THE HUMAN GENOME

Hikoyuki Yamaguchi
University of Tokyo, Japan

Keywords: human genome, human chromosomes, genetic diseases, mutation, pulsed field gel electrophoresis (PFG), DNA variation, DNA polymorphisms, restriction fragment length poly-morphisms (RFLPs), DNA fingerprinting repetitive elements, minisatellite, microsatellites, tandemly repeated sequences, simple sequence repeats (SSRs), variable number of tandem repeats (VNTRs), minisatellite variant repeat (MVR), telomeric repeats, subtelomeric repeats, long-period interspersed sequences (LINES),short-period interspersed sequences (SINES), *Alu* repeat, Kpn repeat, endogenous retroviruses, disease gene mapping, body expression map, cDNA library, pharmacogenomics, genetic therapy, mitochondrial DNA, modern human origin

Contents

Summary

The human DNA packaged in 23 chromosomes contains the entire genetic code, called the “human genome.” An understanding of human DNA is certainly a key to

understanding a host of human diseases. With advances in genetic research, a more rapid and accurate diagnosis of genetic disease is now possible, and genetic methods can even be used as a form of therapy. Genomic physical mapping experiments were facilitated by the development of pulsed field gel electrophoresis and accompanying techniques. Arrays of tandemly repeated sequences predominantly locate in the centromeric chromatin and the physical ends, or telomeres, of different chromosomes. The human genome contains a large number of relatively short sequences that are related. The *Alu* sequences are most common, and can cause genetic disease in various ways. Also, the human genome contains multiple classes of endogenous retroviruses.

Revolutionary improvements in our ability to map human disease genes have come from molecular biology. Comparing expression profiles of normal and disease systems offers new insights into the molecular mechanisms of disease, identifying new diagnostic markers and potential therapeutic targets. The uniparental inheritance of mtDNA means that all the human mtDNAs in the world today must trace back to a single common ancestor who most likely lived in Africa about 150 000–200 000 years ago.

1. Introduction

The genomic material of most organisms and all cells is double-stranded (ds) DNA. Each separate piece of genomic nucleic acid, in all cases, can be termed a chromosome. Eukaryotic genomes are usually encoded by several separate, linear pieces of dsDNA, and the term “nuclear genome” signifies the genetic information sequestered as chromosomes, sometimes also exchromosomal DNA fragments, in the nucleus. The extranuclear genome of eukaryotic cells refers to the genomes of mitochondria and chloroplasts, which, like those of prokaryotes, are organized as single circular pieces of dsDNA.

The characteristics of an organism are specified by its genetic information represented as a precise nucleic acid sequence, and the sum total of this sequence information is its genome. The genome organization is necessarily both led by and related to the functional requirements placed upon it by the biological characteristics of the organism. In general, genome size (c-value) is proportional to an organism’s phenotypic complexity as dictated by the number of gene products necessary for the replication and functional maintenance of new individuals, although departures from this principle are found. A definition of the complete nucleic acid sequence of an organism’s genome is expected to identify its every gene product, their relative arrangements and organization, and the sequences required for controlled gene expression and replication, thus leading to a much greater understanding of the organism.

2. Human Genome

The human genome is very complex, and alterations causing genetic diseases can occur in various ways. With the advances in genetic research, a more rapid and accurate diagnosis of genetic disease is now possible, and genetic methods can even be used as a form of therapy. Some diseases occur by the mutation of a single gene, while others are of a more complex origin, involving multiple genetic and environmental factors.

The human genome consists of approximately 3×10^9 nucleotides and includes up to 10^5 genes. Each gene usually has a coding region that determines the sequence of amino acids in the final protein product. These genes are monocistronic, although a few can produce poly-proteins, or possess alternatively transcribed regions.

The coding region is split into exons, with intervening sequences called introns. A promoter contains DNA sequences that direct the synthesis of an RNA copy for the entire gene, and splicing signals then control the precise removal of the introns. Each end of the spliced RNA, now called mRNA, is further modified chemically before the molecule is used as a template for protein synthesis.

3. DNA Variation

Except for identical twins, no two people have exactly the same sequence of DNA. Most of the variations in DNA among individuals have no apparent effect, since less than 5% of the genome consists of protein-coding sequence. On the other hand, phenotypic differences such as hair and eye color are found, but most of these affect normal, variable characteristics. A small but important fraction of DNA variation results in a phenotype that is not considered "normal," such as a genetic disease. In DNA analysis for the diagnosis of a genetic disease, it is important to distinguish between normal DNA variations such as polymorphisms and the alteration that results in disorders such as mutation.

3.1. Mutations

Several thousand of the genes in the human genome can cause disease if disrupted in some way. A mutation in DNA may be a deletion, insertion, duplication, inversion, or other rearrangement, and the extent of a mutation may range from a single nucleotide to a whole chromosome. If the mutation occurs in the protein-coding region of a gene, the amino acid sequence may be changed, or even shortened if a premature stop codon is created. A mutation can also affect other portions of a gene, such as the promoter or processing signals. A mutation in a particular gene may cause it to produce an altered protein, no protein at all, the wrong quantity of protein, or protein at the wrong time in development.

3.2. Polymorphisms

Polymorphisms, that is, DNA variations among individuals, are usually one of two types. Single-nucleotide changes, if they happen to fall within the recognition site of a restriction endonuclease, may be detected as restriction fragment length polymorphisms (RFLPs). If single-nucleotide alterations do not occur within restriction endonuclease recognition sites, other detection methods are needed.

The other major type of DNA polymorphism is a variation in the number of tandem repeat sequences at a particular location in the genome. In general, eukaryotic genomes can contain relatively high proportions of repetitive DNA, much of which has no coding function, although the extent of such repetition is highly variable among genomes, even those of phenotypically closely related organisms. The repeated sequence may be 10–60

nucleotides long, in which case the tandem array is called a minisatellite. They are also known as variable number of tandem repeats (VNTRs). Groups of shorter repeats of di-, tri-, and tetra-nucleotides are called "microsatellites." Both mini- and microsatellites are numerous in the genomes of higher eukaryotes, and are the most important source of polymorphic markers in the population genetics and genetic mapping of diseases.

4. Physical Maps of Human Chromosomes

In humans, mapping has progressed relatively slowly for two interdependent reasons. First, for obvious ethical reasons, the human species cannot be used for experimentation in classical genetics. Observations were initially limited to studying the transmission of phenotypes and genotypes in families who had given their informed consent to participate in such studies. Second, because of this lack of experimentation, the identification of Mendelian phenotypic characteristics has been extremely difficult. A useful trait for genetic transmission studies must have at least two allelic forms. For many years such Mendelian characteristics in humans were limited to hereditary diseases and a few proteins, such as those involved in blood group polymorphisms. For these two reasons, the development of a human genetic map was a laborious task. With the application of recombinant DNA technology to human genetics in the early 1980s, however, this situation improved remarkably. Now, a number of molecular techniques have enabled the construction of physical maps for virtually any chromosome. This means that genome analysis of an uncharacterized organism can begin with a physical dissection.

The placement of genes on various physical maps has been referred to as "genetic mapping," although only physical distances and locations were known. A physical map can consist of objects located along the chromosome, and includes cytogenetic maps, chromosomal breakpoint maps, genomic restriction fragment maps, and overlapping clone libraries.

4.1. Human Chromosomes

Human chromosomes are large, complex and condensed; hence, they can be seen under the microscope. A method that allows a finer division of these large genomes is the differential staining of condensed chromosomes into regions that are estimated to be 5 to 10 Mb in size and are presumed to reflect regional differences in GC content. The division of the genome into chromosomes and chromosomal bands is convenient for top-down mapping approaches. In top-down mapping, the genome is divided into units that are convenient to study. Molecular approaches to genome analysis made it possible to characterize molecules up to about only 0.05 Mb. The resolution gap, 0.05 to 10 Mb, was exactly the size range that was most amenable to pulsed field gel electrophoresis (PFG). Thus, the entire size range of DNA molecules can now be analyzed.

The application of PFG analysis has been particularly useful for positional cloning of disease genes. First, with genetic analysis the region of the genome containing a gene of interest can be narrowed down to 1 to 10 Mb. Then PFG analysis can be used to determine the size of the region and the number of putative gene candidates it contains. This certainly facilitates the isolation of genetic markers and candidate genes. Thus,

genomic physical mapping was facilitated by the development of PFG and accompanying techniques. These techniques allow the isolation, characterization, and manipulation of large pieces of DNA, including intact chromosomes, which can range in size to at least 10 million base pairs, from virtually any organism. For instance, intact megabase (Mb) chromosomes may be analyzed directly; larger chromosomes are analyzed after cleavage into specific Mb pieces, or after cloning as Mb segments.

4.2. Human Repetitive Elements

Repetitive elements, repetitive (reiterated) sequences, and repeats are a broad variety of DNA fragments present in multiple copies in the genome. They can be divided into similar or sometimes identical elements. Although gene families with known functions, such as rRNA, tRNA, and histone genes, are included, the biological functions of repetitive elements (which may represent as much as 60% of the human genome), have not been clearly assigned.

4.2.1. Tandem Arrays of Repeats

The term “satellites” describes arrays of tandemly repeated sequences predominantly located in the centromeric chromatin of different chromosomes. They account for up to 30% of the genomic DNA. Originally, satellites were separated from the bulk DNA by centrifugation in density gradients. The separated major fractions were defined as classical satellites I, II, III, and IV. Satellite I contains a basic unit of 42 base pairs. The major simple components of classical satellites II and III are named satellites 2 and 3, both containing 5’-ATTCC as a basic repeat unit. So-called α-satellites are the other components of classical satellites II and III, and are composed of basic units, monomers, approximately 171 bp in length. The sequences of alphoid monomers differ by 20% to 40%. Diverse monomers can form multimeric units, which then repeat again. Microsatellites, also known as simple sequence repeats (SSRs), consist of tandem arrays. Unlike satellites, however, they are not localized in any particular chromosomal region, but rather are quite randomly interspersed with other genomic DNA. They are also known as variable number of tandem repeats (VNTRs). Minisatellites, which are less randomly distributed than microsatellites, tend to increase toward chromosome ends.

-
-
-

Bibliography

Brown P.O. and Hartwell L. (1998). Genomics and human diseases-variations on variation. *Nature Genetic*s **18**, 91–93. [Presents a multistrategy approach to the mapping of complex diseases and traits.]

Cantor C.R., Smith C.L., and Mathew M.K. (1988). Pulsed-field gel electrophoresis of very large DNA molecules. *Annual Review of Biophysics and Biophysical Chemistry* **17,** 287–307. [Presents a method of electrophoresis exposing nucleic acids to alternating electrical fields.]

Cavalli-Sforza L.L. (1998). The DNA revolution in population genetics. *Trends in Genetics* **14**, 60–65. [Discusses comprehensively the origin of modern humans and the history of human populations.]

Goff S.P. (1992). Genetics of retroviral integration. *Annual Review of Genetics* **26**, 527–544. [Presents a discussion the genetics of viral integration in cell and in vitro integration reactions.]

Jurka J., Walichiewicz J., and Milosavljevic A. (1992). Prototypic sequences for human repetitive DNA. *Journal of Molecular Evolution* **35**, 286–291. [Represents a collection of prototypic sequences of repetitive elements from the human genome.]

Matsubara K. and Okubo K. (1993). Identification of new genes by systematic analysis of Cdna and database construction approach. *Current Opinions of Biotechnology* **4**, 672–677. [Presents a method to map the expression of all human genes.]

Reed P.W., Davies J.L., Copeman J.B., Bennett S.T., Palmer S.M., Pritchard L.E., Gough S.C.L., Kawaguchi Y., Cordell H.J. Balfour K.M., Jenkins S.C., Powell E.E., Vignal A., and Todd J.A. (1994). Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature* **7**, 390–395. [Provides chromosome-specific sets of microsatellite marker loci suitable for use with an automated DNA fragment analyzer.]

Sheffield V.C., Stone E.M., and Carmi R. (1998). Use of isolated inbred human populations for identification of disease genes. *Trends in Genetics* **14**, 391–396. [Presents a comprehensive discussion of the genetic mapping of disease loci using patient phenotype and genotype data.]

Biographical Sketch

Hikoyuki Yamaguchi, a Professor Emeritus at the University of Tokyo, has served as Executive Director of the Shonai Regional Center for Plant Biotechnology since 1989. He is the author of several books on general genetics, radiation biology, and biotechnology. From 1970 to 1989, he was a Professor of the University of Tokyo. He became a Visiting Professor at Sao Paulo State University in 1968 and 1980, and served as Director of the Research Center of Nuclear Sciences and Technology, University of Tokyo, from 1976 to 1978. Professor Yamaguchi was a trustee of the International Center of Maize and Wheat Improvement (CIMMYT) from 1986 to 1992. He was Professor of Komazawa University from 1989 to 1999 and was appointed to chair Komazawa Junior College during 1994 and 1998.