Introduction to Molecular Biology

CS 4364/5364
Spring 2021
At the highest level

Organism are made up of one or multiple cells

inside the cell is the nucleus, which contains the DNA

humans are *diploid* meaning we have 2 copies of each chromosome (one from each parent)
The Central Dogma

**DNA**
- double stranded
- contains all of the information for "you"
- only about 1.5% of the human genome encodes proteins

**RNA**
- {A, C, U, G}

**Proteins**

**DNA**
- Transcription
- Translation
The Central Dogma

Transcription
- process of uncoiling, separating, and copying DNA into RNA
- first stage is called "pre-mRNA" in the case of protein coding genes
The Central Dogma

RNA
- pre-mRNA undergo splicing to remove the *introns* and leave only (some) *exons*
- some RNA perform functions on their own and are not spliced, called ncRNA (non-coding RNA)
The Central Dogma

RNA
• pre-mRNA undergo splicing to remove the *introns* and leave only (some) *exons*
• some RNA perform functions on their own and are not spliced, called ncRNA (non-coding RNA)
The Central Dogma

RNA

- pre-mRNA undergo splicing to remove the *introns* and leave only (some) *exons*
- some RNA perform functions on their own and are not spliced, called ncRNA (non-coding RNA)
# The Central Dogma

## Translation

- 3-letter groups of RNA characters, **codons**, are converted to amino acids, the building blocks for proteins

### Table

<table>
<thead>
<tr>
<th>First Char.</th>
<th>Second Character</th>
<th>Third Char.</th>
<th>DNA</th>
<th>RNA</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>N</td>
<td>S</td>
<td>AAC</td>
<td>ACC</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>U</td>
<td>U</td>
<td>AUC</td>
<td>AUC</td>
<td>A</td>
</tr>
<tr>
<td>AAA</td>
<td>A</td>
<td>I</td>
<td>AUA</td>
<td>AUG</td>
<td>G</td>
</tr>
<tr>
<td>AAG</td>
<td>K</td>
<td>M/start</td>
<td>AGC</td>
<td>U</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>H</td>
<td>R</td>
<td>CUC</td>
<td>CGC</td>
<td>G</td>
</tr>
<tr>
<td>CAA</td>
<td>Q</td>
<td>R</td>
<td>CCU</td>
<td>CGU</td>
<td>C</td>
</tr>
<tr>
<td>CAG</td>
<td>Q</td>
<td>A</td>
<td>CUA</td>
<td>CGA</td>
<td>A</td>
</tr>
<tr>
<td>CCA</td>
<td>P</td>
<td>A</td>
<td>CUG</td>
<td>CCG</td>
<td>G</td>
</tr>
<tr>
<td>CAU</td>
<td>P</td>
<td>S</td>
<td>UUC</td>
<td>UGC</td>
<td>C</td>
</tr>
<tr>
<td>CAU</td>
<td>P</td>
<td>F</td>
<td>UUU</td>
<td>UGU</td>
<td>C</td>
</tr>
<tr>
<td>CAA</td>
<td>P</td>
<td>F</td>
<td>UUA</td>
<td>UGA</td>
<td>R</td>
</tr>
<tr>
<td>CAG</td>
<td>P</td>
<td>F</td>
<td>UUG</td>
<td>UGG</td>
<td>G</td>
</tr>
<tr>
<td>U</td>
<td>Y</td>
<td>C</td>
<td>UCC</td>
<td>UGC</td>
<td>C</td>
</tr>
<tr>
<td>UUA</td>
<td>Y</td>
<td>C</td>
<td>UCU</td>
<td>UGU</td>
<td>C</td>
</tr>
<tr>
<td>UAU</td>
<td>Y</td>
<td>C</td>
<td>UCU</td>
<td>UGU</td>
<td>C</td>
</tr>
<tr>
<td>UAA</td>
<td>Y</td>
<td>C</td>
<td>UCA</td>
<td>UGA</td>
<td>R</td>
</tr>
<tr>
<td>UAG</td>
<td>Y</td>
<td>C</td>
<td>UCG</td>
<td>UGG</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>D</td>
<td>A</td>
<td>GCC</td>
<td>GGC</td>
<td>C</td>
</tr>
<tr>
<td>GAA</td>
<td>D</td>
<td>A</td>
<td>GCC</td>
<td>GGC</td>
<td>C</td>
</tr>
<tr>
<td>GAG</td>
<td>D</td>
<td>A</td>
<td>GCA</td>
<td>GGA</td>
<td>A</td>
</tr>
<tr>
<td>GAU</td>
<td>D</td>
<td>A</td>
<td>GCG</td>
<td>GGG</td>
<td>G</td>
</tr>
<tr>
<td>GAA</td>
<td>D</td>
<td>A</td>
<td>GCA</td>
<td>GGA</td>
<td>A</td>
</tr>
<tr>
<td>GAG</td>
<td>D</td>
<td>A</td>
<td>GCG</td>
<td>GGG</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>E</td>
<td>V</td>
<td>GUC</td>
<td>GGC</td>
<td>G</td>
</tr>
<tr>
<td>GAA</td>
<td>E</td>
<td>V</td>
<td>GUA</td>
<td>GGU</td>
<td>G</td>
</tr>
<tr>
<td>GAG</td>
<td>E</td>
<td>V</td>
<td>GUG</td>
<td>GGU</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>E</td>
<td>V</td>
<td>GUC</td>
<td>GGC</td>
<td>G</td>
</tr>
<tr>
<td>GAA</td>
<td>E</td>
<td>V</td>
<td>GUA</td>
<td>GGA</td>
<td>A</td>
</tr>
<tr>
<td>GAG</td>
<td>E</td>
<td>V</td>
<td>GUG</td>
<td>GGG</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>E</td>
<td>V</td>
<td>GUC</td>
<td>GGC</td>
<td>G</td>
</tr>
<tr>
<td>GAA</td>
<td>E</td>
<td>V</td>
<td>GUA</td>
<td>GGA</td>
<td>A</td>
</tr>
<tr>
<td>GAG</td>
<td>E</td>
<td>V</td>
<td>GUG</td>
<td>GGG</td>
<td>G</td>
</tr>
</tbody>
</table>

### Diagram

- **DNA** \(\{A,C,T,G\}\)
- **RNA** \(\{A,C,U,G\}\)
The Central Dogma

Proteins
• Do stuff in the cell, including help with translation and transcription
When copying a genome "errors" may occur, these changes are what make people different

- 99.99% of our genomes are identical
- Single Nucleotide Polymorphism (SNP) -- a change at a single base
- Structural Variants (SV) -- large scale changes
Genetic Variants

When copying a genome "errors" may occur, these changes are what make people different

• 99.99% of our genomes are identical
• Single Nucleotide Polymorphism (SNP) -- a change at a single base
• Structural Variants (SV) -- large scale changes
Genetic Variants

- **Deleterious Mutations** -- changes that are harmful (lethal) to a cell
- **Germline Mutations** -- changes passed to offspring
- **Somatic Mutations** -- those not passed down
- **Heterozygous** -- different between copies
- **Homozygous** -- same on both copies
- **Allele** -- specific position on a chromosome

Copy 1

- homozygous allele

Copy 2

- heterozygous allele
Genetic Variants

• **Deleterious Mutations** -- changes that are harmful (lethal) to a cell
• **Germline Mutations** -- changes passed to offspring
• **Somatic Mutations** -- those not passed down
• **Heterozygous** -- different between copies
• **Homozygous** -- same on both copies
• **Allele** -- specific position on a chromosome

- Copy 1
  - homozygous allele
  - heterozygous allele

- Copy 2
  - ...TACACCGTACGATCG...
  - ...ATGTGGCATCTAGC...
  - ...TACACCGTACGATCG...
  - ...ATGTGGCATCTAGC...
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.

figure adapted from Gibson and Muse, 3rd Edition (2009)
Sanger Sequencing

The basis of all modern sequencing.
Second Generation Sequencing

Also called next generation sequencing

Based on the same principles, but at a much larger scale

Improvements were made in the amplification and reading with better microscopes

With this came shorter sequences
  • Sanger could do >1,000 bases (characters) at once but all done by hand, so 10s of sequences, very accurate
  • Illumina (current standard) ~250 base reads, 1,000,000s of sequences, some errors
Second Generation Sequencing

stack of NY Times, June 27, 2000

stack of NY Times, June 27, 2000 on a pile of dynamite

this is just hypothetical

BOOM

so, what did the June 27, 2000 NY Times say?
Second Generation Sequencing

NextGen sequencing also introduced paired-end reads
- take a long piece of sequence (much longer than the read size, but predictable size)
- sequence both ends but keep them together
- gives two reads that you know are a certain distance from each other
Recently Pacific Biosciences and Oxford Nanopore have introduced new technologies that:

• have long reads
• with high(er) error rates
<table>
<thead>
<tr>
<th></th>
<th>Sanger</th>
<th>Next-Generation</th>
<th>Third-Generation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Launched</strong></td>
<td>1977 Basic chemistry</td>
<td>2005 with significant improvements</td>
<td>2010 with significant improvements</td>
</tr>
<tr>
<td></td>
<td>1998 Modern form</td>
<td>since</td>
<td></td>
</tr>
<tr>
<td><strong>Estimated Error Rate</strong></td>
<td>0.001% - 1%</td>
<td>0.46% - 2.4%</td>
<td>11% - 14% (but decreasing)</td>
</tr>
<tr>
<td><strong>Cost</strong></td>
<td>🏷️ 🏷️ 🏷️</td>
<td>🏷️</td>
<td>🏷️ 🏷️ 🏷️</td>
</tr>
<tr>
<td><strong>Throughput</strong></td>
<td>📉</td>
<td>📉</td>
<td>📉</td>
</tr>
<tr>
<td><strong>Currently Available</strong></td>
<td>Applied Biosystems*</td>
<td>Illumina Ion Torrent*</td>
<td>Pacific Biosciences</td>
</tr>
<tr>
<td><strong>Platforms</strong></td>
<td></td>
<td>Qiagen (Europe)</td>
<td>Oxford Nanopore</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complete Genomics (China)**</td>
<td></td>
</tr>
<tr>
<td><strong>Clinical Uses</strong></td>
<td>Many (but dwindling)</td>
<td>Many (and growing)</td>
<td>Niche uses (today)</td>
</tr>
</tbody>
</table>

*Part of Thermo Fisher  **Part of BGI

Error rate figures courtesy of G. Corey Shan from https://shanguangyu.com/articles/comparison-of-sanger-NGS-and-SMS/
Sequencing Applications

whole genome sequencing

reads

adapted from figure 1.2 in Mäkinen, et al. 2015
Sequencing Applications

methylation

bisulphite sequencing

adapted from figure 1.2 in Mäkinen, et al. 2015
Sequencing Applications

targeted sequencing

adapted from figure 1.2 in Mäkinen, et al. 2015
Sequencing Applications

DNA

pre-mRNA

mRNA

RNA sequencing

adapted from figure 1.2 in Mäkinen, et al. 2015
Sequencing Applications

chromatin immunoprecipitation (ChIP) sequencing

adapted from figure 1.2 in Mäkinen, et al. 2015
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1866</td>
<td>Gregor Mendel discovers genetics using pea plants</td>
</tr>
<tr>
<td>1869</td>
<td>DNA was discovered</td>
</tr>
<tr>
<td>1871</td>
<td>Avert and McCarty show DNA carried genetic information</td>
</tr>
<tr>
<td>1953</td>
<td>Watson and Crick discovered the 3D structure of DNA</td>
</tr>
<tr>
<td>1961</td>
<td>Nirenberg maps DNA to proteins</td>
</tr>
<tr>
<td>1968</td>
<td>Discovery of restriction enzymes</td>
</tr>
<tr>
<td>1970s</td>
<td>Development of the first sequencing techniques</td>
</tr>
<tr>
<td>1985</td>
<td>Development of PCR</td>
</tr>
<tr>
<td>1986</td>
<td>Discovery of RNA splicing</td>
</tr>
<tr>
<td>1980-1990</td>
<td>Complete sequencing of genomes of small organisms</td>
</tr>
<tr>
<td>1990</td>
<td>Launch of the Human Genome Project</td>
</tr>
<tr>
<td>1998</td>
<td>Discovery of post-transcription RNA interference</td>
</tr>
<tr>
<td>2000</td>
<td>Announcement of the draft human genome</td>
</tr>
</tbody>
</table>
Major Ongoing Projects

ENCODE (The Encyclopedia of DNA Elements)
• Effort to identify all functions elements in the human genome

1000 Genomes Project
• Large sample size will hopefully show all (most) of the variation within the population

UK BioBank
• 500,000 UK genomes in great details

SRA (Sequence Read Archive)
• Public repository of all types of sequencing data

GWAS Catalog (Genome Wide Association Studies)
• Multiple studies for many possible purposes (i.e. cancer, disorders, etc.)