Current strategies for mapping the genes for Mendelian traits

Girisha KM

Department of Medical Genetics, Kasturba Medical College, Manipal University, Manipal-576104

Email: girish.katta@manipal.edu

Mendelian traits or disorders refer to a group of phenotypes that exhibit one of the characteristic modes of inheritance: autosomal dominant, autosomal recessive and sex linked. These are also called 'single gene disorders' or 'monogenic traits' as it is usually a single gene that has a major effect on the phenotype. Though individually rare, as a group, there are more than 7,500 disorders known to be inherited in Mendelian fashion, and probably many more traits (normal variants) and so called private syndromes (affecting single families) exist. These are expected to affect about 5% of the general population. Hence they contribute to an important subgroup of human diseases and understanding them is important for any physician.

We now know that every nucleated cell of the human body has about 22,000 genes. At least 3,125 genes have been described to cause about 5,115 Mendelian phenotypes and we can expect many more genes to be annotated soon. Table 1 gives the current list of monogenic traits and the genes characterized (http://omim.org/statistics/entry accessed on 27 December 2013). We are likely to see the discovery of genes underlying single gene disorders progress exponentially over the next few years and some genes that cause private syndromes may then be characterized more slowly. This article reviews the current strategies that enable researchers to pin point a gene involved in the causation of a disease or phenotype.

Why should we map human diseases?

Understanding the genetic basis of human Mendelian disorders first of all provides an explanation for the phenotype. Simultaneously, this also enables us to understand the function of this gene in health or in other words the pathophysiology. Often a biological pathway in which the protein product is a component gets elucidated. This not only helps in understanding the causation of the disease, but also paves way for treatment of the condition as exemplified by the use of ivacaftor in cystic fibrosis caused by the G551D mutation in the *CFTR* gene and therapy for S447X mutation-related lipoprotein lipase deficiency.[1–3] In the clinic, this is translated to diagnosis, genetic counseling, predictive testing and prenatal diagnosis. The management is better guided by knowledge of the underlying genetic mechanism and preventive strategies can then be offered for the affected families.

Traditional gene identification strategies

Most of the traditional gene characterization strategies relied heavily on Sanger sequencing. Though it still remains the gold standard, next generation sequencing techniques have eased the burden on researchers. Candidate genes can be selected by the knowledge of the function of the involved protein (or similarity to a known protein function), a strategy called functional mapping. The more widely used positional cloning is discussed in the next section. An abnormal karyotype was often an important clue to the location of a genetic defect.[4] Routine cytogenetic analysis has taken a back seat with the entry of cytogenetic microarray, though even now we often resort to karyotyping in the clinic, when affordability is an issue.

Linkage analysis and positional cloning

Genome wide linkage analysis was first proposed in 1980.[5] This is one of the earliest and yet robust ways of identification of a gene for a Mendelian trait. Positional cloning simply refers to identification of the position of the gene along the human

chromosome and then selecting the specific gene for the disease and is probably the most successful approach.[6] Linkage assumes a specific mode of inheritance that often is inferred from the families selected for analysis. Several markers spread across the human genome are then typed and recombination events then define the boundaries of the position of the gene in question. Some of the successes of this approach lead to the discovery of genes for hemochromatosis, cystic fibrosis and Duchenne muscular dystrophy.[7–10] Cystic fibrosis is the most widely accepted example of early success of this approach.[8] Some other important milestones are identification of genes for lactose intolerance, chronic granulomatous disease, neurofibromatosis I, retinoblastoma and breast cancer.[11–16] Prior to the publication of the results of the Human Genome Project in 2003, it often used to be a mammoth task to clone these large segments of the genome before the gene could be identified. The Human Genome Project is now credited with making the information of all the genes in any region of the chromosome known for such a search and has accelerated the pace of gene discovery. If characterization of the first one thousand genes took two decades, the next decade saw more than 3,000 genes being identified.

Homozygosity mapping and autozygosity mapping

Autosomal recessive disorders often are precipitated by consanguinity. Identifying the regions of homozygosity in families affected with an autosomal recessive monogenic disorder can be an approach to identify the location of the gene.[17] The data from several families can be combined to narrow down the critical region to search for the candidate genes.[18,19] A similar strategy is autozygosity mapping that focuses on regions of homozygosity by descent in a single, usually large, family.[19,20] The current techniques of next generation sequencing and SNP microarray facilitate detection of the gene even in small kindreds and sporadic cases. Homozygosity mapping is the method of choice for gene mapping in the current era as the newer technologies require only a small number of affected individuals or families for research.[6]

Cytogenetic microarray

Cytogenetic microarray (CMA) is an important tool in the evaluation of individuals with intellectual disability and multiple congenital anomalies. Many sporadic or dominant Mendelian traits are known to be associated with copy number variations. The use of cytogenetic microarray has often led to further evaluation of the locus for a causative gene as illustrated by the thrombocytopenia-absent radius (TAR) syndrome.[21,22] CHARGE, once the best known example of an association, is now redefined as a syndrome after the identification of the causative gene.[23] CMA still remains an important tool for diagnosis in the clinic and for research by providing clues to the location of genes that cause Mendelian traits.[24]

SNP microarray in gene identification

The publication of a map of single nucleotide polymorphisms across the human genome has just made linkage analysis and homozygosity mapping easier than ever.[25] This has obviated the need for use of short tandem repeats as markers. The additional advantage is that the entire genotyping is now automated. The current platforms which often combine various oligonucleotide probes with SNPs make detection of copy number variants and linkage analysis possible in one experiment thus helping the clinicians in a dual way: in both diagnosis and research.[26]

Contribution of next generation sequencing techniques for gene mapping

Whole exome and whole genome sequencing have made gene discovery quicker and less expensive and have resulted in dramatic acceleration of gene identification in the last two to three years. The successful sequencing of human exomes was first reported in 2009 and the first identification of a gene for a Mendelian trait in 2010.[27,28] The same group further identified the gene *MLL2* for Kabuki syndrome that resided beyond what was perceived as the exome then.[29] Whole exome sequencing has now emerged as both a gene discovery and a diagnostic tool.[30,31]

Several paradigms or filters can be used in combination with the next generation sequencing

strategies to maximize the yield.[32] These include analysis of linkage, homozygosity, *de novo* occurrence of mutations and candidate genes.[33–35]

Whole exome sequencing has not only added tremendous pace to the discovery of genes for Mendelian phenotypes but also the variety of ways in which it can be applied. As illustrated by several researchers, this technique can be used to identify the gene for mosaic conditions like megalencephaly-capillary malformation syndrome and Proteus syndrome.[36,37] This strategy can also identify more than one gene, often involved in the same pathway, in a cohort of patients with similar phenotypes.[38–40] A recessive gene for Charcot-Marie-Tooth disease was identified from a single family and a gene for mental retardation could be identified from sporadic cases using this technique.[41,42] It has been successfully used as a diagnostic as well as a research tool in intellectual disability and to identify the genetic basis of novel syndromic mental retardation.[43,44]

Non–traditional strategies

Often sheer brilliance in analysis of a phenotype can identify the genetic basis of disease. *TRPV4* was postulated as a candidate gene for metatropic dysplasia by Ralph Lachman and his co-workers and was tested and confirmed because of the resemblance of the radiological features of this condition with spondylometaphyseal dysplasia, Kozlowski type.[45] Based on the phenotype in mouse models, researchers have shown that the genetic basis of human diseases can be identified.[46] Earlier knowledge of pathogenesis or components of a pathway was also used in identification of the *SMC1* gene for Cornelia de Lange syndrome.[47] The known gene *NIPBL* mediates its action through sister chromatid cohesion, of which the *SMC1* gene is also a component.

Confounding factors in gene mapping

Almost all gene mapping strategies rely heavily on exact phenotyping in the clinic. The selection of patients and families is very critical for the success of linkage which assumes that the phenotype is defined accurately. A detailed pedigree should be drawn and all possible modes of inheritance should be taken into consideration. Due consideration should be given to gonadal mosaicism and occurrence of new sporadic mutations. In addition, biological variations like reduced penetrance and variable expressivity of the mutation can be important confounding factors. Often, the phenotypes need to be re-examined to verify the accuracy. Etiological heterogeneity also needs to be kept in mind as some diseases that appear to be genetic may just be multifactorial (with genetic predisposition contributing only to a fraction of the phenotype) or environmental or teratogenic in causation.

DNA banking

A repository of human phenotypes with information on pedigree and DNA from the affected and unaffected family members has proven to be a vital part of gene discovery strategies. This way, collaborators across the world can share the clinical information and biological material to put together larger numbers of families for confirmation and validation of results and establish the causation. Once the genes for common Mendelian disorders are identified, these repositories will only gain more importance to identify the genetic basis of left-over private syndromes that occur only in one or two families or individuals. It is important that ethical implications of such DNA banking are given due importance to prevent misuse of such an effort.[48]

Current national and international scenario

Several efforts are underway to map the genes for the Mendelian traits. These include Finding of Rare Disease Genes (FORGE) in Canada, International Rare Disease Research Consortium in Europe and the Centers for Mendelian Genetics in United States.[49] It appears that it may take just a few years to identify most of them, as pointed out recently (the 'Mendeliome').[50] It is not surprising that an issue of a journal often carries articles on the discovery of a gene by two independent research groups, as for Cornelia de Lange syndrome and opsismodysplasia.[51–54]

India, with its huge population and practice of inbreeding in some select regions and communities, is a rich source of genetic material for research in this area. It is also likely that some of the Mendelian diseases manifest as 'private syndromes' in one or

only a few families. Though several centers now have the equipment, we are yet to see good collaborations that can be successful in identifying many genes. Hopefully the wait is not too long!

References

1. Eckford PD, et al. *J Biol Chem* 2012; 287: 36639-49.
2. Ramsey BW, et al. *N Engl J Med* 2011; 365: 1663-72.
3. Yla-Herttuala S. *Mol Ther* 2012; 20: 1831-2.
4. Kurotaki N, et al. *Nat Genet* 2002; 30: 365-6.
5. Botstein D, et al. *Am J Hum Genet* 1980; 32: 314-31.
6. Botstein D, et al. *Nat Genet* 2003; 33 Suppl: 228-37.
7. Koenig M, et al. *Cell* 1987; 50: 509-17.
8. Kerem B, et al. *Science* 1989; 245: 1073-80.
9. Feder JN, et al. *Nat Genet* 1996; 13: 399-408.
10. Riordan JR, et al. *Science* 1989; 245: 1066-73.
11. Enattah NS, et al. *Nat Genet* 2002; 30: 233-7.
12. Fung YK, et al. *Science* 1987; 236: 1657-61.
13. Miki Y, et al. *Science* 1994; 266: 66-71.
14. Royer-Pokora B, et al. *Nature* 1986; 322: 32-8.
15. Wallace MR, et al. *Science* 1990; 249: 181-6.
16. Wooster R, et al. *Nature* 1995; 378: 789-92.
17. Lander ES, et al. *Science* 1987; 236: 1567-70.
18. Alkuraya FS. *Genet Med* 2010; 12: 236-9.
19. Alkuraya FS. *Curr Protoc Hum Genet* 2012; Chapter 6: Unit6 12.
20. Alkuraya FS. *Genet Med* 2010; 12: 765-71.
21. Klopocki E, et al. *Am J Hum Genet* 2007; 80: 232-40.
22. Albers CA, et al. *Nat Genet* 2012; 44: 435-9, S1-2.
23. Vissers LE, et al. *Nat Genet* 2004; 36: 955-7.
24. Vissers LE, et al. *Hum Mol Genet* 2005; 14 Spec No. 2: R215-23.
25. Sachidanandam R, et al. *Nature* 2001; 409: 928-33.
26. Palmer E, et al. *Am J Med Genet A* 2013.
27. Ng SB, et al. *Nat Genet* 2010; 42: 30-5.
28. Ng SB, et al. *Nature* 2009; 461: 272-6.
29. Ng SB, et al. *Nat Genet* 2010; 42: 790-3.
30. Ku CS, et al. *Ann Neurol* 2012; 71: 5-14.
31. Yang Y, et al. *N Engl J Med* 2013; 369: 1502-11.
32. Gilissen C, et al. *Eur J Hum Genet* 2012; 20: 490-7.
33. Becker J, et al. *Am J Hum Genet* 2011; 88: 362-71.
34. Veltman JA, et al. *Nat Rev Genet* 2012; 13: 565-75.
35. Pierce SB, et al. *Am J Hum Genet* 2010; 87: 282-8.
36. Lindhurst MJ, et al. *N Engl J Med* 2011; 365: 611-9.
37. McDonell LM, et al. *Nat Genet* 2013; 45: 556-62.
38. Poirier K, et al. *Nat Genet* 2013; 45: 639-47.
39. Sharma VP, et al. *Nat Genet* 2013; 45: 304-7.
40. Twigg SR, et al. *Nat Genet* 2013; 45: 308-13.
41. Lupski JR, et al. *N Engl J Med* 2010; 362: 1181-91.
42. Vissers LE, et al. *Nat Genet* 2010; 42: 1109-12.
43. de Ligt J, et al. *N Engl J Med* 2012; 367: 1921-9.
44. Schuurs-Hoeijmakers JH, et al. *Am J Hum Genet* 2012; 91: 1122-7.
45. Krakow D, et al. *Am J Hum Genet* 2009; 84: 307-15.
46. Liegel RP, et al. *Am J Hum Genet* 2013; 93: 1001-14.
47. Deardorff MA, et al. *Nature* 2012; 489: 313-7.
48. Godard B, et al. *Eur J Hum Genet* 2003; 11 Suppl 2: S88-122.
49. Bamshad MJ, et al. *Am J Med Genet A* 2012; 158A: 1523-5.
50. Alkuraya FS. *Hum Genet* 2013; 132: 1197-211.
51. Krantz ID, et al. *Nat Genet* 2004; 36: 631-5.
52. Tonkin ET, et al. *Nat Genet* 2004; 36: 636-41.
53. Below JE, et al. *Am J Hum Genet* 2013; 92: 137-43.
54. Huber C, et al. *Am J Hum Genet* 2013; 92: 144-9.