Molecular Diagnostics of Polygenic Human Disease: Novel Strategies, Novel Perspectives

Molekulare Diagnostik polygener menschlicher Erkrankungen:
neue Strategien, neue Perspektiven

A. Pfeufer

Summary: Since the introduction of cellular pathology, pathobiochemistry and immunology into medicine, no other single methodology has advanced the knowledge of medical pathophysiology more than molecular genetics with its ability to identify disease genes without initially having to know about disease mechanisms. Molecular genetic research has proven enormously efficient for any mendelian – i.e. monogenic – disease. As today's research in laboratory medicine is mostly concerned with polygenic and complex diseases, molecular genetics so far has only been of limited value to the field.

Recent scientific progress including the near-finishing of the human genome sequence and the cataloging of human genetic variability in the form of an SNP map has greatly advanced our understanding of human molecular genetics. This knowledge will in return very likely increase our future ability to successfully apply the methodology of molecular human genetics in order to promote our insight into causation and pathophysiology of polygenic and complex disease.

Keywords: polygenetic disease; molecular genetics; linkage study; association studies; SNPs; haplotypes; genetic diagnostic testing.

Zusammenfassung: Seit Einführung der Zellulärpathologie, der Pathobiochemie und der Immunologie in die Medizin hat keine andere Methodologie das Verständnis der medizinischen Pathophysiologie stärker vorangetrieben als die molekulare Genetik. Das liegt vor allem an ihrer Fähigkeit, Krankheitsgene auch ohne eine initiale Kenntnis der Krankheitsmechanismen zu identifizieren. Molekulargenetische Forschung hat viel zu unserem Verständnis der mendelsch vererbten, d. h. monogenen Erkrankungen beigetragen. Da sich die Forschung auf dem Gebiet der Laboratoriumsmedizin heutzutage jedoch am intensivsten mit polygenen und komplexen Erkrankungen beschäftigt, war die molekulare Genetik hier bisher nur von geringem Nutzen.

Der aktuelle wissenschaftliche Fortschritt einschließlich der Vollendung der humanen Genomsequenz und dem Katalogisieren von humaner genetischer Variabilität in der Form von SNP-Karten hat unser wissenschaftliches Verständnis von molekularer Humangenetik stark vertieft. Dieses Wissen wird mit hoher Wahrscheinlichkeit im Gegenzug unsere zukünftigen Fähigkeiten substantiell verbessern, die molekulare Humangenetik erfolgreich bei der Erforschung der Ursachen und Mechanismen polygener Erkrankungen einzusetzen.

Schlüsselwörter: polygene Erkrankung; Koppelungsanalyse; Assoziationsstudien; SNPs; Haplotypen; molekulargenetische Diagnostik.

Introduction

One major and characteristic focus of modern biomedical research has always been directed towards understanding the physiology and pathophysiology of a cell, a tissue, an organ, or the whole organism in healthy and in diseased states. It is generally deeply believed that fundamental knowledge gained from this kind of research will increase our ability to cure, alleviate, or prevent diseases, and this idea has proven correct innumerable times.

The introduction of cellular pathology by Rudolf Virchow in the second half of the 19th century has put cells in the focus of the pathophysiologist. It reflects a milestone change of concept in medical research. The introduction of molecular biology into medical research starting about three decades ago has led to coining the term "molecular medicine" to communicate the view that molecular biology had added the molecular perspective to medicine and had established the next milestone after Virchow's. In fact this is only in part the case.

Laboratory medicine from the early investigations into disease carried out at the bench by physicians, chemists, and pharmacists* up to the success story of

Corresponding address: A. Pfeufer, Institute of Human Genetics, University Clinic Rechts der Isar, Technical University Munich, and NGFN Research Network, Munich Alliance for Genomic Research in Cardiac Arrhythmias, Trogerstr. 32, D-81675 Munich. Tel.: +49 89 2890-6381, E-Mail: arne.pfeufer@web.de

*All scientists pursuing molecular research and diagnosis in the field of medical pathophysiology, be they clinical chemists, clinical biochemists, laboratory physicians, molecular pathophysiologists or any other, will in this article be jointly referred to as "laboratory physicians" solely for the reason of brevity.

clinical enzymology initiated in the 1950s was a first true molecular kind of medicine. It has always pursued research into disease by the scientific method at the level of the molecule, be it either a gene product (e.g. enzyme, globulin, antibody, or other protein) or one of its metabolites. By this, laboratory medicine has, much earlier than molecular biology, played a pioneering role in shaping the concept of molecular pathophysiology and has proven the effectiveness of research based on individual molecules. The scientific success of laboratory medicine has led to the identification of many causal factors of disease together with disease-associated parameters, which have often translated into valuable diagnostic tests. This heritage of laboratory medicine has now placed it in a leading role in investigating the molecular causes and mechanisms of such common ailments as lipid disorders in atherosclerosis, inflammation parameters in infection and autoimmunity, glucose metabolism in diabetes, and hormone levels in endocrinopathies.

Novel Methods of Disease-Oriented Research

Independently of the methodical differences, all research in classical histopathology, cellular pathophysiology, molecular biology, or laboratory medicine critically depends on the analysis of affected tissues – and unaffected controls – or the molecules released from them into the circulation.

About two decades ago a revolutionary molecular research method has entered the scene of medical research, which abolished this need and thereby established a truly new milestone. It set out to identify the causes of disease not by analyzing their pathophysiology but their pattern of inheritance. Because working from the disease to the gene and not from the gene product to the disease first appeared somewhat like doing research “the other way round”, the method was initially termed “reverse genetics”. It has later become known as positional cloning [1]. It identifies the position of those stretches of the DNA molecules that are shared by all affected family members but not by unaffected ones, a procedure known as linkage analysis. On those stretches of DNA a mutant form of the disease gene will reliably be present, which can subsequently be “positionally cloned”, i.e. identified and isolated. The whole discipline has been termed molecular genetics, not to be confused with molecular biology.

The positional cloning approach proved overwhelmingly successful for any disease with a clear cut (i.e. mendelian) autosomal or X-chromosomal, dominant or recessive inheritance, a high disease penetrance, and little modification to it by other genes or environmental factors; in other words, a typical monogenic disease. Through positional cloning, molecular genetics has strongly influenced large fields of medical research in the last decade.

Not surprisingly, the fields that have benefited the most from positional cloning methods have been those

in which laboratory medicine methods were the least effective some decades earlier. In these fields most disease proteins are structural or regulatory in function and are either insoluble or present in miniscule amounts, as in neurological, muscular, or neoplastic disease. So clinical chemistry methods can only be of limited value here.

Laboratory medicine is among those medical specialties that have only slowly embarked on this new track of research. This is mainly due to laboratory medicine's past success story in elucidating the etiologies of most monogenic diseases in its field (e.g. recessive enzymopathies and other inborn errors of metabolism) by its very own methods. Clinical chemists and laboratory physicians, unlike many other medical specialists, did not have to await positional cloning to identify the causal molecules in the diseases they were interested in, but had accomplished the task ahead of time.

Because most of its monogenic diseases have been elucidated long ago, today's laboratory medicine research is primarily concerned with diseases known to be polygenic or rather multicausal and complex in origin, sharing obscure patterns of inheritance, a multitude of exogenous modifying factors, and incomplete penetrance. Up to now positional cloning had little to offer for researchers involved in such projects. This article is about the fact that this may change in the very near future.

Association Studies in Molecular Genetics

Molecular genetics does not only offer to the researcher the positional cloning approach via the combination of family study method and linkage analysis. Another molecular genetic way of proving the involvement of a genomic locus in a disease is called genetic association study. Similar to linkage analysis, it can lead to the positional cloning of the responsible gene in that locus.

The rationale of association studies

Genetic association study is in principle a straightforward concept that has much in common with other ways of doing medical research and practice. It has traditionally been a general method used by physicians to look for specific signs and symptoms occurring “together” with a certain medical problem in order to find out a little more about a patient's disease or prognosis. Association is in fact the basic concept of obtaining any medical diagnosis. To find that a certain parameter is highly indicative for (or in other words, strongly associated with) the presence of a disease, e.g. serum troponin levels and acute myocardial infarction or sleep apnea and snoring, is well accepted. (And, hey, wearing glasses is associated with being a little smarter, isn't it?)

So association is not new in the medical field and has for long times been applied in all of its disciplines. It is also well known by physicians that none of the as-

sociative diagnostic tests or procedures gives a perfect result (i.e. the association is never 100 percent), but any can return a false-positive or a false-negative value. Knowing this and still obtaining correct medical diagnoses by taking the appropriate probabilities (the positive and negative predictive values of each test) into account is one of the aspects that makes medicine an art, not a craft.

Genetic association is very similar. It is doing basically the same thing but taking genetic markers instead of lab values (e.g. troponin T), results of apparative diagnostic tests (e.g. ECG), or signs and symptoms named to honor some masters of the art (e.g. Kussmaul's sign).

The major rationales behind genetic association studies are first the observation that some people contract disease while others do not, even if both are exposed to the same environmental challenges, and second the hypothesis that this effect may be, at least in part, due to the genetic variation existing between those different individuals.

The history of genetic association studies

Genetic association is also not new but has been used for several decades: In the days of protein biochemistry, only those genetic markers that gave rise to a typable protein polymorphism were available for association (e.g. blood groups, HLA antigens, haptoglobin variants [2], glucose 6 phosphate dehydrogenase variants [3], some of which are still perceived important in disease-oriented research [4]). Later, restriction length polymorphisms (RFLPs, most of them being in fact single nucleotide polymorphisms, SNPs) entered research and were among the first markers typed directly at the level of DNA. But back then the method suffered from the limited number of known markers compared to the large size of the human genome, as there were only some 350 known polymorphic protein systems and initially not too much more RFLPs.

That changed with the introduction of genome-wide, easily typable genetic marker sets in the 1990s, and the widespread use of the method became feasible. Since then, the number of genetic association studies has soared. At the same time, the criticism on the method has soared, too. Critics have accused the method of obtaining false or at least unreliable results. An exemplary case is that of the ACE I/D polymorphism, which was discovered in 1992 to be associated with cardiovascular risk and has probably since then been the genetic marker most often typed in disease-oriented association studies (not taking blood groups and HLA antigens into account). The multitude of studies about this polymorphism with conflicting results reporting or remitting association with a wide variety of traits have led to a controversial and in part even emotional discussion in the literature.

General problems of association studies

Critics of genetic association studies complain about the problems of obtaining spurious, false-positive, or

irreproducible results. These accusations have a well-founded basis that needs a thorough consideration by anyone who is setting out to conduct or interpret such studies.

The problems of genetic association studies can be divided in two main categories, which need separate consideration. First, problems of a general nature with respect to association studies are encountered, which occur no matter whether the associated parameters are genetic, environmental, sociological, or other. Second, problems occur that are specific to genetic association studies.

The general problems of association studies have to do with their intrinsic statistical requirement of maximum sample sizes. It is intuitively clear that the larger the number of individual observations is, the more valid the association results will be. The statistical framework of the issue of significance and related topics in association studies is given by the well-known χ^2 statistics, only to name the most popular one.

Another problem is the requirement for sample homogeneity. It is also intuitively clear that if the investigated characteristic can be caused by several factors independently, it will be much harder to identify any one of those in samples where all those causal factors are present to a varying degree. So maximum care should be taken that the samples used for association are homogeneous for what is called etiology in the medical field.

In addition – and well known to any laboratory physician – all association studies have to cope with the problems concerning correctly and reproducibly measuring the parameters that will be used in the investigation. This is especially a problem in association studies as their intrinsic statistical requirement is to include the largest possible number of observations, individuals, or events, which makes preanalytic, analytic, and postanalytic error estimation and avoidance crucial, no matter whether the measured entities are sociological parameters or lab values.

Specific problems of genetic association studies

Problems specific to genetic association studies are additive to the general problems just considered.

Number of markers investigated

Traditionally genetic association studies have been carried out with single, or in the case of HLA-antigens, few polymorphisms at one locus. In the age of molecular biology and genetics, now a multitude of genetic markers is available. The dbSNP database of human genomic base variation, termed single nucleotide polymorphisms (SNPs), now comprises more than 2 million entries of the total of 3 billion basepairs in the human genome. This multitude of easily typable markers gives rise to a new quality of problem concerning significance level.

Significance level

This problem has in part already been mentioned as a general problem but has become a lot more important

now. The generally accepted significance level for finding an association is a probability of the null hypothesis of less than five percent ($P < 0.05$). It is clear that when genetic association is conducted with a multitude of SNP markers with this P value, statistically one in 20 markers will give a false-positive result just by chance. Such a large false-positive (or alpha) error cannot be accepted.

To circumvent this problem, corrections for multiple testing are known. The most straightforward one of them is to divide the maximum acceptable P value by the number of tests done in one study (e.g. $0.05/20 = 0.0025$) and demand this significance level for the individual test. Another one is the somewhat more sophisticated Bonferroni correction, which gives almost similar P values for a small number of tests. Unfortunately, for larger numbers of markers these corrections demand P values that are known by practical experience to be unlikely to achieve in the field of polygenic disease association studies because of the rather small sizes of the genetic effects studied. This problem may make genome-wide association studies impossible due to too many false-positive signals [5].

Kinds of markers investigated

Markers suitable for association study have to fulfill some common criteria that differ strongly from those applying to markers used in linkage analysis (Table 1).

The mutational stability of the markers used has to be much higher than in linkage analysis, which is concerned with a family pedigree (typically < 5 generations). In contrast, association studies include many more generations and meioses (in principle, all generations since the associated allele appeared on the stage of evolution, typically > 100 generations ago), therefore increasing the chance of an additional mutation.

Also, the marker typing reliability in association studies should be higher than in linkage studies, as the higher number of marker typings introduces an increased likelihood of errors, and in linkage studies typing errors are much more easily detected by using the marker information from the relatives.

The single nucleotide markers (SNPs) recently identified in large numbers [6] perfectly match these requirements. They are biallelic, come in high num-

bers, and are relatively straightforward to type. Once created by mutation, SNPs are much more stable against further mutational events than microsatellites.

In addition, recent research indicates that for association analysis there should not be a large difference in allele frequency between the investigated markers (e.g. SNPs) and the disease variant looked for, because that decreases the statistical power to detect the disease variant [7]. Because the frequencies of the disease variants are by principle unknown, SNP markers with various allele frequencies, or a large number of SNP markers, may have to be included in the investigation. Fortunately, SNP markers are abundantly available in almost any allele frequency required.

Poor irreproducibility of association studies

Some researchers have complained about the poor reproducibility of association studies, which has made them feel the approach was incompatible with the scientific method, and as a consequence they have deemed the whole field to be of limited trustworthiness.

Fortunately, there are some scientific explanations for this problem. The first lies in the statistics of the association method itself, which has been thoroughly reviewed [8]. The statistical problem is that the initial positive association result has not been found with certainty, but always with some luck, i.e. with a false positive (alpha) error of 5% ($P < 0.05$) and a false-negative (beta) error depending on the sample size and in practice mostly being around 20% to 50% (beta = 0.5 to 0.8). The beta error causes the problem. Reproducing the study, even under entirely the same conditions (population, hospital, patients, etc.), will come with the same beta error again, so reproduction can never be guaranteed. For a higher chance of reproduction you will definitely need higher numbers of individuals, which increase beta but still don't guarantee anything.

Another problem is that, unfortunately, the medical literature is strongly biased towards the publication of positive association results. So all the many negative ones that are not known to the public distort the view. As a consequence of this it has been proposed to create an archive of all conducted association studies found to be negative to get a better view of the picture. This is highly desirable but has never been realized so far.

The heterogeneity of different populations is also one major reason for limited reproducibility. Many populations investigated for the same trait and markers have been knowingly different or – even worse – have been erroneously supposed to be similar. In fact, nothing guarantees that the same marker allele variants will be associated with the same diseases in different populations, as there is a lot of evolutionary history between all these populations and alleles. It is not even for sure that variants in the same genes will cause the same – or to be exact – will cause phenocopy versions of what appears to be the same disease in different populations.

Table 1 Comparison of marker requirements for association and linkage analysis

Markers	Association analysis	Linkage analysis
Number of alleles	low (2–3)	high (3–10 or more)
Heterozygosity	high	high
Mutational stability	very high	less high
Marker density	very high (> 1 per 0.1 cM)	lower (1 per 5 cM)
Typing reliability	very high	high

The only thing for sure is that the physiology of any human individual will be generally the same and that the variability of "wildtype" alleles across populations is limited by the requirement that "wildtype" gene products have to be functional within that context.

Comparison of genetic association and linkage studies

The simplest way of describing the differences between a linkage and an association study is to look at the knowledge each method obtains about genotypes of the tested individuals.

Linkage studies are conducted on related individuals. If the genomic segments investigated for linkage are sufficiently small so that no or only little recombination will occur between the investigated individuals (e.g. genetic distance < 10 cM), the genotypic information gained from each individual is not only that of the typed markers. It is no less than the *entire* genomic segment between the markers, including all untyped genetic variants. This is because they all have been co-inherited together on the same physical DNA segment. Thus linkage analysis identifies genomic segments that are *identical by descent* and all such analyses are therefore termed IBD analyses.

This is much different in association studies. Here only the state of the typed markers is known and nothing can be inferred about any variants in between them, as the population history of the investigated individuals' alleles is not known. Such analysis, identifying only *identical by state* alleles, is termed IBS analysis correspondingly.

IBD methods are therefore much more powerful than IBS analysis. However, they suffer one major disadvantage: During the transmission of a genomic segment through some generations in a family (i.e. some meiotic events), only very few recombinational events will occur. But, recombination is necessary to delimit the length of the causal segment shared by the affected persons and to circle into the causal gene. Practical experience has shown that it is very unlikely to collect families of over 50 informative individuals. So the genomic resolution of IBD methods is limited by this fact.

This is where IBS methods join in. During its evolutionary population history, an ancient disease-predisposing allele has gone through many meiotic events in lots of different individuals, a process that has significantly shaped (i.e. reduced) the physical size of the segment on which the actual causal variant is found. In theory this translates into a much higher resolution of IBS methods compared to IBD methods as the "collected" information of all individually shaped alleles adds to their positional informativity in the association analysis.

This comparison makes clear that both methods have their individual limitations [9]. Concerning the significance of genetic association studies, two facts should also be kept in mind. The first is that IBS studies for genetic association – if conducted properly – can prove the involvement of a genetic locus in the development of a trait at the given significance level.

On the other hand, and unlike in linkage studies, where exclusion of linkage disproves a locus because a set of IBD markers is representative for the whole locus, exclusion of association between a set of IBS markers and a trait in a population does not generally rule out the involvement of the locus in the trait, as investigated IBS markers will not with certainty be representative of the whole locus and the investigated population will not with certainty be representative for any other population.

The second fact is that because the concept of association is to identify features that occur together more commonly than expected by average, it does by itself never prove a direct causal link between those features. Causality always has to be established separately by subsequent functional analysis following a positive association result.

Linkage Disequilibrium in the Genome

Linkage disequilibrium (LD) is a way to describe the nonrandom association of alleles on the DNA strand. LD is due to the neighborhood of marker alleles on the physical DNA strand. Neighboring markers are said to be physically *linked* on the DNA molecule, but you could also term the markers' alleles to be *associated*. (I know, terminology is a bit confusing here.)

LD is created by mutation. When a mutation happens to introduce a new marker in a genomic region where one marker already exists, it will only strike a DNA strand with one of the two existing marker's alleles on it. So at the moment of mutation the new and the old marker are in what is called complete linkage disequilibrium. When inherited from now on, the new marker's allele will only occur on those DNA molecules with that specific one of the existing marker's alleles on them.

LD may remain but may also break down over the generations. That happens mainly by meiotic recombination. It is clear that LD will not exist very long if the two markers are far apart on one chromosome. But this changes, the physically closer the two markers come to be.

Overall, the factors determining the extent of LD between markers include the age of the markers, recombination rates between them, population size and history, as well as selection and genetic drift. LD is not a monotonic function of the physical distance in the genome but rather depends on the genetic distance that is shaped by the recombination rate between the markers.

The most important feature is that formulas for LD can quantify the degree to which neighboring marker alleles in the genome are occurring more frequently together than would be expected on average in a given population.

Researchers interested in LD have made predictions on its degree in isolated and in general (so-called outbred) populations. One important question is how

quickly LD breaks down over physical distance and time. Recent research indicates that conserved "blocks" of increased LD seem to be interspersed in the genome with regions of low LD even in outbred populations depending on the particular location, which may in part be due to locally decreased or increased recombination probabilities.

Recombination and LD

In recombination two homologous chromosomes cross over during meiosis. The rate of recombination per nucleotide, which affects the conservation of LD, can be calculated by comparing the genetic to the physical map. The human genetic map, which is made up of marker distances in terms of recombination frequency, has been established as a prerequisite for conducting linkage analysis. The ultimate human physical map is the nucleotide sequence of the genome, which is currently in the status of near completion.

Both maps can now be compared to estimate recombination rates. Research has shown that these vary greatly along each chromosome, from 0 to at least 9 centimorgans (cM) per megabase (MB) [10]. Several chromosomal regions up to 6 MB in length extend with particularly low or high recombination rates. The LD was noted to be higher in the former and lower in the latter. Only along the metacentric chromosomes in males was the recombination rate and hence the LD strongly correlated with the physical marker distances.

Quantification of LD

Several formulas have been proposed to quantify LD. An ideal measure of LD should definitely be independent of the size of the investigated sample of alleles. It should also be rather independent from any one single marker in a locus and from the marker allele frequencies and rather reflect a more general characteristic of the multitude of markers in a genomic locus. Table 2 lists the proposed formulas for LD.

A recent analysis of these measures on genomic and simulated data has been conducted that suggests that the "association" (ϕ) may be among the more useful parameters to quantify LD [11].

Table 2 Different metrics of LD: f_{ab} are the frequencies of the four possible haplotypes. The haplotypes are defined in a way that $f_{11}f_{22} \geq f_{12}f_{21}$ and $f_{12} \leq f_{21}$.

Name	Symbol	Formula
Covariance	D	$= f_{11} f_{22} - f_{12} f_{21}$
Association	ϕ	$= \frac{f_{11} f_{22} - f_{12} f_{21}}{(n^2 (f_{11} + f_{12}) (f_{12} + f_{22}))}$
Correlation	r	$= \frac{f_{11} f_{22} - f_{12} f_{21} \cdot x}{(n^2 (f_{11} + f_{12}) (f_{12} + f_{22}))}$
Frequency difference	f	$= \frac{f_{11} f_{22} - f_{12} f_{21}}{(n^2 (f_{11} + f_{22}) (f_{12} + f_{21}))}$
$x = \sqrt{\frac{(f_{11} + f_{12}) (f_{12} + f_{22})}{(f_{11} + f_{21}) (f_{12} + f_{22})}}$		

The necessity of LD for association studies

Some LD is necessary for all genetic association studies but the larger the LD in the investigated region the better it is.

This is due to the fact that few of the genetic markers (i.e. SNPs) that will be selected for conducting genetic association studies will be the actual causal disease variants searched for by themselves. They are much more likely to be in LD with those. Thus LD among markers is essential for the positional identification of susceptibility genes by genetic association [12]. The number of SNPs needed will strongly depend on the yet to be determined size of the individual LD blocks mentioned above.

LD-based fine mapping

LD has been very efficiently used in studies where a disease gene had already been prelocalized by linkage analysis. As this method does not have a very good positional resolution (see "Comparison of genetic association and linkage studies"), a positional method to close in on the gene is necessary in those studies where no obvious functional candidate gene is available and which cannot afford to use a brute force sequencing strategy.

In this case multiply affected individuals without pedigrees large enough to conduct a linkage analysis can be subjected to a combined approach, where via the measurement of the LD between markers and the disease variant the latter can be identified. A recent example of the success of this secondary LD mapping in complex diseases is the identification of the NOD2 gene's involvement in Crohn's disease [13, 14] where a locus had previously been identified by linkage analysis. Variation in a leucine-rich repeat in that gene appears to be responsible for some 15 % of the heritability of this frequent and complex inflammatory bowel disease. Whether other success stories are soon to follow we will soon witness. This result has been termed a milestone in the elucidation of genetic causes of complex diseases.

LD-based association analysis of candidate genes

In LD-based fine mapping, loci are investigated because they turned up as positional candidates from linkage studies. The same can be done on the rationale that some genes may be functional candidates for a disease. This method is then called LD-based association analysis of candidate genes, and the haplotype-based association approach described below (see "Haplotype-based association studies") is the method of choice.

LD-based whole genome association studies

If fine mapping of positional or functional candidate gene loci is the craft of LD-based association analysis, whole-genome LD mapping of common disease genes is the art, in that it does not require any other method to get the work done. It just needs an affected and a control population and a tight enough genome-wide

set of markers (SNPs preferred) to start and look for association. Unfortunately, up to now this is just an art by theory. Why this is so and what limitations have to be overcome to perhaps make it possible has been critically outlined elsewhere [15] although some assumptions in this work especially about the extent of LD in the genome are by now known to be too pessimistic. In addition, statistical methods to perform multilocus LD calculations (instead of the two-marker methods described above) have to be developed [16] in analogy to the multipoint linkage analysis methods developed for genome-wide linkage studies.

Haplotypes

Molecular basis of haplotypes

Linkage disequilibrium is also the basis of what has long been known to geneticists as a haplotype. A haplotype is best to be imagined as one physical DNA strand with several particular alleles on it. These alleles are, of course, in some form of LD, because they are physically close together: LD is what keeps the alleles together as a haplotype over the generations and prevents them from independently segregating, which on the contrary would be called linkage equilibrium.

Conserved haplotypes as a consequence of LD

The mentioned "blocks" of increased LD give rise to rather conserved haplotypes. In regions of strong LD, analysis of prevailing haplotypes in some populations has shown that only very limited numbers of different haplotypes occur. A recent comprehensive effort to determine the occurring haplotypes in 313 genes has come to the conclusion that the number of haplotypes in a locus does not increase by $h = 2^n$ (n being the number of markers typed in the locus) as expected in complete linkage equilibrium, but only roughly by $h = 1.1n$ [17]. This finding makes haplotype-based methods of candidate gene association analysis appear much more feasible in the near future.

Haplotype construction

Haplotype construction is not straightforward, as humans like any other diploid beings have on each of the two chromosomes two versions of each haplotype. It is therefore not straightforward to infer the haplotype information (the "phase" of the marker alleles) directly from an individual person's genotype.

Classical haplotype construction

The classical way of establishing haplotype patterns is by the analysis of genotype inheritance patterns. In meiosis, the chromosomes with the haplotypes on them become haploid in the gametes and combine to form the zygote. This strategy is expensive, as it requires the genotyping of relatives not normally necessary in genetic association studies.

Empirical haplotype construction

Empirical haplotype construction works by sequencing haploid DNA, which means cloned haploid stretches of human DNA. A referential haplotype of the long-studied human HLA antigen region has, for example, been completed by haplotype sequencing, which will now allow for detailed haplotype analysis in this highly polymorphic region [18]. This method is the most costly and laborious.

Statistical haplotype construction

The third way to establish haplotypes is from genotyping of individuals by statistical inference either via a parsimony method [19] or via an expectation maximization algorithm [20]. Recently, a novel, more advanced statistical method for haplotype construction termed the PHASE algorithm has been introduced that takes additional genomic information into account [21]. All three methods of course cannot resolve any haplotype in a locus but only the more frequent ones. The actual difficulties encountered in statistical haplotype analysis and the percentage of haplotypes resolvable by statistics in each locus are yet unknown and will become more clear as they are conducted and compared to real data.

Haplotype-based association studies

While the construction of haplotypes from diploid SNP data may be cumbersome, haplotype association analysis is straightforward. It can be best achieved by the well-established means of statistical analysis of contingency tables. The most regarded haplotype-based association study so far has been able to associate in asthmatics the response to inhaled sympathomimetic aerosol sprays to the haplotypes of the β_2 -adrenoreceptor [22].

Combining Linkage and Association Analysis

Because uncertainties remain about the success rates of association studies and that neither linkage nor association studies are sufficient or optimal on their own, joint linkage and association approaches promise greater power in understanding the genetic aspects of complex disorders.

Sequential combining of linkage and association studies

Linkage and association studies can be combined in a very suitable way as described in the section "LD based fine mapping". First a linkage test is done followed by a confirmatory association test, which also allows for a finer mapping of the disease gene.

This approach to some extent resembles a two-step diagnostic laboratory medicine test where an initial screening test with a high sensitivity is combined with a very specific confirmatory test with high specificity. This design accepts some initial false positives, in order not to miss any cases.

In the genetic two-step design, sensitivity and specificity are reverted in order. The initial linkage test has a high specificity, and the subsequent association test has a high sensitivity. The latter is necessary to be in principle able to confirm the initial linkage test result with a high probability. The association test also has a high false-positive rate, but this is ameliorated as the test is only applied to the genomic region tested positive in the linkage test, the size of which usually is about one percent of the genome.

This design is optimized to prevent false-positive genomic associations, as it encourages high specificity from the beginning, at the cost of some sensitivity. This makes much sense in research (it would not make sense in diagnosis) as it keeps the enormous costs for further testing genomic regions low, which would explode in the case of too many initial false-positive results. In this way, some significant genetic associations will be missed initially but the ones found will be highly trustworthy with high positional resolution due to the second association step.

The feasibility and sensitivity of this method has been tested by investigating 60 SNPs in a 1.5-Mb region surrounding the well-established susceptibility gene for late-onset Alzheimer's disease, APOE, where two SNPs in absolute LD give rise to three alleles APOE2, ϵ 3, and ϵ 4, which have known biochemical and pathophysiological differences. The authors found that single SNP association analysis gave significant positive signals in several other neighborhood SNPs that were up to 16 KB distant in the genomic sequence. Haplotype analysis did increase significance over that seen in single-locus tests for some of the markers and in addition improved the localizability of the – in this case known – causal gene variant [23].

Nevertheless, this approach always depends on the availability of suitable families to do the initial linkage test. For many interesting traits no such families are available, as there is neither a monogenic model disease for the trait nor is the heritability in the general population high enough to find affected families of sufficient size. In this case, a novel test invented less than a decade ago offers promise.

The transmission disequilibrium test

The transmission disequilibrium test (TDT) was developed in 1993 initially to confirm the association of the 5' region of the insulin gene on chromosome 11p to noninsulin-dependent diabetes mellitus [24] and has been continuously refined since then.

The TDT is the way of doing a linkage and association analysis in a combined fashion in one same statistical procedure. Its classical requirements in terms of patients are so called "trios", which consist of an affected child and both of its parents; the procedure has been somewhat modified statistically to circumvent this limitation in case parents are not available for analysis. In principle, the TDT compares the two haplotypes of an affected child to the two other haplotypes it could have but did not inherit, and then recognizes

if this set of alleles in affected people is somewhat distorted from those expected. This distortion is taken as indicative for a positive test result. In a way, the two non-transmitted parental alleles in the TDT serve the same purpose as the alleles of the control group individuals in a classic genetic case-control association study.

Conclusion

LD-based genetic association studies hold great future promises for all fields of medical research that address the differences in disease susceptibility and predisposition between individuals to find out about the mechanisms and etiological causes of disease. The main reason for this is, on the one hand, that those studies are directly targeted to and can reliably measure the genetic variance between individuals and within populations, and on the other hand, that the measured genetic variance does not only reflect but is in itself the molecular cause of interindividual differences. The promise of genetic association studies is especially pronounced if the experimental, statistical, logistic, and financial problems towards the expansion of the method can be overcome and genome-wide association scans become possible.

Interesting is, in addition, that in this emerging field quantitative traits from laboratory medicine such as the APOE gene variants or the insulin gene locus serve as pioneering examples. So laboratory medicine should step up and join this game as the methods and its ensuing diagnostic procedures will add a whole important new field to the endeavor towards understanding the mechanisms of disease at the very basis of its causal roots called molecular medicine.

As a resume it has to be said that molecular genetics, be it in the form of linkage or association studies, shall not be regarded as "one more very special discipline of its own" but as no less and no more than another highly valuable tool for elucidating the physiology or pathophysiology of a condition at the molecular level. This view comes along with molecular medicine as an art of immediate interdependence and collaboration between all medical and scientific disciplines centered on pathophysiology, clinical chemistry, immunology, molecular biology, molecular genetics, and many others. The added value that you may want to call synergy does not lie in the differences of the disciplines but in their integration.

References

1. Collins FS. Positional cloning: let's not call it reverse anymore. *Nat Genet* 1992;1:3–6.
2. MacDonald JL, Papiha SS. A segregation analysis of the association of haptoglobin types and ABO blood groups. *Hum Hered* 1974;24(1):45–52.
3. Saha N, Patgurnarajah N. Phenotypic and quantitative relationship of red cell acid phosphatase with haemoglobin, haptoglobin, and G6PD phenotypes. *J Med Genet* 1981;18:271–5.
4. De Bacquer D, De Backer G, Langlois M, Delanghe J, Kesteloot H, Kornitzer M. Haptoglobin polymorphism as a risk fac-

- tor for coronary heart disease mortality. *Atherosclerosis* 2001;157:161-6.
5. Emahazion T, Feuk L, Jobs M, Sawyer SL, Fredman D, St Clair D, Prince JA, Brookes AJ. SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. *Trends Genet* 2001;17:407-13.
 6. Wang D, Lander E. Large-scale identification, mapping and genotyping of single nucleotide polymorphisms in the human genome. *Science* 1998;280:1077-82.
 7. Ohashi J, Tokunaga K. The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J Hum Genet* 2001;46:478-82.
 8. Suarez BK, Hampe CL, van Eerdewegh P. Problems of replicating linkage claims. In: Gershon ES, Cloninger CR (eds.) *Genetic approaches to mental disorders*. Washington, London: American Psychiatric Press Inc., 1994: 23-46.
 9. Baron M. The search for complex disease genes: fault by linkage or fault by association? *Mol Psychiatry* 2001;6:143-9.
 10. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL. Comparison of human genetic and sequence based physical maps. *Nature* 2001;409:951-3.
 11. Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A. The optimal measure of allelic association. *Proc Natl Acad Sci (USA)* 2001;98:5217-21.
 12. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single nucleotide polymorphisms. *Proc Natl Acad Sci (USA)* 1999;96:15173-7.
 13. Hampe J, Cuthbert A, Croucher PJ, Mirza MM, Mascheretti S, Fisher S, Frenzel H, King K, Hasselmeier A, MacPherson AJ, Bridger S, van Deventer S, Forbes A, Nikolaus S, Lennard-Jones JE, Foelsch UR, Krawczak M, Lewis C, Schreiber S, Mathew CG. Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* 2001;357:1925-8.
 14. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, Binder V, Finkel Y, Cortot A, Modigliani R, Laurent-Puig P, Gower-Rousseau C, Macry J, Colombel JF, Sahbatou M, Thomas G. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001;411:599-603.
 15. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics* 1999;22:139-44.
 16. Ott J, Hoh J. Statistical multilocus methods for disequilibrium analysis in complex traits. *Hum Mutat* 2001;17:285-8.
 17. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH, Duan J, Carr JL, Lee MS, Koshy B, Kumar AM, Zhang G, Newell WR, Windemuth A, Xu C, Kalbfleisch TS, Shaner SL, Arnold K, Schulz V, Drysdale CM, Nandabalan K, Judson RS, Ruano G, Vovis GF. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001;293:489-93.
 18. The MHC sequencing consortium. Complete sequence and gene map of a human major histocompatibility complex. *Nature* 2001;401:921-3.
 19. Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990;7:111-22.
 20. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995;12:921-7.
 21. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001;68:978-89.
 22. Drysdale CM, McGraw DW, Stack CB, Stephens JC, Judson RS, Nandabalan K, Arnold K, Ruano G, Liggett SB. Complex promoter and coding region β_2 -adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci (USA)* 2000;97:10483-8.
 23. Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM. SNPping away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. *Am J Hum Genet* 2000;67:383-94.
 24. 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.

Erscheinungsweise:
4 Hefte jährlich;
2002: 6. Jahrgang
ISSN 1432-9123

Bezugspreise und -bedingungen
erfahren Sie unter:
www.blackwell.de

Auch als Online-Version bei
Synergy erhältlich.

Somnologie

Organ der Deutschen Gesellschaft für
Schlafmedizin (DGSM) • Erforschung der
Physiologie, Pathophysiologie und Klinik von
Schlafen und Wachen • Theorie und Praxis
einer Vielzahl medizinischer Fachdisziplinen
• Kongressberichte • Buchbesprechungen

Somnologie veröffentlicht begutachtete
deutsche Originalarbeiten und wird von drei
Literaturdiensten ausgewertet.

Herausgegeben von:

Dr. G. Mayer, Schwalmstadt-Treysa, D
Prof. Dr. Dr. J. H. Peter, Marburg, D
Prof. Dr. E. Rüther, Göttingen, D
Prof. Dr. M. E. Schläpke, Bochum, D
(Schriftleitung)

Blackwell
Synergy

www.blackwell.de/synergy

Blackwell Wissenschafts-Verlag · Kurfürstendamm 57 · 10707 Berlin · Germany
Telefon: +49 30 / 32 79 06-18 · Fax: +49 30 / 32 79 06-10
e-mail: journalsmarketing@blackwis.de · Internet: www.blackwell.de

Blackwell
Berlin