Module 2 - Basic Genetic Background

(Fundamentals of) Statistical Genetics

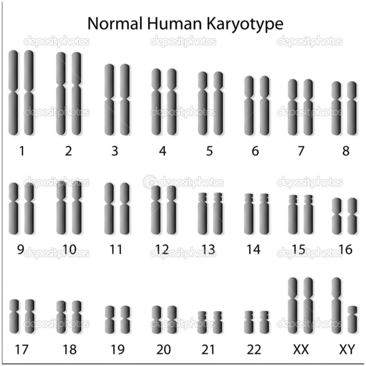
Lei Sun

Department of Statistical Sciences, FAS
Division of Biostatistics, DLSPH
University of Toronto

Genetic Terminologies

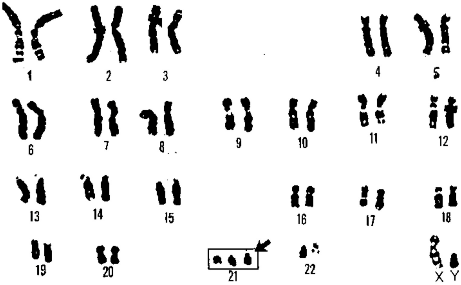
- From statistical point of view; sufficient for the purpose of this course.
- Check genetic books for precise molecular biological interpretations.
- Confusions and inconsistencies exist, e.g.
  - "Gene" was introduced long before the discovery of DNA and the understanding of its structure and functions.
  - "Interaction" might be interpreted differently by statisticians and geneticists.

Human Genome and Chromosomes



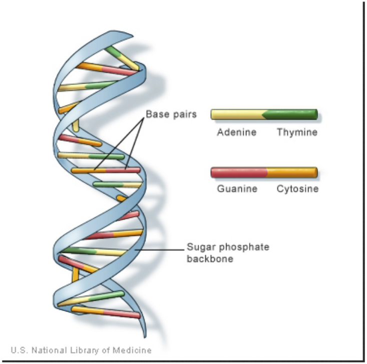
- **23 pairs of chromosomes**: 22 homologous pairs (**autosomes**), and 1 pair of sex chromosomes (**XX female, XY male**).
- Where genetic material is stored and in the nucleus of every cell.
- The 22 autosomes are numbered in order of decreasing length from 1 to 22 (except that 21 is slightly shorter than 22).
- **p-arm** (short-arm)
- **centromere** (links sister chromatids and "open up" during meiosis/cell division/reproduction)
- **q-arm** (long-arm).

Karyotype



- It is a microscopic examination of chromosome size and banding patterns.
- Allows medical laboratories to identify genetic diseases.
- The extra copy of chromosome 21 in this karyotype identifies this individual as having Down's syndrome.

DeoxyriboNucleic Acid (DNA) I



- *DNA is the basic biological material of inheritance; it determines how proteins are manufactured in the body.*
- **Each** chromosome has a **double** helix structure: two long strands of DNA, bounded to each other lengthwise.
- Each strand of DNA is a long molecule made up of a linear sequence of subunits/**base pairs**: ATGC.

  ('Size' of the genome: $\approx$ 3 billions of DNA base pairs)
- **A-T and G-C matching**: information on one strand is sufficient.

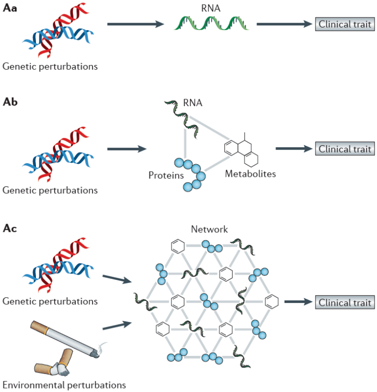
DeoxyriboNucleic Acid (DNA) II

- How DNA store information: language based on 4 letters ATGC, e.g.

```
     translated into              control over
AGA ---------------> protein  ------------> cellular
                     product                functions
```

- The translation is much more complex than this, involving RNA (the messenger) and network etc.

Complex Phenotype and Genotype Relationship



- Figure 1 of Civelek and Lusis (2014). Systems genetics approaches to understand complex traits. Nature Review.
- **Systems biology** *is an approach to understand the flow of biological information that underlies complex traits.*
- *It uses a range of experimental and statistical methods to quantitate [how to standardize] and integrate intermediate phenotypes, such as transcript, protein or metabolite levels, in populations that vary for traits of interest.*

Double Helix Structure $\neq$ Paired Chromosomes

- On one chromosome:

```
  AAGTCCAGA
-------------  one DNA strand
-------------  the other DNA strand
  TTCAGGTCT
   |
a base pair (bp)
```

- On the other **paired chromosome** of the **SAME** individual:

```
  TGAGGCTGC
-------------  one DNA strand
-------------  the other DNA strand
  ACTCCGACG
   |
a base pair (bp)
```

Strand Issue Important in Practice I

- Individual 1 (A-T,G-C) → AG

```
   AAGTCCAGA
-------------           strand 1
-------------           strand 2
   TTCAGGTCT
     |
   TGAGGCTGC
-------------           strand 1
-------------           strand 2
   ACTCCGACG
```

- Individual 2 (A-T,G-C) → TC

```
   AAGTCCAGA
-------------           strand 1
-------------           strand 2
   TTCAGGTCT
     |
   TGAGGCTGC
-------------           strand 1
-------------           strand 2
   ACTCCGACG
```

- e.g. In **meta analysis** different studies may use different technologies and collect data from different strands.

- Suppose **two** individuals have the **same** data at this bp location: A-T, G-C.
- If study 1 picks up signal from strand 1, then individual 1 will have genotype data: AG.
- If study 2 picks up signal from strand 2, then individual 2 will have apparent **different** genotype data: TC.

- Careful Quality Control (QC) and matching the strand between studies is practically important!!!

Strand Issue Important in Practice II

A high-profile controversial example due to strand issue!

- Genetic Signatures of Exceptional Longevity in Humans, a longevity study published online July 1 2010 in Science and later retracted.

- Check the article from Newsweek: The Little Flaw in the Longevity-Gene Study That Could Be a Big Problem.

  Kári Stefánsson, the Icelandic geneticist who founded deCode Genetics is convinced that the reported association between exceptional longevity and most of the 33 variants found in the Science study, including all the variants that other scientists hadn't already found, is due to genotyping problems.

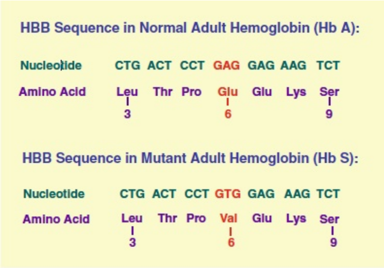
Mutations I

- Introduce genetic diversity in populations.
  (Reshuffling/recombination of chromosome during meiosis or reproduction process is also responsible.)

- Different types of mutations:
  (from 1 DNA base pair to larger segment up to a entire chromosome, e.g. Down's syndrome: triple chromosome 21)
  - Substitution: say TCT replaced by TCA.
  - Deletion: one base pair or a segment missing.
  - Insertion: adding an extra segment.
  - Translocation: segment from another region.

- Effect of mutations:
  - May have no, unknown or undetectable genetic effect on individuals.
    (e.g. TCT was replaced by TCA due to substitution, but both TCT and TCA specify the same protein product).
  - May cause certain protein(s) to malfunction; cells rely on the proteins may not function properly.

Mutations II



HBB Sequence in Normal Adult Hemoglobin (Hb A):

Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

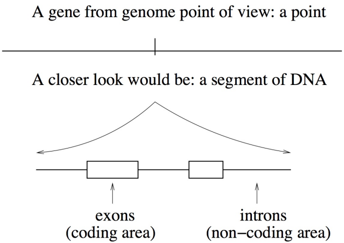
Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9

- **Codons**: three bp sequences coding amino acids.
- Most codons have a many-to-one relationship with an amino acid, e.g. both TCT and TCA (silent or **Synonymous mutation**) encode Serine.
- Sickle cell anemia: an example of **Missense mutation**, coding for a different amino acid.
- **Nonsense mutation**: the changed sequence result in too little protein being produced.

Genes I

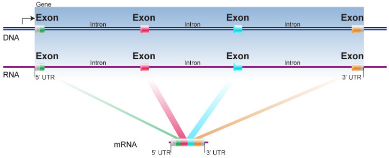
A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that **encodes a specific functional product** *(a protein or RNA molecule).*

- About **20,000 - 30,000 genes** (evolving estimate) throughout the genome.
- **Gene sizes vary** from about 1K DNA base pairs to more than 1 million bp.
- A gene is a segment of DNA consists of several coding segments (**Exons**), separated by non-coding sequences (**Introns**).



A gene from genome point of view: a point

A closer look would be: a segment of DNA



Genes II

- A few notes on Introns:
  - Introns do not code for specific proteins, BUT, they are not junk and may regulate exons.
  - There is an *increased appreciation of the role that non-coding DNA plays in* **gene regulation and expression**.
  - Impact of functional information on understanding variation: *Strikingly, 88% of associated SNPs [from recent years of GWAS] are either intronic or intergenic.*
    - *suggest that complex disorders may be influenced by genetic variants in non-coding regions as well.*
    - Associated SNPs are not necessarily the functional variants.

- Some unfortunate “synonyms” in practice (they are actually different): gene, genetic marker, locus, location.

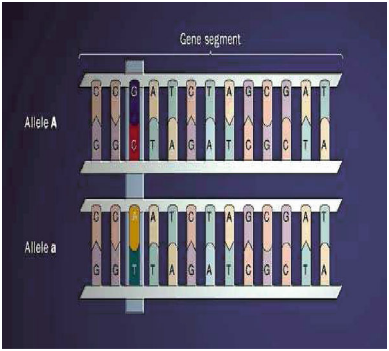
Genetic Markers/Polymorphisms, Alleles

- Human genome
  - $\approx$ 3 billion DNA base pairs.
  - $\approx$ 99% identical across individuals (genetic monomorphism).
  - However, variation (genetic diversity across individuals) is the most interesting part of the human genome!
- **Polymorphism** is any sequence of DNA that differ among individuals (caused by mutations accumulated over the history of mankind.

  Polymorphic means that the data at that locus can have more than one possible variant.)
- The different variants/states of a polymorphism are called **Alleles**
- In statistical term: a polymorphism is a (discrete) random variable and an allele is one of the outcomes/elements in the sample space.

Different Types of Generic Markers/Polymorphisms I

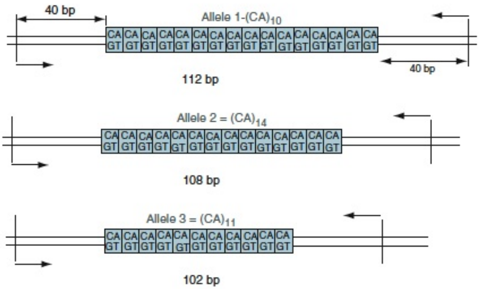
- **(Bi-Allelic) Single Nucleotides Polymorphisms (SNPs)** (simplest):
  - The third base pair of each chromosome shows variation.
  - The two outcomes/alleles can either be G-C or A-T (remember the matching).
  - Could code them
    - *A* (say for G-C) and *a* (for A-T); *A* and *a* not be confused with the A base.
    - 1 and 2, or other formats; the baseline or reference allele vs. the alternative allele


Gene segment
Allele A
C C G A T C T A G C G A T
G G C T A G A T C G C T A
Allele a
C C A A T C T A G C G A T
G G T T A G A T C G C T A

- Most common polymorphisms in human genome.
- Occur throughout the genome every ≈ 300 bp.
- ≈ 10 million SNPs (later on the concept of common vs. rare SNPs).
- Modest cost of collecting SNP data making them, attractive markers for current large scale genetic (association) studies.

Different Types of Genetic Markers/Polymorphisms II

- **Variable Number of Tandem Repeats (VNTRs)** and **Microsatellites**
  - Variation in the number of a repetitive sequence, e.g. 16, 14 and 11 repeats of CA.
  - **Microsatellites** are an important class of VNTRs which have a small (1-6) number of base pairs/sequence which are repeated.
  - Usually 3-30 repeats/alleles of the sequence, could code them $1, 2, \ldots,$
  - More expensive to collect the data, but contain more 'information' than SNP per marker/polymorphism. Basis of linkage mapping.



Different Types of Genetic Markers/Polymorphisms III

Overview of Structural Variation

Structural variation (SV) is generally defined as a region of DNA approximately 1 kb and larger in size and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs). These CNVs often overlap with segmental duplications, regions of DNA > 1 *kb present more than once in the genome, copies of which are* $> 90\%$ *identical. If present at* $> 1\%$ *in a population a CNV may be referred to as copy number polymorphism (CNP).*

- Copy Number Variants (CNVs),
- Insertions and deletions (INDELs),
- ...

Genotypes of Genetic Markers

Genotype: the two alleles at each chromosomal location (a pair of chromosome) for a given individual.

- e.g. a marker with two alleles denoted as A and a has 3 possible (unordered) genotype: AA, Aa/aA, aa.
- **Homozygous** genotype: same allelic type (AA or aa);
- **Heterozygous** genotype: different allelic type (Aa/aA).

- How many unordered genotypes for a marker with n alleles (e.g. denoted as m_i, $i = 1, ..., n$)?

$$n + \binom{n}{2} = n(n+1)/2$$

  - n homozygous genotypes ($m_i m_i$), and
  - $\binom{n}{2} = n(n-1)/2$ heterozygous genotypes ($m_i m_j$).

Characteristics of Genetic Markers

- **Population frequency**: genotype and allele frequencies.

  For a given population, different genotypes and alleles may appear with different frequencies (more on estimation of frequency later)

- **Hardy Weinberg Equilibrium(HWE)**: specifies a particular relationship between allele frequency and genotype frequency (more on testing HWE and applications of HWE later)

Recap

- Human genomes and **paired** chromosomes
- DNA and double helix structure (**strand issue**; not the same as paired chromosomes)
- A complex language from a simple 4-letter (A-T, G-C) system.
- Variations/Mutations
  - polymorphisms/genetic markers $\equiv$ discrete random variables
  - alleles $\equiv$ outcomes of a random variable
  - **SNP** $\equiv$ a r.v. with two outcomes
  - Microsatellite $\equiv$ a random variable with typically 3-30 outcomes
- **Genotype** data of a polymorphism/genetic marker: paired alleles (from the paired chromosomes.

What's Next

- Chapter 3: Some Basic Concepts from Population Genetics
- Review of Some Basic Concepts from Statistical Inference