Epidemiology

Objectives

- Identify the cause of a disease and its risk factors
- Determine the extent of disease found in communities and populations
- Study the natural history and prognosis of disease
- Evaluate new preventive and therapeutic measures
- Provide foundation for developing public policy and regulatory decisions
Genetic Epidemiology (GE)

<table>
<thead>
<tr>
<th>Genetic epidemiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>The fusion of epidemiology and genetics provides the foundation for genetic epidemiology</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study of genetic factors and their interaction with environmental factors as they relate to disease distribution in human population</td>
</tr>
<tr>
<td>Ultimate goal of controlling and preventing diseases</td>
</tr>
<tr>
<td>GE is related to: genetics (genomics), epidemiology, population genetics, public health and biomedical sciences</td>
</tr>
<tr>
<td>GE uses tools from epidemiology, statistical genetics and bioinformatics.</td>
</tr>
</tbody>
</table>
Genetic epidemiology questions

1. Is there familial clustering?

2. Is there evidence or a genetic effect?

3. Is there evidence for a particular genetic model?

4. Where is the disease gene?

5. How does this gene contribute to disease in the general population?
Genetic epidemiology questions

1. Is there familial clustering?
   Familial aggregation studies

2. Is there evidence or a genetic effect?
   Heritability studies

3. Is there evidence for a particular genetic model?
   Segregation analyses

4. Where is the disease gene?
   Linkage, association studies

5. How does this gene contribute to disease in the general population?
   Variant frequency, risk magnitude, environmental interactions...
Why study genetic factors?

- Understand biological process leading to a disease
- Diagnostic and prognostic
- Prevention: screening and genetic counseling
- Treatment:
  - Develop new treatment
  - Personalized medicine (pharmacogenetics)
- Strengthen/confirm epidemiological inference for modifiable environmental factors
Simple mendelian diseases

- Controlled by a single gene (monogenic disorder)
- Follow a simple pattern of inheritance (recessive, dominant,...)
- No influence of the environment
- Relatively rare (1% of live births)
- Eg: Huntington’s disease (autosomal dominant), cystic fibrosis (autosomal recessive), Rett syndrome (X-linked dominant), Hemophilia A (X-linked recessive), male infertility (Y-linked)

GE and simple diseases

Many of the great successes of GE have been with simple mendelian disorders
Complex diseases

- Also called multifactorial diseases
- Disorder is caused by several genes (from two to hundreds !)
- Genes may interact together
- Environment also play a role in the disease occurrence
- Environment $\times$ Genes interactions
- Eg: cardiovascular diseases, cancer, psychiatric disorders...

Current trend

GE is increasingly focusing on complex diseases
Complex diseases

ROCHE Genetic Education (www)
GE: Flow of research

- Phenotype definition
- Familial clustering: family aggregation studies
- Segregation analysis
- Find disease susceptibility loci: linkage analysis
- Find disease susceptibility markers: association analysis
PHENOTYPE DEFINITION
- Familial clustering: family aggregation studies
- Segregation analysis
- Find disease susceptibility loci: linkage analysis
- Find disease susceptibility markers: association analysis
Phenotype

First step in GE process: carefully define and measure phenotype

Definition

A phenotype is an **observable** trait that must be measurable

- Eg: hair color, presence/absence of a condition, blood pressure, serum levels, IQ, score at a math test...
- Quality of phenotype measurement is essential to study success (validity/reliability/reproducibility)
Phenotype

- Relationship between genotype and disease-related phenotypes can be simple or very complex.
- Phenotype can be:
  - **Dichotomous** (0 or 1): affected or not by the disease
  - **Quantitative**: continuous or categorical (e.g., BMI, Blood pressure...)

![Histogram of Males BP](image)
Quantitative phenotypes: examples

- Identification of genetic risk factors for cardiovascular disease
- One can measure the cholesterol metabolism or blood pressure rather than the presence or absence of cardiovascular disease itself.
- Cholesterol metabolism is an example of an intermediate trait or endophenotype for cardiovascular disease.
- It is related to the disease and may be useful as a "proxy measure" of the disease.
GE: Flow of research

- Phenotype definition
- **FAMILIAL CLUSTERING: FAMILY AGGREGATION STUDIES**
- Segregation analysis
- Find disease susceptibility loci: linkage analysis
- Find disease susceptibility markers: association analysis
What is familial aggregation?

- First step in pursuing a possible genetic etiology of the disease
- Based on phenotypic data only (don’t need DNA)
- Demonstrate that the disease tends to run in families more than what would expect by chance
- Examine how that familial tendency is modified by the degree or type of relationship, age or environmental factors
- Familial aggregation does not separate genetic from environment
Rational of aggregation studies

**Rational**

Identify a group of individuals with a specific disease and determine whether relatives have an excess frequency of the same disease when compared to an appropriate reference population.

- Often, phenotype of interest is a disease (i.e. affected vs non-affected). But it can also be a physiological trait that has a continuous distribution (e.g. cholesterol levels).
Neurally Mediated Hypotension

(Lucas et al., 2006)
Figure Pedigrees of five familial PSP cases with pathologic confirmation of the proband.

Family 1

Family 2

Family 3

Family 4

Family 5

Familial aggregation based on Family History (FH)

- Positive FH = presence of disease in one or more first degree relatives
- FH should not be considered as a simple attribute of a person, comparable to age or cigarette smoking
- Depends on many factors:
  - No of relatives and types of relatives
  - Biologic relationship with the case/control
  - Age distribution of relatives
  - Disease frequency in the population
## FA designs: Twin studies

<table>
<thead>
<tr>
<th>Concordance rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of twinships in the population who are concordant (eg: both are affected)</td>
</tr>
</tbody>
</table>

- If MZ twins have a higher concordance than the DZ twins, there is suggestive evidence for a genetic basis of the disease.
- Since DZ twins and full sibs share on average 50% of their gene, a higher concordance rate in DZ twins compared to full sibs further points to a role for shared environmental factors.

### Example: Multiple Sclerosis (Willer et al., 2003)

- MZ: 25.3%
- 5.4%
- Non twin siblings: 2.9%
Recurrence risk ratio ($\lambda_R$)

Prevalence of the disease in relatives of type $R$ of affected cases, divided by the prevalence in the population

Prevalence

Proportion of a population affected by the disease

Examples

- $\lambda_{sibs} = 2,500$ for Phenolketonuria (Mendelian disease)
- $\lambda_{sibs} > 10,000$ for Huntington chorea (Mendelian disease)
- $\lambda_{sibs} = 10$ for schizophrenia (complex disease)
In addressing whether there is phenotypic aggregation within families, no attempt is made to determine the cause of any aggregation.
GE: Flow of research

- Phenotype definition
- Familial clustering: family aggregation studies
- **SEGREGATION ANALYSIS**
- Find disease susceptibility loci: linkage analysis
- Find disease susceptibility markers: association analysis
Segregation analysis moves beyond aggregation of disease and seeks to more precisely identify the factors responsible for familial aggregation.

Questions
- Is the aggregation due to environmental, cultural or genetic factors?
- What proportion of the trait is due to genetic factors?
- What mode of inheritance best represents the genetic factors?
- Does there appear to be genetic heterogeneity?
After documenting familial aggregation for a trait or disease, the next logical step is to ask *how much of the familial aggregation can be attributed to genetic causes*. 

**Heritability** is typically used to answer this question.

Heritability is computed from phenotypic data only (don’t need DNA), measured on *relatives*. 
Heritability is defined as the proportion of the trait variation directly attributable to genetic differences among individuals relative to the total variation in a population.

- A high heritability constitutes circumstantial evidence for genetic control of a trait.
- A high heritability means that a large proportion of the phenotypic variation among relatives follows patterns predicted by simple genetic factors.
Heritability and phenotype variation

- Genotypic difference makes 3-unit difference on phenotype.
- Environmental difference also makes 3-unit difference.
- Heritability = 50%
Heritability and phenotype variation

- Genotypic difference makes 3-unit difference on phenotype
- Environmental difference just makes 1-unit difference
- Heritability $\gg 50\%$
### Examples

<table>
<thead>
<tr>
<th>Condition</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADHD</td>
<td>80%</td>
</tr>
<tr>
<td>Childhood delinquency</td>
<td>20%-40%</td>
</tr>
<tr>
<td>Fingerprint ridge count</td>
<td>98%</td>
</tr>
<tr>
<td>Height</td>
<td>66%</td>
</tr>
<tr>
<td>IQ</td>
<td>34%</td>
</tr>
<tr>
<td>Social maturity score</td>
<td>16%</td>
</tr>
<tr>
<td>Alzheimer disease</td>
<td>60%</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>80%</td>
</tr>
</tbody>
</table>

**Note:** these estimates may vary a lot from one study to another.
Heritability values

- If heritability $= 0$, all observed phenotypic variation is attributable to non-genetic factors.
- If heritability $= 1$, there is no phenotypic variation NOT due to genetic differences.
After estimating heritability, one need to determine the **mode of inheritance** that best represents the genetic factors.

Mode of Inheritance is the manner in which a particular genetic trait or disorder is passed from one generation to the next.
In the mid 1800’s, Gregor Mendel demonstrated the existence of genes based on the regular occurrence of certain characteristic ratios of dichotomous characters among the offspring of crosses between parents of various characteristics and lineages.
These ratios are known as segregation ratios
The analysis of segregation ratios remains an important research tool in human genetics.
The demonstration of such ratios for a discrete trait among the offspring of certain types of families constitutes strong evidence that the trait has a simple genetic basis.
Mendel’s laws

- Simple Mendelian disorders or traits can be adequately modeled using Mendel’s laws. Generally, these traits are close to completely penetrant.

### Law of Segregation
The alleles at a gene segregate (separate from each other) into different gametes during meiosis. An individual receives with equal probability one of the two alleles at gene from the mother and one of two alleles at a gene from the father.

### Law of Independent Assortment
The segregation of the genes for one trait is independent of the segregation of genes for another trait, i.e., when genes segregate, they do so independently.
Segregation Analysis

Dominant disease allele

Recessive disease allele

Segregation
Consider a disease that is believed to be caused by a fully penetrant rare mutant allele at an autosomal locus.

Let $D$ be the allele causing the disorder and let $d$ represent the normal allele.

There are 6 possible mating types: $DD \times DD$, $DD \times Dd$, $DD \times dd$, $Dd \times dD$, $Dd \times dd$, $dd \times dd$.

Each of these mating types will produce offspring with a characteristic distribution of genotypes and therefore a distribution of phenotypes.
Segregation of an autosomal dominant disease

- The proportions of the different genotypes and phenotypes in the offspring of the six mating types are known as the segregation ratios of the mating types.
- These specific values of the segregation ratios can be used to test whether a disease is caused by a single autosomal dominant gene.
Marfan syndrome, a connective tissue disorder, is a rare disease that is believed to be autosomal dominant (and actually is!).

112 offspring of an affected parent and an unaffected parent are sample

52 of the offspring are affected and 60 are unaffected

Are these observations consistent with an autosomal dominant disease?
This is a simple statistic test that uses the Binomial distribution.

The p-value is 0.5085 (do not reject the hypothesis of autosomal dominant)

If only 42 of the offspring were affected, the p-value would be 0.0104
- Phenotype definition
- Familial clustering: family aggregation studies
- Segregation analysis

**FIND DISEASE SUSCEPTIBILITY LOCI: LINKAGE ANALYSIS**

- Find disease susceptibility markers: association analysis
Linkage analysis: objective

- Localise the disease gene with respect to genetic markers

Note: Coarse mapping (> 1cM)
Ingredients

- Families (pedigrees) ascertained using an affected proband
- For each subject in family:
  - Affection status (yes/no)
  - Genotypes at a set of markers (usually microsatellites)
Recall: recombination

Crossing-over and recombination during meiosis
Linkage: basic idea

If the disease gene is "close" to the marker $M$, recombination will occur with a small probability.

If the disease gene is "far" from marker $M$, recombination will occur with high probability.

Linkage analysis searches for **non random segregation** between the disease locus (to be located) and markers whose locations are already known.
Linkage analysis

- Two main types:
  - **Parametric** linkage analysis: estimate the recombination fraction between the causal locus (to be located) and the marker locus (known location)
  - **Non-parametric** linkage analysis: test whether affected relatives share more alleles *identical by descent* than expected by chance.
Genome-wide linkage analysis

Gene location (inferred)
GE: Flow of research

- Phenotype definition
- Familial clustering: family aggregation studies
- Segregation analysis
- Find disease susceptibility loci: linkage analysis
- **FIND DISEASE SUSCEPTIBILITY MARKERS: ASSOCIATION ANALYSIS**
Association versus linkage

**Approaches to Identifying Susceptibility Genes for Common Diseases**

**Linkage Studies:**
- A relation between loci
- Done within families

**Association Studies:**
- A relation between alleles
- Done in populations

Type 1 Diabetes Patients vs. Controls

ROCHE Genetic Education (www)
# Association versus linkage

<table>
<thead>
<tr>
<th><strong>Linkage</strong></th>
<th><strong>Association</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Linkage is a property of loci</td>
<td>Association is a property of alleles</td>
</tr>
<tr>
<td><strong>Role:</strong></td>
<td><strong>Role:</strong></td>
</tr>
<tr>
<td>* To identify a biological mechanism for</td>
<td>* To identify association between an allelic</td>
</tr>
<tr>
<td>transmission of a trait</td>
<td>variant and a disease</td>
</tr>
<tr>
<td>* To locate the gene involved</td>
<td>* To identify linkage disequilibrium between a</td>
</tr>
<tr>
<td></td>
<td>disease allele and a marker</td>
</tr>
<tr>
<td>Coarse mapping (&gt;1cM)</td>
<td>Fine mapping (&lt;1cM)</td>
</tr>
<tr>
<td>No information about which allelic variant</td>
<td>Case-control or family based approach</td>
</tr>
<tr>
<td>associated with higher risk of disease</td>
<td></td>
</tr>
<tr>
<td>Require family pedigrees</td>
<td></td>
</tr>
<tr>
<td>Use very polymorphic markers</td>
<td>Usually bi-allelic markers</td>
</tr>
</tbody>
</table>
Association analysis

- Population based
  - Cases and unrelated population controls from the same study base
- Family-based association
  - Child-family trios is the most common
Population-based association

Association Studies

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Type 1</th>
<th>Controls</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA DR4</td>
<td>17</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td>NON-HLA DR4</td>
<td>20</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>37</td>
<td></td>
</tr>
</tbody>
</table>

$X^2 = 5.377$

$\text{p} < 0.025$

ROCHE Genetic Education (www)
Genomewide association studies

HapMap
Select SNPs to tag haplotypes

Genotyping
300,000–500,000 SNPs typed on high-density arrays

Case–control study
Compare SNP allele frequencies in disease cases and controls

Genome scan result
Significant differences in SNP allele frequencies indicate possible new disease genes and loci

Replication test
Confirm scan findings

Nature Reviews | Genetics
90’s - early 2000: predominance of linkage studies and candidate gene association studies

2003 - GWAs

... hoping that covering the genome with hundreds of thousands of SNPs in thousands of subjects would be the solution to the complexities of diseases.
GWAs

- In the last years, it was hoped that GWAs would bring definitive evidence for gene effects.
- GWAs revealed much less than hoped.
- GWAS papers have reported a couple of hundred genetic variants that show statistically significant associations with a few traits.
- But the genes typically do not replicate across studies.
- Even when they do replicate, they never explain more than a tiny fraction of any interesting trait.
- In fact, classical Mendelian genetics based on family studies has identified far more disease-risk genes with larger effects than GWAS research has so far.
The missing heritability may reflect limitations of DNA-chip design.

GWAS methods so far focus on relatively common genetic variants in regions of DNA that code for proteins.

They under-sample rare variants and DNA regions translated into non-coding RNA, which seems to orchestrate most organic development in vertebrates.

At worst, each human trait may depend on hundreds of thousands of genetic variants that add up through gene-expression patterns of mind-numbing complexity.

Next generation sequencing...