BIOC4004 - Industrial Biochemistry
Lecture 18 - Wed Mar 10, 04

Topics for the Day:

- Novel applications of the microarray concept
- SNPs
  - genetics
  - importance
  - detection

Single Nucleotide Polymorphisms (SNPs)

- Humans share 99.9% of their DNA
- The 0.1% variation mostly at the level of single nucleotide differences

    ie. SNPs

- Some SNP stats:
    - ~ 3 million SNPs in the human genome
    - an SNP found every ~ 1250 bp of sequence
    - most SNPs occur outside gene coding or regulatory sequences

        > 99 %

    - only a couple of thousand SNPs occur within coding or regulatory sequences

        synonymous sites vs. non-synonymous sites

Why should we care ?

- Because big pharma "cares" !!!!

Single Nucleotide Polymorphisms (SNPs)

The SNP Consortium Ltd.

- Formed by ten of the worlds biggest pharma companies(AstraZeneca, Roche, Bayer, Pfizer, SmithKline Beecham, Novartis…), and industrial and academic genome centres (Celera, Sanger centre, TIGR, Incyte…)
- "Non-profit consortium", public release of the data
- The goal: map 300,000 SNPs ASAP !!!
- The rationale:
    - a catalogue of important human SNPs is a necessity
    - share the data and share the cost
    - otherwise, everyone produces redundant data at ten times the cost
- Why are SNPs important ?
    - Genetic markers that are densely scattered around the genome
    - heritable genetic markers
        - "guilt by association" and disease

Linkage Disequilibirum and Gene Discovery

Concept of "Linkage Disequilibrium" (LD)

- a polymorphic site means more than one variant exists
- each of these variants represents an allele
  - each allele is present at a certain frequency in the population
- what happens when we look at two different polymorphisms ?

Marker 1	Marker 2
Allele 1 = 25%	Allele 1 = 40%
Allele 2 = 75%	Allele 2 = 60%

In theory, we should observe the following combinations in the following frequencies:

Marker 1-1 + Marker 2-1 = 0.25 x 0.40 = 0.10
Marker 1-1 + Marker 2-2 = 0.25 x 0.60 = 0.15
Marker 1-2 + Marker 2-1 = 0.75 x 0.40 = 0.30
Marker 1-2 + Marker 2-2 = 0.75 x 0.60 = 0.45

• LD occurs when combinations of polymorphic markers are co-inherited at a frequency greater than would be expected from their frequency in the population

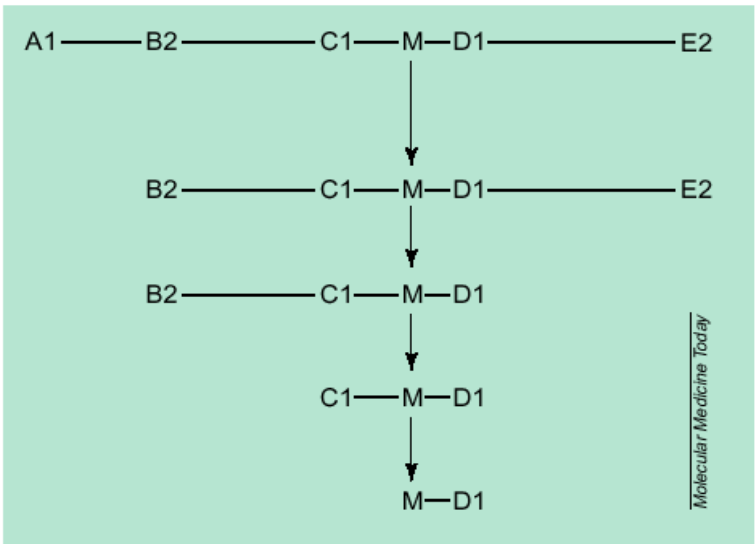

Figure 1. **Linkage disequilibrium** is the basis of all association studies. When a mutation (M) first arose in an individual, it was found on a chromosome with the A1-B2-C1-D1-E2 **haplotype**. As a result of recombination, association of M with the markers is gradually lost in successive generations. After many generations, only the markers closest to the mutation remain linked to it.

Linkage Disequilibirum and Gene Discovery

Why is LD important ?

• If LD is observed between "two traits", that means the traits must be on the same chromosome, and physically close together
    • the physical proximity (ie. linkage) prevents separation of the traits by either segregation or recombination

Linkage Disequilibrium and Disease:

On Guilt by Association…

• Common disease-common variant (CD-CV) hypothesis:

"a finite number of common DNA variants in the human population make a significant contribution to genetic risk for common diseases"

• Remember, a disease is a polymorphic trait
    diseased vs non-diseased

• Suppose we find a correlation (or linkage disequilibrium) between a polymorphic marker and a disease state:
    • Marker and disease gene must be close to each other !!!
    • this is the way traditional genetics has been used to clone disease genes
        • positional cloning
        • use statistical methods to determine LD between disease and genetic markers

• Simple genetic disorders are "easy" to track because transmission is easily followed from parent to offspring

• Complex diseases require lots more markers to tighten up the statistics
    1. Get "normal" and "diseased" patients
    2. Determine genetic profile of each group using widely scattered markers
    3. Determine if any haplotypes are associated with the disease

Why are SNPs cool ???

• The reason big pharma is interested in SNPs is because they provide THE LARGEST source of polymorphic markers scattered around the genome

• The idea behind The SNP Consortium is to generate all of the SNP data, make it available to everybody, then each pharmaceutical goes off and does its own thing with SNP data

• Big Pharma can use SNPs to look for "disease" genes
    • look for a "correlation" between diseased state and the presence of a given SNP variant (or combination of variants)

SNPs are cool so let's detect them !!!!

• The Old Way:
    • PCR amplify a region containing the SNP
    • sequence to determine which SNP allele is present
    • pretty slow and tedious --> could only do a few SNPs at most

• Since SNPs have tremendous potential:
    • we need a way to perform accurate SNP genotyping
    • we need high throughput !!!

New Ways of Detecting SNPs

Melting Curve Analysis:

- amplification of SNP region
- allele determination by examining melting curve of PCR product
    - each allele has its own characteristic melting temperature
    - A260 of DS-DNA vs SS-DNA
- faster than running a gel or sequencing

Hybridization using microarrays:

- make an SNP-based oligonucleotide array
- contains SNP variants for a number of different SNPs
- hybridize to genomic DNA or PCR amplified SNPs
- hybridization signal determines which SNP allele is carried
- can assay thousands of SNPs at the same time

Mass Spectrometry:

- amplification of SNP region
- microspotting of SNP amplicons onto MALDI-TOF matrix
- MALDI-TOF determination of SNP fragment