A Brief Introduction to Genetics

Darlene R. Goldstein

Abstract

This very brief introduction to genetics is included to provide greater accessibility to the papers in this volume. More extensive details are available in genetics textbooks and the literature.

Keywords: DNA sequencing; genetic map; genome; microarray; molecular biology; physical map

1 Introduction

What follows is a very brief introduction to genetics concepts to provide greater accessibility to the papers in this volume. More extensive details are available in genetics and molecular biology textbooks, *e.g.* [2, 3, 4, 6, 8].

2 Genomes

The *genome* of an organism consists of the biological information content of a cell. This information is necessary for all cellular processes required by the organism. With the exception of some viruses, genomes are comprised of *deoxyribonucleic acid*, or *DNA*. DNA is a double-stranded, linear polymer consisting of a sugar-phosphate backbone attached to subunits called *nucleotides*. There are four nucleotides: the *purines*, *adenine (A)* and *thymine (T)*, and the *pyrimidines*, *cytosine (C)* and *guanine (G)*. Although DNA can form other tertiary structures, the best known is that of the double helix. The two strands of the double helix are held together by weak hydrogen bonds between complementary bases on the strands. Base pairing occurs as follows: A pairs with its complementary base T, and G pairs with C. The sequence complementarity provides a mechanism for DNA replication: each strand may serve as a template for sythesis of a new DNA molecule. *Ribonucleic acid*, or *RNA*, is similar to DNA but (i) contains the sugar ribose rather than deoxyribose, (ii) uses the base *uracil* (U) instead of thymine (T), and (iii) is usually single-stranded rather than double-stranded.

Genomic DNA is distributed along *chromosomes* in the cell nucleus. A *gene* is a segment of DNA that codes for a *protein*. Proteins perform a large number of diverse functions, serving as enzymes or antibodies, providing storage or transportation for other molecules, and providing structure (*e.g. collagen*). Proteins are made up of

subunits called *amino acids*; there are 20 amino acids. The set of rules associating the DNA sequence of a gene with the amino acid sequence of a protein is called the *genetic code*.

Genes occur at particular sites, or *loci* along a chromosome. A gene may exist in multiple versions, called *alleles*. The two alleles at a genetic locus comprise the *genotype* at that locus. If both alleles are the same, the genotype is *homozygous*; otherwise, the genotype is *heterozygous*. A locus may refer more generally to genetic units other than genes; for example, to sequences of DNA smaller than genes. Genetic entities such as these following normal hereditary laws are referred to as *markers*. Loci represented by more than one allelic variant in a population are said to be *polymorphic*. Examples of types of polymorphic marker systems include *restriction fragment length polymorphisms*, or *RFLP*s, and *single nucleotide polymorphisms*, or *SNP*s.

Gene expression is the process whereby the genetic information in a gene is made available to the cell. When a gene is expressed, it is said to be "turned on." Gene expression occurs in two major steps: *transcription* of the gene DNA sequence into *messenger RNA* (*mRNA*), followed by *translation* of the *mRNA* into protein. A number of intermediate steps also occur during expression, the details are omitted here.

Gene expression depends on not only allele status (genotype), but also chromosomal structure, DNA modifications, and gene-gene interactions (*epistasis*). An example of these other effects is *imprinting*, the phenomenon that genes are differently expressed depending on whether they came from the mother or father.

3 Molecular Laboratory Techniques

Advances in genetic knowledge have been made possible by innovative techniques in the molecular biology laboratory. Important techniques involve manipulations of DNA: hybridization, copying, cutting or binding, labeling and visualization.

Hybridization refers to the annealing of complementary strands of DNA. The two strands of DNA can be *denatured* (separated) by heating; upon cooling, the strands bind, restoring the original molecule.

This hybridization property of DNA can be exploited to *amplify* (copy) sequences of DNA. The process for amplifying DNA is the *polymerase chain reaction*, or *PCR*. PCR is used to amplify specific DNA sequences when the ends of the sequence are known. In PCR, the source DNA is denatured into single strands, short sequences complementary to one end of each strand are added in great excess, then the temperature is lowered so that the short sequences hybridize with their complementary sequences. The genomic DNA remains denatured, because the complementary strands are at too low a concentration to encounter each other during the period of incubation. The hybridized ends then serve as *primers* for DNA synthesis, which begins upon addition of a supply of nucleotides and a temperature resistant polymerase, an enzyme for synthesizing DNA. When the synthesis cycle is complete, there are approximately twice as many DNA molecules as there were at the start. Repeated cycles (25 – 30) of denaturing and

synthesis quickly provide many copies of the original DNA.

When a bacterium is invaded by a DNA-containing organism (*e.g.* a virus), it can defend itself with *restriction enzymes*, also called *restriction endonucleases*. Restriction enzymes recognize a specific short sequence of DNA and cut both strands at that sequence. They are used in the laboratory as "molecular scissors" for cutting large DNA molecules into smaller fragments. Restriction fragments may also be joined with the enzyme *DNA ligase*. This ability to join fragments is an important step in the creation of artificial *cloning vectors*, DNA molecules able to replicate inside a host. *Bacterial artificial chromosomes* (*BAC*s) and *yeast artificial chromosomes* (*YAC*s) are high capacity cloning vectors capable of cloning large fragments of DNA. These are used in large scale DNA sequencing projects.

Southern [9] showed that is possible to detect a specific DNA fragment within a mixed pool of fragments. Crucial to this process is DNA labeling, which enables the location and visualization of a particular DNA molecule. DNA may be labeled radioactively, then visualized by X-ray (autoradiography). Labeling for many procedures is also done with nonradioactive alternatives, most commonly with fluorescent markers.

4 Linkage and Genetic Maps

Most cells of *diploid* individuals, that is individuals whose cell nuclei contain two of each chromosome, contain a homologous pair of each *autosome*, or non-sex chromosome, and two sex chromosomes. However, gametic cells (sperm or egg) are an exception to this general rule, as they contain only one chromosome of each autosomal pair and one sex chromosome (*haploid*). Gametes are produced via a reduction divison process called *meiosis*. During meiosis, a diploid gametic precursor cell replicates DNA once and divides twice, producing four gametes.

It is also during meiosis that *crossing over* occurs. For each chromosome, after DNA replication, the two sets of chromosomal pairs (the four *chromatids*) become aligned, at which time pairs of nonidentical homologous chromosomes form regions of contact (*chiasmata*). Because physical exchange of chromosomal DNA occurs in these regions, the gametic chromosomes of an individual are generally not exact copies of the originals, but rather are combinations of the original pair.

It is one of these new combinations which is then passed on to offspring. The combination of alleles (at different loci) an offspring receives from one parent is called a *haplotype*. A *recombination* between two loci has occurred when the exchange of DNA is such that the resulting haplotype passed to an individual contains alleles at the two loci contributed by different grandparents. It is on the basis of haplotypes passed from parent to offspring that recombinations can be recognized. However, recombinations can only be distinguished from nonrecombinations when at least one parent is heterozygous at each locus.

Pairs of (gene or marker) loci on different chromosomes, or so far apart on the same chromosome that there is the same chance of recombination as nonrecombination, are

said to be *unlinked*. Two loci are *linked* when they are *not* passed on independently. When loci are in *linkage equilibrium*, the haplotype frequency is the product of the individual allele frequencies in the population; when this rule does not hold, the loci are in *linkage disequilibrium*.

The probability of recombination, or *recombination fraction*, measures the extent of linkage between loci, thereby providing a means for creating a *genetic map*. A measure of genetic distance is given by the expected number of crossovers on a single strand between two loci; the unit for this distance is the *Morgan* (*M*); distances are more commonly specified as centimorgans (*cM*; 100 *cM* = 1 *M*).

5 Physical Maps and Genome Sequencing

Genetic maps show the position of genes and other types of genetic loci in terms of genetic distance. These maps are constructed using techniques such as cross-breeding experiments and analysis of *pedigrees* (families). Prior to large scale, whole genome level sequencing, a genetic map should be supplemented by a *physical map*, constructed by direct examination of DNA molecules.

Techniques for physical mapping include: *restriction mapping*, which locates relative positions of cut sites for restriction enzymes on a DNA molecule; *fluorescent in situ hybridization* (FISH), whereby marker locations are mapped by hybridizing the marker to intact chromosomes; and *sequence tagged site* (STS) mapping, where positions of short sequences are mapped by PCR and hybridization analysis of genome fragments. Since STS is quick and not too technically demanding, it has been used for creating detailed maps of large genomes.

A single experiment is capable of directly sequencing DNA molecules with lengths of up to around 750 bp (base pairs, or nucleotides). Therefore, the sequence of an entire chromosome, which has length measured in mega-base pairs (Mb), must be constructed from smaller sequences. *Shotgun sequencing* is the standard approach used for smaller genomes. With this method, long DNA molecules are broken into fragments of sizes that can be sequenced directly. The fragments are individually sequenced, and the entire original sequence is reconstructed using computational algorithms to search for overlaps between contiguous DNA sequences (*contigs*). This approach does have some problems, though. When a genetic map is available, sequencing can proceed using variations of shotgun method: the *clone-contig* approach [7] or *directed shotgun* [10].

6 Microarray Technologies

Measuring the amounts of *mRNA* can provide information on which genes are being expressed, or used by, a cell. Microarrays provide a means to measure gene expression. Common areas currently under study with microarray experiments include: differential gene expression, that is, which genes are expressed differently between two

(or more) sample types; similar gene expression patterns (profiles) across treatments; tumor sub-class identification using gene expression profiles; classification of malignancies into known classes; and identification of genes associated with clinical outcomes, such as response to treatment or survival time. There are several microarray technologies in current use, but the two most widely used are *cDNA (complementary DNA) microarrays* and *high-density (short) oligonucleotide gene chips* produced by the company Affymetrix.

cDNA microarrays consist of thousands of individual cDNA *probe* sequences printed in a high-density array on a glass microscope slide. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples (*targets*) are reverse-transcribed into cDNA, labeled using different fluorophores ("dyes"), usually Cyanine 5 (Cy5), which fluoresces at red wavelengths, and Cyanine 3 (Cy3), which fluoresces at green wavelengths. The labeled samples are then mixed in equal proportions and hybridized with the arrayed DNA sequences. After this competitive hybridization, the slides are scanned and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of red to green fluorescence intensity for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. See [1] for a more detailed introduction to the biology and technology of cDNA microarrays and oligonucleotide chips.

Affymetrix gene chip arrays use a photolithography approach to synthesize probes directly onto a silicon chip. In addition to a number of short sequences used to probe each gene, the *perfect match* (*PM*) probes, there is an equal number of negative controls, the *mismatch* (*MM*) probes. A single labeled sample is hybridized to the array, so that absolute rather than relative measures of gene expression are obtained. Further details are available in [1, 5].

Darlene R. Goldstein, Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer, 1066 Epalinges, Switzerland; and Institut de mathematiques, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland,
darlene.goldstein@isrec.unil.ch

References

[1] *The Chipping Forecast*, volume 21, January 1999. Supplement to Nature Genetics.

[2] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson. *Molecular Biology of the Cell.* Garland Science, New York, 4th edition, 2002.

[3] T. A. Brown. *Genomes.* BIOS Scientific, Oxford, 1999.

[4] A. J. F. Griffiths, J. H. Miller, D. T. Suzuki, R. C. Lewontin, and W. M. Gelbart. *An Introduction to Genetic Analysis*. W. H. Freeman and Company, New York, 6th edition, 1996.

[5] D. J. Lockhart, H. L. Dong, M. C. Byrne, M. T. Follet tie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonu cleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[6] H. Lodish, A. Berk, L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. *Molecular Cell Biology*. W. H. Freeman, New York, 4th edition, 2000.

[7] S. G. Oliver, Q. J. van der Aart, M. L. Agostini-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J. P. Ballesta, P. Benit, and (128 others). The complete DNA sequence of yeast chromosome III. *Nature*, 357:38–46, 1992.

[8] J. Ott. *Analysis of Human Genetic Linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.

[9] E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98:503–517, 1975.

[10] J. C. Venter, M. D. Adams, G. G. Sutton, A. R. Kerlavage, H. O. Smith, and M. Hunkapiller. Shotgun sequencing of the human genome. *Science*, 280:1540–1542, 1998.