Mapping Multifactorial Traits

Source: **Human Molecular Genetics**
By
Strachan & Reed

PKT

The Multifactorial Characters

- Parallel to Mendel's rediscovery in 1900,
- A new School of Thought, by Francis Galton,

Published (1865): 'Hereditary Talent and Characters'

Family resemblance (Anthropometry): Degree of correlation among relatives in various attributes, termed

Biometrics: Most traits are continuous and quantitative and can not be explained by Mendelism- Dichotomous, Yes or No

R A Fisher (1918): Resolved the issue, as such traits to be polygenic (independent Mendelian factors),

Later**, Falconer D S,** Extended this model to Dichotomy

Polygenic Theory

- Such quantitative characters governed by polygenes, follow
- Normal (Gaussian) Distribution in the population,

  No simple one-to- one genotype-phenotype relation exist,

  With increase in the participating loci + environment, develops , Good Gaussian curve

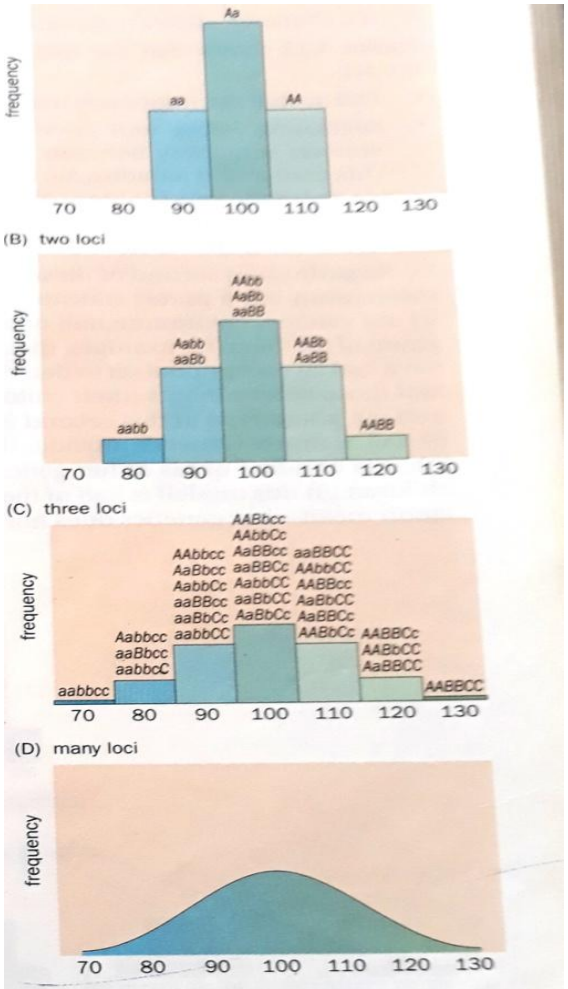

Figure 3.24 Successive approximations to a Gaussian distribution. The charts show the distribution in the population of a hypothetical character that has a mean value of 100 units. The character is determined by the additive (co-dominant) effects of alleles. Each upper-case allele adds 5 units to the value, and each lower-case allele subtracts 5 units. All allele frequencies are 0.5. (A) The character is determined by a single locus. (B) Two loci. (C) Three loci. (D) The addition of a minor amount of 'random' (environmental or polygenic) variation produces the Gaussian curve.

Regression to the Mean

- Suppose, IQ is entirely genetic,
  In a two locus model,

For each class of mothers,
  the average IQ of their children is,
  half way between mother's value and the population mean,
  termed **regression to the mean**

There is, however, a hidden assumption, that is,
  there is **random mating**

For each class of mothers, the husbands average IQ is assumed to be 100.

Thus, the average IQ of children is mid-parental value.

But, this is not the general observation.

Assortative mating or selective marriages with respect to IQ in husbands (above average),
  the regression half-way to the population mean is not expected.

The other assumption is, the absence of dominance, or it will influence mid-parental value, since dominant alleles will mask over the other (recessive) alleles.

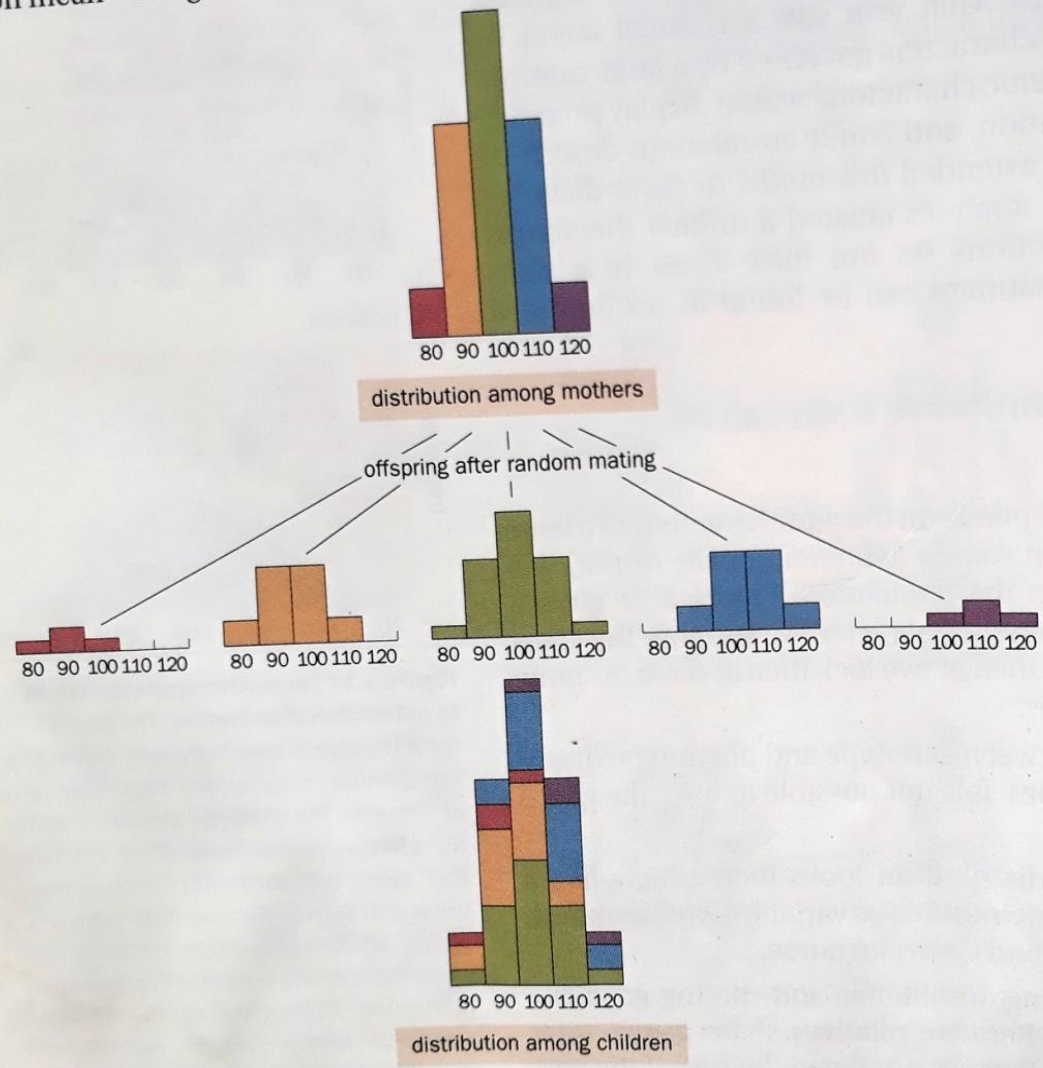

Figure 3.25 Regression to the mean. The same character as in Figure 3.24B: mean 100, determined by co-dominant alleles *A*, *a*, *B*, and *b* at two loci, all gene frequencies = 0.5. Top: distribution in a series of mothers. Middle: distributions in children of each class of mothers, assuming random mating. Bottom: summed distribution in the children. Note that: (a) the distribution in the children is the same as the distribution in the mothers; (b) for each class of mothers, the mean for their children is halfway between the mothers' value and the population mean (100); and (c) for each class of children (bottom), the mean for their mothers is halfway between the children's value and the population mean.

Heritability

Gaussian curve considers- mean and the variance (is additive) (or SD=√ variance)

The overall variance of the Phenotype V_P is the sum of the variance due to individual causes of variation, i.e., Environmental variance V_E + Genetic variance V_G , which is variance due to simply additive genetic effect V_A + variance due to dominance effect V_D

Thus, Heretibility (h^2) of a trait is

the proportion of total variance, that is genetic, $[V_G / V_P]$

$[V_G / V_P]$ is the Broad heritability, while,

since dominance variance can not be fixed, selection response of a breeding experiment is determined by,

$[V_A / V_P]$, the Narrow heritability

Heritability...

- In human behavioural traits, separation of genetic and environmental shares is not applicable, since children in a family get both genes and environment similarly, so both are correlated (genetic and environmental factors), hence, V_P does not equal $V_G + V_E$; there are other interaction variances.

- Heritability differs from the mode of inheritance; In different social circumstances the heritability of IQ will differ. The more equal a society is (with equal opportunity), the higher the heritability of IQ should be. Any variation from each other should be genetic.

Discontinuous Characters: Polygenic Concept

- Several diseases that tends to run in families, but do not show Mendelian pedigree.
- **Falconer** proposed, **Susceptibility**, a continuous variable parameter, e.g., Cleft lip palate- may or may not be?

  Every embryo has certain susceptibility to cleft lip, low or high, is polygenic, follow a Gaussian distribution.

Falconer suggested,

Along with susceptibility, threshold, which actually cause cleft palate to appear or not.

Those, whose susceptibility exceeds the critical threshold value, develop cleft lip palate, but, if less, do not.

Polygenic Concept...

All embryo start with a cleft palate,

During early development the palate shelves must become horizontal and fuse together, within a certain specific time period. Many factors (environmental + genetic) influence this fusion. A natural threshold superimposed on a continuously variable process, is what determines this.

Recurrence risk of several diseases that vary in families can be explained by this threshold theory.

Affected people must have an unfortunate combination of high susceptibility alleles, which is shared by relatives to varying degrees,

tend to run in the families.

The threshold is fixed and only in the increase or decrease in the susceptibility alleles decides the recurrence. Parents with several affected children must have more high risk alleles and the chance (recurrence-risk) increases in every birth.

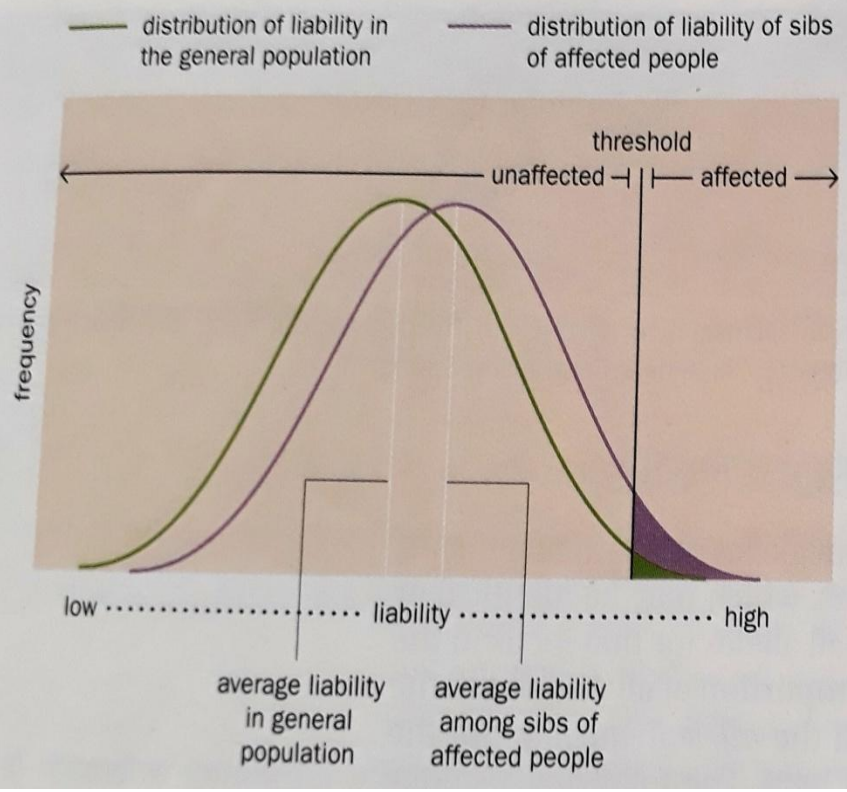

Figure 3.27 A polygenic threshold model for dichotomous non-Mendelian characters. Liability to the condition is polygenic and normally distributed (green curve). People whose liability is above a certain threshold value (the balance point in Figure 3.26) are affected. The distribution of liability among sibs of an affected person (purple curve) is shifted toward higher liability because they share genes with their affected sib. A greater proportion of them have liability exceeding the (fixed) threshold. As a result, the condition tends to run in families.

Polygenic concept...

- Sex-specificity in the threshold' e.g., Congenital Pyloric Stenosis, is five times more common in boys that in girls, i.e., the threshold is more for girls than boys. Thus the relatives of an affected girl have a higher average susceptibility than those of a boy.
- Emperic risks or population based risk is calculated in such non-Mendelian conditions.

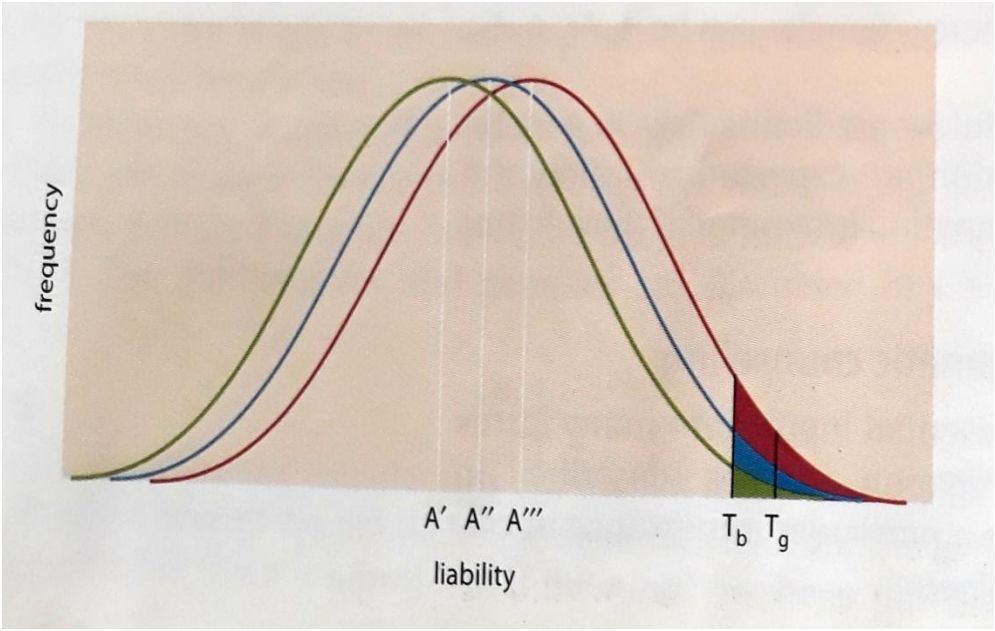

Figure 3.28 A polygenic dichotomous character with sex-specific thresholds. The figure shows a model that explains data such as those in Table 3.2. For all three curves, girls with a liability above the threshold value T_g and boys above the threshold value T_b manifest the condition, and are represented by shaded areas under the curves. As in Figure 3.27, the general population (green curve) displays a liability to this polygenic disease condition that is normally distributed, with an average liability of *A′*. The liability among siblings of affected boys (blue curve) is higher, with average *A″*, and a greater proportion of these brothers and sisters have a liability that exceeds the respective threshold levels. Among siblings of affected girls the liability is still higher (red curve, average liability *A‴*), and an even greater proportion of these brothers and sisters will be affected because they have a liability that exceeds their sex-specific threshold levels.

Mapping Genes for Complex Traits

The sequence of study may be,

- Family, twin or adoption study- for genetic susceptibility identification
- Segregation analysis- types and frequencies of susceptibility alleles
- Linkage analysis- mapping of susceptibility loci
- Population association- narrowing candidate gene
- DNA sequence variation (polymorphism/mutation) and biochemical analysis

Role of Family, Twin and Adoption studies

- Risk ratio- λ value
- For non-Mendelian characters – continuous (quantitative) or discontinuous (dichotomous)?
- If runs in the family?
- What are the genetic determinants?
- Family clustering of the diseases expressed by [λ_R], i.e., the risk to relative R of an affected proband compared to the population.
- The value of λ increases with closeness, but drop back towards 1 for more distant relatives.

TABLE 15.1 RISK OF SCHIZOPHRENIA AMONG RELATIVES OF SCHIZOPHRENICS: POOLED RESULTS OF SEVERAL STUDIES

Relative	No. at risk	Risk (%)	λ
Parents	8020	5.6	7
Sibs	9920.7	10.1	12.6
Sibs, one parent affected	623.5	16.7	20.8
Offspring	1577.3	12.8	16
Offspring, both parents affected	134	46.3	58
Half-sib	499.5	4.2	5.2
Uncles, aunts, nephews, nieces	6386.5	2.8	3.5
Grandchildren	739.5	3.7	4.6
Cousins	1600.5	2.4	3

Numbers at risk are corrected to allow for the fact that some at-risk relatives were below or only just within the age of risk for schizophrenia (say, 15–35 years). λ values are calculated assuming a population incidence of 0.8%. [Data from McGuffin P, Shanks MF & Hodgson RJ (eds) (1984) The Scientific Principles of Psychopathology. Grune & Stratton.]

Shared Family Environment

- Many characters run in families because of shared environment, e.g., IQ, Schizophrenia, physical attributes or even birth defects (for varying extent, if not fully).

Twin studies: Francis Galton

- MZ twins are concordant for many inherited characters except post-zygotic somatic genetic changes.
- DZ twins share less concordance.
- MZ twins are of same sex, hence, share most family environment similarly.
- DZ could be different sexes or even same.

TABLE 15.2 TWIN STUDIES IN SCHIZOPHRENIA

Study	Country	Concordant pairs	
		MZ	DZ
Kringlen et al. (1968)	Norway	14/50 (0.28)	6/94 (0.06)
Fischer et al. (1969)	Denmark	5/21 (0.23)	4/41 (0.10)
Tienari et al. (1975)	Finland	3/20 (0.15)	3/42 (0.07)
Farmer et al. (1987)	UK	6/17 (0.35)	1/20 (0.05)
Onstad et al. (1991)	Norway	8/24 (0.33)	1/28 (0.04)

The numbers show the total number of twin pairs ascertained and the number that were concordant (both twins diagnosed as schizophrenic). Diagnostic and inclusion criteria varied between studies; despite the heterogeneity there is a clear tendency for more monozygotic (MZ) than dizygotic (DZ) pairs to be concordant. [For references, see Onstad S, Skre I, Torgersen S & Kringlen E (1991) *Acta Psychiatr. Scand.* 83, 395–401.]

Shared Family Environment

MZ Twins separated at Birth:

Brought up in entirely different family environment, yet found similar in many attributes even after decades of separation.

But due to,

-non-availability of few but exceptional people,

-completely separated twins (by the time they are born and not after birth or brought up by relatives)

-bias of ascertainment (biasness to similar but not dissimilar twins)

-intrauterine environmental causes, maternal hormones or sexual orientation (gay gene)

Thus, twin studies have remain less informative.

Shared family environment...

Adoption Studies:

- Finding adopted people having same disease that runs in families-in – (whether?) biological or adoptive family? Enquire.
- Affected parent with adopted children (in other family). Enquire if that saved the children?
- Rosenthal & Kety, Kendler et al, Kety et al,
- Proved the significance of adoption studies in genetic dissection of schizophrenia.
- However, lack of information about the biological family, and
- Selective placement (adoption)
- Have remain a hindrance to some extent.
- Still adoption studies are gold-standard.

TABLE 15.3 AN ADOPTION STUDY IN SCHIZOPHRENIA

Case types	Schizophrenia cases among biological relatives	Schizophrenia cases among adoptive relatives
Index cases (47 chronic schizophrenic adoptees)	44/279 (15.8%)	2/111 (1.8%)
Control adoptees (matched for age, sex, social status of adoptive family, and number of years in institutional care before adoption)	5/234 (2.1%)	2/117 (1.7%)

The study involved 14,427 adopted persons aged 20–40 years in Denmark; 47 of them were diagnosed as chronic schizophrenic. The 47 were matched with 47 non-schizophrenic control subjects from the same set of adoptees. [Data from Kety SS, Wender PH, Jacobsen B et al. (1994) *Arch. Gen. Psychiatry* 51, 442–455.]

Segregation Analysis of Susceptibility Loci

- Traits could be

  Mendelian (Dichotomous) --- Oligogenic --- Polygenic

Segregation analysis is a useful statistical tool, which determines the major players (loci).

Bias of ascertainment:

Collection of cases and families ---- gives raw data

The method of ascertainment of the carrier and affected families/cases, may some time cause biases and thus, in segregation analysis, e.g., for a Mendelian recessive disease,

In a collection of cases and families, the proportion of cases should be ¼. In fact, the expected proportion of ¼ in the samples collected is not so,

Cause bias of ascertainment

Segregation analysis...

- For example in Figure below: Two children families,

  The families are identified with the affected children, hence, many will not be ascertained if do not have affected children.

  The observed segregation ratio would be 8/14 and not ¼

  -For three children families, the ratio would be 48/111

For any given family size it can be calculated from the truncated binomial distribution $(1/4 + 3/4)^n$

omitting the term, 'no affected children'.

The data is corrected for the bias by the method of Li & Mantel

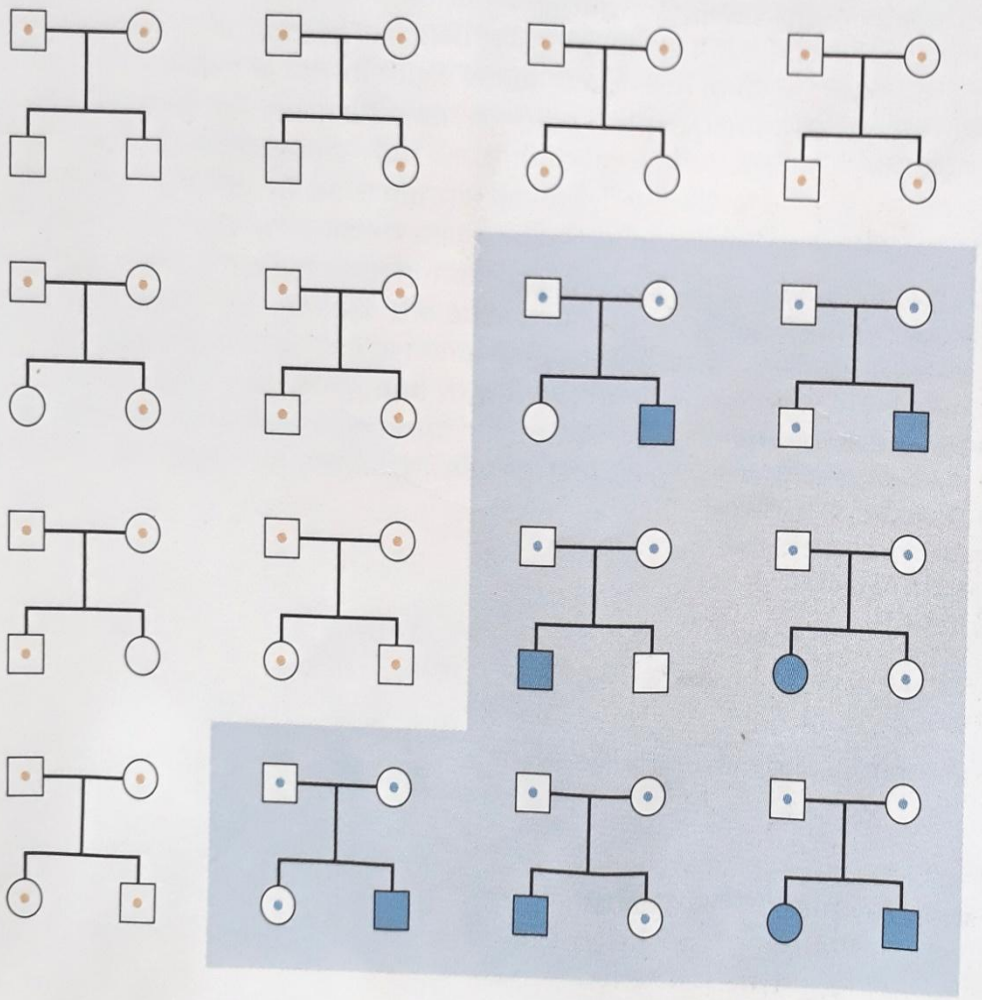

Figure 3.11 Biased ascertainment. In each of these 16 pedigrees, both parents are carriers of an autosomal recessive condition. Overall, their 32 children show the expected 1:2:1 distribution of genotypes (*AA*, 8/32; *Aa*, 16/32; *aa*, 8/32). However, if families can be recognized only through affected children, only the families shown in the shaded area will be picked up, and the proportion of the children in that sample who are affected will be 8/14, not 1/4. Statistical methods are available for correcting such biased ascertainment and recovering the true ratio.

BOX 3.4 CALCULATING AND CORRECTING THE SEGREGATION RATIO FOR A RECESSIVE CONDITION

The segregation ratio for a family with n children, under complete truncate ascertainment, can be estimated from the following formulae, which give the proportions of sibships with different numbers of affected children:

$(1/4)^n$	all children affected
$(n,1)\ (1/4)^{n-1}\ (3/4)$	all except one child affected
$(n,2)\ (1/4)^{n-2}\ (3/4)^2$	all except two children affected
etc.	

(n,x) means $n!/[x!(n-x)!]$, where $n!$ (n factorial) means

$$n \times (n-1) \times (n-2) \times (n-3) \times \ldots \times 2 \times 1$$

For the mathematically inclined, this is an example of a truncated binomial distribution, a binomial expansion of $(1/4 + 3/4)^n$ in which the last term (no affected children) is omitted.

The overall proportion of affected children can now be calculated. For example, for every 64 three-child families we have:

1 family with 3 affected	total 3 affected, 0 unaffected
9 families with 2 affected	total 18 affected, 9 unaffected
27 families with 1 affected	total 27 affected, 54 unaffected
[27 families with no affected—but these families will not be ascertained]	
Overall (among ascertained families)	48 affected, 63 unaffected

Apparent segregation ratio $48/(48 + 63) = 48/111 = 0.432$

Correcting the segregation ratio

If: p = the true (unbiased) segregation ratio
R = the number of affected children
S = the number of affected singletons (children who are the only affected child in the family)
T = the total number of children
N = the number of sibships

For complete truncate ascertainment, $p = (R - S)/(T - S)$.

For example, for the three-child families illustrated above, $p = (48 - 27)/(111 - 27) = 21/84 = 0.25$.

For single selection, $p = (R - N)/(T - N)$.

Segregation analysis...

- Complete truncated ascertainment:
- Collect all families from one defined population with at least one affected child.
- Alternatively,

  Take the first 100 to be seen in a clinic, so that many more can be ascertained. This increases the likelihood of collecting a family with affected children twice that of one affected child and four times more for a family with four affected children.

  In single selection, the probability of being ascertained is proportional to the number of affected children in the family.

Complex Segregation Analysis

- For familiar non-Mendelian disease, families and relatives are analysed.
- Both genetic and environmental factors are at work,

  Genetic factors could be polygenic, oligogenic or Mendelian with any mode of inheritance or a mix of these.

  Environmental factors may be familial or non-familial

Thus, a whole range of possible mechanisms, gene frequencies, penetrances, etc., are considered, followed by computer analyses to know the maximum likelihood.

The most likely hypothesis is tested for **models**, and compared:

Sporadic, polygenic, dominant, recessive and general or mixed model that freely optimizes all.

Cont...

- In the example (Table) best fit model is the major dominant susceptibility to the disease (on the argument that simple explanations are preferable to complex explanations).
- However, if a major factor is omitted, segregation analysis may lead to wrong interpretation, e.g., the example of McGuffin & Huckle Table),
- Thus caution is required to take all variables (including family environment) into consideration to avoid spurious genetic effects.

TABLE 15.5 A RECESSIVE GENE FOR ATTENDING MEDICAL SCHOOL?

Model	*d*	*t*	*q*	*H*	χ^2	*P*
Mixed	0.087	4.04	0.089	0.008		
Sporadic					163	$<10^{-5}$
Polygenic				0.845	14.4	<0.005
Major recessive locus	0.00	7.62	0.88		0.11	n.s.

The data are taken from a survey of medical students and their families. The meaning of the symbols is explained in the foornote to Table 15.4. Affected is defined as attending medical school. The analysis seems to support recessive inheritance, because this accounts for the data equally well as the unrestricted model (but see the text). n.s., not significant. [Data from McGuffin P & Huckle P (1990) *Am. J. Hum. Genet.* 46, 994–999.]

Major dominant locus

Data are for families ascertained through a proband with long-segment Hirschsprung disease (OMIM 142623). Parameters that can be varied are as follows: *d*, the degree of dominance of any major disease allele; *t*, the difference in liability between people homozygous for the low-susceptibility and the high-susceptibility alleles of a major susceptibility gene, measured in units of standard deviation of liability; *q*, the gene frequency of any major disease allele; *H*, the proportion of total variance in liability that is due to polygenic inheritance, in adults; *z*, the ratio of heritability in children to heritability in adults; *x*, the proportion of cases due to new mutation. The values shown are those that best account for the family data using the stated model. The χ^2 statistic is a standard test that compares the performance of each model with the mixed model, in which a mix of all mechanisms is allowed. A single major locus encoding dominant susceptibility explains the data as well as the mixed model. [Data from Badner JA, Sieber WK,

Linkage Analysis

- **Parametric (Standard LOD score analysis) and non-parametric linkage analysis:**

Parametric: Requires a precise model (autosomal/ sex-linked, dominant/recessive)

Mode of inheritance

Gene frequencies

Penetrance of each genotype

All are best suited for Mendelian characters

Linkage analysis...

Diagnostic criteria: In case of psychiatric diseases,

The most relevant criteria should be that, which are valid for agreeable by two independent investigators. They are often biologically arbitrary, e.g., in behavioural or psychiatric phenotypes.

It helps make comparison between different species but does not guarantee for their genetic informativeness.

For Mendelian characters/ syndromes:

Features of the patients are part of the syndrome, and that the components of the syndrome co-segregate.

Linkage analysis...

Near Mendelian Families:

After fixing diagnostic criteria, look for a subset of families where the condition segregate in a near Mendelian manner.

Perform segregation analysis to define the parameters of a genetic model.

Perform standard (parametric) linkage analysis

For example, Near Mendelian families arise in following ways,

i. Any complex disease is likely to be heterogenous, thus, collected families may include some Mendelian conditions, non-separable from non-Mendelian majority. E.g., breast cancer, Alzheimer

ii. Such families mar represent cases where, by chance, many determinants of the disease are already present in most people, so that Mendelian segregation of just one the many susceptibility factors make the critical difference. E.g., Hirschsprung disease (loci mapped are also susceptibility factors)

iii. Near Mendelian pattern may be spurious, just chance aggregation of affected people within one family. E.g., Schizophrenia (where some early workers have found LOD score to be 6, now known to be spurious)

Linkage analysis: Shared Segment Method

- The above discussions led to shifting to non-parametric or model-free linkage analysis,
- Which looks for alleles or chromosomal segments that are shared by affected individuals.
- **Shared segment method** is used for both family and population.
- Alleles: IBD, Identical by Descent- copies of the same ancestral (parental) allele

  IBS, Identical by State- look identical, or may be so, but their common ancestry is not demonstrable, treated in terms of population frequency rather than Mendelian probability of inheritance from the common ancestor.
- For rare alleles, two independent origins are unlikely, hence, IBS implied IBD, but is not true for common alleles.
- Multiallelic microsatellite markers are better than two allele markers in defining IBD,
- Multilocus multiallele haplotypes are still better, as any one haplotype is likely to be rare.
- Shared segment analysis is conducted using IBS or IBD data.
- IBD is more powerful, but requires more relative's samples.

Shared Segment Analysis in Families

- **Affected Sib Pair (ASP) and Affected Pedigree Member (APM) Analysis:**

  Selecting a chromosomal segment at random from a pair of sibs --- expected to share: 2 (¼), 1 (½), 0 (¼) alleles,

when both sibs are affected	2, 1 alleles, or share both parental alleles

  They are likely to share the relevant chromosome segment, one, if disease is dominant and two, if recessive (fig.)

ASP...

- **Affected sib pair (ASP) analysis** :
- ASPs are typed for markers for identifying chromosomal regions where sharing is above the random 1:2:1 ratios of sharing 2,1 or 0 haplotypes identical by descent (IBD). If the sib pairs are tested for IBS, the sharing is calculated as frequencies of the shared genes/alleles.
- ASP analysis does not asks for the genetics of the disease, and collection of affected sib pairs is easier thqan extended families (pedigree making).
- Multipoint analysis is carried out, which provide information about IBD sharing more efficiently.
- MAPMAKER/SIBS program of Kruglyak & Lander (1995)
- Produce non-parametric LOD (NPL) scores.

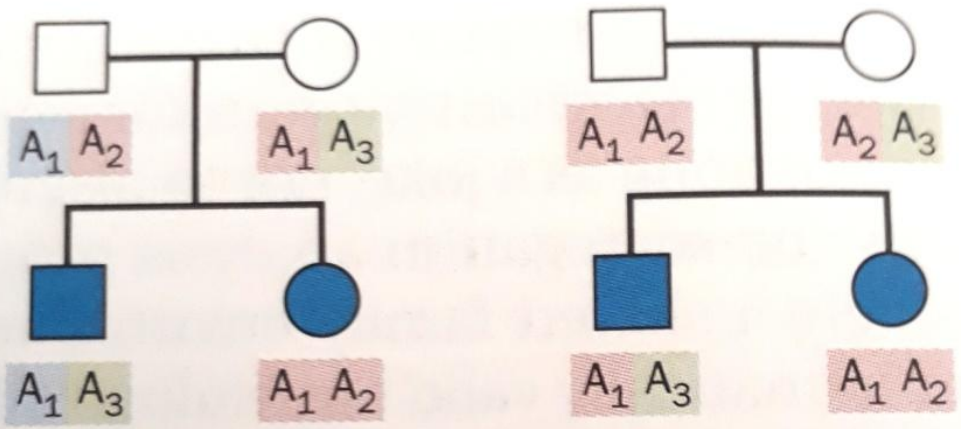

Figure 15.1 Identity by state and identity by descent. Both sib pairs share allele A_1: The first sib pair have two independent copies of A_1 (colored red and blue), indicating identity by state but not by descent. The second sib pair share copies of the same paternal A_1 allele (red), showing identity by descent. The difference is only apparent if the parental genotypes are known.

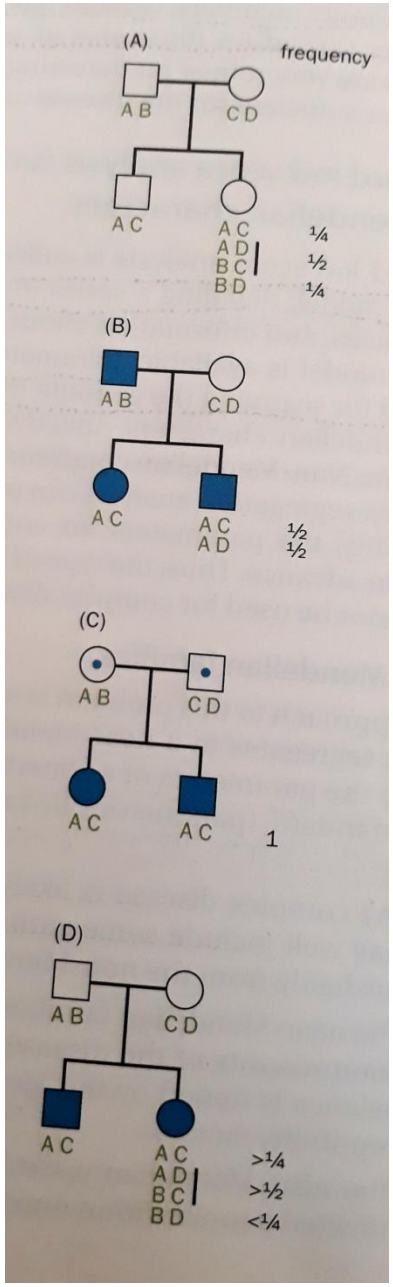

Figure 15.2 Affected sib pair analysis. (A) By random segregation, sib pairs share 2 (both *AC*), 1 (*AC* and either *AD* or *BC*), or 0 (*AC* and *BD*) parental haplotypes ¼, ½, and ¼ of the time, respectively. (B) Pairs of sibs who are both affected by a Mendelian dominant condition must share the segment that carries the disease allele, and they may or may not (a 50:50 chance) share a haplotype from the unaffected parent. (C) Pairs of sibs who are both affected by a Mendelian recessive condition necessarily share the same two parental haplotypes for the relevant chromosomal segment. (D) For complex conditions, haplotype sharing above that expected to occur by chance (as in panel A) identifies chromosomal segments containing susceptibility genes.

ASP...

- **Limitations of ASP**: The candidate regions are large enough (large parental segments are shared between affected sibs either by chance or because of a shared susceptibility) for positional cloning.
- If a susceptibility factor is neither necessary nor sufficient for disease, not all sib pairs will share the relevant segment. Yet ASP analysis is simple and robust, hence, used in mapping susceptibility genes.
- GENEHUNTER program extend shared segment analysis to other relationships/ affected relatives,
- Calculate the extent of sharing allele descent.
- Results are compared across all affected pedigree members with the null hypothesis of simple Mendelian segregation, i.e., markers should segregate according to Mendelian ratios unless segregation is distorted by linkage or association.

- **Threshold of Significance**:
- As compared to Mendelian genes, the result of analysis of complex diseases has been irreproducible across the studies made in different laboratories.
- Thus, appropriate threshold at which the results are to be said as significant, need to be decided.
- For a whole genome study,
- The appropriate significance threshold is a value where the probability of finding a false positive anywhere in the genome is 0.05.
- Theoretically, the genome-wide LOD score threshold is 3.6 for IBD and 4.0 for IBS testing.

Association Studies & Linkage Disequilibrium

- **Association:** Is not specifically a genetic phenomenon, but a statistical statement about the co-occurrence of alleles or phenotypes.
- Allele ‘A’ is associated with disease ‘D’, if people with ‘D’ also have ‘A’ significantly more often (or less often) than would be predicted from the individual frequencies of ‘D’ and ‘A’ in the population.
- E.g., HLA DR4 is found in 36% of the general UK mpopulation, but in 78% of people with rheumatoid arthritis.

Association...

- **Why association happens?**
- Direct causation: Allele ‘A’ makes susceptible to ‘D’, but neither necessary nor sufficient for anyone to develop ‘D’, but increases the likelihood.
- Natural selection: People with disease ‘D’ might be likely to survive and have children, if also have allele ‘A’ (protective)
- Population stratification: Population contains several genetically distinct subsets and both the disease and allele ‘A’ happen to be particularly frequent in one subset. E.g., Association of HLA A1 and ability to eat with chopsticks in San Francisco Bay area. HLA A1 is more frequent among Chinese than among Caucasians.
- Type1 error: Association studies normally test a large number of markers for association with a disease. Even without any true effect, 5% of results will be significant at the p=0.05 level and 1% at the p=0.01 level. The raw p values need correction for the number of questions asked, as past studies failed to replicate in subsequent studies.
- Linkage disequilibrium (LD): Detecting association due to LD bteween the marker and the disease.

Association vs. Linkage

- **Association**: - is a relation between alleles or phenotypes.
  - is simply a statistical observation that might have various causes.
- **Linkage**: - is a relation between loci.
  – is specifically a genetic relationship.
  – it does not produce any association on its own in the general population.

  e.g., STR45 locus /marker is linked to Dystrophin locus,

  Nevertheless, the distribution of STR45 among unrelated DMD cases is similar as in general population.

  However, within a family having DMD cases, the affected people are expected to share the same allele of STR45, since they are tightly linked.

Thus, linkage creates association within families, but not among unrelated people, unless they are ancestrally related, hence, share ancestral alleles at loci closely linked to the disease.

Association...

- **Common ancestor**: We all share them. All humans are related in evolutionary history, thus, a population is an extended family, ... the population level association due to LD should exist between ancestral disease susceptibility genes and closely linked markers.
- Suppose, the two unrelated people each inherit a disease allele from their common ancestor.
- After many generations / meiosis/recombination... There would be reduction in the shared segment considerably, e.g., for a locus with Rf= θ with the susceptibility locus,

  a proportion 'θ' of ancestral chromosome will loose association each generation and '1- θ' proportion will retain the association.

  After 'n' meiosis, a fraction $(1- \theta)^n$ of chromosomes will retain the association.

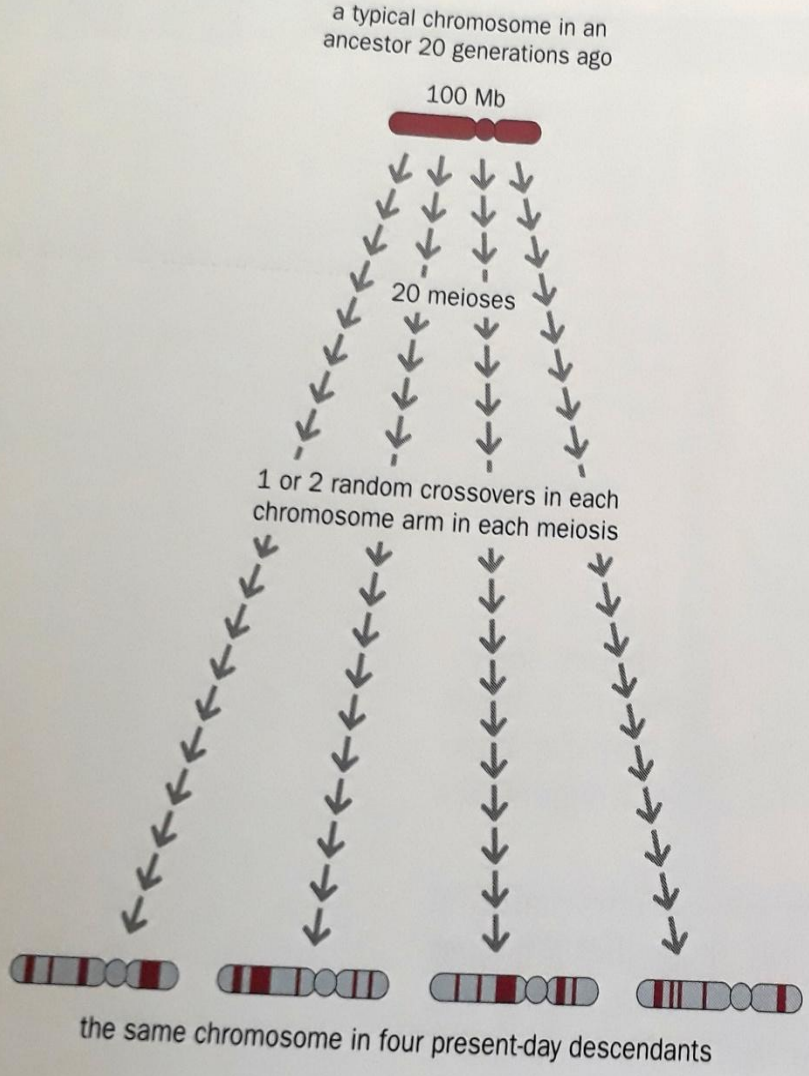

Figure 15.6 The size of shared ancestral chromosome segments. A typical chromosome is shown in a common ancestor, 20 generations ago, of four present-day individuals. There will be one or two random crossovers in each chromosome arm in each of the 20 meioses linking each present-day person to their common ancestor. Only a small proportion of the sequence of the ancestor's chromosome will be inherited by descendants after 20 generations (red segments). The ancestral segments that are shared by a significant proportion of all descendants are very small, typically 5–15 kb.

Association...

- **Islands of LD separated by recombination hotspots:**
- Although in cystic fibrosis analysis of gradients of LDs helped reach to the nearest, it could not be found true for others.
- For HD, there are cases of strong association with a more distant marker and a weak association with a closure marker. E.g., the marker D4S95, closely linked to HD locus detected RFLP with three Res, Taq1, Mbo1 & Acc1, strong association was noted in various studies with particular Acc1 and Mbo1 alleles, but not with either of Taq1 alleles,
- Could probably have occurred due to chance recombination event in small founder populations, and may be due to an origin of some marker polymorphisms more recently than some disease mutations.

Association...

- **LD...**
- Several studies show that LD does not decay smoothly with distance. Instead, chromosomes contain a series of islands of relatively long-range LD that can sharply separated from each other.
- Within the islands, useful LD may extend for 50KB (in European, less in African), but even very closely spaced markers in different islands show no LD with each other.
- It is reported that the island boundaries are indeed recombination hotspots, presumably presenting the mosaic of ancestral chromosomal segments forming our common heritage.
- If LD island structure of a population could be defined across the whole genome, then a set of markers (“hap-SNPs”) could be defined to establish haplotypes at each island that could then be tested for association with any disease. It is suggested that most islands would have only 4-6 different common haplotypes in any population.

Linkage Disequilibrium

- Is defined as the non-random association between two alleles at two different loci on the same chromosome.
- When a gene mutation occurs, it forms part of a unique haplotype. However, this particular linkage relationship dissipates in successive generations, as recombination events scramble the alleles of the loci surrounding the disease gene.
- After many generations, the disease gene and only those very closely linked loci from the original chromosome remains together. These closed-linkages represent examples of LD and in these instances, the difference between loci is often less than 1-2 cM, which is the lower limit that can be resolved by conventional genetic linkage studies.
- LD is an extremely useful phenomenon for mapping genes.
- First, detection of LD between linked loci produce a finer scale map than could be constructed by genetic linkage.
- Second, the sites that show LD are often within 0.10cM or 100Kb of a disease gene, in comparison to 1cM or 10^3 Kb for sites that are genetically mapped.
- The closure the distance between a disease gene and a marker locus, the easier it is to discover and isolate the disease gene.

LD...

- LD mapping is carried out after genetic linkage between a polymorphic locus and the disease gene is determined.
- Then members of families with a genetic disease within a founder population are haplotyped with a number of additional polymorphic markers, on the same chromosome, and statistical tests are run to determine which loci are in LD with the disease.
- The genetic distance (θ) between a marker locus and the disease gene that are in LD is calculated from the following equation:

$$P_{excess} = \frac{P_{affected} - P_{normal}}{(1 - P_{normal})}$$

LD...

Or,

$P_{excess} = (1-\mu g q^{-1}) (1-\theta)^g$

Where,

Θ is the recombination fraction between the marker and disease loci

μ is the mutation rate for the disease gene

g is the number of generations since the common ancestor

q is the world-wide frequency of the disease allele

P normal, is the proportion of the marker allele in normal (non-disease) chromosomes

P affected, is the proportion of marker allele in chromosomes with the disease gene

P excess, is a measure of disequilibrium, is the fraction of excess occurrence of a chromosome with the disease gene and the marker allele in comparison to the chromosomes with the non-disease gene and the marker allele.

LD can map only those diseases which occur in founder populations.

Design of Association Studies

- Association studies do not require multicase families or special family structure, hence, easier than linkage analysis. Even weak association can also be detected.
- **Choice of method to test for association**:

  Choice of control group is important.

  A likely possibility of something overlooked is always there.

  When association is found, the doubt is,

  could it be due to inadequately matched controls and not by LD with a susceptibility locus?

  These and irreproducibility of the result hindered its choice for a long time.

Association studies with internal controls circumvented most of the above limitations. The methods are: TDT, ETDT, sib-TDT

Association...

- **Transmission Disequilibrium Test (TDT):**
- TDT starts with couples with one or more affected offspring. The condition of the couple is not important.
- To test,

  If , A particular M1 is associated with the disease, parents who are heterozygous for M1 are selected.

  The test compares the number of cases where such a parent transmits M1 to the affected offspring with the number when their other allele is transmitted.

  First, Affected probands are ascertained. Then the probands and their parents are typed for the marker.

  Next, those parents who are heterozygous for the marker allele M1 are selected. They may or may not be affected.

Association...

- Let ‘a’ be the number of times a heterozygous parent transmits the M1 to the affected offspring, and

  ‘b’ the number of times the other allele is transmitted,
- The TDT test statistics is $(a-b)^2 /(a+b)$

  This has a X^2 distribution with 1 degree of freedom when the numbers are large enough.

ETDT: Extended TDT

Developed for data on multi-allelic markers, like microsatellites, etc.

TDT can even be used when only one parent is available, though with some bias.

Sib-TDT: When parent are alive (in late onset diseases)

used to look into differences in marker allele frequencies between affected and unaffected sibs.

Since, TDT asks for alleles and not loci, it is a test of association. The associated allele may itself be a susceptibility allele at a nearby locus. It can not detect linkage if there is no disequilibrium when whole genome is scanned.

However, now conventional case-control studies are being considered as an alternative to TDT, since it requires fewer samples and easier for late-onset diseases (no parents).

Optimally, (Risch & Teng 1988). Affected sib-pairs as cases with two unrelated control are chosen.

Association...

- **Selection of markers**:
- SNPs are more common due to

  being numerous (≥ 1 per Kb on average) to define LD islands and can be scored.

  are rather stable over a long time scale, appropriate for identifying the ancestral haplotypes.

The older a disease allele is in human history, the higher is the density of SNPs needed to detect it.

Some favours SNPs to be located within genes, particularly in coding regions (CSNPs), which are likely to be the actual susceptibility determinants.

However, in view of the diversity in the population histories for different diseases, a custom made strategy may be required.

Association...

- **Choice of Population**:
- Is an isolated population more appropriate for association study?
- Population derived from a small number of founders are expected to show limited haplotype diversity and high LD.
- The DeCode project of Iceland justifies it. Such populations show strong and long range LD around loci for the rare Mendelian diseases common to a population, e.g., 'Finnish' disease in Finland.
- But, that LD exist only on chromosomes carrying the disease allele, which are presumably all derived from a single common ancestor.
- One, therefore, needs large numbers of potential subjects with good medical records.
- This originated 'BioBank' Project in UK, which aims to collect medical and life style data and DNA from 500,000 British people aged 45-69 years and follow their health prospectively.

Association...

- **Genotype or Haplotypes**:

  When individuals are studied, genotype is the raw data, but association studies require haplotypes.

  Which are inferred from genotypes by an expectation-maximization computer analysis,

  Which is not fully reliable.

  Typing on somatic cell hybrids are more reliable, which contains haploid chromosomes.

Thus, the most reliable design (which one of the above?) is to be tested.

- **Linkage vs. Association:**
- Risch & Merikangas (1996), when compared with linkage (ASP) analysis, found association to be more powerful than linkage for detecting weak (association) susceptibility alleles.

- ASP requires unfeasibly large samples to detect susceptibility loci, conferring a relative risk of >3, where as TDT might detect alleles giving a relative risk below 2 with smaller sample size.

- **Linkage & Association**: Complementary methods
- Linkage operates over a long chromosomal range and can scan the entire genome in a few hundred tests,
- e.g., a typical study of 250 ASPs with 300 markers would require 1.5 – 3 x 10^3 genotypes to be generated. A few weeks time for a well establish lab.
- But,
- In a whole genome TDT scan of 300 trios, even testing only one SNP in each island (since LD is a short range phenomenon, with island of LD of about 20-50 Kb in size), would require 10^8 genotypes, assuming island average 25Kb in size.
- Highly costly
- Thus, association studies should focus on predetermined candidate regions, defined by linkage analysis or by reference model organism or the known genes.
- Hence, a good study design, starts with a genome-wide screen by linkage (e.g., ASP) to initially localize a candidate region.
- Then narrowed down to the candidate region by LD mapping (e.g. TDT)