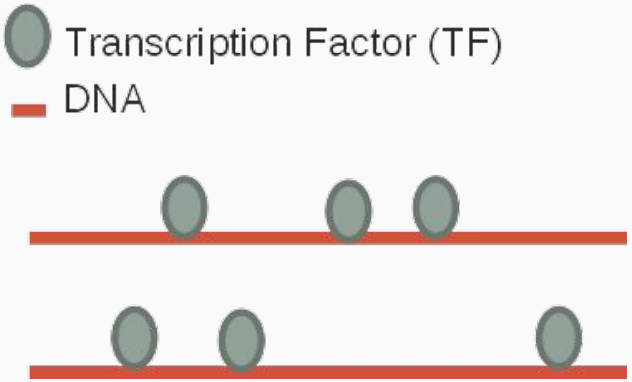


Millions of DNA sequences 30-150 nucleotides long

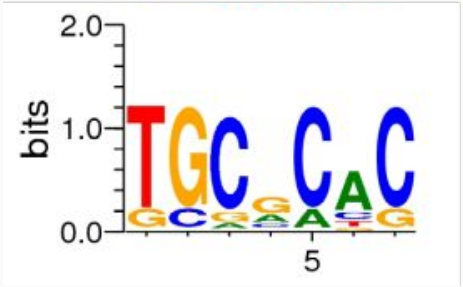
Count the number of sequences that map to individual regions (e.g. genes)

Example: DNA Binding Site Discovery

Identify genomic regions where a particular TF is bound across the entire genome



...TGCTCAC...
...TGCACAC...
...GGCGCAC...
...TGAGCAC...
...TCGCCTC...
...TGCAACG...
...TCCCACG...
...GGTACTC...



By extracting and aligning the DNA sequence corresponding to these binding events, we can identify which DNA sequences this TF tends to bind

Human Genomics and Gene Expression

The Human Genome Project

- Planning begins 1984, launched 1990, “completed” 2001, “finished” 2004
- Championed by Dr. Charles DeLisi
- Overview of the Human Genome Project: http://www.genome.gov/12011238

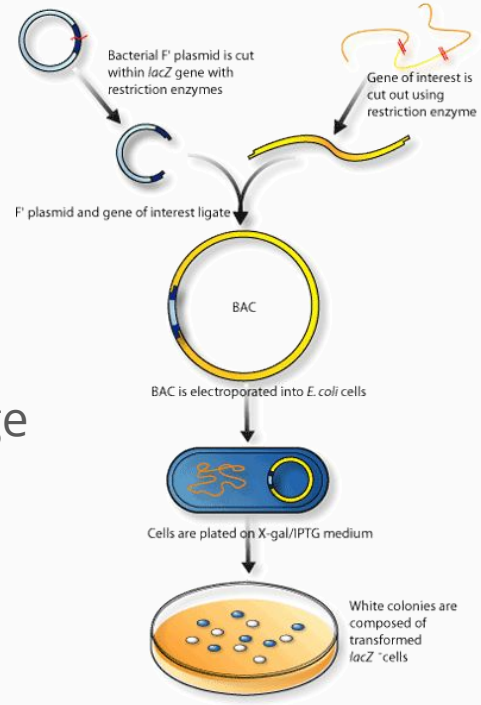
Human Genome Composition

- Key findings:
  - ~20k genes
  - More segmental duplications than expected
  - Fewer than 7% of protein families vertebrate specific
  - ~3% of sequence codes for protein coding genes
  - >85% of the genome is transcribed
  - Repetitive elements may comprise >66% of genome

How The Genome Was Determined

International Human Genome Sequencing Consortium

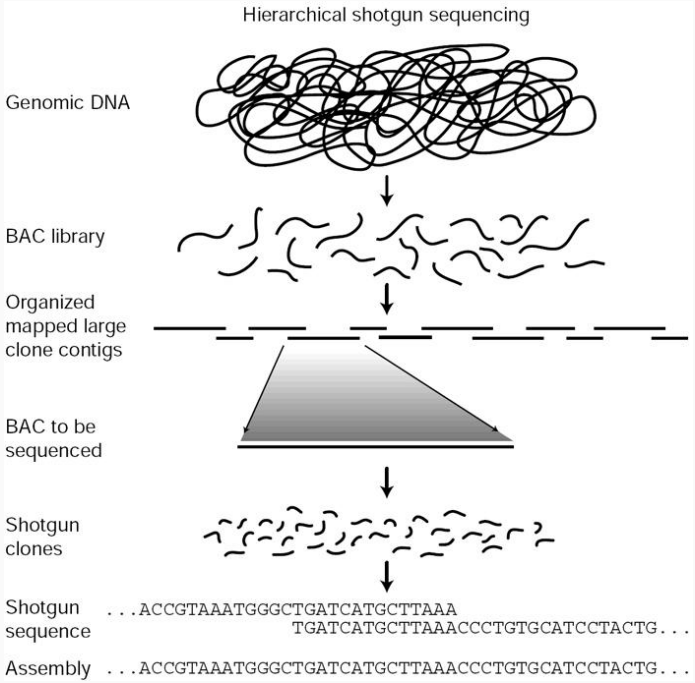
- Fragment DNA with restriction enzymes
- Ligate fragments into bacterial artificial chromosomes (BACs)
- Amplify BACs with tagged DNA fragments
- Fragment isolated BAC vectors
- Sequence via Sanger-style sequencing to 4x coverage
- Finished draft genome in ~10 years



How The Genome Was Determined

Celera Technologies: shotgun sequencing

- Used public BACs contigs from the Human Genome Project and theirown
- Much shorter DNA reads, assembled later in silico using the HGP BAC clones as a scaffold
- Finished draft genome in ~3 years



The Genome Is All About Genes

- Genic sequences
- What do our genes do?
- How are genes controlled?
- What genes are different between humans?
- How are genes associated with disease?

Gene Expression

Gene Expression

"**Gene expression** is the process by which information from a gene is used in the synthesis of a functional gene product."

- Wikipedia

But What Is A Gene?

- A specific DNA sequence
- A fundamental unit of inheritance
- A molecule created by transcription of an RNA product (then translated into a protein) which has a function
- A “gene” is an abstract concept

But What Is A Gene?

- DNA?
- RNA?
- Protein?
- Informational molecule?
- Functional molecule?

Yes, all of them

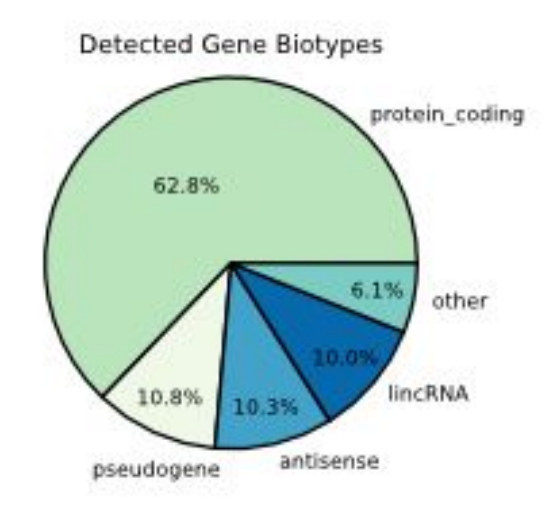
What Is Gene Expression?

- Active mRNA transcription?
- mRNA abundance?
- mRNA translation?
- RNA function?
- Protein abundance?
- Protein function?

Yes, all of them

The Gene Expression Landscape

- mRNA - protein coding genes
- Functional non-coding RNA (ncRNA) biotypes:
  - microRNA (miRNA)/small interfering RNA (siRNA)
  - Long (intergenic) non-coding RNA (lncRNA/lincRNA)
  - Ribosomal RNA (rRNA)
  - Transfer RNA (tRNA)
  - Many more (30+)
- Antisense: transcript initiated from TSS in opposite direction of primary gene
- Pseudogenes



How We Measure Gene Expression

- mRNA transcription/translation
  - Fluorescent tagging + microscopy
  - ribosomal capture
- mRNA abundance
  - Northern blots
  - Quantitative polymerase chain reaction (qPCR)
  - Microarrays
  - High-throughput sequencing

How We Measure Gene Expression

- Protein abundance
  - Western blots
  - Fluorescent tagging + microscopy
  - Mass spectrometry
  - Protein arrays
- mRNA/Protein localization
  - Fluorescent tagging + microscopy

mRNA Measurement Considerations

- Most mRNA quantification techniques measure *steady state abundance*
- mRNA measurements are *snapshots*
  - Measure large populations of cells to quantify “average” abundance
- Poor concordance between mRNA and corresponding protein abundance