Molecular Genetics of Susceptibility to Infectious Diseases

Prapat Suriyaphol, Ph.D.

Bioinformatics and Data Management for Research Unit, Office for Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand.

Siriraj Med J 2008;60:301-304
E-journal: http://www.sirirajmedj.com

Introduction

By definition, infectious disease is a disease caused by pathogens such as bacteria or virus that enters a host. However, not every encounter results in a disease and not every individual responds in the same way to pathogens. This has been unfortunately demonstrated in an accidental incidence in Lubeck, Germany in 1927, where babies were given BCG vaccine contaminated with virulent *Mycobacterium tuberculosis*. The outcome was death of 67 of 249, others with severe disease, but some were left unaffected[1]. In a tuberculin skin test survey of health care workers in highly endemic countries, approximately 20% of individuals showed a negative test result throughout their lives, despite a high chance of repeated exposure[2]. The finding suggests that natural resistance to tuberculosis infection exists in some fraction of the population. Another well-known example is the resistance to *Plasmodium vivax* malaria in negative Duffy-blood group individuals, which explains the almost absence of this specie in West Africa, where the majority of people are Duffy negative[3]. The outbreak of the SARS coronavirus in 2003 displayed three major host phenotypes of infection – death, recovery and superspreaders[4]. Different responses despite their first encounter to pathogens in a shared environment provides strong evidence to support the involvement of human genetic variation in determining the outcome of an infection process.

Measurement of genetic effects

In fact, the observation that genetics modulate the susceptibility or resistance to infectious diseases is not new. During the 18th century, tuberculosis and leprosy were believed to be inherited within a family. The familial aggregation of diseases suggests that genetics may contribute to the response of the host, although it does not exclude the pure effect of a shared environment within a family. A twin study is an effective tool to examine the genetic contribution by observing disease concordance. Disease concordance is the probability that one twin develops disease when the other is affected. Taking advantage that identical twins share 100% of their genetic materials while non-identical twin share only 50% on average, one would expect more disease concordance in identical twins if genetics exert an effect on the susceptibility of disease. Tuberculosis, leprosy, poliomyelitis, and hepatitis B have been shown to have higher disease concordance in identical twins. An adoption study is an approach trying to exclude the effect of a shared environment. An adoption study in Denmark showed that the risk of death of children, who were adopted early in life, from an infectious disease was nearly six folds when their biological parents died before the age of 50 from the same infectious disease. Whereas the death of their adoptive parents from infectious diseases had no significant effect of the relative risk of their adopted child[5].

The difference in disease prevalence or severity among different ethnic groups can also lead to the discovery of genes involved in the process. The almost absence of *Plasmodium vivax* malaria in sub-Saharan Africa leads to the discovery of a variation of the FY gene which results in a Duffy-negative blood group. The Fulani people, showing more resistance to malaria parasitemia then other ethnic groups in the same area in West Africa, have a higher level of antimalarial antibodies. This leads to the discovery of variation of IL-4 (Interleukin-4) that is associated with the increased level of antibodies[5]. During Dengue outbreaks in Cuba, there was an observation that the Nigroid race has less severity than the Caucasoid race. While dengue hemorrhagic fever is common in Southeast Asian countries, very low incidence of cases were reported in African countries. Monocytes isolated from African individuals showed resistance to Dengue infection. Comparison analysis of genomes from African and Asian origin may provide clues on genes regulating the severity of Dengue infection[6].

Sibling risk ratio (λ_s) is an estimated ratio of prevalence of the disease or an interested phenotype in siblings to those in the general population. The higher sibling risk ratio (λ_s) means a higher chance of siblings to develop disease or have the same phenotype. Similar

to familial aggregation, sibling risk includes the risk produced by a shared environment as well as the genetic part. The genetic component from λ_s thus has a tendency to be overestimated. Nonetheless, λ_s is useful for genetic assessment, statistical power determination and study design. Small λ_s indicates a weak genetic background on the phenotype and a large amount of samples are needed to be statistically powerful enough to pick up the genes. Several attempts are made to increase the λ_s, which, in turn, increases the involvement of genetic components, and reduces the number of required samples in the study to a reasonable figure. One way is to carefully choose phenotypes such that the recurrent risk in sibling increases, the prevalence in general population decreases or both[7].

Phenotype consideration

The obvious and natural choice for a phenotype of an infectious disease is affected and non-affected status when the primary focus is on finding resistant or susceptible genes. One needs to be careful when making assessment because many infections may have similar manifestation and could be misdiagnosed. For example, it is difficult to distinguish between Dengue fever and other acute febrile illness, unless specific immune responses or viral detection are confirmed in a laboratory. For dengue, the most beneficial and interested phenotype is not the affected status but the two entities of manifestation – dengue fever and dengue hemorrhagic fever. Since dengue fever is relatively mild and self-limited, no hospital admission is needed except in rare cases with unusual hemorrhage, whereas dengue hemorrhagic fever is life-threatening and needs intensive monitoring. Predictive markers such as a patient's genotypes that could pinpoint definite cases of dengue fever from the rest would return scarce hospital beds to those who really need them and relieve the patient's family from unnecessary expense – hence saving a large amount of the national health budget. However, the grey zone exists when clinicians make decisions on the diagnosis based on the rigid definition for dengue hemorrhagic fever set up by WHO in 1997[8]. All 4 criteria – fever, hemorrhagic tendencies, thrombocytopenia and evidence of plasma leakage – must be fulfilled to define a case of dengue hemorrhagic fever. The general sense of clinicians is that dengue fever is mild whereas dengue hemorrhagic fever is severe. There are some severe cases that do not fit all 4 criteria for dengue hemorrhagic fever but clinicians are reluctant to put them into dengue fever because of the severe outcome. Many options are possible to resolve the issues. One is to exclude those samples in the grey zone and study only the two extreme and distinct set of samples. This approach reduces the number of valid samples, but tends to increase the power of detection. Another approach is to put those 'severe' cases into the dengue hemorrhagic fever group despite their failure in criteria fulfillment in order to focus on genes related to the severity. This sounds logical, but may cause some trouble. First, one must make a clear definition of severity or another grey zone will be introduced. Second, this design may reduce statistical power from the admixture of samples between dengue fever and dengue hemorrhagic fever if the set of genes involved are not common between the two manifestations. Choosing phenotypes with a clearer definition and less subjectiveness seems to be a preferable choice. The laboratory findings that could reflect severity are good candidate phenotypes such as the number of platelets (related to thrombocytopenia), and albumin/globulin ratio (related to leakage). Though these phenotypes are further away from the normal definition of clinical severity, this approach could increase the power of gene detection since the genes involved in these 'intermediate' phenotypes are likely to be a subset of the whole set of disease severity-related genes.

In malarial infection, susceptibility can be measured in several ways. The number of infection episodes, level of parasitemia (maximum, minimum or average), and severity are among common phenotypes chosen for genetics study. Each phenotype seems to have its own set of modulating genes. Severe malaria phenotype (i.e. cerebral malaria, severe anemia and respiratory distress) currently is under the focus of a large active research network, Malaria GEN, funded by the Grand Challenges in Global Health. One reason is that severe malaria is the major cause of death in children especially those in Africa. The other is that there is a stronger genetic effect when concentrating on severe malaria phenotypes. HbS heterozygote reduces the risk of malaria-fever episode only by 2-fold but reduces the risk of severe malaria by 10-fold. Therefore, there is higher chance of being able to identify the underlying genes for severe malaria[3].

Methodology and study design

The modern approach of genetic studies – both linkage analysis and association — are based on the basic understanding of how the genetic material is passed from parents to offspring. The human cells are diploid with one set of chromosomes inherited from the mother and the other from father. In the meiosis division process of egg and sperm generation, maternal and paternal chromosomes are randomly passed into the gamete cells where the numbers of chromosomes were reduced by half. There is approximately an equal chance of maternal and paternal chromosomes to pass onto the gamete. This random distribution of chromosomes explains the segregation of traits according to Mendel's law and we say that there is an independent assortment and no linkage between chromosomes. In addition, genetic loci on the same chromosome are not necessary to be transmitted together. During the meiosis process, the homologous chromosomes pair up and exchange parts of their material in the process called crossing-over, resulting in a recombination of genetic materials of a homologous chromosome (Fig 1). Such a process occurs at random. If there is a large distance between two genetic loci, it is likely that recombination will occur between them and the chance of transmittal together is approaching 50%. In contrast, if two loci are close, the chance of recombinant to occur between them is less than 50%; they are transmitted together more often than randomly. Those two loci are said to be linked.

In linkage analysis, we follow the genetic markers and phenotype in families. An example of a family is given in Fig 2a. The gene predisposing to an infectious disease is mark as band 'm', which we would like to identify. Four genetic markers were used in this study and their locations on the chromosome are marked as *a, b, c* and *d*, respectively. Different alleles of each

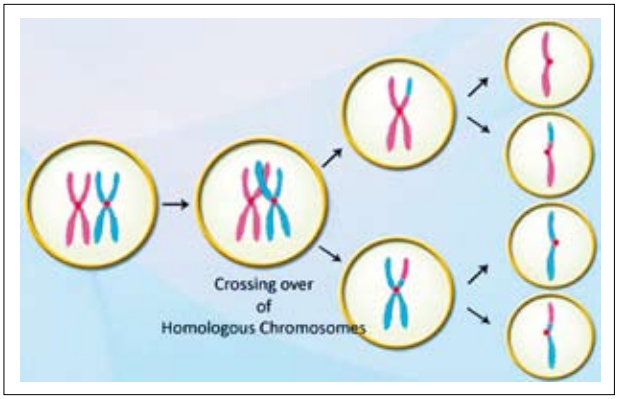

Fig 1. Gamete formation by meiosis division. In the process, homologous chromosomes pair up and exchange part of their body between their pairs (crossing-over). Ultimately, the 4 daughter cells resulting from the meiosis process contain half but randomly allocated genetic material. As a result, the chromosomes of the daughter cells are not genetically identical with one another and their original cell.

marker are painted in different colors. Marker analysis in this family indicates that 'purple' alleles of markers b and c are linked with the affected phenotype. This suggested that the gene lies between these two markers. Investigation in more families may find that different color alleles are linked with the affected status. In other words, for linkage study, it is not necessary that the same allele of markers will be linked with the gene. It is possible that a different allele is linked in a different family, but nonetheless the linked markers are likely to be the same. Because we follow the markers in just a few generations, there are only a few opportunities for recombination to occur between markers. As a result, the 'marker-linked' region often spans across many megabases of DNA. Another approach such as fine mapping or candidate gene study is required to narrow down the region into genes of interest.

Association analysis is to test whether some markers are transmitted with the phenotype more often than expected, which means that the markers are in linkage disequilibrium with the gene related to the interested phenotype. Association study works at the level of the population, in which, over many generations, recombination has occurred and eliminated linkage between loci that are far apart. Only markers within close proximity can retain the linkage disequilibrium. In Fig 2b, over many generations, marker *a, b, c* and *d* are not close enough to be in linkage disequilibrium with the band 'm'. We need a denser marker for association study and in this example, we have marker e in which the 'purple' allele of marker e is associated with the gene of interest.

Therefore, by technique, there are essentially two approaches: - linkage and association analysis. Linkage is a method of choice if susceptibility to infection or interested phenotype follows the Mendelian pattern of inheritance and DNA samples from large families with multiple affected cases can be collected. A few hundred microsatellite markers can be used in linkage analysis to cover the whole human genome. This method is less favorable if the inheritance pattern is complex. The reason is that linkage analysis is insensitive to detect genes with a small effect which generally is the case with a complex pattern of inheritance. In addition,

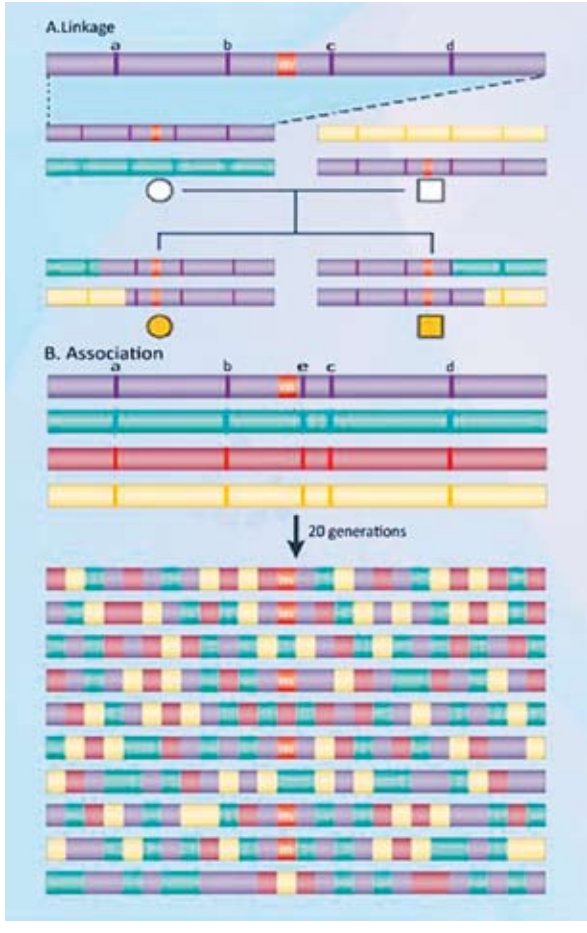

Fig 2. Band 'm' marks the susceptible gene. A) Linkage study —markers (*a, b, c* and *d*) are used to follow between phenotypes and genotypes. Markers which are closer to the gene have higher probability to be transferred together. Linkage study is to follow markers in family generation. Few opportunities for recombination relatively occur, therefore the region identified by linkage will often be large, covering many megabases of DNA. In panel a, the region around marker b and c (purple color) are linked with the gene. B) Association study — through many generations, more recombination occur and eliminate the linkage between markers; the further markers, the higher chance of losing link. Only markers closed to the band 'm' remain transmitted together. This non-random association of markers is called linkage disequilibrium. In panel B, only marker e (purple color) remains associated with the gene [adapted from Ref 14].

collection of samples from large families with multiple affected cases is difficult especially for infectious disease. Without family data, one cannot perform linkage analysis. Nonetheless, genome-wide linkage studies have successfully identified genomic regions related to many infectious diseases including the intensity of infection with *Schistosoma mansoni*, susceptibility to *H. pylori*, and to leprosy[9].

The discovery of single nucleotide polymorphisms (SNPs) from the Human Genome project serves very well as required dense markers for association study. Association study does not require family data, though we can use them to calculate the segregation ratio in a transmission disequilibrium test (TDT)[10]. It is arguable that conducting a thousands case-control cohort is easier

than collecting samples from hundreds of families. Currently, a large cohort of 2000 cases and 3000 controls under the Wellcome Trust Case-Control Consortium is conducting genome-wide association studies (GWAS) on eight complex diseases including tuberculosis and malaria[9]. The fast advance in genotyping technology and rapid drop in cost make genome-wide association study no longer restricted to large genome centers, but available to medium-to-small research groups. We would expect a flood of data from GWAS in the very near future. Whether any interesting genes can be identified remains to be seen.

Most infectious diseases have a modest λ_s and follow a complex mode of inheritance. Considering that genes involved in the immune response are the most abundant and diverse in the human genome, it is no surprise to see a modest contribution of many genes. It is generally agreed that the genetics contribution to sus-ceptibility to infection are complex. Identifying genes with modest power requires a large sample size. Almost all genetic studies to date have been underpowered. Therefore replication and validation are highly recom-mended to prevent spurious genetic associations. One of the most common causes of spurious genetic associa-tions is unmatched case-control genetic background – cases come from one ethnic group whereas controls come from others. The identified variations from such studies are likely to be unrelated to the disease of interest.

On the other hand, replication study may be unsuccessful despite true association. The rapid progression of AIDS is associated with an HLA class I allele only when the patient carries a specific variant of the KIR3DS1 gene[9]. This gene-gene interaction (epistasis) may cause a failure in replication if a replication study of such HLA class I is carried out in a different population with a different variant of the KIR3DS1 gene. The association of polymorphisms of inducible nitric oxide synthase (NOS2) with malaria disease severity was found in studies in Gabon and Gambia, but cannot be replicated in other populations. The association may be true, but failure in replication may come from differences in the genetic background of each population, differences in the strains of malaria, or differences in the epidemiology of malaria infection[11].

At least, 25 genes associated with common infectious diseases such as malaria, tuberculosis, AIDS, and leprosy have been successfully replicated. Some genes exert a very strong effect on the disease outcome including the heterozygote of HbS and malaria severity, and Duffy negative and *P. vivax* infection. The homozygote of 32-bp deletion in the CCR5 chemokine receptor is almost completely resistant to HIV-1 virus infection. The similar protection was found in the homozygote of valine at codon 129 in the PRPN gene against Creutz-feldt-Jakob prion disease[12]. The discussion of these genes and others in details can be found in Ref 12 and 13.

Concluding remarks

We believe that we have moved into the era of modern medicine, but infectious disease still remains a major health problem worldwide. The modern genetic studies enable us to investigate the contribution of our genetics in response to pathogens. How can this be translated into practical benefits? The prediction for disease outcome such as dengue infection could save time and money. The knowledge of the CCR5 gene as a receptor for the HIV-1 virus leads to the functional study of anti-CCR5 antibodies in the protection of HIV-1 infection. The finding of a protective HLA leads to a new field of epitope-based vaccines. As more and more genes related to diseases come out from rapidly-growing genetic studies, hopefully we can make a better prediction, develop new vaccines and other routes of prevention, and new interventions. Perhaps, adding these new 'weapons' into our arsenal could help us to end the saga of the host-vs-pathogen war.

ACKNOWLEDGMENTS

A part of the content in this article was presented at the **Thailand Research Fund-Senior Research Scholar Academic Conference 2008 – *'Human Molecular Gene-tics Conference: Molecular Genetics of Complex and Common Genetic Diseases'***, September 29, 2008, Royal River Hotel, Bangkok, Thailand.

REFERENCES

1. Andersen P, Doherty TM. The success and failure of BCG - implications for a novel tuberculosis vaccine. Nat Rev Microbiol 2005;3:656-62.
2. Kassim S, Zuber P, Wiktor SZ, Diomande FV, Coulibaly IM, Coulibaly D, et al. Tuberculin skin testing to assess the occupational risk of Mycobacterium tuberculosis infection among health care workers in Abidjan, Cote d'Ivoire. Int J Tuberc Lung Dis 2000;4:321-6.
3. Kwiatkowski DP. How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 2005;77: 171-92.
4. Kellam P, Weiss RA. Infectogenomics: insights from the host genome into infectious diseases. Cell 2006;124:695-7.
5. Segal S, Hill AV. Genetic susceptibility to infectious disease. Trends Microbiol 2003;11:445-8.
6. de la CSB, Kouri G, Guzman MG. Race: a risk factor for dengue hemorrhagic fever. Arch Virol 2007;152:533-42.
7. Burton PR, Tobin MD, Hopper JL. Key concepts in genetic epidemiology. Lancet 2005;366:941-51.
8. World Health Organization. *Dengue Haemorrhagic Fever: Diagnosis, Treatment, Prevention and Control* (World Health Organization, Geneva, 1997).
9. Burgner D, Jamieson SE, Blackwell JM. Genetic susceptibility to infectious diseases: big is beautiful, but will bigger be even better? Lancet Infect Dis 2006;6:653-63.
10. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.
11. Clementi M, Di Gianantonio E. Genetic susceptibility to infectious diseases. Reprod Toxicol 2006;21:345-9.
12. Hill AV. Aspects of genetic susceptibility to human infectious diseases. Ann Rev Genet 2006;40:469-86.
13. Casanova JL, Abel L. Human genetics of infectious diseases: a unified theory. EMBO J 2007;26:915-22.
14. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001;2:91-9.