Genome Biology: The *Second* Modern Synthesis

Jun Yu and Gane Ka-Shu Wong
Beijing Institute of Genomics, Chinese Academy of Sciences

A scientific dream proposed some 20 years ago has been realized – the completion of the DNA sequence for the Human Genome Project (HGP) in 2004. As a result, an entirely new field of biological research has arisen: genome biology or genomics is celebrated for its unprecedented scale, intrinsically digital output, and systematic approach to getting all the data. Its sequel, the HapMap Project, will reach fruition later this year. These projects established new precedents for international collaborations and open data access. Chinese scientists contributed to 1% of the HGP and 10% of the HapMap. They also initiated and completed several projects of their own, including the Chinese Superhybrid Rice Genome Project, the Silkworm Genome Project, the Chicken Genome Polymorphism Project, and a Genome Survey of the Porcine Genome. These projects will benefit fields as diverse as agriculture, medicine, and the economy in general.

Biologists have celebrated every new genome that has been sequenced and deposited into the public databases. With each new data set, the clamor for more data grows. This is possible because of a relentless focus on technology development. DNA sequencing costs have decreased from $3 billion USD, some 20 years ago, to $30 million USD today, for a typical human-sized genome. Indeed, one of the stated goals of the HGP was to provide a compelling vision to motivate this technology development. The most recent vision is the Human Cancer Genome Project, which hopes to reduce these costs to $100,000 USD, and eventually $1,000 USD. Follow up activities made possible by having these copious data are popularly called “omics” (*e.g.* proteomics and metabolomics). Despite the excitement and obvious practical benefit of having so much data, it has been asked if there are deeper fundamental questions that can be answered. We address that issue here.

Consider the information that has been acquired. For any given species, this includes but is not limited to: (1) a complete DNA sequence; (2) list of genes and encoded proteins tentatively annotated as regards to their cellular components (*e.g.* membrane and cytosol), molecular functions (*e.g.* enzyme and transcription regulator), and biological processes (*e.g.* cell cycle and lipid metabolism); (3) RNA and protein expression data, sampled over different tissues and cell types, under different physiological, pathological, and ecological conditions; (4) deduced regulatory motifs; and (5) a compendium of the genetic variation observed in selected populations from that species.

Information is not knowledge. The annotations must be carefully validated by follow up experiments. Determining the function of a gene is never easy. Most knockouts exhibit no obvious phenotypes. Having many polymorphisms does not indicate which ones might be responsible for susceptibility to a specific disease. This is a widely misunderstood fact

of genomics. It is about building infrastructure for further investigations, not the ultimate end point of biological sciences. It does not close the door on any existing field of inquiry, but it does free the investigators from the more mundane job of gathering this information and lets them to focus on the deeper questions. Chief among these is the question of how subtle changes in an organism's DNA sequence leads to the diversity of observed species, with vast differences in behavior and response to environmental cues.

Every organism raises different questions. For example, in the area of plant genomics, one can address several questions. One is the evolution of new genes through duplication of individual genes, chromosomal segments, and even entire genomes. Signatures of past duplication events can be discerned in the sequence of the rice genome, with implications for understanding differences between cereals (*e.g.* rice and maize). Indeed, polyploidy is a major factor in agriculture, as over 80% of crop species are polyploid, some ancient and others recent. Another major factor is hybrid vigor or heterosis, the phenomenon whereby the progeny of a breeding experiment is often more fit than either progenitor. This too has been used, to immense effect, in agriculture. How it arises, and why it is so universal, is a question of long-term interest to plant geneticists. Third, the process of domestication has resulted in dramatic changes in the crops that we grow, vis-à-vis their wild ancestors. One of the signatures of domestication is the lack of polymorphism among domestic cultivars, in regions under selective sweep for adaptive traits of domestication. Genes so discovered are good candidates for further crop improvement.

From the very beginning, human biology has been the driving force behind genomics, even though it was not the first genome to be sequenced. With over a thousand Mendelian disease genes identified along the way, it is clear that the next challenge in genomics will be in studying the genetic basis of the common diseases that involve a complex interplay of multiple genetic and environmental factors. This was in fact the motivation behind the HapMap Project, and more recently, the Human Cancer Genome Project, which hopes to identify targets for drug intervention and biomarkers for early detection. A new discipline called "systems biology" has arisen face up to the challenge of biological complexity, but even the proponents cannot agree on what their ultimate goals are. We believe the history of biology in the first half of the 20th century can offer some guidance.

Darwin in his theory of speciation said nothing of Mendelian genetics, and Mendel in his theory of inheritance said nothing of Darwinian selection. It took the combined efforts of Thomas Hunt Morgan, Ronald Fisher, Theodosius Dobzhansky, J.B.S. Haldane, Sewall Wright, Julian Huxley, Ernst Mayr, and others to link the concepts of Darwin and Mendel into a unified theory of evolution now known as the *Modern Synthesis*. Genetic variation in a population arises by chance, through mutation and recombination. Evolution consists primarily of changes in the frequencies of alleles for the population, as a result of genetic drift, gene flow, and natural selection. Speciation occurs when populations are isolated by

geographic barriers. Everything is formulated around the concept of a population, in stark contrast to molecular studies on one individual.

Classical concepts in genetics like epistasis and pleiotropy are formulated as a sum of variances in a population. On one side of the equation is V_P (phenotype). On the other are V_G (genotype), V_E (environment), V_{GxG} (gene-gene interaction), V_{GxE} (gene-environment interaction), and V_N (noise). Systems biology is more focused on the interactions between the molecules of the cell (*e.g.* protein-protein interactions and RNA-binding proteins) and how they interact with the environment. It is not clear how these views will be reconciled. For example, epistasis can arise from multiple levels of protein-protein interaction, and it is not clear how many levels of interaction must be considered. It is however clear that, at some level, every protein interacts with every other protein. Perhaps even more assiduous, population considerations must still be factored into the molecular data.

The days of simple relationships between genotype and phenotype are over. Much as biologists in the first half of the 20th century managed to reconcile the concepts of Darwin and Mendel, today's biologists will need to reconcile the molecular biology of individuals with the classical concepts of genetics that are based on variance in a population, leading to a *Second* Modern Synthesis. So, although the proverbial ink on the recently completed HGP is not yet dry, more data will be needed to understand what we already have. This is the heart of genomics, not DNA sequencing *per se*, but the ability to acquire, analyze and eventually comprehend massive amounts of data.

By learning how to convert genotype to phenotype, in the presence of environmental cues, we can finally begin to understand the diverse and complex phenomena that are the essence of biology. We are particularly intrigued by phenotypic plasticity, where radically different phenotypes are activated in response to environmental cues. The transformation from a worker to queen bee is a spectacular example of phenotypic plasticity. Queen bees are larger and live much longer. They fly and lay eggs. Their sisters, worker bees with the same genome, cannot. Understand that, and genomics will have arrived.