Statistics for Human Genetics and Molecular Biology
Lecture 1: Review Basic Terminology of Genetics

Dr. Yen-Yi Ho (yho@umn.edu)

Sep 09, 2015
Logistics

Lectures & Labs: M W F & Labs: 1:25 to 2:15
Office Hours: Yen-Yi MW 2:30-3:30
Cavan MW 2:30-3:30
Zhiyuan (Jason) Xu Tue 3-4p in Mayo A446
Hahne, Huber, Gentleman, and Falcon (2008): Bioconductor Case Studies
John Verzani’s SimpleR notes
Website: http://www.biostat.umn.edu/~cavanr/pubh7445.html
Goals for the Course

• Basic knowledge of R
• Basics of statistics for human genetics
• Basics of genetic data analyses using R/Bioconductor
• Interpreting results and simple diagnoses
Objectives of Lecture 1

- Review basic terminology of genetics
  - Central dogma of molecular biology
  - Chromosomes, genes, DNA, RNA, and proteins
  - Gene expression
  - Genetic variation
  - Mutations

- Technologies for Genome Analysis
Mendelian Genetics (1866)

Segregation of alleles in the production of sex cells
1. the principle of segregation
2. the principle of independent assortment
Mendelian Genetics Translates to Modern Genetics

- A parent contributes only a single chromosome within a pair to the offspring.
- A fixed location on a chromosome pair is called a **locus**, and only those loci coding (for proteins or functional RNA) are typically called **genes**.
- An **allele** is the state or type of genetic info at a locus on a single chromosome. Thus there are two alleles at each locus in an individual (for autosomes, and for sex chromosomes in females).
Example: A particular disease locus has two possible allele types in the population: d (the disease allele) and D (normal).

Genotype: the joint (unordered) state of the two alleles. Could be dd, DD (called **homozygous** genotypes), or Dd (**heterozygous** genotype).

Alleles that are common in the population are often called **wild type** while disease alleles are called **mutant**.

Phenotype: an observed trait we care about, such as disease status, etc.
Mendelian Genetics Translates to Modern Genetics

Huntington's Disease
- Dominant
  - Allele: D
  - Genotypes:
    - DD: Homozygous
    - Dd: Heterozygous
    - dd: Homozygous
  - Phenotypes:
    - DD: Affected
    - Dd: Dominant
    - dd: Unaffected

Sickle Cell Anemia or Cystic Fibrosis
- recessive
  - Allele: D
  - Genotypes:
    - DD: Homozygous
    - Dd: Heterozygous
    - dd: Homozygous
  - Phenotypes:
    - DD: Unaffected
    - Dd: Recessive
    - dd: Affected

Hemophilia
- Parents
  - Father: X  Y
  - Mother: X\(^1\) \ X\(^2\)
  - Son: X\(^1\) Y
  - Daughter (Carrier): X Y
  - Father (with Hemophilia): X Y
  - Mother: X\(^1\) \ X\(^2\)
  - Son: X\(^1\) Y
  - Daughter (Carrier with Hemophilia): X Y

Adapted from NHGRI Talking Glossary
Central Dogma of Biology: Classic View

DNA

transcription

RNA

translation

Protein
Example: Human genome

DNA is organized into **chromosomes**: 22 pairs of autosomes (1-22) and 1 pair of sex chromosomes (X,Y).

**Genes**, the functional units of heredity, are carried on chromosomes.

Plus the mitochondrial DNA
Humans have \( \approx 3 \times 10^9 \) base pairs in their nuclear genome.

<table>
<thead>
<tr>
<th>IUPAC code</th>
<th>Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>adenine</td>
</tr>
<tr>
<td>c</td>
<td>cytosine</td>
</tr>
<tr>
<td>g</td>
<td>guanine</td>
</tr>
<tr>
<td>t (or u)</td>
<td>thymine (or uracil)</td>
</tr>
<tr>
<td>r</td>
<td>a/g</td>
</tr>
<tr>
<td>y</td>
<td>c/t</td>
</tr>
<tr>
<td>s</td>
<td>g/c</td>
</tr>
<tr>
<td>w</td>
<td>a/t</td>
</tr>
<tr>
<td>k</td>
<td>g/t</td>
</tr>
<tr>
<td>m</td>
<td>a/c</td>
</tr>
<tr>
<td>b</td>
<td>c/g/t</td>
</tr>
<tr>
<td>d</td>
<td>a/g/t</td>
</tr>
<tr>
<td>h</td>
<td>a/c/t</td>
</tr>
<tr>
<td>v</td>
<td>a/c/g</td>
</tr>
<tr>
<td>n</td>
<td>any base</td>
</tr>
<tr>
<td>./-</td>
<td>gap</td>
</tr>
</tbody>
</table>
Gene

Gene: a functional and inheritable element in the genome, usually codes for a protein; human genome ≈20,000 genes. The gene consists of three major structures:

- Regulatory segment
- Exons
- Introns

source: http://www.nobelprize.org/educational/medicine/dna/a/replication/gene.html
Transcription is the process of making RNA from DNA.
Translation

Translation is the process of translating the sequence of nucleotide bases in DNA/RNA into a sequence of amino acids in a protein.
### SECOND POSITION

<table>
<thead>
<tr>
<th>FIRST POSITION</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>phenylalanine</td>
<td>U</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>leucine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tyrosine</td>
<td>U</td>
<td>C</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>cysteine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stop</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stop</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tryptophan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>leucine</th>
<th>proline</th>
<th>histidine</th>
<th>arginine</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>isoleucine</th>
<th>threonine</th>
<th>asparagine</th>
<th>serine</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>G</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* methionine

<table>
<thead>
<tr>
<th>G</th>
<th>valine</th>
<th>alanine</th>
<th>aspartic acid</th>
<th>glycine</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>C</td>
<td>A</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>

* and start

**Primary protein structure** is sequence of a chain of amino acids

**Amino Acids**

**Amino Acid**

### Diagram

- Primary protein structure:
- Amino acid structure:
  - Amino group: \( \text{NH}_2 \)
  - Carboxyl group: \( \text{COOH} \)
  - R group
- Main chain of amino acids
Gene Expression

Gene expression is a highly specific process. Only a small fraction of the genes are expressed, or turned "on," in any particular type of cell.

gene expression in different tissues
gene expression in the same tissue, but different points in time
Putting it all together

- **DNA:**
  Info on chromosome is static, and essentially the same across cells within the individual

- **mRNA:**
  Not as relevant as protein, but easier to quantify

- **Protein:**
  Difficult to quantify globally, though very relevant

source:

http://www.nobelprize.org/educational/medicine/dna/index.html
Source of Variation
Environment Vs. Gene

Any two individuals are 99.9% identical in their DNA
Genetic Variations (Polymorphisms)

That 0.1 % is very important in defining our differences

- single nucleotide polymorphisms (SNPs, every 300 nucleotide on average)
- small-scale mutation, insertions, deletions
- copy number variations (AAGAAGAAGAAG)

Mutations

**Micro**

DNA

Mutagenic event

Deletion

...GTGGAGTCTA CAGCTTATCGCT...

...CAGCTCAGAT GGCTATCGCT...

Insertion

...GTGGAGTCTA GCGCTATCGCT...

...CAGCTCAGAT CGGCTATCGCT...

Substitution

...GTGGAGTCTA CGCTATCGCT...

...CAGCTCAGAT GGCTATCGCT...

**Macro**

Deletion

Duplication

Inversion

Substitution

Translocation
Genome Analysis Technologies

1. DNA
   - Microarrays:
     SNP, Copy number variation (CNV), Methylation
   - DNA sequencing:
     SNP, Insertion, Deletion, Mutation, CNV, Methylation

2. mRNA
   - Microarrays
   - RNA sequencing

3. Protein
   - 2-D electrophoresis
   - Maldi-Tof mass spec
General Steps in Obtaining Gene Expression Data
General Steps in Next-Generation Sequencing
Next Lecture

- Review basic terminology of population genetics
  - Crossing Over
  - DNA Recombination
  - Genetic Markers
  - Genetic Association Analysis
- Structures of Genetic Data