November 2004

Gene markers and complex disorders: A review

M S. Mahmood
Aga Khan University

N Kureshi
Aga Khan University

P M. Frossard
Aga Khan University

Review Article

Gene Markers and Complex Disorders: a Review

M. S. Mahmood, N. Kureshi, P. M. Frossard
Department of Biological and Biomedical Sciences, The Aga Khan University, Karachi.

Historical Background

For centuries human traits and diseases have been associated with genetic markers. The association of specific hair color, nail contour and facial structure with certain personality types has a long tradition in folklore and occult sciences. The rate-limiting factor has been the availability of measurable genetic markers. The first genetic marker was reported in 1901[1] and it took a generation to identify the second one.[2] When the ABO blood groups were identified as early biological markers under genetic control in the early part of the twentieth century, they were promptly used for association studies. The descriptions by Aird and his colleagues of the associations of blood-group O with peptic ulcer[3,4] were reported in the 1950s and recently Davies et al.[5] confirmed the existence of a locus for type 1 diabetes mellitus in the same region of chromosome 18 as the Kidd blood group locus.[6,7] There were 17 polymorphic blood groups recognized in 1960s[8] when isozymes took over. By the mid-1970s there were 40-50 blood group antigens and serum protein polymorphisms available as genetic markers. Yet, these markers covered only 5-15% of the human genome, and their genotyping was laborious, time-consuming and technically demanding.

About 150 protein polymorphisms were known in 1980s[9], when they were superseded by nucleotide markers identified as restriction fragment length polymorphisms (RFLPs) by Botstein and colleagues.[10] This discovery eased the work of the geneticists greatly. RFLPs covered 85-90% of the human genome and their genotyping was far more simple and rapid than protein polymorphisms. These powerful tools quickly led to major discoveries and the gene for Huntington's disease was the first one successfully mapped using RFLPs in 1983 by Gusella and colleagues.[11] RFLP analysis which required large amounts of DNA and 1-2 weeks of experimental work was soon overshadowed by sequence polymorphisms revealed through the polymerase chain reaction (PCR).

There are several types of DNA sequence variations, including insertions and deletions, differences in the copy number of repeated sequences, and single base pair differences. They are widely distributed across the genome which has brought mapping of disease genes within the realms of practicality.

Genetic markers

The DNA sequence variations function as genetic markers. A useful marker is polymorphic, i.e., exists in several states. Genetic markers may not necessarily have functional significance and therefore may not have any effect on the phenotype. They can be derived from either coding or noncoding regions of the genome. Most markers for genetic studies have been derived from repetitive sequences in the non-coding regions of the genome. Variable Number Tandem Repeats (VNTRs) or minisatellites are 15 to 100 base pairs long sequences repeated several times at a particular locus in tandem. This variation at a locus is utilized in DNA fingerprinting to identify individuals. In 1989, a novel type of polymorphic marker called microsatellite was first described.Error! Bookmark not defined. Microsatellites or Short Tandem Repeats (STRs) are repetitive runs of DNA similar to VNTRs but much shorter, usually di-, tri-, or tetra- nucleotide repeats, flanked by unique sequences. The human genome is estimated to have about 105 microsatellite repeats distributed widely throughout the genome. The repeat length of STRs is variable within populations and the mutation rate is low, providing a rich source of multi-allelic markers. These markers[12,13], can produce genotyping data within hours instead of weeks, using PCR technology. Another class of polymorphisms involves the deletion or insertion of DNA into an existing sequence. Deletions or insertions may be as small as one base pair or may involve the entire gene.

Single base pair differences are the most common of DNA sequence variations.[14] They are termed Single Nucleotide Polymorphisms (SNPs) when the variant sequence type has a frequency of at least 1% in the general population. SNPs are generally biallelic genetic markers or dimorphisms, such as a substitution C>T at position 825 in exon 10 of GNB3 gene (GNB3 825 C>T), located on chromosome 12p13.[15]

Recently there has been considerable interest in the use of SNPs for understanding the genetics of complex human diseases.[16] SNPs have many properties that make them attractive for such investigations. Most importantly, the genetic sequence variations that are responsible for functional changes will often be SNPs, since they are the most frequent and can occur in coding regions. SNPs within protein-coding sequences are of particular interest because they are more likely than a random SNP to have functional significance. These may result in critical changes in the gene-products resulting in pathogenic alterations. SNPs in non-coding DNA also have functional conse-

non-coding DNA also have functional consequences, such as those in sequences that regulate gene expression and intronic SNPs which create splice variants. Discovery of SNPs that affect biological function will become increasingly important over the next several years. The GNB3 825 C>T dimorphism causes alternative splicing in exon 9 resulting in deletion of 126 nucleotides and a subsequent loss of 41 amino acids which constitutes one complete domain of the Gi protein.[15] This change leads to enhanced signal transduction activity of the mutated Gi protein.[17] Genotyping for possible functional polymorphisms is of extreme significance and is known as 'direct' association study strategy.[16]

Secondly, SNPs are far more abundant than microsatellite polymorphisms (about once every 500-1000 base pairs) and hence are more powerful in fine mapping chromosomal regions potentially harboring the disease genes.[18] In addition to their frequency of occurrence in the genome, they are stable as they have much lower mutation rates than do repeat sequences, such as STR. Thus, non-functional SNPs lying in close physical proximity to disease-causing mutations can be effectively used to map chromosomal regions of pathogenic significance. This is known as the 'indirect' association study strategy.[16]

Thirdly, detection methods for SNPs are potentially more amenable to being automated and used for large-scale genetic analysis, due to the binary nature of SNPs. Thus, high throughput genotyping of large number of markers is feasible with the advent of microarray technologies.[18] Therefore, SNPs will be particularly important for mapping and discovering the genes associated with complex diseases.

Complex Disorders

Complex disorders refer to the characteristics or traits of organisms that are determined by multiple genes, usually with a significant environmental component as well. They are thus polygenic and multifactorial. Examples of complex disorders include cardiovascular diseases, diabetes, Alzheimer's, cancers and autoimmune disorders. These traits do not follow a clear pattern of Mendelian (monogenic) inheritance, where only single genes are involved.

In complex disorders, the interaction of multiple genetic loci and their resulting effect on phenotypic expression and disease risk is far more complex than that of monogenic disorders. Genes for about 10,000 monogenic disorders have been identified most of them resulting from single, highly penetrant mutations. Complex disorders are not the result of independent mutations in multiple genetic loci rather it is their interaction, which in combination results in the occurrence and variability of complex diseases.

Complex disorders probably result from genetic mutations that create only subtle changes in the function of the gene-products. Due to this subtlety, gene hunters have found it difficult to get a single "snapshot" of all the genes that may be involved in these diseases.[19] The genes involved in complex disorders are often referred to as 'susceptibility genes,' since they increase the risk and predispose individuals towards a particular disease. This distinguishes them from causative genes in which a mutation directly leads to a particular phenotype, as in most Mendelian disorders.

The goal is to determine the genetic components of complex disorders. To reach this goal several genetic mapping techniques have to be utilized to determine the location of susceptibility genes. It is estimated that the human genome has about 40,000 genes[20] and trying to locate the small number of susceptibility genes is like looking for the proverbial needle in the haystack! This has led to the development of several laboratory techniques coupled with elaborate theoretical and statistical armamentarium to help in the search for disease-genes.

Linkage Analysis

Linkage analysis is just one of the several methods that can be used to map genes, and is a part of a strategy known as positional cloning. This term initially referred only to the physical mapping and actual cloning of disease-causing mutations. It originated from a general paradigm that emerged from the study of single gene disorders, over the past two decades. More recently positional cloning refers to a complete process of identifying disease-genes, from the initial study design to mutational analysis.[21] Conventional methods of mapping genes such as fluorescent in situ hybridization, have been successfully used for mapping genes that have already been identified. However, the underlying biochemical defects for most diseases remain unknown. Thus statistical linkage techniques which do not require this information have become the key strategy in positional cloning.

Linkage is a method that allows the determination of regions of chromosomes that are likely to contain a risk gene, and rule out areas where there is a low chance of finding a risk gene. Linkage works by using markers such as the microsatellites, known to exist at specific chromosomal locations. A family having members with the disease of interest as well as healthy individuals, known as a pedigree, is the first and foremost requirement for linkage analysis. This method attempts to search for a marker that is consistently present in those with the disease, and not present in those without the condition. When a marker is found consistently in the affected individuals of the family more commonly than in their healthy counterparts, the marker and the

marker and the disease-causing gene are said to be linked, and are assumed to be very close together. Thus, this approach identifies disease genes by their chromosomal location, gradually zeroing in on them among their neighbors in the genome.

The co-segregation of the marker with the disease-causing mutation is essentially observed for its conformity with Mendel's law of independent assortment. That is, whether the alleles at the disease and marker loci segregate independently according to Mendel's laws or tend to be inherited together, violating Mendel's laws. While there may be several possible explanations for this effect, the most likely is that the two loci reside in close proximity to each other. Alleles on the same chromosome should segregate together at a rate that is proportional to the distance between them.

Linkage analysis has been used to find genes for many disorders, including Huntington's disease Error! Bookmark not defined. and Alzheimer's disease.[22] There are several downsides to using linkage. First, it requires the use of DNA from both affected and unaffected individuals from the same family. Second, linkage requires the use of DNA from several pedigrees harboring the disease. Using mathematical models Risch (1996) showed that linkage analysis requires a huge sample size to derive the same volume of information which can be obtained with a smaller sample size, using association studies.Error! Bookmark not defined. Lastly, linkage is a very time and labor intensive method.[23]

Association studies

Apart from linkage analysis the other major approach to mapping disease susceptibility genes, is association studies. Association is a statistical statement about the co-occurrence of alleles and phenotypes. The simplest and the most widely practiced model of association is a case-control study. A large number of cases of a complex disorder, such as stroke, are compared with age and gender matched controls from the same population for the frequency of the alleles of a specific marker in a candidate gene. Allelic association is said to occur when there is a significant change in the frequency of a marker allele for a disease trait (phenotype) compared to the control population. This is observed as deviations from the expected random occurrence of the alleles with respect to the disease phenotype. Association can occur if the allele is itself directly responsible for a biological function, such as in the case of APOE-4 allele in Alzheimer's disease.[24] This, as earlier mentioned, is referred to as the direct-association study strategy.

The indirect-association study strategy depends on the concept of linkage disequilibrium (LD). LD occurs when a marker allele lies in close proximity with the disease susceptibility allele, such that these two alleles are inherited together over several generations in the population studied. Thus the same allele will be detected in affected individuals from multiple unrelated families, but belonging to the same population and thus to the same genetic pool. The concept is similar to that which underlies linkage analysis, except that linkage is studied within families and that, the chromosomal region mapped is much larger, owing to greater genetic homogeneity of families compared to populations.

The phenomenon which leads to linkage and linkage disequilibrium is recombination, first postulated by the American geneticist, Thomas Hunt Morgan. During meiosis homologous chromosomes exchange genetic material. The alleles at chromosomal loci undergoing crossing-over are altered. Allelic recombination occurs at variable distances along the chromosomes. Those alleles that lie in close physical proximity have a much lower chance of undergoing recombination than those that lie further away from each other.

If a new mutation occurs in a gene leading to a disease phenotype and the mutation-harboring chromosome is transmitted to several generations, the marker alleles at loci further away from the mutation will undergo recombination faster than those that are in close physical proximity with the mutation. The closer the marker is to the disease gene, the longer the allele-disease association will persist. Thus the markers in close proximity with the mutation will appear to cosegregate with the mutation. In such a case the marker allele and the mutation are said to be linked and, there is said to be an association between the marker allele and the disease phenotype.

Thus, in 'direct' association studies a candidate gene approach is used to test the hypothesis that an allele of a given SNP increases disease susceptibility. This is achieved by comparing the frequency of the candidate gene marker (SNP) allele in affected versus unaffected individuals. If reasonable candidate genes can be proposed for a disorder, then association studies can be undertaken on these genes, without the requirement of the presence of linkage disequilibrium. The probability of observing an association for a reasonable candidate gene is larger than for a randomly selected marker allele. If an association is found, it could still be due to linkage disequilibrium, and so determination of the causative mutation may not be straightforward.

In 'indirect' association studies multiple SNPs are used to search for evidence of an ancestral allele that is enriched in affected individuals. Indeed, the chances of a randomly selected marker to be of direct biological significance and strongly associated with the complex phenotype are low. Given the small distance over which LD extends for biallelic SNPs[25], the size of the human genome and the low penetrance of the polymorphisms for complex traits. A more reasonable role of indirect-association studies is in

extends for biallelic SNPs[25], the size of the human genome and the low penetrance of the polymorphisms for complex traits. A more reasonable role of indirect-association studies is in fine-mapping, once linkage studies have indicated a region of interest for follow-up analysis.

Signals derived from genetic association studies tend to be stronger than those produced by linkage analysis. However, such investigations are experimental 'gambles', since they are prone to a range of potential pitfalls (appropriate patients/controls and populations, fundamental genetic complexity, technology limitations), many of which cannot be controlled for. Large-scale association studies have become feasible, because of the recent development of genotyping strategies and a wide array of statistical tools.

Example

The renin-angiotensin system (RAS) is one of the most appealing candidate pathways for cardiovascular diseases (CVD). Associations of various SNPs in the renin, angiotensinogen, angiotensin-converting enzyme (ACE) and angiotensin receptor genes could possibly be searched for. There are several known polymorphisms in each of these genes. The ACE insertion/deletion (I/D) polymorphism has attracted much attention of the scientific community for its possible involvement in CVD. Since its discovery in 1990[26], there has been an upsurge of association studies of the ACE I/D polymorphism and CVD. The D allele was first shown to be associated with myocardial infarction (MI).[27] Studies in the rat showed increased rate of angiotensin II production in the myocardium due to increased ACE gene expression.[28] If the relation between the I/D polymorphism and serum ACE concentration was extrapolated to the myocardium it would be logical to expect higher angiotensin II generation in the heart and coronaries of D/D individuals. Based on such possible pathophysiological relations - the candidate gene approach - a fascinating wave of studies was started in search of associations between the D allele and CVD including stroke, renovascular disease and MI.[29] Even though after several years of scientific endeavour and hundreds of association studies of the I/D polymorphism and CVD resulted in conflicting results[30,31], the ACE locus continued to remain of much interest especially since linkage analysis reported two QTL on long arm of chromosome 17.[32] Recently a polymorphism in exon 17, ACE G2350A was reported to have the most significant effect on plasma ACE concentration. After adjustment for the effect of ACE G2350A polymorphism, the I/D polymorphism was no longer associated with plasma ACE concentration, indicating that it is in LD with ACE G2350A and unlikely to be a functional mutation.[33] We recently reported an association of this dimorphism with hypertension. Thus, the search for the functional mutation continues and association studies are a major tool geneticists have to undertake such an immense task.

Ethical Issues

The recent access to genome knowledge constitutes a major scientific breakthrough, the implications of which are still today difficult to grasp. This revolution has led to legitimate worries. Genetic technology has great potential to improve treatments and prevent human disease. On the other hand, it also has the potential to cause great harm.

The basic principles of ethics apply to all aspects of medical practice, including medical genetics. They give a guide as to what is ethically important and provide a conceptual framework for discussing ethical issues. They do not provide hard and fast rules. In genetics, as in other aspects of medicine, ethical decision-making is complex requiring sensitivity to context.

Thus, research programmes aimed at better defining the ethical consequences of studies of the human genome are greatly encouraged and supported. Moral questions, but also those related to economy and public care are brought about by the quick pace of genetic progress, and they lead to very involved societal debates. Laws and regulations on the applications of the new genetics are regularly adopted in many countries.[34,35]

As we cross the threshold of the new millennium, we simultaneously cross a threshold into an era where the human genome sequence is largely known. The genetic markers thus discovered, may not only help in identifying susceptibility genes for complex disorders but may also have a huge impact on genetic diagnostic services, helping to predict and prevent the occurrence of disease. We must commit ourselves to exploring the application of these powerful tools to the alleviation of human suffering, a mandate that, undergrids all medicine. At the same time, we must be mindful of the great potential for misunderstanding of a field that is developing very quickly, and make sure to advance the ethical consideration of genetics with just as much vigor as the medical research.

Glossary

Genetic Markers DNA sequence variations distributed widely across the genome.

Locus, a chromosomal region of DNA sequence variation

Alleles are detectable variations occurring at a single genetic locus

VNTR Variable Number Tandem Repeats (also called minisatellites).

STR Short Tandem Repeats (also called microsatellite markers).

SNP Single Nucleotide Polymorphism. The alleles have a

SNP Single Nucleotide Polymorphism. The alleles have a frequency of at least 1% in the general population.

Mutation The allele has a frequency less than 1% in the general population and leads to a change in the protein produced (functional genetic alteration).

Dimorphism is a single base variation having two alleles at corresponding chromosomal loci.

Low Penetrance a few individuals in the population carrying the allele causing the disease.

Mendel's law of independent assortment states that genes present on homologous chromosomes segregate from each other and assort independently with other segregating chromosomes during gamete formation.

'Direct' association study is carried out for diseases using possible functional polymorphisms.

Complex disorders characteristics determined by multiple genes, usually with a significant environmental component as well.

Positional Cloning the complete process of identifying disease-genes, from the initial study design to mutational analysis

Controls subjects enrolled in the study from the same population as the cases but free of disease. They constitute a comparison, rather than a 'control' group, since it is not known whether they will develop the disease in the future or will continue to remain unaffected.

Candidate Gene A gene known to be involved in the pathophysiology of the disease and thus hypothesized to be harboring the disease causing mutation.

Linkage Disequilibrium occurs when a marker allele lies in close proximity with the disease susceptibility allele, such that these two alleles are inherited together over several generations in the population.

'Indirect' association study multiple SNPs are used to search for evidence of an ancestral allele (disease-causing) that is enriched in affected individuals. They are based on linkage disequilibrium mapping.

Acknowledgements

We thank all members of the "Gene Markers and Complex Disorders" Research Program, Department of Biological and Biomedical Sciences, The Aga Khan University, whose input and support made this work possible. We also thank Mrs. Aneela Saeed, Writer-in-Residence, Research Office, The Aga Khan University, for reviewing this manuscript.

References

1. Landsteiner, K. Uber agglutination serscheinungen normalen menschlichen bluttes. Wien. Klin. Wschr 1901;14:1132-35.
2. Landsteiner, K, Levene P. Further observation on individual differences of human blood. Proc Soc Exp Biol Med 1927;24, 941-2.
3. Aird I., Bentall HH., Mehigan JA, et al. The blood groups in relation to peptic ulceration and carcinoma of colon, rectum, breast and bronchus: an association between the ABO groups and peptic ulceration. BMJ 1954;4883:315-21.
4. Aird I., BentalI H, Roberts JA. A relationship between cancer and stomach and the ABO blood groups. BMJ 1953;1:799-801.
5. Davies JL, Kawaguchi Y, Bennett ST, et al. A genome-wide search for human type 1 diabetes susceptibility genes. Nature 1994;371:130-6.
6. Le Beau MM, Overhauser J, Straub RE, et al. Report of the first international workshop on human chromosome 18 mapping. Cytogenet Cell Genet 1993;63:78-96.
7. Hodge SE, Anderson CE, Neiswanger K, et al. Close genetic linkage between diabetes mellitus and kidd blood group. Lancet 1981;2:893-5.
8. Race RR, Sanger R. Blood groups in man. 6th ed. Oxford: Blackwell, 1975.
9. Roychoudhury AL, Nei M. Human polymorphic genes. New York: Oxford University Press, 1988.
10. Bostein D, White RL, Skolnick, MH, et al. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Gennet 1980;32:314-31.
11. Gusella JF, Wexler NS, Noneally PM, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature 1983;306:234-8.
12. Weber JL, May PE. Abundant class of human DNA polymorphism which can be typed using polymerase chain reaction. Am J Hum Genet 1989;44:388-96.
13. Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle action gene. Am J Hum Genet 1989;44:397-401.
14. Taillon-Miller PZ, Gu Q, Li L, et al. Overlapping genomic sequences: a treasure trove of single nucleotide polymorphisms. Genome Res 1998;8:748-54.
15. Rosskopf D, Busch S, Manthey I, et al. Protein b3 gene. Promoter and additional polymorphism. Hypertension 2000;36:33-41.
16. Risch N, Merikangas K. The future of genetic studies of complex human disease. Science 1996;273:1516-17.
17. Siffert W, Rosskopf D, Moritz A, et al. Protein activation in immortalized lymphoblasts from essential hypertensive patients. J Clin Invest 1995;96:759-66.
18. Chakravarti A. Population genetics - making sense out of sequence. Nat Genet 1999; (Suppl 21):56-60.
19. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;264:2037-48.
20. McPherson JD, Marra M, Hillier L, et al. International Human Genone Seequencing Consortium: a physical map of human genome. Nature 2001;409:934-41.
21. Nelson DL. Positional cloning reaches maturity. Current opinion. Genet Develop 1995;5:298-303.
22. Pericak-Vance MA, Bebout JL, Gaskel PC, et al. Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage. Am J Hum Genet 1991;48:1034-50.
23. John Hopkin's University. What is genetic mapping? Epidemiology-genetics Program in Psychiatry, 2001.
24. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 1993;261:921-3.
25. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 1999;22:139-44.
26. Rigat B, Hubert T, Alhenc-Gelas F, et al. An insertion/deletion polymorphism in the angiotensin 1 converting enzyme gene accounting for half the variance of serum enzyme levels. J Clin Invest 1990;86:1343-6.
27. Cambien F, Poirier O, Lecert F, et al. Deletion polymorphism in the gene for angiotensin converting enzyme in a potent risk factor for myocardial infarction. Nature 1992;359:641-4.
28. Muller DN, Bohlender J, Hilgers KF, et al. Vascular angiotensin-coverting enzyme expression regulated local angiotensin II. Hypertension 1997;30:128-33.
29. Staessen JA, Wang JG, Ginocchio G, et al. The deletion/insertion polymorphism of angiotensin coverting enzyme gene and cardiovascular-renak risk. J Hypertens 1997;15:1579-92.
30. Keavney B, McKenzie C, Parish S, et al. Large-scale test of hypothesized

associations between the angiotensin-coverting-enzyme inserton/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. International Studies of Infarct Survival (ISIS) Collaborators. Lancet 2000;355:434-42.

31. O'Malley JP, Maslen CL, Illingworth DR. Angiotensin-converting enzyme and cardiovascular disease risk. Curr Opin Lipidol 1999;10:407-15.

32. Villard E, Tiret L, Visvikis S, et al. Identification of new polymorphisms of the angiotensin I converting enzyme (ACE) gene and study of their relationship to plasma ACE levels by two-QTL segregation linkage analysis. Am J Hum Genet 1996;58:1268-78.

33. Zhu X, Bouzekri N, Southam L, et al. Linkage and association analysis of angiotensin I converting enzyme (ACE) gene polymorphisms with ACE concentration and blood pressure. Am J Hum Genet 2001;68:1139-48.

34. Herzog AV, Frankel MS: A model ethical protocol as a guidance document for human genome diversity research. Law Hum Genome Rev 1999; 10: 21-40.

35. Dodson M, Williamson R: Indigenous peoples and the morality of the Human Genome Diversity Project. J Med Ethics 1999; 25: 204-208.