Chapter 14
Molecular Epidemiology

Hui Wang

Key Points

- Describe the concept of molecular epidemiology
- Familiar with the concept and classification of biomarkers and know how to select biomarkers.
- Understand the relationships between molecular epidemiology and traditional epidemiological methods
- Discuss, apply and interpret application of molecular epidemiology in disease control and prevention.

In molecular epidemiology, the study of the determinants of disease will pay attention to the causative, protective, and predisposing factors (including infectious agents and various environmental exposures, e.g., chemical or physical agents and lifestyle habits) and host characteristics such as genetic susceptibility. Most of these studies are performed via molecular techniques within the molecular biology.

14.1 Introduction

14.1.1 Concept

Molecular epidemiology is defined as the study of the epidemiology of human diseases by application of the techniques of molecular biology at the population level. Molecular investigations can contribute to the elucidation of diseases' etiology.

H. Wang (✉)
School of Public Health, Peking University, Beijing, China
e-mail: huiwang@bjmu.edu.cn

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_14

Proposing the clear definition of biomarkers is the vital step in the study of molecular epidemiology. Biomarkers are referred to any markers which could be detected and represented the changes from the exposure to the onset of disease on a population scale. The range of biomarkers is quite broad, including cellular, biochemical, and immunological elements. Generally, all biomarkers (e.g., nucleic acid, protein, lipids, and antibody) can be investigated in the molecular epidemiology. According to the process of disease, biomarkers applied in the molecular epidemiology are divided into three categories: markers of exposure, markers of biological effects, and markers of susceptibility.

14.1.2 Characteristic

Compared with conventional epidemiology, molecular epidemiology lay emphasize on the knowledge of the pathogenesis of diseases by elucidating specific molecular pathways and pointing out specific molecules or genes that influence the risk of developing diseases. For instance, genetic biomarkers rather than family history might be more precise in characterizing host susceptibility. Molecular epidemiology can enhance the validity and reduce bias in the assessment of environmental exposures. It can predicate the onset of disease at the subclinical level and provide tools to discern heterogeneity within a disease, such as the development of breast cancer subtypes (i.e., basal, luminal A, luminal B, normal breast-like, and $ERBB2^{+}$).

Not all biomarkers are suitable for molecular epidemiological studies due to expensive cost or intensive labor. It is necessary to examine those laboratory techniques used in the studies of molecular epidemiology, for their sensitivity, validity, specificity, and variability within and between laboratories before using them in any epidemiological research. Meanwhile, the acquisition of appropriate biological specimens, costs, and ethical issues need to take into considerations in molecular epidemiological research design as well.

This chapter shows the introduction of main characteristics of molecular epidemiology, the description of three major categories of biomarkers, most commonly used research methods and application and prospective of molecular epidemiology. This chapter will present an overview of molecular epidemiology. The authors refer the readers to more articles published recently and update the knowledge of molecular epidemiology.

14.2 Classes of Biomarkers

14.2.1 Biomarkers of Exposure

Molecular epidemiological studies intend to establish the causal and biological associations between exposures and diseases. The exposure is defined as any contact

with physical, chemical, or biological agents by the International Programme on Chemical Safety. Exposure assessment should be continuous which requires that biomarkers of exposure should be continuous as well. Thus, this assessment will provide more exact knowledge with regard to the exposure-disease association.

Molecular epidemiology implements biomarkers to improve exposure assessments in the complexity of distinguishing respect between the effect of individual and environmental factors to diseases etiology. Validated biomarkers could measure the disease process at the individual level which leads to the causal inference or biological plausibility of an exposure-disease association. For instance, in the area of virology, antibodies have been used to identify what kind of virus which a person has been infected with. Furthermore, by measuring the accumulation of chemical agents or metals in biospecimen, such as arsenic in hair or mercury in fingernails or toenails, it can directly measure an exposure at the individual level.

Using sensitive laboratory techniques, low levels of exposure can be detected by trace analysis of biomarkers. Most biomarkers usually represent the exposure of environmental toxicants, nonetheless, they could identify the crude amount of ingested dietary components as well, such as bacterial or viral infections. Moreover, they also serve as terminuses for a determination of the success of interventional strategies.

Biomarkers of exposure are classified relying on what they measure, either an internal dose (i.e., serum vitamin D) or a biological effective dose (i.e., a dose that causes DNA damages). Biomarkers of internal dose estimate the presence of environmental chemicals and their metabolites in human tissues, excretions, and/or exhaled air. Further, measurements of dietary biomarkers either as "recovery" biomarkers, for example, measurements of sucrose and fructose in 24-hour urine samples for direct assessment of sugar consumption, or as "concentration" biomarkers, such as serum carotenoids, which indirectly indicates dietary intake since they are the results of complex metabolic processes. Exposure biomarkers also serve as evaluation indicators for the effects of interventional strategies. The utility of such biomarkers is restricted to the availability of detectable levels of the compound. Although biomarkers of exposure are often described separately from biomarkers of effect, many actually overlaps exist, which can be partially attributed to the fact that the biomarkers provide information related to both of the exposure and the effect. For example, lymphocytes could be a surrogate for exposure and also a target for the exposure's effect.

14.2.2 Biomarkers of Effects

Biomarkers of effects measure the interaction between an agent and/or its metabolites and target cell(s) or molecule(s); they are defined as "measurable changes in the organism." This definition consists of markers of effects that can indicate a preclinical response and is not always detected by using traditional clinical diagnostic techniques. The early effects of an individual can be used as informative markers

of disease risk. Several molecular-based assays have been developed to identify cellular response(s) activated by such exposures.

The amount of the agent that reaches a crucial cellular target can be detected by biomarkers of a biologically effective dose, for instance, DNA adducts or the amount of a chemical agent bound to a cellular receptor. Other biomarkers of effect measure the damage caused by the agent. For instance, under situation of a biological effective of UV dose, DNA damage induced by UV is measured and UV length can be further used to classify the type of damage. UVA exposure induces base excision repair, while UVB exposure induces nucleotide excision repair. Biomarkers of DNA damage include mutations, DNA strand break, adducts, micronuclei, sister chromatid exchanges, and chromosomal aberration.

14.2.3 Biomarkers of Susceptibility

Molecular epidemiology provides tools to identify the genetic and acquired susceptibility (such as DNA repair capacity). At present, association studies are the most common genetic epidemiological studies.

Abundant studies have been carried out on candidate genes based on biochemical hypotheses in terms of DNA repair, carcinogen metabolism, or cell cycle. Multiple GWAS and meta-analyses (see below) are currently evaluating large numbers of SNPs for an overview of methods and genetic loci that seems to correlate with diseases. Although these studies include thousands of cases and controls, which tend to be huge, some smaller studies with very well-designed selection of subjects have contributed to the understanding of genetic susceptibility as well. For instance, in the study by Klein et al., a genome-wide screen of 96 cases and 50 controls shown that an intronic and common genetic variant in the complement factor H gene (CFH) played a crucial role in age-related macular degeneration (OR = 7.4, 95%CI = 2.9~19).

14.2.4 Biomarker Selection

Before starting the experiments, several issues should be considered when identifying candidate biomarkers. For example, it should be known the prevalence of biomarkers of interest in the population. The ability of the biomarker representing the agent of interest and the sensitivity and specificity of the biomarker measuring low dose of exposure should be considered as well. Additionally, the validity of the biomarker and reliability must be determined. Epidemiological studies implementing biomarkers must take into consideration that environmental exposures will vary qualitatively and quantitatively over time. It also should be taken into account that part of biomarkers decay over their lifetime, thus, when selecting an appropriate design of a molecular epidemiological study, the half-life of biomarkers

must be considered. Most biomarkers are transient with a relatively short half-life period. In conventional epidemiology, case-control studies have great advantages in research of which the disease of interest is rare and the exposure is frequent and easy to identify. For instance, when investigate cancers or other chronic diseases, studies are usually focused on events that happened many years before the disease onset and often involve chronic exposures. However, implementing molecular epidemiology in chronic diseases, the method of case-control is restricted if the biomarker has a short half-life time. Therefore, it indicates an acute exposure that occurred in a short time before disease onset. However, such studies are often confined in the sense that they are using a "one-time" biological sample, which cannot certainly represent the common exposure or of changing exposures. For those biomarkers, prospective molecular epidemiological studies are more suitable.

In molecular epidemiological study, biomarker validation contains several issues that need to be considered when assessing the utility of a biomarker. Analytical validity, clinical validity, clinical utility, and ethical, legal, and social implications and safeguards are often regarded as the *ACCE* evaluation of a biomarker. The analytical validity points at the ability of a test to reliably and accurately measure the genotypes/markers of interest which includes its sensitivity and specificity. The ability of a genetic test to detect or predict the phenotype is the clinical validity or the positive predictive value. The clinical utility component of this assessment considers the risks and advantages related to the incorporation of the test into routine clinical practice. Recently, the legal, ethical, social implications and safeguards are other issues that need to consider when assessing the utility of a biomarker.

Betsou and colleagues recommended several methods to evaluate the vulnerability of a biomarker to pre-analytical variation. These evaluations can be conducted to ensure that association with clinical end points is not because of uncontrolled pre-analytical variation. For example, the characteristics could change rapidly (e.g., vitamin C is light-sensitive) for serum is not processed rightly. Other reasons of pre-analytical variations include fasting conditions, specimen collection's time, the position of the patient when collecting, the patient's diet, or other life habits, all of which are necessary consideration when choosing appropriate biomarkers. The use of inappropriate biomarkers may partly due to the publication bias which causes false-positive associations. Many biomarkers were tested but never published because of the unfavorable results. Publication bias might be a result of time consuming and costly assays, such that the positive findings of manuscripts are more likely to be published than negative findings, although they may have been acquired by chance.

14.3 Main Research Methods Used in Molecular Epidemiology

14.3.1 Study Design in Molecular Epidemiology

In the past several years, it might be the most revolutionary changes in molecular epidemiology for the emerging of discovery technologies that can been put into use in many study designs, such as genome-wide scans of common genetic variants, messenger RNA (mRNA) and microRNA expression arrays, proteomics, and metabolomics (also referred to as metabonomics). These approaches are helping investigators to explore biological responses to exogenous and endogenous exposures, to evaluate potential modification of those responses by variants in essentially the entire genome, and to define tumors at the chromosomal, DNA, RNA, and protein levels. Biomarkers of genetic and environmental factors referred to human disease have been applied to cross-sectional studies, case-control studies, and cohort studies.

14.3.1.1 Cross-Sectional Studies

Cross-sectional studies can be used to assess allele and genotype frequencies, exposure levels in the population, and the relationships among genotypes, exposures, and phenotypes. Although cross-sectional studies cannot infer causality between incidence and natural history, they can provide information on genetic variants and environmental exposures that may help to guide research and health policy at population level. For example, a population-based prevalence study analyzes two common mutations in the hemochromatosis gene (HFE) (C282Y and H63D variants of HFE) in the U.S. population. Steinberg et al. genotyped 5,171 samples from the Third National Health and Nutrition Examination Survey (NHANES III) of Center of Disease Control and Prevention (CDC), a nationally representative survey conducted in the United States from 1992 to 1994. Genotype and allele frequency data were cross-classified by sex, age, and race/ethnicity. The CDC provides an ongoing assessment of the U.S. population's exposure to environmental chemicals by the analysis of NHANES surveys. The first National Report on Human Exposure to Environmental Chemicals was issued in 2001 and presented exposure data for 27 chemicals from NHANES 1999–2001. In 2003, The second report presented exposure data for 116 environmental chemicals stratified by age, gender, and race/ethnicity. Furthermore, by cooperating with the National Cancer Institute (NCI), the CDC applied the NHANES III survey to measure prevalence of variants in 57 genes and correlate the resulting genotypes with clinical, medical history, and laboratory data. When completed, such studies will provide valuable information on the association between genetic variations and numerous health end points.

14.3.1.2 Case-Control Studies

The case-control approach is especially well suitable to study genetic variants in that (a) unlike other biologic markers of exposures such as DNA adducts and hormonal levels, genetic markers are stable indicators of host susceptibility; (b) case-control studies can implement an all-sided search for the effects of several genes, along with other risk factors, and look for gene-environment interactions; and (c) case-control studies are suited for plentiful unusual disease end points (e.g., specific cancers and birth defects). Furthermore, because the environmental exposures change over time, cohort studies with repeated biomarkers of exposures and intermediate outcomes may be preferable to case-control studies, unless case-control studies are nested within an underlying cohort of a well-defined population for which biological samples stored at the start of the study are later analyzed for exposures. Case-control studies can synchronously support gene discovery and population-based risk characterization. For instance, registries of population-based incident disease cases and their families offer a platform to conduct family-based linkage and association studies. The reflection of this philosophy is the NCI sponsors Cooperative Family Registries for Breast and Colorectal Cancer Research. Population-based case registries can support many study designs, including extended family studies, case-parent trios, and case-control family designs. One type of family-based association study is the kin-cohort design in which researchers access the genotype-specific risk of disease occurrence in first-degree relatives of study participants (probands), inferring genotypes of relatives from genotypes measured in probands.

14.3.1.3 Cohort Studies

Efforts are now being done to integrate genomics into cohort studies started in the pregenomic era to study disease incidence and prevalence, natural history, and risk factors. Well-known cohort studies include the Framingham study, the Atherosclerosis Research in Communities study, the European Prospective Investigation on Cancer, and the newly designed National Children Study, a planned U.S. cohort study of 100,000 pregnant women and their offspring to be followed from before birth to age 21 years. In addition, the genomics era is enlightening the development of very large longitudinal cohort studies and even studies of entire populations to set up repositories of biologic materials ("biobanks") for discovery and characterization of genes relevant to common diseases. There are adequate number of studies could be listed, which range from large random samples of adult populations such as the UK Biobank ($N = 500{,}000$) and the CartaGene project in Quebec ($N = 60{,}000$) to populations of entire countries such as Iceland ($N = 100{,}000$) and Estonia ($N = 1{,}000{,}000$; Estonian Genome Project), to a cohort of twins in multiple countries (GenomeEUtwin). It is worth mentioning that the China Kadoorie Biobank was launched in 2004, which recruited 0.5 million people with blood data and then collect their health information for at least two decades. These biobanks can help

epidemiologists to quantify the occurrence of diseases in multifarious populations and to understand their natural histories and risk factors, including gene-environment interactions.

Longitudinal cohort studies allow for repeated phenotypic and outcome measures of individuals over time, including intermediate biochemical, physiologic, and other precursors and sequels of disease. Cohort studies can also be applied to nested case-control studies or even as an initial screening method for case-only studies (as explained before). Such studies will generate abundant data on disease risk factors, lifestyles, and environmental exposures, and make preparations for data standardization, sharing, and joint analyses. An example of data standardization across international boundaries is the global P3G (Public Population Project in Genomics), which, to date, includes three international studies from Europe and North America. "Harmonization" is vital for creating comparability across sites on measures of genetic variation, environmental exposures, personal characteristics and behaviors, and long-term health outcomes.

14.3.2 Main Molecular Methods Used in Molecular Epidemiology

Numerous molecular biological techniques are implemented in the epidemiological studies. This section listed main methods used in the molecular epidemiological studies.

14.3.2.1 Electrophoretic Mobility Shift Assay (EMSA)

The EMSA or mobility shift electrophoresis referred as a gel mobility shift assay, gel shift assay, gel retardation assay, or band shift assay as well, a usual affinity electrophoresis techniques, is used to study protein-DNA or protein-RNA interactions. This procedure can confirm if a protein or mixture of proteins is able to combine with a given DNA or RNA sequence. Sometimes, it can be used to indicate if more than one protein molecule take part in the binding complex. Gel shift assays are often performed in vitro concurrently with DNase footprinting, primer extension and promoter-probe experiments when studying transcription initiation, DNA replication, DNA repair or RNA processing and maturation. Precursors can be found in earlier literature, but most present assays are based on methods described by Garner and Revzinand Fried and Crothers.

The EMSA technique is based on the observation that protein-DNA complexes migrate more slowly than free linear DNA fragments when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis. Because the rate of DNA migration is shifted or retarded when bound to protein, the assay is also defined as a gel shift or gel retardation assay. The ability to resolve protein-DNA complexes

depends greatly on the stability of the complex during each step of the procedure. During electrophoresis, the protein-DNA complexes are quickly resolved from free DNA, providing a "snapshot" of the equilibrium between bound and free DNA in the original sample. The gel matrix provides a "caging" effect that contribute to stabilize the interaction complexes: even if the components of the interaction complex dissociate, their localized concentrations remain high, promoting positive reassociation. Additionally, the relatively low ionic strength of the electrophoresis buffer helps to stabilize transient interactions, permitting even labile complexes to be resolved and analyzed by this method.

Protein-DNA complexes formed on linear DNA fragments lead to the characteristic retarded mobility in the gel. However, if circular DNA is used (e.g., mini-circles of 200–400 bp), the protein-DNA complex may actually migrate faster than the free DNA, analogous to what is observed when supercoiled DNA is compared to nicked or linear plasmid DNA during electrophoresis. Gel shift assays also help to resolve altered or bent DNA conformations that induce by the binding of certain protein factors. Also, gel shift assays are suited for protein-RNA and protein-peptide interactions by using the same electrophoretic principle as well.

14.3.2.2 Dual Luciferase Reporter Assay

The wide applications of genetic reporter systems help to study eukaryotic gene expression and cellular physiology, including the study of receptor activity, intracellular signaling, transcription factors, mRNA processing and protein folding, and so on. Dual reporters are usually applied to enhance experimental accuracy. The term "dual reporter" refers to the simultaneous expression and measurement of two individual reporter enzymes within a single system. Generally, the "experimental" reporter has relation to the effect of specific conditions of experiment, while the activity of the co-transfected "control" reporter offers an internal control that act as the baseline response. Normalizing the activity of the experimental reporter to the activity of the internal control minimizes experimental variability caused by differences in cell viability or transfection efficiency, which also can effectively eliminate other sources of variability, including differences in pipetting volumes, assay efficiency and cell lysis efficiency, and so on. Hence, dual-reporter assays often permit more reliable interpretation of the experimental data by reducing extraneous influences. The Dual-Luciferase Reporter (DLR™) Assay System offers an efficient method of performing dual-reporter assays. In the DLR™ Assay, the activities of firefly (Photinuspyralis) and Renilla (Renillareniformis, also known as sea pansy) luciferases are measured sequentially from a single sample. The firefly luciferase reporter is measured first by adding Luciferase Assay Reagent II (LAR II) to generate a stabilized luminescent signal. After quantifying the firefly luminescence, this reaction is quenched, and the Renilla luciferase reaction is simultaneously initiated by adding Stop & Glo Reagent to the same tube. The Stop & Glo Reagent also produces a stabilized signal from the Renilla luciferase, which decays slowly over the course of the measurement. In the DLR™ Assay System, both reporters

yield linear assays with subattomole sensitivities and no endogenous activity of either reporter in the experimental host cells. Furthermore, the integrated format of the DLR™ Assay provides rapid quantitation of both reporters either in transfected cells or in cell-free transcription/translation reactions. Promega provides the pGL4 series of firefly and Renilla luciferase vectors designed for use with the DLR™ Assay Systems. These vectors may be used to co-transfect mammalian cells with experimental and control reporter genes.

14.3.2.3 The Comet Assay

The "comet" assay was developed in the late 1980s/early 1990s and used only some lymphocytes. The lymphocytes are frozen at a very low temperature to ensure their viability, and then treated and run out on a gel that was spread on a glass slide. DNA from the cell "migrates" to form a "tail." If DNA is "broken" (i.e., single-strand breaks), then the length of the tail is relative to the amount of breakage. This assay tends to measure DNA single-strand breaks, cross-links, base damage, and apoptotic nuclei. Cells could be subject to damaging agents first, then allowed to repair, and placed on the gel on the glass slides. In this situation, this assay measures DNA repair "capacity" by the length of the comet tail. The comet assay is commonly used in assessment environmental toxicant-induced DNA damage. The application of this assay exponentially increased based on its high sensitivity and specificity. This method also enables researchers to detect increased risk for different health outcomes. Massive validation efforts have been taken on optimizing standardization and reliability of the comet assay by the European Standards Committee on Oxidative DNA Damage.

14.3.2.4 Micronucleus (MN) Assay

MN assay, which is used to detect MN, extracellular bodies, after the cells go through first cell cycle, has the ability to discern chromosome breaks from aneuploidy (abnormal number of chromosomes) and can detect chromosome loss. Since MN are formed from acentric chromosomal fragments or chromosomes that are not involved in either daughter nuclei, they are classified relying on whether they include chromosomal fragments or whole chromosomes.

This assay is suited for use in molecular epidemiological studies for the relative ease of scoring, limited costs and personnel requirements, and the precision that scoring larger numbers of cells provides. The MN assay can be proceeded in peripheral blood lymphocytes, alveolar macrophages, erythrocytes, epithelial cells, and fibroblasts. In this assay, the cells under investigation must survive at least one round of nuclear division, so some of the damaged cells are lost before the analysis begins, and the survivability of the damaged cells is not known with this assay.

A review of published evaluated the occurrence of MN and the influence of genotoxic exposures on MN frequency in children and adolescents. This review

indicated that this cytogenetic assay is a helpful and sensitive tool which is suitable for biomonitoring studies of children including those with low-dose exposures to environmental agents. The confounding effects of age, sex, and chronic and infectious diseases on MN levels were evaluated in these studies, and the only variable irrelevant to MN frequency was sex.

14.3.3 Genome-Wide Association Studies (GWAS)

GWAS are designed to identify the entire human genetic associations with detectable traits or the presence or absence of a disease of interest. The precondition is the entire genome can be assessed for variation and a few SNPs would stand out as key risk factors of disease. The comparison is carried out between individuals with and without the disease of interest. Since the genome is large and the number of SNPs is countless, participants by the thousands are required to suitably investigate the associations. The method of GWAS takes advantages over candidate gene studies and it enlarges the potential of exploration of genetic analyses. GWAS recruit numerous study subjects with a disease or phenotypic trait of interest. The study subjects usually originate from ongoing collaborative scientific work including different institutions or over all the continents. These studies take benefits of high-throughput genotyping technologies, DNA isolation, automated collection of biospecimen, and high-quality-control practices, and then employ statistical analyses to determine associations between qualified SNPs and diseases or phenotypic trait of interest. Great efforts from the laboratory and biostatistical have contributed to thousands of GWAS so far, which, no doubt, conduce to the knowledge base of molecular epidemiology around the world. Accurate GWAS would replicate their results in different populations or in experimental animals, when the biological pathways have mechanistic modeling. Regarding the "common disease, common variant" hypothesis, GWAS depending on SNPs as markers of allelic variants that indicates over 1–5% of each human genome. By genetic characterization, and then fine mapping and analyses, researchers are capable of determining common genetic variations of chronic diseases.

Generally, genome-wide scanning is conducted on an initial group of cases and controls, and then a smaller standout SNPs are assessed to replicate findings in a second and a third set of cases and controls. The possibility of false-positive or false-negative findings will be reduced by the performance of such multistage study design. Furthermore, it reduces the genotyping costs as well. Additionally, with employed quality controls, the replicated genotyping provides the essential validation, particularly for SNPs of intron or unknown functional region.

Biases are inclined to happen in GWAS. Especially, population stratification is one of the most crucial confounders. For instance, a potential population structure leads to false-positive associations when the detected SNPs are also linked with unknown factors which reflect geographical origin or ethnicity of study individuals. The vast data produced by GWAS is prone to false-positive associations. Effective

statistical skills must be used to decrease the possibility of false positives raised by a lot of multiple comparisons.

Another common limitation is that current statistical methods used in GWAS capture a large number of common variants, which derived from the concept of linkage disequilibrium or other statistical algorithms validated according to the Human Genome International HapMap databases. These methods establish on the theory of human genome is constituted by blocks of nucleotides named haplotypes. Haplotypes are inherited together. Some SNPs within a given block define and explain or "tag" within block variability. These tagging SNPs get popular in GWAS. However, certain variants may not be captured by the current genotyping chips while they potentially represent crucial but unknown function. Besides, it is necessary to consider that some genetic variants might be influenced only when combining with exposures that initiate or modify expression of that gene. Without considering exposures to assess risks of chronic disease, we cannot successfully reveal the complicated patterns of gene-environment or gene-gene interactions which contribute to a great degree chronic disease risk. "Next-generation GWAS" probably should combine with more detailed analyses of common exposures (e.g., smoking, alcohol, dietary patterns, air pollutants, over-the-counter medications (like common non-steroidal anti-inflammatory drugs (NSAIDs)), and recreational drugs), which influenced the chronic disease etiology and pathogenesis.

14.3.4 Mendelian Randomization (MR)

Confounding, selection bias, and reverse causation are major problems in building causal relationships between exposures and diseases, which may lead to spurious associations. MR is a method by using genetic variations of known function to detect the causal effect of a modifiable exposure on disease in nonexperimental situation. A vital characteristic of observational epidemiology is to identify the causes of common diseases which public health takes interest. For the purpose of confirming the favorite effects of a recommended public health intervention, the association of observation between the certain risk factor and a disease must prove that the risk factor indeed causes the disease. Well-known successful examples are that causal relationships are identified between smoking and lung cancer, and between blood pressure and stroke. However, there are failures when identified exposures were later demonstrated by randomized controlled trials (RCTs) to be noncausal. For example, hormone replacement therapy (HRT) was previously thought prevent cardiovascular disease. However, it did not and may even have other adverse effects in health. In observational epidemiological studies, the confounders such as social, behavioral, or physiological factors result commonly in such spurious findings. They are easy to uncontrol and especially difficult to measure accurately. Furthermore, many findings repeat unlikely by RCTs for ethical reasons.

MR allows one to test for a causal effect from observational data in the presence of confounders by taking common genetic polymorphisms with well-understood

effects on exposure patterns. Necessarily, the genotype must only affect the disease process directly through its effect on the exposure. Since genotypes are assigned randomly when inherit from parents to offspring during meiosis, if we hypothesized that option of mate is unrelated with genotype (panmixia), the genotype distribution among population should be irrelevant to confounders that commonly trouble observational epidemiological studies. Therefore, MR can be considered as a "natural" RCT. From a statistical perspective, it is a use of instrumental variables, with genotype serving as an instrument/proxy for the exposure.

The same with all studies of genetic epidemiology, trouble exists in the requirement for large sample sizes, the non-replicable results, and the lack of functional proof on genetic variants. In addition to these limitations, genetic findings could be confounded by other genetic variants by linkage disequilibrium with the variant under study or by population stratification. Moreover, pleiotropy of a genetic variant may contribute to null associations on account of canalization of genetic effects. If correctly performed and carefully interpreted, MR studies can offer valuable evidence to identify causal hypotheses between environmental exposures and common diseases.

14.4 Application and Prospection

14.4.1 Control and Prevention of Infectious Diseases

The aim of molecular epidemiology of infectious diseases is to apply molecular (amino acid or nucleotide) sequences to study the ecology and dynamics of pathogens. For infectious diseases, it includes the transmission system (source of infection, transmission route, and susceptible population), pathogenesis and virulence of the microbe, the interaction between microbe and the human (or other) host(s), and the microbiota of the host (the area microbes usually live on and in the human body).

14.4.1.1 Outbreak Investigation

In all outbreak investigations, setting the definition of a case is a key step. Molecular techniques are the standard tool in an outbreak investigation for clarifying case definitions, enhancing specificity, and decreasing misclassification. During an outbreak of disease, it is commonly assumed that a single microbe causes the clinical symptoms. A microbe of the same genus and species but different strains is possible cause of disease during the same period. Case definitions can be refined by including the molecular typing which would increase the specificity, reduce misclassification of non-outbreak cases with outbreak cases, and potentially increase the possibility to identify the outbreak source. Only based on clinical symptoms, we are hard to distinguish between diseases. This could make outbreak investigations complicated, especially if the symptoms are not very typical. For instance, lots of viruses could

cause flulike symptoms; however, classification of influenza based on clinical symptoms is specific only during an epidemic when a large number of flulike patients suffer from influenza. Even during an epidemic, the confirmation from laboratory is required as well, since there may be not only one strain of influenza in transmission. In 2008, there were two predominant influenza A strains in circulation: H1N1 and H3N2. Laboratory test is particularly helpful for identifying individuals with mild or atypical symptoms, and determining the specific type. A variety of methods of laboratory tests could provide a molecular fingerprint based on the microbial genotype. For example, pulsed-field gel electrophoresis (PFGE) is applied as the standard method for foodborne outbreaks investigations.

14.4.1.2 Trace Dissemination of a Specific Subtype of Pathogen Across Time and Space

Microbes that cause human disease are constantly emerging and reemerging. In order to prevent and control the spread of infection, we must be capable to trace the origin and source of entry of pathogens into the population. By comparing strains, we can determine if there have been single or multiple points of entry, and if emerging resistance is from multiple spontaneous mutations or from dissemination of a single clone. For example, Streptococcus pneumonia (S. pneumoniae), a major human pathogen and one of the most common indications for antibiotic use, results primarily in pneumonia, but also gives rise to meningitis and otitis media. However, resistance to penicillin emerges relatively slowly, once it emerged it was widely disseminated in relatively few clones as defined by multilocus sequence typing (MLST). By contrast, the recent emergence of S. pneumoniae resistant to fluoroquinolones has been due to various genetic mutations, suggesting spontaneous appearance after treatment. Because the resistance of S. pneumoniae to fluoroquinolones rapidly followed the introduction of fluoroquinolones, alternative antibiotics will be needed in relatively short order to treat S. pneumoniae infections.

14.4.1.3 Determine the Origin of an Epidemic

Molecular tools help us to trace an outbreak or epidemic return to its origin in time, and return to its reservoir in space. Knowing the origin in time is critical to predict future spread, and identification of the reservoir for infection is the key to control disease spread. For instance, the prevalence of methicillin-resistant staphylococcus aureus (MRSA) has been a steady increase in America's hospitals. In 2004, among some intensive care units, the prevalence was as high as 68%. Nevertheless, in the early 2000s, the emerging of new strains of MRSA in population from community could not be traced back to hospitals. Genetic typing of the strains verified that strains isolated from those who had no linkage with hospitals on epidemiology were genotypically different from hospital strains. More recently, community-acquired MRSA has been transmitted into hospitals. When comparing with hospital-acquired

MRSA, community-acquired MRSA has different virulence factors and different patterns of antibiotic resistance so there is a clinical benefit in enabling us to distinguish between the two.

14.4.1.4 Follow the Emergence and Spread of New Infections

Severe acute respiratory syndrome (SARS) is anew infectious disease emerging firstly this century. Before its identification, coronaviruses were not regarded as primary pathogens in that only 12 known coronaviruses can be able to infect humans or other animals. The identification of SARS resulted in a search for other coronavirus pathogens, and horseshoe bats were identified as the reservoir and civets as the amplification hosts at last. The time from the initial observation to the sequencing of the virus and development of a diagnostic test was 5 months. The story of the rapid isolation, identification, and sequencing of the coronavirus causing SARS is illustrative of the synergistic effects of the combination of molecular methods with epidemiology. This effective combination enables scientists to follow the emergence and spread, and to identify ways to prevent transmission and further introductions of the virus into human populations.

14.4.1.5 Identify Previously Unknown or Uncultivable Infectious Microbes

The most microbes could not be cultured using standard laboratory techniques. The ability of replication of genetic material and determination of genetic sequence, which can then be compared to known genetic sequence, has brought about a fundamental revaluation of vast life around, in, and on us. Noncultural techniques have enabled us to describe the microbial communities living in the mouth, vagina, gut, and other body sites, and the body sites thought to be sterile by previous detection, such as the blood. Epidemiological data may suggest an infectious origin for a disease. Previously, if an organism cannot be cultured, it remained only a suggestion. Molecular tools have altered this by the achievement of detecting uncultivable microbes. It is now known that human papillomavirus (HPV) types 16 and 18 can cause cervical and other cancers, and vaccines are licensed to prevent acquisition. HPV 16 was first identified in 1983 before the virus could be grown. When discovering HPV 16, we realize that papillomavirus could give rise to cancers in cows, rabbits, and sheep, but it was unclear whether the HPV can lead to human cancers. HPV was a suspected cause of genital cancer for the similarity to Kaposi's sarcoma, and the epidemiology suggested that an infectious agent was involved. But other genital infections, especially herpes simplex virus, were also suspects. HPV had been excluded by many, but a new molecular technique, the hybridization assay, detected in cancerous tissue a new subtype, HPV 16, which was specifically

associated with cervical and other cancers. The correlation of HPV 16 with cancers was verified by comparing presence of HPV 16 between cancer patients and controls. Notwithstanding this evidence can be very suggestive, it does not differentiate temporal order, because the cancer might happen before the infection of HPV 16. Demonstrating temporal order required large-scale prospective cohort studies. These studies also offered crucial perspectives supporting the possibility that vaccination could protect against HPV because of rare occurrence of reinfection with the same HPV subtype, and antibody could prevent reinfection and persistence of low-grade lesions.

14.4.2 Control and Prevention of Chronic Diseases

With the development of economics and the implementation of vaccines, most infectious diseases were controlled, while the incidence and mortality of chronic diseases were increased dramatically, such as cancer, cardiovascular disease, and type 2 diabetes mellitus. Molecular epidemiology played an important role in the discovery of the cause, mechanism of pathogenesis, and individual susceptibility.

14.4.2.1 Improving the Understanding of Mechanism of Pathogenesis

Previously, most cancer epidemiological studies were restricted to evaluating possible causal relationships between two types of events: exposure to potential causative "environmental" agents (cigarette smoking, dietary factors, specific chemicals from workplace, etc.) and disease outcome (i.e., clinical cancers incidence or cancers mortality). However, the specific mechanism was unknown. Increasingly, molecular epidemiological studies are combining panels of biomarkers related to exposure, preclinical effects, and susceptibility using samples of exfoliated cells, blood cells, body fluids, or tissues. These biomarkers are now being widely used in cross-sectional, retrospective, prospective, and nested case-control epidemiological studies, for the purpose of improving our cognition to the causes of specific human cancers. For example, the cotinine in serum or urine represented cigarette smoke exposure, which was a valuable supplement to traditional means of evaluating exposure. Moreover, assays have been implemented to measure "biologically effective dose" of a compound, for example, the amount that has reacted with key cellular macromolecules. The metabolite of cotinine could form the carcinogen-DNA adducts which related with lung cancer. Other molecular epidemiological studies in Chinese populations have prospectively linked DNA damage induced by aflatoxin B1 (AFB1) to liver cancer risk.

14.4.2.2 Evaluating the Susceptibility of Individual and Defining the Risk Population

Human beings evidently differ from one another in physical characteristics, personality, and other factors. They are also different in genetically determined susceptibility to disease. When we investigate the etiology of a disease, we cannot help asking the question: How much of the incidence of the disease is due to genetic factors, how much is due to environmental factors, and how do these types of factors interact with each other to increase or decrease the risk of disease? Obviously, not everyone who exposed to an environmental risk factor will necessarily develop disease. Even though the relative risk for exposed to a specific factor is very high, the notion of attributable risk implies that not all occurrence of a disease is due only to the specific exposure in question such as the relationship between cigarette smoking and lung cancer. It is demonstrated that lung cancer does not develop in every smoker, and it does develop in someone who does not smoke.

People often accept a fatalistic approach when they are told that a disease is primarily genetic in origin. But even in diseases originate primarily from gene, a good deal of environmental interaction often occurs. For example, phenylketonuria is characterized by a deficiency of phenylalanine hydroxylase for genetic reason; the child who affected cannot metabolize phenylalanine, an essential amino acid, and the excessive phenylalanine accumulation causes irreversible mental retardation. Can we prevent the genetic abnormality? No, we cannot. Can we decrease the likelihood that a child manifest mental retardation because of this genetic abnormality? Yes, we can do so by providing a diet with low phenylalanine to reduce or eliminate the child's exposure to phenylalanine. As shown in this example, we can prevent the adverse effects of a genetic disease by controlling the affected person's environment so that the manifestations are not expressed. Hence, in viewpoints of both public health and clinical medicine, it is crucial that bear in mind the interrelationships between genetic and environmental factors in disease causation and expression.

14.4.3 Conclusions

Traditional epidemiology has achieved greatly vital goals by means of simple tools such as interviews and questionnaires. Even a difficult issue, for example, the relationship between air pollution and chronic disease, has been successfully disposed by time-series analysis and other means not depended on the laboratory. Hence, it needs to be evaluated carefully for the application of molecular techniques combined with epidemiological designs.

As the examples above demonstrated, molecular epidemiology is not different with conventional epidemiology, but represents an endeavor that commence to achieve specific scientific goals: (1) a better description of exposures, especially when exposure doses are fairly low or different sources of exposure should be

integrated in a single measure; (2) the study of gene-environment interactions; (3) the application of markers of early response, for the purpose of overcoming the main limitations of chronic disease epidemiology, that is, the relatively low frequency of specific forms of disease and the long latency period between exposure and the onset of disease. Also limitations of molecular epidemiology should be acknowledged: the complicacy of various laboratory methods, with scanty knowledge of measurement error or interlaboratory variability; the lacking recognition of some sources of bias and confounding; in some situations, the lower degree of accuracy (such as urinary cotinine compared to questionnaires on smoking habits); and the indefinite biological meaning of markers, like some circumstances of some types of adducts or some early response markers.