Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 1
Concept of Genomes and Genomics

Zhanjiang Liu

When searching for the basic concept of genomics, one may find numerous definitions such as:

- The study of genes and their functions
- The study of the genome
- The molecular characterization of all the genes in a species
- The comprehensive study of the genetic information of a cell or organism
- The study of the structure and function of large numbers of genes simultaneously
- etc., etc.

In order to have a good concept of genomics, let us first explore the concept of genome, and its relationship to genome expression and genome functions.

The Concept of Genome and Genomics

The term genome is used to refer to the complete genetic material of an organism. Strictly speaking, the genetic material of an organism includes the nuclear and mitochondrial genomes for plants and animals, and also chloroplast genomes for plants. Since the mitochondrial and chloroplast genomes are small and contain only a limited number of genes, the focus of genome research is on the nuclear genome. Hence, I will limit this chapter largely to the nuclear genome.

Let us define genomics in its narrowest sense using the genetic central dogma (Figure 1.1) where in most cases, deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), and RNA is translated into proteins. Although genetic information is stored in DNA, it cannot be realized without being transcribed into the intermediate molecules RNA, which with a few exceptions, must be translated into proteins in order to have biological functions. Thus, the entire DNA content of an organism is called the genome; the entire RNA world of an organism is called its transcriptome, and the entire protein content of the organism is called its proteome. The science of studying the genome is called genomics; the science of studying the transcriptome is called transcriptomics; and the science of studying the proteome is called proteomics. In spite of such divisions, the term genomics often is used to cover not only this narrow sense of genomics, but also transcriptomics, and in some cases proteomics as well.

Genomics can be divided into structural genomics, which studies the structures, organization, and evolution of genomes, and functional genomics, which studies expression and functions of the genomes. Since genome functions are reflected in the transcripts and proteins that the transcripts encode, genomics must also study the transcriptome and the proteome.

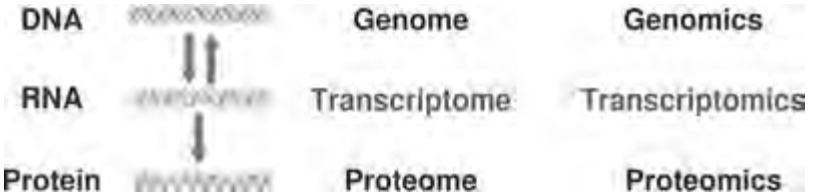

Figure 1.1. The concept of genome and genomics in relation to the genetic central dogma. The entire DNA content of an organism (the genome) is transcribed into RNA (the entire RNA content of the organism is called the transcriptome), and the RNA is translated into proteins (the proteome). Genomics, transcriptomics, and proteomics are sciences that study the genome, transcriptome, and proteome, respectively.

It must be pointed out that while the genome is relatively stable in an organism in most cell types with the exception of gene rearrangements in immune-related cell types, the transcriptome is highly dynamic. The types of transcripts and their relative levels of expression are highly regulated by tissue specificity, developmental stage, physiological state, and the environment. For instance, if an organism has 25,000 genes, not all genes are expressed in every type of cell. Those genes required for the basic cell structure and functions are probably expressed in all tissues, organs, and cell types; whereas each cell type expresses a subset of the genes specific for that cell type. Many genes are expressed throughout development, but certain genes are expressed only at a specific developmental stage. Physiological state can affect gene expression in a fundamental and dramatic way. For instance, gonadotropin genes are expressed only in the pituitary and gonad, and expressed highly during spawning seasons of the reproductive cycle in fish. The environment can insert its effect on gene expression in multiple dimensions. Temperature, pH, water quality, stress, dissolved oxygen, and many other environmental factors can induce or suppress expression of a large number of genes.

In addition to the dynamic nature of the transcriptome, variation of the transcriptome can also be brought about by production of alternative transcripts by the same set of genes. It is now widely believed that the complexity of the transcriptome is much larger than the genome because of alternative transcripts. The largest proportion of alternative transcripts is produced by alternative splicing where a single gene is transcribed into heterogeneous nuclear mRNA (hnRNA); through splicing, more than one mRNA molecule is produced, leading to the phenomenon that introns of one transcript may be exons of another. The second mechanism for the generation of alternative transcripts is through the use of alternative promoters. In a single gene, more than one promoter can be functional leading to the generation of different, but related transcripts. In addition, use of differential polyadenylation sites can also lead to the generation of alternative transcripts. Therefore, it is widely believed that the information stored in the genome is amplified and diversified at the transcriptome level. The genetic information is further amplified and diversified at the protein level. Though each transcript may only encode one protein, the primary protein may be differentially processed to produce more than one active polypeptide; posttranslational glycosylation, acetylation, phosphorylation, and other modifications can result in a much larger complexity leading to drastically different biological functions. Even highly related gene products may encode proteins leading to absolutely opposite biological functions. For instance, an interleukin-1 Type II receptor is a decoy target for

IL-1, whose binding to interleukin 1 intercepts the function of interleukin 1. Therefore, the genetic central dogma is correct in terms of the basic flow of genetic information, and the capacities of the primary functions of transcription and translation, while much larger complexities result from amplification and diversification of the same set of genetic material, lead to the generation of biologically different molecules. Such differences in biological molecules, when considered for the various combinations of many genes, can result in numerous biological outcomes.

As a new branch of science, genomics has its own defined scope of study, its own box of tool kits, and its own unique set of approaches. It is different from traditional molecular genetics which looks at single genes, one or a few genes at a time. Genomics is trying to look at all of the genes as a dynamic system, over time, to determine how they interact and influence biological pathways, networks, physiology, and systems in a global sense. Genome technologies, the focus of this book, have been developed to cope with the global scope of tens of thousands of genes as a snapshot. Much like dealing with a globe, landmarks (or as we have called them molecular markers) are needed to mark the position within the huge genome. Genetic and physical maps have been developed to understand the structure and organization of genomes, and to understand genomic environs and genome evolution in relation to genome expression and function. Specific approaches have been developed to cope with the large number of genes, regardless if it is for gene discovery, cloning and characterization, or for analysis of gene expression. Thus large-scale analysis of expressed sequence tags using highly normalized complementary DNA (cDNA) libraries allows rapid gene discovery and cloning in the scale of tens of thousand of genes. Such operations have also been supported by other genome technologies such as powerful automated sequencing to allow gene discovery and identification in a streamlined industrial fashion. Expression of genes is determined in an entire genome scale, or sometimes referred to as genome expression, to relate complex regulation of genes to their functions in terms of systems biology. Expression of tens of thousands of genes can be monitored simultaneously and continuously, allowing their interactions and networking to be detected. Signal transduction is no longer "behind the scene" molecular events, but can be observed with clustering of co-regulated gene expression under specific development, physiology, or environmental conditions. Genes and their functions are studied much in terms of their sociology, networking, and interactions, rather than looking at one or a few genes at a time, as conducted by traditional molecular biology. Such operations demand the development of very powerful gene expression analysis such as microarray technologies. Such technological advances allow the generation of tremendously large data sets that have been beyond the comprehension capacities of biologists. Assistance is needed from all areas of biology, and more so from disciplines outside biology that can handle large amounts of information. Computer sciences and mathematics are among the first disciplines genomics has demanded cooperation from. While handling large data sets from the genome, genome expression, and genome function, much confusion has emerged regarding whether the observed phenomenon is real or if it is just a fluctuation of the systems biology. As such, statisticians are also called upon to join computer scientists, mathematicians, and the biologists. Because these scientists speak different languages (e.g., English for one group, French for the second, Chinese for the next, and so on), understanding all of the languages and being able to function among these different disciplines is becoming the goal of a large group of scientists who define themselves as bioinformaticians working in the new area of bioinformatics. It is clear that genomics cannot be a science without

bioinformatics. Clearly, the definition of genomics is becoming more complex with this discussion. Now, you can certainly come up with your own definitions.

The excitement and success of genomics has brought the emergence of numerous '–omics' sciences (http://genomicglossaries.com/content/genomics_glossary.asp). Subbranches of genomics are emerging in large numbers. The following list includes some of those subbranches:

- agricultural genomics
- applied genomics
- behavior genomics
- biochemical genomics
- chemogenomics
- clinical genomics
- combinatorial genomics
- comparative genomics
- computational genomics
- deductive genomics
- ecotoxicogenomics
- environmental genomics
- evolutionary genomics
- forward genomics
- functional genomics
- immunogenomics
- industrial genomics
- intergenomics
- inverse genomics
- lateral genomics
- nanogenomics
- network genomics
- oncogenomics
- pharmacogenomics
- phylogenomics
- physiological genomics
- population genomics
- predictive genomics
- reverse genomics
- structural genomics
- toxicogenomics
- translational genomics
- and so on

Cells, Nucleus, Chromosomes, Genomes, and Genomic DNA

Genomes can exist in various forms. A genome can be either RNA or DNA, single-stranded or double-stranded. For example, the human immunodeficiency virus (HIV) is a retrovirus whose genome contains a single-stranded RNA molecule. However, such unusual genomes are mostly found within viruses and bacteriophages. In prokaryotes such as bacteria, by definition they do not have a nucleus; the genomes are made up with double-stranded DNA in either circular or linear forms. For instance, the *Escherichia coli* genome is made of a single circular DNA molecule, whereas the genome of *Borrelia burgdorferi* is composed of a linear chromosome approximately one megabase (million base) in size. Eukaryotic genomes contain two or more linear molecules of double-stranded DNA in the form of chromosomes.

Within each eukaryotic cell, there is a nucleus in which chromosomes are located. Individual species harbor a fixed number of chromosome pairs (2n) with fixed shapes, sizes, and centromere location. These chromosome morphologies are commonly known as the karyotypes. All somatic cells in a diploid organism harbor identical chromosome pairs that are randomly shared into a single chromosome set during meiosis to produce eggs and sperms. Upon fertilization of an egg (n) by a sperm (n), the embryo recovers the diploid state with two sets of chromosomes.

Chromosomes are threadlike structures containing genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes.

Humans have 46 chromosomes—44 autosomes and 2 sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers. This is important in sexual reproduction where the gametes (i.e., sperms and eggs) are haploid cells, and upon fertilization of an egg by a sperm, the embryo recovers the diploid state. The number of chromosomes is usually constant for each organism, but may vary greatly from species to species. For instance, the fruit fly *Drosophila melanogaster* has four chromosomes whereas *Ophioglossum reticulatum*, a species of fern, has the largest number of chromosomes with more than 1,260 (630 pairs). The minimum number of chromosomes found in a species occurs in a species of ant, *Myrmecia pilosula*, in which females have one pair of chromosomes and males have just a single chromosome. This species reproduces through a process called haplodiploidy, in which fertilized eggs (diploid) become females, while unfertilized eggs (haploid) develop into males.

Each chromosome is a portion of the genome and all the chromosomes compose the entire genome. Although all chromosomes maintain their own integrity, they each can be viewed as a segment of the genome. The total length of genomic DNA thus is equal to the sum of all chromosomal DNA. In their natural existence, the physical pieces of DNA in each cell are equal to the number of chromosomes. It must be emphasized that such entire chromosomal DNA is essentially impossible to obtain for routine molecular analysis. Chromosomal DNA is randomly broken during genomic DNA extraction even under the most sophisticated preparation by the most skilled researchers. Most often, millions of cells are used in a single DNA extraction. Therefore, genomic DNA used in molecular analysis represents multiple copies of the genome with multiple overlapping segments, simply because the breakage points are random and different in each cell genome.

Genome Sizes

Genome sizes of organisms vary greatly, spanning a range of almost 100,000 fold. The bacterial genomes are commonly at the range of a million base pairs (Mbp), while the largest animal genome reported to date is 133 picograms (pg) (or about 1.3×10^{11} base pairs; 1 pg DNA = 1×10^{-12} g = 978 Mbps) for a species of lungfish, *Protopterus aethiopicus*, which is some 40 times larger than the human genome, followed by a number of amphibians, *Necturus lewisi* and *N. punctatus* at 120 pg, *Necturus maculosus* and *Amphiuma means* and the lungfish *Lepidosiren paradoxa*, all at roughly 80 pg. In general, the genome size is correlated with biological complexities, but many exceptions exist. For instance, some plant species and amphibians can have very large genomes, dozens of times larger than the human genome.

The largest teleost genome size is 4.4 pg in the masked *Corydoras metae*, and the smallest teleost genome size is approximately 0.4 pg in several puffer fish of the family *Tetraodontidae*. Fish as a whole have the largest ranges for genome sizes.

Crustaceans also have a wide range of genome sizes from 0.16 pg to 38 pg with an average of 3.15 pg. The smallest crustacean genome size (0.16 pg) is in a water flea, *Scapholeberis kingii*, and the largest crustacean genome size (38 pg) is in *Hymenodora* sp., a deep-sea shrimp. The most important crustacean species for aquaculture involves several major species of the shrimps. Their genome sizes are approximately 2.5 pg.

The molluscan genome sizes are more uniform ranging from the smallest molluscan genome size of 0.4 pg in the owl limpet *Lottia gigantean*, to the largest molluscan genome size of 5.9 pg in the Antarctic whelk *Neobuccinum eatoni*. Many aquacultured shellfish belong to the molluscans. The most important of these species in aquaculture include the oysters, such as the Pacific oyster (genome size 0.91 pg), the eastern oyster (genome size 0.69 pg), and the scallops (genome size between 0.95 to 2.1 pg).

The size of the genome of an organism is a constant. However, the ploidy of organisms varies. For instance, channel catfish are believed to be a diploid organism, whereas most salmonid fish used in aquaculture are believed to be tetraploid. In cultivated wheat plants, various ploidies exist including diploid, tetraploid, and hexaploid. In order to standardize the genome size so that they can be compared, genome sizes are presented in C-values, which is the haploid genome size in picograms.

Several excellent databases exist for genome sizes. The Animal Genome Database (http://genomesize.com/) is a comprehensive catalogue of animal genome size data. It includes haploid genome sizes for more than 4,000 species including approximately 2,750 vertebrates and 1,315 invertebrates compiled from 5,400 records from more than 425 published sources (Gregory 2005; Animal Genome Size Database, http://www.genomesize.com/). The Database Of Genome Sizes (DOGS) (http://www.cbs.dtu.dk/databases/DOGS/) is also a very useful database that includes a number of links to genome size and genome research resources such as the following:

- the Plant DNA C-Value Database (http://www.rbgkew.org.uk/cval/)
- the Genome Atlases for Sequenced Genomes (http://www.cbs.dtu.dk/services/GenomeAtlas/)
- the DBA mammalian genome size database (http://www.unipv.it/webbio/dbagsdb.htm)
- several other useful databases and resources.

Knowledge of genome size is not only important for genome studies in relation to genome structure, organization, and evolution, but also for a number of practical reasons such as genome mapping, physical mapping, and genome sequencing. As listed in Table 1.1, the primary methods for the determination of genome sizes are Feulgen densitometry (Hardie et al. 2002), flow cytometry, and Feulgen image analysis densitometry (Lamatsch et al. 2000). These three methods account for over 81% of all methods used for the estimation of genome sizes (Table 1.1). Readers with an interest in methodologies for the determination of genome size are referred to the literature list of the Animal Genome Size Database (http://www.genomesize.com/).

Number of Genes

The number of genes in a given organism is fixed, but discovering it is a daunting task. For the best characterized human genome, the number of genes now is believed to be approximately 25,000. In the 1980s, the number of human genes was believed to be 100,000 to 125,000. In the early 1990s, the human genome was believed to include 80,000 genes. Although the final completion of the Human Genome Project was celebrated in April 2003 and sequencing of the human chromosomes is essentially "finished," the exact number of genes encoded by the genome is still unknown. In

Table 1.1. Methods and their frequencies used for the determination of genome sizes. The table was adapted from the Animal Genome Size Database (http://www.genomesize.com/).

Methods	Abbreviation	Number of genomes	Percentage of methods used for genome size determination
Feulgen Densitometry	(FD)	2,480	45.93%
Flow Cytometry	(FCM)	1,075	19.91%
Feulgen Image Analysis Densitometry	(FIA)	839	15.54%
Bulk Fluorometric Assay	(BFA)	471	8.72%
Static Cell Fluorometry	(SCF)	303	5.61%
Biochemical Analysis	(BCA)	142	2.63%
Not Specified	(NS)	63	1.17%
Ultraviolet Microscopy	(UVM)	13	0.24%
Gallocyanin Chrom Alum Densitometry	(GCD)	11	0.20%
Complete Genome Sequencing	(CS)	2	0.04%
Methyl Green Densitometry	(MGD)	1	0.02%

2000 when the human genome project was originally declared as being completed, the human genome was believed to contain 35,000 to 40,000 genes. Now in 2006, the total number of human genes is believed to be around 25,000. Clearly, many of the "gene-like" reading frames were proved not to be genes.

It could still take years before a truly reliable gene count can be assessed. The uncertainty is derived from different methods used for the assessment of genes. Some prediction programs detect genes by looking for distinct patterns that define where a gene begins and ends. Other programs look for genes by comparing segments of sequence with those of known genes and proteins. The first tends to overestimate, while the second tends to underestimate, the gene count. No matter which programs are used, the bottom line is that evidence to support a gene model has to come from expression information. In spite of some 7 million Expressed Sequence Tags (EST) obtained from humans, they cannot support all of the gene models yet because many gene products have not been found. Although the ballpark range of the number of human genes should not change dramatically, finer tuning for the total number of genes is still expected.

The number of genes an organism has is correlated with the biological complexity of the organism. With this belief, the number of human genes came as a shock to many scientists because even the *E. coli* has 4,377 genes with 4,290 protein encoding genes. Saying that we are only six times more complex than a bacteria is truly a humiliation to many, but it is probably worse to say that the human gene count is only one-third greater than that of the simple roundworm *C. elegans* which has about 20,000 genes (Claverie 2001). Nonetheless, the unique number of gene products (proteins) is likely correlated with biological complexities, though the absolute number of genes may vary depending on the level of gene duplications. With such assumptions, it is reasonable to believe that many fish genomes will have a similar number of unique

genes as the human genome, but their total number of genes could even be slightly larger, considering high levels of gene duplications in teleosts.

A basic understanding of the genome, genome size, the number of chromosomes, and the number of genes is important before the start of a genome project. Not only the efforts required to characterize the genome are affected by the genome size and complexity, but also proper methodologies should be taken according to the circumstances as well.

References

Claverie JM. 2001. Gene number. What if there are only 30,000 human genes? Science, 291, pp. 1255–1257.

Gregory TR. 2005. Genome size evolution in animals. In: *The Evolution of the Genome*, edited by TR Gregory. Elsevier, San Diego, CA, pp. 3–87.

Hardie DC, R Gregory, and PDN Hebert. 2002. From pixels to picograms: a beginners' guide to genome quantification by feulgen image analysis densitometry. J Histochem Cytochem, 50, pp. 735–749.

Lamatsch DK, C Steinlein, M Schmid, and M Schartl. 2000. Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: detection of triploid *Poecilia formosa*. Cytometry, 39, pp. 91–95.