CHAPTER 1

Gene Discovery: From Positional Cloning to Genomic Cloning

WEIKUAN GU and DANIEL GOLDOWITZ

Contents

Despite the highly significant advances in studying the genetics and genomics of human populations, there are still large gaps in our understanding of the molecular genetic mechanisms involved in the pathogenesis of many human diseases. The mutated genes in many human diseases remain unknown. Identification of these mutations is crucial for correlating disease pathology and biology to the molecular basis of the disease. Discovery of new gene functions depends on the identification of the mutated genes responsible for disease in humans and other species. The techniques of positional cloning have oftentimes discovered new functions of known genes or new genes for known diseases. The goal of this book is to provide illustrations of the strategy in the post-genomic era for the identification and initial characterization of mutated genes in inherited human diseases and animal models.

1.1 CONCEPT OF CLASSIC POSITIONAL CLONING

Positional cloning, also called reverse genetics, is the identification and cloning of a specific gene, with its chromosomal location being the only available

Gene Discovery for Disease Models, First Edition. Edited by Weikuan Gu, Yongjun Wang.

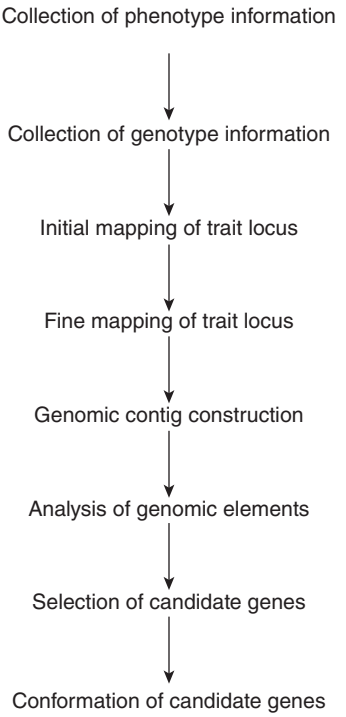

Figure 1.1. Procedure of identification of a mutated gene using strategy of classic positional cloning.

information about that gene (Collins, 1990). The identification of the X-linked gene for chronic granulomatous disease in 1986 was the first report employing such a strategy (Baehner et al., 1986; Royer-Pokora et al., 1986). For the past several decades, positional cloning has been widely used in humans, animals, and plants to isolate genes known only by their phenotypic effects. Underlying positional cloning is the assumption that a gene's location can be pinpointed with sufficient precision to narrow down its location to a DNA segment that is small enough to be sequenced and/or subjected to transformation/complementation experiments.

The classic procedure for positional cloning usually includes several steps as shown in Figure 1.1. It starts with the phenotype collection from a genetically mappable population. The population genetics necessary for creating the mappable population is beyond the scope of this chapter (Holsinger and Weir, 2009; Zou, 2009). Briefly, however, a mutant phenotype can be genetically mapped when (1) the phenotype shows Mendelian inheritance, (2) the phenotype is differentially distributed among individuals within the population, and (3) a population is large enough to reach a statistical significance when the phenotype is analyzed using mapping software. Parallel to the phenotype collection, genotype information of the same individuals in the same popula-

tion is collected. Usually, molecular markers that segregate in the population along each and every chromosome are analyzed.

The collected phenotype and genotype data from the population are used in conducting linkage analysis by one of a variety of softwares to define the chromosomal regions that the locus is likely to occupy. If a trait is controlled by a single gene or locus, the linkage analysis should point to a single chromosomal region. For traits regulated by multiple genes, multiple loci, or quantitative trait loci, multiple chromosomal regions are identified. To actually identify the gene underlying the trait of interest, fine mapping has to be conducted to narrow down the chromosomal regions so that genomic searching is practical. The next step, then, is to construct a genomic contiguous region (contig), which is defined as a set of overlapping segments of DNA, to connect and cover all the genomic elements in the targeted area. After a precise contig is constructed, it will be sequenced and analyzed by a technique termed *chromosomal walking*. This is a lengthy procedure that involves the recognition of potential genes, noncoding genes, and/or coding and noncoding regions. Finally, potential candidate genes should be confirmed using a variety of genetic and biochemical methods.

Because all of these procedures require a large amount of work, positional cloning typically requires a team effort and positional cloning projects have been known to take many years. First, the genetic region needs to be narrowed down as precisely as possible by means of initial linkage analysis and fine mapping. Second, linkage analysis requires both the availability of a large pedigree and PCR-based analysis of microsatellite markers of that pedigree to allow a whole-genome search for linkage. Fine mapping is a particularly difficult task consisting of breaking the linkage and identifying useful markers in the targeted region. Contig construction entails identification of a large insert genomic library, either BAC (bacterial artificial chromosomes) or YAC (yeast artificial chromosomes), with known markers. Analysis of genetic elements within a contig can be very difficult because of the lack of knowledge of both genes and gene organization.

However, the recent completion of the human and mouse genome projects (e.g., Mouse Genome Sequencing Consortium. 2002), along with other new technology, such as mutation analysis and microarrays, allows unprecedented progress in positional cloning of mutant genes. There are four major changes in the technique of positional cloning (Hinkes et al., 2006): (1) Contig construction is no longer needed because of the availability of whole genomes that have been sequenced. (2) Sequencing of an entire region—usually 10 Mbp of the genome, is no longer necessary, as those sequences are now readily available through public (Ensembl) and private databases (Celera). (3) Sequence analysis requires much less time and effort since annotations of whole genomes have been done (e.g., we now know that the majority of the mouse genome is made up of repetitive sequences, such as transposons, that are easy to identify and, therefore, can be eliminated from further analysis). (4) Because of the availability of whole genome sequences and high-throughput

technologies, we can now work on a much larger genomic regions, which eliminates fine mapping. (5) Annotations of genomes and bioinformatic algorithms has paralleled the rapid acquisition of genomic data and has permitted an in silico assessment of candidate genes. This is the major theme of this book. As a result of new high-throughput technologies and whole-genome libraries, a genome-based integrative strategy is the most practical method for gene discovery in our current post-genome era (Gu et al., 2002; Jiao et al., 2005a, 2005b, 2007, 2008).

Consequently, pure positional cloning in humans, animals, or plants is no longer necessary. The definition of positional cloning is cloning or identifying a gene with specific function purely according to its position. In humans, mice, and rats it is rare to localize mutations to a gene or the expression of that gene is unknown. For example, microarray technology has arrayed every gene into their chips. As a result, microarray analysis of gene expression profiles has become routine in many laboratories. Therefore, soon we may find out that expression data of every gene in every tissue is available to public. Thus, for any gene, even if nothing else is known about that gene, its expression level in a tissue can be assessed. As such, the classic positional cloning method is of little utility in the rapidly evolving arena of functional genomics. A new procedure that integrates both genomic and high-throughput technology has been created and will be, and should be, the next generation's tool of choice.

1.2 CONCEPT OF GENE DISCOVERY IN THE POST-GENOME ERA

The strategy for gene discovery using positional cloning depends on the availability of genetic-based data and technology. The new approach for gene discovery is highly integrative and is based on the availability of genome resources and biotechnology (Rintisch et al., 2008). There are three distinct and significant differences between new gene discovery strategies and classical positional cloning. The first one is the elimination of fine mapping. Rather than narrowing down the genomic regions using several approaches, a large number of genomic regions can be searched to discover the genes of interest all at once. The second is the direct investigation of genetic elements within the targeted region, without construction of contig or sequencing, because of the availability of genomic sequences and annotation of genomic elements. The third one is the high-throughput screening of candidates within the targeted region. The high-speed analytical methods include mutation screening, resequencing, and both gene expression profiling and functional predictions (Jiao et al., 2008). The following chapters provide detailed information on each of those aspects. The first part of this book introduces the technologies and resources used in gene discovery in our post-genome era. The second part of this book provides experimental procedures and methodologies for gene discovery using both genome resources and high-

throughput technologies. The third and final part of this book predicts the future direction of gene discovery based on the elucidation of genomes and developing technologies.

We are living in an era of both technology explosion and unparalleled expansion of biological resources. of the advances in gene discovery, however, are rooted in the technology of genome sequencing. Without the completion of whole genome sequences for humans and other species, gene discovery would still be stuck in the classic positional cloning approach. Therefore, gene discovery in every chapter is based on the fact that genomic sequences are available for the subjects of interest. Parallel with the necessity of completed genomes is the demand for, and rapid development of, high-throughput technologies necessary for mutation screening, genome analysis, and bioinformatics. Without these tools, there would be no effective method for capitalizing on the completion of whole genomes and for allowing our current rapid methods for gene discovery. Due to the significance of these various technologies, Chapters 2–4 introduce these technologies.

Chapters 2–6 illustrate a variety approaches, including SNP analysis, DNA methylation, protein turnover rate measurement, microarray analysis, and bioinformatic tools. Finally, the integrative analysis of data from a variety procedures provides clues for potential candidate genes for the follow-up experiments, such as RT-PCR, DNA sequencing of the potential mutation(s), and/or northern or western blot analysis to determine the significance of the mutated gene.

An important reminder to readers is that although this book mainly focuses on coding sequences known as genes, mutations in many other genetic elements could be identified using the same or similar technologies or procedures. Those none-gene elements of the genome include not only the introns, 5′ and 3′ ends of the genes, but also many others (Chen et al., 2008), such as transcription factor binding sites, microRNAs, *cis*-acting elements, palindromic motifs, and/or conserved k-tuples (phylogenetic footprints) (Hui and Bindereif, 2010). Readers should keep in mind that gene regulation is a complicated process and regulators are not necessarily near the genes that they influence. They can be located at long distances, called distant regulatory elements (REs) (Gotea and Ovcharenko, 2008), such as enhancers, repressors, and silencers. In addition, repetitive sequences sometimes play unexpected roles in gene regulations (Hui and Bindereif, 2005).

1.3 STRATEGIES FOR GENE DISCOVERY IN THE POST-GENOME ERA

Current experimental procedure strategies for mutation screening have been summarized (Jiao et al., 2008) and are shown in Figure 1.2. Individual chapters in this book focus on one or more steps or different approaches of this strategy. We briefly touch on screening for mutations in DNA in this introduction using

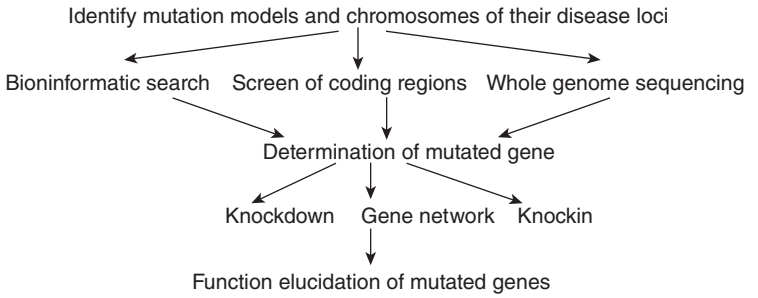

Figure 1.2. Strategy of gene discovery through mutation models.

the mouse as the model. Detailed procedure and methodologies are presented in Chapters 7–13.

The first step is to determine the total number of genes/transcripts within the targeted region. Chapter 7 describes the genetic markers and methods for determine the genomic location of target genetic loci. Any of the many recently developed software programs (see, for example, www.genediscovery.org/pgmapper/index.jsp; Xiong et al., 2008a) can be used to identify every candidate gene from a defined genomic region. The next step is to evaluate candidate genes to reduce the number of genes in the list to a more workable and feasible amount (Chapters 8–13). At this step, obvious candidate genes are first evaluated. We believe that a large number of differences exist between the gene of interest (GOI) in mutation and in wild type (control). Our current knowledge of gene function and bioinformatics should allow us to eliminate most of the unlikely candidate genes. Series of comparisons and function analyses should be made to rule out the candidacy of variation in introns sequences, if those sequences do not affect the phenotype (Chapters 11–13) At the end, a short list of candiate genes are expected or, in the best case senario, only one gene will remain. Finally, mutation evaluation or testing is carried out (Chapters 14–20). This evaluation considers differences between the GOI and control, sequence differences in these genes, potential gene function changes due to these differences, and whether other strains or populations have similar differences. Information on differences is combined with gene expression profiling and possible gene function to determine a list of candidate genes. Finally, selected candidate genes are tested and confirmed using a variety of experimental approaches, such as gene knockout and/or knockin.

1.4 FUTURE DIRECTION

Gene discovery or mutation identification has gone through two stages, as we have discussed: the classical and the post-genome era. The next stage of gene discovery will depend on development of high-throughput technology and

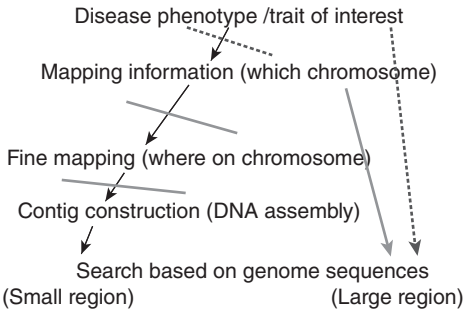

Figure 1.3. Different stages of positional cloning (from left to right): classic, post-genome era, and future (dashed blue line).

bioinformatic tools. As shown in Figure 1.3, in the first stage, positional cloning a GOI (the classical approach) has to go through every step, including initial mapping, fine mapping, contig construction, and candidate searching based on genome sequences. Currently at the second stage, in most cases, fine mapping and contig construction are not necessary because of the available information of genomic sequences and genetic elements within the targeted region.

The next stage of genomic cloning will allow researchers to conduct a search of candidate genes without mapping information (shown as dashed lines in Figure 1.3). At that stage, once a phenotype is found from an animal model or an individual, a search of candidate genes can be done based on the annotation of every gene or regulatory element in the genome. To reach the next stage, two critical improvements in our genomic research are needed. The first one is the complete evaluation of potential function of every gene and regulatory element in the whole genome. This seemingly large amount of work is most likely to be done within a decade or even sooner, as technologies for the analysis of gene function, SNP analysis, and proteomics are rapidly developing. The second is the availability of software for rapid automatic high-throughput searching. Currently, some programs such as PGmapper (Xiong et al., 2008a) has provided the capability to search genome regions of several megabases. The capability of searching whole chromosomes and whole genomes within a reasonable time (under an hour) will follow development of computational tools in coordination with genome and literature databases.

1.5 REFERENCES

Baehner RL, Kunkel LM, Monaco AP, Haines JL, Conneally PM, Palmer C, Heerema N, Orkin SH. (1986). DNA linkage analysis of X chromosome-linked chronic granulomatous disease. *Proc Natl Acad Sci U S A* 83(10):3398–401.

Chen HP, Lin A, Bloom JS, Khan AH, Park CC, Smith DJ. (2008). Screening reveals conserved and nonconserved transcriptional regulatory elements including an E3/

E4 allele-dependent APOE coding region enhancer. *Genomics* 92(5):292–300. Epub Sept. 3.

Collins FS. (1990). Identifying human disease genes by positional cloning. *Harvey Lect* 86:149–64.

Gotea V, Ovcharenko I. (2008). DiRE: identifying distant regulatory elements of co-expressed genes. *Nucleic Acids Res* 36:W133–39. Epub May 17.

Gu W, Li X, Lau KH, Edderkaoui B, Donahae LR, Rosen CJ, Beamer WG, Shultz KL, Srivastava A, Mohan S, Baylink DJ. (2002). Gene expression between a congenic strain that contains a quantitative trait locus of high bone density from CAST/EiJ and its wild-type strain C57BL/6J. *Funct Integr Genomics* 1(6):375–86.

Hinkes B, Wiggins RC, Gbadegesin R, Vlangos CN, Seelow D, Nürnberg G, Garg P, Verma R, Chaib H, Hoskins BE, Ashraf S, Becker C, Hennies HC, Goyal M, Wharram BL, Schachter AD, Mudumana S, Drummond I, Kerjaschki D, Waldherr R, Dietrich A, Ozaltin F, Bakkaloglu A, Cleper R, Basel-Vanagaite L, Pohl M, Griebel M, Tsygin AN, Soylu A, Müller D, Sorli CS, Bunney TD, Katan M, Liu J, Attanasio M, O'toole JF, Hasselbacher K, Mucha B, Otto EA, Airik R, Kispert A, Kelley GG, Smrcka AV, Gudermann T, Holzman LB, Nürnberg P, Hildebrandt F. (2006). Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nat Genet* 38(12):1397–405. Epub Nov. 5.

Holsinger KE, Weir BS. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 10(9):639–50.

Hui J, Bindereif A. (2005). Alternative pre-mRNA splicing in the human system: unexpected role of repetitive sequences as regulatory elements. *Biol Chem* 386(12): 1265–71.

Jiao Y, Li X, Beamer WG, Yan J, Tong Y, Goldowitz D, Roe B, Gu W. (2005a). Identification of a deletion causing spontaneous fracture by screening a candidate region of mouse chromosome 14. *Mammal Genome* 16(1):20–31.

Jiao Y, Yan J, Zhao Y, Donahue LR, Beamer WG, Li X, Roe BA, Ledoux MS, Gu W. (2005b). Carbonic anhydrase-related protein VIII deficiency is associated with a distinctive lifelong gait disorder in waddles mice. *Genetics* Epub Aug. 22.

Jiao Y, Yan J, Jiao F, Yang H, Donahue LR, Li X, Roe BA, Stuart J, Gu W. (2007). A single nucleotide mutation in Nppc is associated with a long bone abnormality in lbab mice. *BMC Genet* 8:16.

Jiao Y, Jin X, Yan J, Zhang C, Jiao F, Li X, Roe BA, Mount DB, Gu W. (2008). A deletion mutation in Slc12a6 is associated with neuromuscular disease in gaxp mice. *Genomics* 91(5):407–14.

Koppel I, Aid-Pavlidis T, Jaanson K, Sepp M, Palm K, Timmusk T. (2010). BAC transgenic mice reveal distal cis-regulatory elements governing BDNF gene expression. *Genesis* 48(4):214–19.

Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62.

Rintisch C, Ameri J, Olofsson P, Luthman H, Holmdahl R. (2008). Positional cloning of the Igl genes controlling rheumatoid factor production and allergic bronchitis in rats. *Proc Natl Acad Sci U S A* 105(37):14005–10. Epub Sept. 8.

Royer-Pokora B, Kunkel LM, Monaco AP, Goff SC, Newburger PE, Baehner RL, Cole FS, Curnutte JT, Orkin SH. (1986). Cloning the gene for an inherited human

disorder—chronic granulomatous disease—on the basis of its chromosomal location. *Nature* 322(6074):32–38.

Xiong Q, Qiu Y, Gu W. (2008a). PGMapper: a web-based tool linking phenotype to genes. *Bioinformatics* 24(7):1011–13. Epub Jan. 18.

Xiong Q, Jiao Y, Hasty KA, Stuart JM, Postlethwaite A, Kang AH, Gu W. (2008b). Genetic and molecular basis of QTL of rheumatoid arthritis in rat: genes and polymorphisms. *J Immunol* 181(2):859–64.

Xiong Q, Jiao Y, Hasty KA, Canale ST, Stuart JM, Beamer WG, Deng HW, Baylink D, Gu W. (2009). Quantitative trait loci, genes, and polymorphisms that regulate bone mineral density in mouse. *Genomics* 93(5):401–14.

Zou F. (2009). QTL mapping in intercross and backcross populations. *Methods Mol Biol* 573:157–73.