Ravindranath Duggirala · Laura Almasy
Sarah Williams-Blangero · Solomon F.D. Paul
Chittaranjan Kole *Editors*

Genome Mapping and Genomics in Human and Non-Human Primates

Springer

Genome Mapping and Genomics in Animals

Volume 5

Series editor
Chittaranjan Kole, Mohanpur, India

For further volumes:
http://www.springer.com/series/7518

Ravindranath Duggirala
Laura Almasy · Sarah Williams-Blangero
Solomon F.D. Paul · Chittaranjan Kole
Editors

Genome Mapping and Genomics in Human and Non-Human Primates

Springer

Editors
Ravindranath Duggirala
South Texas Diabetes and Obesity Institute
University of Texas Health Science Center at San Antonio
Edinburg, TX
USA

Laura Almasy
South Texas Diabetes and Obesity Institute
University of Texas Health Science Center at San Antonio
San Antonio, TX
USA

Sarah Williams-Blangero
South Texas Diabetes and Obesity Institute
Brownsville, TX
USA

Solomon F.D. Paul
Faculty of Biomedical Sciences, Technology and Research
Sri Ramachandra University
Chennai
India

Chittaranjan Kole
Bidhan Chandra Krishi Viswavidyalaya
Mohanpur, West Bengal
India

Genome Mapping and Genomics in Animals
ISBN 978-3-662-46305-5 ISBN 978-3-662-46306-2 (eBook)
DOI 10.1007/978-3-662-46306-2

Library of Congress Control Number: 2015932067

Springer Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer-Verlag GmbH Berlin Heidelberg is part of Springer Science+Business Media (www.springer.com)

Preface

In recent years, there has been phenomenal progress in the understanding of the genetic architecture of normal and disease-related complex phenotypes. The progress has been fueled by an explosion of research activities related to the Human Genome Project and subsequent sequencing projects, and the nonhuman primate comprehensive sequencing projects. Advances in molecular genetics, statistical genetics, medical genetics, and bioinformatics have accompanied this progress.

Tracing its roots back to the laws of inheritance established by Mendel, which continue to be the basic tenets underlying modern genetics, the field of genetics has expanded tremendously and has richly diversified over the years. Gene mapping efforts and genomic research on humans and nonhuman primates have generated an enormous amount of information relevant for studies of evolution, phylogenetics, human genetics, anthropological genetics, and for biomedical research. Elucidation of gene function, expression, and regulation and of genetic variation and conservation among primate species has exciting potential for informing research in the areas of biology, evolution, population genetics, anthropological genetics, and biomedicine. The huge increase in the amount of available genomic information and advances in the tools available to analyze that data have already had a tremendous impact on disciplines such as evolutionary biology, bioinformatics, genetic epidemiology, medicine, pharmacogenetics, pharmacogenomics, and anthropology.

This volume is an attempt to provide researchers and academicians with a review of advanced methodologies and applications in gene mapping and genomics of humans and nonhuman primates, with an emphasis on genetics of complex phenotypes and diseases. As a part of the "Genome Mapping and Genomics in Animals" series (Dr. C. Kole, Editor), this volume is designed to illustrate ongoing research activities related to gene mapping and genomics in human and nonhuman primates. The topic of this volume is broad and a full coverage of such a huge area of research would be impossible. Therefore, we limited the volume to 16 chapters that illustrate the amazing changes in genomic studies that have occurred since the Human Genome Project. From the initiation and expansion of the Human Genome Project to revolutionary next generation sequencing approaches, we have seen dramatic improvement in the understanding of the genetic architecture of complex phenotypes in human and nonhuman primates.

This volume constitutes an overview of the impact of the genomic revolution on research related to human and nonhuman primate populations. It also reviews the state-of-the-science with respect to the molecular, statistical genetics, and genetic epidemiologic techniques that are used to dissect the genetic architecture of normal and disease-related complex phenotypes using data from human and nonhuman primates. We present examples of successful applications of genomic methods to traits of particular interest in biomedical research and evolutionary biology, and provide discussions of future directions in human and nonhuman primate genomics.

Since genetic investigation of complex phenotypes is by nature multidisciplinary, efforts were made to provide readers with review papers which illustrate the full range of methodological and analytical approaches being applied to human and nonhuman primate population data sets. Examples and applications were drawn from diverse areas including evolutionary genetics, population structure, genetic epidemiology, transcriptomics, copy number variation, molecular ecology, comparative genomics, and gene mapping for phenotypes related to behavior, skeletal biology, and cardio-metabolic disease in human and nonhuman primate populations.

We are in the midst of an exciting scientific era with constantly changing technology revolutionizing genomic research approaches over and over again. The advances will ensure continued interest in explorations of genomics and other "omics" approaches as they relate to normal variation and disease-related traits in human and nonhuman primate populations. Progress in gene mapping and genomic sequencing will add further momentum to progress in comparative genomics, evolutionary genomics, and biomedical research as it corresponds to disease prevention and treatment, pharmacogenomics, and personalized medicine.

We are grateful to the contributors to this volume who have prepared comprehensive and informative reviews of advanced, complex genomics-related topics. We thank Drs. Vidya S. Farook, Sobha Puppala, Geetha Chittoor, and Laura Cox for reviewing one or more chapters of this volume. The editors also express their gratitude to Ms. Maria Messenger whose expert skills in proofreading and formatting greatly improved the quality of this volume.

Ravindranath Duggirala
Laura Almasy
Sarah Williams-Blangero
Solomon F.D. Paul
Chittaranjan Kole

Contents

The Utility of Genomics for Studying Primate Biology

1

Sarah Williams-Blangero and John Blangero

1.1 Genomics of Primate Populations Writ Large

This volume was organized with the intent to review progress in primate genomics and, in particular, to show the value of studying primate genomics for understanding the determinants of risk for disease in human populations. Humans, our hominid ancestors, and nonhuman primates (and their ancestors) share most of their genetic material. The evolutionary proximity of nonhuman primates to humans provides us with a particularly valuable set of tools to make inferences about the causes of human phenotypic variation using experimental techniques applied to our close animal relatives. There is a remarkable amount of anatomical and physiological similarity across all primates that justifies the use of nonhuman primate models rather than more phylogenetically remote animal models, such as the mouse, for many types of studies. However, the use of nonhuman primates for modeling the basis of human phenotypic variation is associated with considerable costs due to the comparatively large size of the animals, their relatively long generation times, and the general expense of working with nonhuman primates.

The utility of considering primate biology writ large as a major source for inference about human biology stems from the close genetic relationship between humans and nonhuman primates. With the advent of large-scale genome sequencing, we now know precisely the extent of genetic similarity among the phylogenetically most proximate relatives.

The chimpanzee genome was the first nonhuman primate genome to be sequenced and was completed in 2005 (Chimpanzee Sequencing and Analysis Consortium 2005). From these data, we know that chimpanzees and humans diverged about 6 M years ago and share ~98 % sequence identity. About 29 % of orthologous proteins are identical between human and chimpanzees with most proteins differing by an average of only two amino acids (Chimpanzee Sequencing and Analysis Consortium 2005). This protein similarity greatly facilitates cross-inference of biological mechanism between the humans and chimpanzees species.

Even in the presence of substantial protein similarity, the genetic differences between the primate species clearly lead to striking phenotypic divergence. For example, comparative quantitative proteomic and metabolomics studies using chimpanzee and human biomaterials are finding fascinating and unexpected differences that may be of utility for understanding the substantial differences in brain and muscle

S. Williams-Blangero (✉) · J. Blangero
South Texas Diabetes and Obesity Institute, Regional Academic Health Center, University of Texas Health Science Center at San Antonio, 2102 Treasure Hills Blvd, Harlingen, TX 78550, USA
e-mail: WilliamsBlan@uthscsa.edu

J. Blangero
e-mail: blangero@uthscsa.edu

R. Duggirala et al. (eds.), *Genome Mapping and Genomics in Human and Non-Human Primates*, Genome Mapping and Genomics in Animals 5, DOI 10.1007/978-3-662-46306-2_1,

function between the species (Bozek et al. 2014). Similarly, advanced neurophenotyping methods have revealed profound differences in synaptic phenotypes such as synaptic density between the humans and chimpanzees that appears to correlate with brain function (Liu et al. 2012). For many reasons, including smaller numbers of available colony-managed animals, greater expense, and extreme regulatory burden, chimpanzees are now little used in biomedical research despite being the most potentially useful of all primate species for making inferences about human health.

The rhesus macaque is the most widely used nonhuman primate model for human biology and its genomic sequence was first obtained in 2007 (Rhesus Monkey Sequencing and Analysis Consortium 2007). The sequence data shows that humans and rhesus monkeys have ~93 % total sequence identity, the reduction over that with chimpanzees correlating with the earlier divergence time of about 25 M years ago.

The baboon also is well utilized in biological research designed to be informative for human health (as evidenced in the chapters by Cox and Sherwood and Duren in this volume), as is the vervet (with examples provided by Bailey et al. in this volume). The genomes of the baboon and the vervet are in the process of being sequenced.

1.2 What Are We Trying to Explain?

In this volume, almost every chapter ultimately focuses on trying to explain the causal sources of human quantitative phenotypic variation. In many cases, the phenotypes under consideration are related to complex disease risk. In general, we would argue that the principle role of modern human genetics is to identify the causal sequence variants responsible for quantitative phenotypic variation. Most of the phenotypes in which we are interested exhibit complex causal pathways unlike those seen for simple monogenic traits. An overriding challenge in the analysis of complex traits as compared to the analysis of simple monogenic traits is that multiple loci may be contributing to the phenotype and, as a result, the effect size of any one locus is likely to be relatively small. In addition, there may be multiple types of sequence variation in play, ranging from substitutions of single nucleotides to rearrangements of chromosomal structure (sequence deletion, duplication, or inversion), which may not be equally detectable by any one analytical technique. Finally, even if the majority of a genetic effect was confined to a single locus, this could be either due to a single variable site or multiple rare alleles segregating in the population (s) under study.

1.3 Measuring Genetic Variation in Quantitative Traits

Many of the chapters in this volume at least implicitly involve characterizing how much of the observed phenotypic variation in primates is due to the action of genes. *Heritability* is the proportion of the total variance of a phenotype that is attributable to the additive effects of alleles. It represents an estimate of the relative extent of genetic variation in a given phenotype. Thus, heritability provides us with a single measure of how important a role genes likely play in the causal determination of a variable human trait. In a classical variance-components-based approach to quantitative genetic analysis, heritability is readily estimated by decomposing the phenotypic covariance between pairs of individuals based on their relatedness:

$$\begin{aligned} \mathrm{Cov}(i,j) &= 2\phi_{ij}\sigma_g^2 \\ \mathrm{Cov}(i,i) &= \sigma_g^2 + \sigma_e^2 \\ h^2 &= \sigma_g^2/\sigma_P^2 \end{aligned} \tag{1.1}$$

where $\mathrm{Cov}(i,j)$ is the covariance between different individuals i and j, and $\mathrm{Cov}(i,i)$ represents the variance for the ith individual, ϕ_{ij} is the kinship coefficient between i and j, σ_g^2 is the additive genetic variance, σ_e^2 is the error (sometimes termed

environmental) variance, σ_P^2 is the total phenotypic variance of the trait given by $\sigma_P^2 = \sigma_g^2 + \sigma_e^2$, and h^2 is the additive genetic heritability which measures the relative contribution of additive genetic factors to the overall observed phenotypic variance.

Twice the 'kinship' coefficient in this context refers to the expected proportion of alleles shared *identical by descent* (IBD) by two individuals given their degree of relatedness: siblings share half their alleles IBD, half-siblings one-fourth of their alleles, and so on. Classically, we estimate this relatedness by knowledge of pedigree records. However, with the advent of high density assays of genetic variants (such as whole genome sequencing), we can now empirically estimate genetic relatedness in the absence of knowledge of the pedigree relationships among individuals. This latter development opens up vast opportunities for the study of wild primate populations. The genetic variance is a cumulative variance; it represents the summation over all additive genetic factors for the phenotype. Hence, depending upon the phenotype, it may represent the influence of a single genetic variant or that of many hundreds of genetic variants.

If variation in a phenotype were entirely due to genetic causes (and these could be clearly discerned), the heritability of the trait would, of course, be 1. Due to multifactorial causation and measurement error, a typical range of heritability for many quantitative traits is between 30 and 80 % of the total variance, and estimates may differ widely from one study to another due to sampling error. The heritability is a critical measure of the importance of within-population genetic variation. This single metric conveys whether or not the search for the individual contributing genes is merited for a given phenotype.

There have been thousands of studies of heritability of human phenotypes but relatively few of nonhuman primate phenotypes. The lack of nonhuman primate studies presumably is due to the paucity of pedigreed populations. However, the studies that have been conducted show that a substantial amount of genetic variation relevant for complex phenotype variation is segregating in nonhuman primate colonies.

Life history traits such as life span (Martin et al. 2002) and age at first birth (Williams-Blangero and Blangero 1995) show significant heritable components in pedigreed baboons. Standard hematological parameters in chimpanzees show significant heritable components (Williams-Blangero et al. 1993), as do many different lipid parameters in baboons (Blangero et al. 1990). Anatomical phenotypes, in particular, show high heritabilities (Mahaney et al. 1993; Rogers et al. 2007). Other even more complex phenotypes, such as the response of liver enzymes to experimental infection with hepatitis C virus in chimpanzees (Williams-Blangero et al. 1996) and longitudinal changes in fetal baboon morphometrics (Jaquish et al. 1997) also are significantly heritable in nonhuman primates. The evidence for substantial additive genetic variation in fundamental anatomical, physiological, and biochemical phenotypic dimensions in nonhuman primates parallels that seen in human studies and further confirms the utility of nonhuman primate models for studying human biological problems.

1.4 Functional Genetic Variation Determines Heritability

What is the biological source of heritability in humans? Ultimately, it comes from observable functional genetic variation at the sequence level. A functional variant is one that influences the focal phenotype via some molecular mechanism. Thus, functional variants can be considered to be phenotype-specific in this context. If a variant influences a quantitative trait (such as performance), we term it a *quantitative trait nucleotide* variant (QTN). The effect of a functional variant on the phenotype can be quantified by the QTN-specific variance which is given by $\sigma_q^2 = 2p(1-p)\alpha^2$, where p is the minor allele frequency of the QTN and α is one-half the difference between phenotypic means of the two

homozygotes. Biologically, we expect α to be determined by biophysical molecular properties of the QTL and relatively constant across populations. The term, $2p(1-p)$, is also known as the expected heterozygosity of the underlying genotype and measures the variance of a trait that is scored as the number of minor alleles in the diploid genotype. The relative genetic signal intensity for this QTN is given by the QTN-specific heritability $h_q^2 = \sigma_q^2/\sigma_P^2$ where σ_P^2 is the total variance of the phenotype. The relative genetic signal for the QTL is determined by the sum of the QTN-specific heritabilities (although these must be corrected for possible linkage disequilibrium amongst variant sites) in the immediate region of the QTL and thus will be influenced by all of the relevant functional variants in the region. In algebraic form, the QTL-specific heritability is $h_q^2 = \frac{\sum 2p_i(1-p_i)\alpha_i}{\sigma_P^2} = \sum h_{qi}^2$ where the summation is over the functional variants in the regions of the QTL. Similarly, the total heritability of the phenotype is given by the sum of all of the QTL-specific heritabilities over the whole genome or $h^2 = \sum h_{qi}^2$.

1.5 Identifying Functional Sequence Variants Is the Critical Problem in Primate Biology

One of the main reasons that we study nonhuman primate species is to aid in the identification of the function of sequence variants. The four chapters that specifically utilize nonhuman primate genomics in this volume ultimately point to ways to better identify human genes and their sequence variants that influence human phenotypic variation.

In Chap. 7, Kelaita provides an overview of genomic methods applied to studies of wild primate populations. Besides the obvious utility of genomic methods for aiding our understanding of primate microevolution and primate population structure, she also suggests that phylogenetic inference can aid our interpretation of human adaptations (which are ultimately about human phenotypic variation).

Information on sequence variation across species can be used to make evolutionary inferences about genes likely to be under the influence of natural selection. Such selection only will occur for functionally relevant sequence variation (and nearby variants that are in linkage disequilibrium). Information on selection can be accumulated and used to aid studies of human sequence variation when attempting to determine which variants are most likely to be functional.

There is even more potential value likely to come from studies of wild nonhuman primate populations. Using our quantitative genetic model as described above, we would further suggest the great potential to better assess the importance of genetic variation for complex phenotypes observed in wild nonhuman primate populations. Now that accurate and direct molecular assay of genetic relatedness can be performed via sequencing technology, the potential to better understand the genetic basis of many complex phenotypes that are only observable in wild populations is greatly enhanced.

In Chap. 8, Cox directly shows how causal gene discovery of relevance for human disease risk can be directly performed in captive pedigreed nonhuman primate colonies. She highlights a clear benefit of using nonhuman primates for making inferences about human biology which is the access to tissues that are extremely difficult to obtain on a large scale in human studies. In studies of nonhuman primates, it is possible to safely obtain tissues such as liver or kidney that may be of critical value in understanding the biological mechanisms underlying functional sequence variation. Indeed, the potential for deep cellular phenotyping of many different nonhuman primate samples represents one of the major benefits of the primate animal model. Although the general paradigm Cox utilizes is that of complex phenotype gene discovery widely used in human populations, she shows that the ability to manipulate the other main component of the

causal players, the environment, is possible in such experimental situations. By carefully controlling the environment, it is feasible to truly test for such complexities such as genotype-by-environment interaction. In her case, she focuses on genotype-by-diet interaction effects on lipid variation. This type of experiment involving the rigorous control of diet is extremely difficult to directly perform in humans and thus the benefit of doing such in nonhuman primates is obvious.

In Chap. 14, Sherwood and Duren examine the genetic determinants of variation in the primate craniofacial complex. This phenotypic dimension also is of obvious utility for understanding a large number of human disorders. Again, they employ standard gene discovery approaches to captive primate colonies. The ability to deeply phenotype animal models becomes a substantial benefit when dealing with potentially high dimensional imaging-derived phenotypes.

Finally, in Chap. 15 Bailey and colleagues apply similar approaches to nonhuman primate behavioral phenotypes. These are the most complex of all phenotypes and also among the most difficult to study. They review a number of studies including work on pedigreed vervets that show a consistent heritable component for complex primate behaviors. Other groups working on vervets also have used advanced genomic and transcriptomic methods to identify likely genes involved in complex phenotypes (Jasinska et al. 2012).

Advanced phenotyping relevant for behavior and psychiatric disease risks such as brain imaging are possible in large numbers of related primates to act as potential endophenotypes of relevance for the focal behaviors/disease risks. For example, an advanced brain imaging study that identified substantial genetic variation in the neural basis of anxious temperament in pedigreed rhesus macaques undergoing brain PET was performed successfully (Oler et al. 2010). Other clear benefits to working with nonhuman primates for psychiatric disease studies include the ability to get cerebrospinal fluid samples from large numbers of animals, a tissue that is exceedingly hard to justify in human studies of normal variation (Rogers et al. 2004).

1.6 Where Are We Going?

Most of the nonhuman primate applications in this volume still focus on causal gene discovery. Like others (Aitman et al. 2011), we believe that the tremendous advances in studies of human genetics will soon eliminate this focus. The field of human complex disease genetics is currently transitioning away from the study of common sequence variants of small effect (such as those that have been found typically in genome wide association studies) to the study of rare variants that are difficult to capture in sufficient numbers for testing except in extended families. These human studies often point to genes that require additional biological investigation in more controlled experimental circumstances.

Given that we can now sequence very large numbers of humans to directly search for likely functional variants, the utility of nonhuman primate genomic studies should transition to our second major question, that of aiding the functional characterizations of sequence variants. One of the most obvious ways to utilize nonhuman primates in this context is the exploitation of existing sequence variation through experimental breeding. For example, in the near future, all nonhuman primate colonies can easily be sequenced and all rare coding variation identified. For a given gene of interest, we can then identify those animals harboring the most likely consequential functional variation and design a breeding plan that will generate sufficient numbers of copies of variant animals for deep phenotyping studies. The ability to get at critical tissues, such as neuronal tissues of relevance for many psychiatric diseases, will greatly facilitate our ability to decide what the most important genes and variants are. Indeed, advances in genome editing and gene therapy are further likely to give us the tools to study the functional consequences of human sequence variants in

nonhuman primates. This should help us identify and prioritize causal genes that may be of greatest importance for human disease pathways with the main advantage of providing an in vivo biological context that is extremely similar to that which we observed in humans.

References

Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TF, Stemple DL (2011) The future of model organisms in human disease research. Nat Rev Genet 12:575–582

Blangero J, MacCluer JW, Kammerer CM, Mott GE, Dyer TD, McGill HC Jr (1990) Genetic analysis of apolipoprotein A-I in two dietary environments. Am J Hum Genet 47:414–428

Bozek K, Wei Y, Yan Z, Liu X, Xiong J, Sugimoto M, Tomita M, Pääbo S, Pieszek R, Sherwood CC, Hof PR, Ely JJ, Steinhauser D, Willmitzer L, Bangsbo J, Hansson O, Call J, Giavalisco P, Khaitovich P (2014) Exceptional evolutionary divergence of human muscle and brain metabolomes parallels human cognitive and physical uniqueness. PLoS Biol 12(5):e1001871

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87

Jaquish CE, Leland MM, Dyer T, Towne B, Blangero J (1997) Ontogenetic changes in genetic regulation of fetal morphometrics in baboons (*Papio hamadryas* subspp.). Hum Biol 69:831–848

Jasinska AJ, Lin MK, Service S, Choi OW, DeYoung J, Grujic O, Kong SY, Jung Y, Jorgensen MJ, Fairbanks LA, Turner T, Cantor RM, Wasserscheid J, Dewar K, Warren W, Wilson RK, Weinstock G, Jentsch JD, Freimer NB (2012) A non-human primate system for large-scale genetic studies of complex traits. Hum Mol Genet 21:3307–3316

Liu X, Somel M, Tang L, Yan Z, Jiang X, Guo S, Yuan Y, He L, Oleksiak A, Zhang Y, Li N, Hu Y, Chen W, Qiu Z, Pääbo S, Khaitovich P (2012) Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome Res 22:611–622

Mahaney MC, Williams-Blangero S, Blangero J, Leland MM (1993) Quantitative genetics of relative organ weight variation in captive baboons. Hum Biol 65:991–1003

Martin LJ, Mahaney MC, Bronikowski AM, Carey KD, Dyke B, Comuzzie AG (2002) Lifespan in captive baboons is heritable. Mech Ageing Dev 123:1461–1467

Oler JA, Fox AS, Shelton SE, Rogers J, Dyer TD, Davidson RJ, Shelledy W, Oakes TR, Blangero J, Kalin NH (2010) Amygdalar and hippocampal substrates of anxious temperament differ in their heritability. Nature 466:864–868

Rhesus Macaque Sequencing and Analysis Consortium (2007) The rhesus macaque genome. Science 316:235–237

Rogers J, Kochunov P, Lancaster J, Shelledy W, Glahn D, Blangero J, Fox P (2007) Heritability of brain volume, surface area and shape: an MRI study in an extended pedigree of baboons. Hum Brain Mapp 28:576–583

Rogers J, Martin LJ, Comuzzie AG, Mann JJ, Manuck SB, Leland M, Kaplan JR (2004) Genetics of monoamine metabolites in baboons: overlapping sets of genes influence levels of 5-hydroxyindolacetic acid, 3-hydroxy-4-methoxyphenylglycol, and homovanillic acid. Biol Psychiatry 55:739–744

Williams-Blangero S, Blangero J (1995) Heritability of age of first birth in captive olive baboons. Am J Primat 37:233–239

Williams-Blangero S, Blangero J, Murthy KK, Lanford RE (1996) Genetic analysis of serum alanine transaminase activity in normal and hepatitis C virus infected chimpanzees: an application of research-oriented genetic management. Lab Anim Sci 46:26–30

Williams-Blangero S, Brasky K, Butler T, Dyke B (1993) Genetic analysis of hematological traits in chimpanzees (Pan troglodytes). Hum Biol 65:1013–1024

The Human Genome Project: Where Are We Now and Where Are We Going?

2

Satish Kumar, Christopher Kingsley, and Johanna K. DiStefano

2.1 The Human Genome Project: Where Have We Been?

An explosion in our understanding of genetics and biochemistry, which began in the 1970s, led to the rapid development of diverse laboratory techniques such as restriction enzymes, cloning vectors, nucleic acid hybridization, and DNA sequencing. Together these methods revolutionized research in molecular biology. It was here, in this fertile atmosphere, that the seeds of genome sequencing were sown. The progressive spirit pervading research in the life sciences at this time consequently helped to fuel the conception of the Human Genome Project (HGP), whose primary aims were to determine the identity of the three billion nucleotides comprising the human genome and characterize the full repertoire of genes encoded therein.

S. Kumar
Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX 78245, USA
e-mail: skumar@txbiomedgenetics.org

C. Kingsley · J.K. DiStefano (✉)
Diabetes, Cardiovascular & Metabolic Diseases Division, Translational Genomics Research Institute, 445 North Fifth Street, Phoenix, AZ 85004, USA
e-mail: jdistefano@tgen.org

C. Kingsley
e-mail: ckingsley@tgen.org

2.1.1 Historical Background of the HGP

The HGP is considered one of the most ambitious and successful international research collaborations in the history of biology. Those individuals and organizations responsible for bringing the HGP to fruition were both visionary and innovative, considering that the technological and computational tools commonplace today were unheard of 20 years ago when the idea of sequencing the human genome was germinated. Because thorough and engaging accounts of the conception, implementation, and completion of the HGP have already been presented elsewhere (Roberts 2001; Choudhuri 2003), we will provide only a brief synopsis of its history here.

The idea of sequencing the human genome was first discussed in 1984 at a meeting in Salt Lake City, Utah, hosted by the Department of Energy (DOE) and the Internal Commission for Protection Against Environmental Mutagens and Carcinogens. Although the purpose of this meeting was focused on mutation detection, the value of a human genome reference sequence was acknowledged, albeit in an oblique manner (Cook-Deegan 1989). The actual merit of sequencing the human genome was brought forward as a focus topic for the first time in 1985 during a conference at the University of California, Santa Cruz. Meeting participants generally supported the idea of such a project, but largely agreed that the endeavor laid outside the then current realms of feasibility and/or practicality.

R. Duggirala et al. (eds.), *Genome Mapping and Genomics in Human and Non-Human Primates*, Genome Mapping and Genomics in Animals 5, DOI 10.1007/978-3-662-46306-2_2,

Enthusiasm for the initiative quickly mounted during the following year at meetings held consecutively at Los Alamos National Laboratory and Cold Spring Harbor Laboratory (Roberts 2001). Debate about the value, expense, and potential consequences of the initiative continued until 1988, when the National Research Council panel officially endorsed the HGP. At that time, the panel refined the initiative, recommending that physical maps of each chromosome be constructed, and genomes of simple organisms be investigated prior to the full-scale sequencing of the human genome. In addition to sequencing the entire human genome, the HGP also aimed to identify all genes in the human genome, store sequence information in publicly available databases, develop and/or improve tools for analyzing sequence data, help transfer technologies resulting from the HGP to the private sector, and address relevant ethical, legal, and social issues (http://www.ornl.gov).

The HGP was officially launched on October 01, 1990, following the initiation of large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabiditis elegans*, and *Saccharomyces cerevisiae*. The International Human Genome Sequencing Consortium (IHGSC), comprised of the National Institutes of Health (NIH), the DOE, and a collaborative of investigators from the United Kingdom, France, Germany, Japan, and China, was formed to implement the goals of the HGP. In 1998 this effort was joined by Celera Genomics, a privately funded venture formed jointly by Dr. J. Craig Venter, from The Institute for Genomic Research (TIGR), and the Perkin-Elmer Corporation. Venter proposed to sequence the human genome in a shorter period of time and at less cost than the publicly funded effort, using the relatively novel technique of whole genome shotgun sequencing. In early 2001, both IHGSC (Lander et al. 2001) and Celera Genomics (Venter et al. 2001) published working draft sequences of the human genome. Although these drafts covered only ~90 % of the euchromatic genome, was interrupted by ~150,000 gaps, had many misassembled segments and errors in the nucleotide sequence, the accomplishment of such a tremendous effort was generally applauded among the scientific community.

Following the publication of these rough draft versions of the genome, the IHGSC initiated efforts to finish sequencing the euchromatic genome and resolve areas containing gaps and misalignments. Results of these efforts were published in 2004 (International Human Genome Sequencing Consortium 2004). This updated version of the human genome covered 2.85 billion nucleotides, corresponding to ~99 % of the euchromatic genome. The near-complete draft was highly accurate: the error rate of the new genome sequence was reduced to <1 event/100,000 bases, a figure that surpassed the original acceptable estimate of the project (International Human Genome Sequencing Consortium 2004). The number of gaps was likewise decreased from ~150,000 to only 341, and most of these remaining gaps were associated with segmental duplications that are not amenable to current methods of sequencing. With the release of the near-complete human genome sequence, the original goals of the HGP were largely achieved (International Human Genome Sequencing Consortium 2004). Despite the incompleteness of this "finished" version, the availability of these sequence data has already had an irrevocable impact on the study of human disease.

2.2 Impact of the Human Genome Project: Where Are We Now?

Completion of the Human Genome Project has provided us with a greatly enhanced understanding of human genetics, including a greater appreciation of how DNA shapes species development and evolution, biology, and disease susceptibility. The HGP has also affected the development and/or maturation of research disciplines such as genome annotation, knowledge of genome evolution and segmental duplication, and comparative genomics, among others. Below we discuss the areas in which completion of the HGP has influenced our basic understanding of genetics, while subsequent sections will address

the impact of the HGP on the manner in which we approach disease risk and development of treatment strategies based on genetic predisposition.

2.2.1 Enhanced Understanding of Human Genetics

2.2.1.1 Genome Annotation

The sequencing portion of the HGP was a significant technological feat, and provided the scientific community with a comprehensive accounting of the working material of the genome. However, acquisition of DNA sequence was only the first step toward the ultimate aim of understanding how the human genome functions at the molecular level. Necessary next steps toward this goal include the systematic identification and characterization of the functional units of the genome. This process of genome annotation is currently a multidisciplinary field, integrating the results of many different analytical approaches, both experimental and computational, to build our understanding of the functional underpinnings of the human genome (Table 2.1).

Prior to the completion of the HGP, the field of genome annotation was largely focused on the comprehensive identification of protein-coding genes, which was primarily achieved through the use of large-scale sequencing of cDNA libraries derived from reverse-transcribed mRNA transcripts. The resulting expressed sequence tags (ESTs) were grouped together based on sequence similarity using multiple sequence alignment algorithms. It was generally held that if the starting material was comprised of a mixture of mRNAs purified from numerous tissue types, then the number of groups produced by this process would provide a rough estimate of the total number of protein-coding genes expressed throughout the body. Prior to the publication of the human genome sequence, estimates on the total number of genes varied widely, from 35,000 to 150,000 (Pennisi 2007).

While cDNA sequencing approaches were fairly open ended in nature, the HGP produced a finite database of sequence information that could be easily searched for the presence of protein-coding genes. Yet, due to the low proportion of coding sequence in the human genome, the large number of exons per genes, and the relatively small exon size, gene annotation presented a much more difficult proposition in

Table 2.1 Experimental and computational methods of genome annotation

Genomic feature	Experimental/computational approach
Gene identification	cDNA and peptide sequencing
	Computational prediction
	Comparative genomics
Transcript identification	Tiling microarray
	cDNA sequencing
	Computational prediction
	Comparative genomics
Regulatory sequence identification	Chromatin Immunoprecipitation and tiling microarray (ChIP-Chip)
	Computational prediction of factor binding sites
	Promoter/enhancer assays
Sequence variation	DNA resequencing
	Copy number microarray
Chromatin structure	*DNaseI* sensitivity assay
	Tiling microarray

A number of methods are currently employed to identify functional regions of the genome. The first column lists several genomic features that are commonly annotated, and the second column lists the experimental or computational approaches that can be used to identify those features in genome sequence assemblies

humans compared to previously sequenced organisms, such as *Drosophila melangaster*, *C. elegans,* or various prokaryotes. Because of this fact, a hybrid approach was taken that incorporated multiple lines of evidence, including homology of genome sequence to ESTs, similarity to other known genes or proteins, and statistical strategies that took into account splice site structure, amino acid coding bias, and known distributions of intron and exon lengths. Using these approaches with the newly available human genome sequence, a surprisingly low estimate of only 30,000–40,000 protein-coding genes was obtained, but the estimate involved considerable guesswork owing to the imperfections of the draft sequence and the inherent difficulty of gene identification (Lander et al. 2001; Venter et al. 2001). In the years following these initial estimates, it was discovered that many open reading frames (ORFs) that occur at random in transcripts are actually nonfunctional, and the total number of protein-coding genes has been steadily revised downward since. Currently, the human genome is estimated to contain approximately 20,000–21,000 protein-coding genes (Clamp et al. 2007; Pennisi 2007). Recent RNA-Seq projects have confirmed the gene catalog, while illuminating alternative splicing, which seems to occur at >90 % of protein-coding genes and results in many more proteins than genes. At this time, the proteome is now known to be similar across placental mammals, with about two-thirds of protein-coding genes having 1:1 orthologues across species and most of the rest belonging to gene families that undergo regular duplication and divergence—the de novo creation of fundamentally new proteins is considered a rare phenomenon (Lander 2011).

The human genome also gives rise to a large number of noncoding RNAs (Kapranov et al. 2007). Oligonucleotide-based tiling microarrays that interrogate every base pair of genome sequence over expansive regions have revealed that a much larger percentage of the human genome is transcribed compared to what was originally presumed (Cheng et al. 2005). While only 1–2 % of the human genome codes for proteins, approximately 15 % of all interrogated bases were able to detect RNA molecules from a single cell line, indicating that the vast majority of transcription from the human genome produces noncoding RNA products. The novel RNA transcripts are often transcribed from both strands, and transcription of coding sequences from the antisense strand is particularly common (Cheng et al. 2005). While the function of most of these products is not yet known, some noncoding RNAs exert regulatory effects on coding transcripts through complementary nucleotide base pairing. This hybridization decreases transcript stability by targeting it for degradation or translational repression (Kim and Nam 2006).

One of the surprising discoveries about the human genome was that the majority of the functional sequence does not encode proteins. Inferring these non-neutral, conserved noncoding elements in humans was a challenge before the HGP. Soon after the first draft the comparative analysis of the human and mouse genomes showed a substantial excess of conserved sequence, relative to the neutral rate in ancestral repeat elements (Mouse Genome Sequencing Consortium 2002).

Research groups working independently of one another have performed most of the approaches applied toward annotating the human genome (Table 2.1). The National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the **ENC**yclopedia **O**f **D**NA **E**lements, in September 2003, to systematically integrate the genome annotation efforts in identifying all functional elements in the human genome sequence. (ENCODE Project Consortium 2004). The project started with two components—a pilot phase and a technology development phase. The pilot phase of the ENCODE project tested and compared the existing arsenal of annotation approaches on a series of 44 genomic regions comprising approximately 30 Mb, or roughly 1 % of the human genome. About half of the targets were chosen to contain extensively characterized genes or functional regions, while the other half were randomly selected (ENCODE Project Consortium 2004). The findings of the pilot project were published in June 2007

(ENCODE Project Consortium 2007) and scores of important information highlighted includes:

- There is abundant transcription beyond the known protein-coding genes both intragenic and intergenic transcription, including noncoding RNA and transcribed pseudogenes. While this has been previously observed in other studies, the ENCODE pilot phase confirmed this phenomenon on a global level.
- At the same time, known protein-coding genes revealed unexpected complexity in distal, untranslated regions (UTRs), exons located as far as 200 kb away, overlapping or interleaved loci, and antisense transcription. Together, these findings challenged the conventional definition of a "gene".
- Patterns of histone modifications and DNase sensitivity revealed domains of packed or accessible chromatin. These accessibility patterns correlate well with rates of transcriptions, DNA replication, and regulatory protein factors binding to the DNA. These results served to underscore the regulatory importance of epigenetic factors.

Combined, the ENCODE findings changed our conceptual framework of the organization and functional aspects of the genome. Two additional goals of the pilot ENCODE Project were to develop and advance technologies for annotating the human genome, with the combined aims of achieving higher accuracy, completeness, and cost-effective throughput and establishing a paradigm for sharing functional genomics data.

In 2007, the ENCODE Project was expanded to study the entire human genome, capitalizing on experimental and computational technology developments during the pilot project period. The genome-wide ENCODE phase is currently in progress focusing on the completion of two major classes of annotations—genes (both protein-coding and noncoding) and their RNA transcripts and transcriptional regulatory regions.

Gene Annotation. A major goal of ENCODE is to annotate all protein-coding genes, pseudogenes, and noncoding transcribed loci in the human genome and to catalog the products of transcription, including splice isoforms. Although the human genome contains 20,000 protein-coding genes (International Human Genome Sequencing Consortium 2004), accurate identification of all protein-coding transcripts has not been straightforward. Annotation of pseudogenes and noncoding transcripts also remains a considerable challenge. While automatic gene annotation algorithms have been developed, manual curation remains the approach that delivers the highest level of accuracy, completeness, and stability (Guigo et al. 2006). This annotation process involves consolidation of all evidence of transcripts (cDNA, EST sequences) and proteins from public databases, followed by building gene structures based on supporting experimental data (Harrow et al. 2006). More than 50 % of annotated transcripts have no predicted coding potential and are classified by ENCODE into different transcript categories. A classification that summarizes the certainty and types of the annotated structures is provided for each transcript. Pseudogenes are identified primarily by a combination of similarity to other protein-coding genes and an obvious functional disablement such as an in-frame stop codon. Ultimately, each gene or transcript model is assigned one of the three confidence levels. Level 1 includes genes validated by RT-PCR and sequencing, plus consensus pseudogenes. Level 2 includes manually annotated coding and long noncoding loci that have transcriptional evidence in EMBL/GenBank. Level 3 includes Ensembl gene predictions in regions not yet manually annotated or for which there is new transcriptional evidence. The result of ENCODE gene annotation "GENCODE" is a comprehensive catalog of transcripts and genemodels. ENCODE gene and transcript annotations are updated bimonthly and are available through the UCSC ENCODE browser, Distributed Annotation Servers (DAS), and the Ensembl Browser (Flicek et al. 2010; ENCODE Project Consortium 2011, 2012).

RNA Transcripts. The work on comprehensive genome-wide catalog of transcribed loci that characterizes the size, polyadenylation status, and subcellular compartmentalization of all transcripts is also ongoing at ENCODE, with transcript data generated from high-density

(5 bp) tiling DNA microarrays (Kampa et al. 2004) and massively parallel DNA sequencing methods (Mortazavi et al. 2008; Wold and Myer 2008; Wang et al. 2009). Because subcellular compartmentalization of RNAs is important in RNA processing and function, such as nuclear retention of unspliced coding transcripts (Schmid and Jensen 2010) or small nucleolar RNA (snoRNA) activity in the nucleolus (Bachellerie et al. 2002), ENCODE is analyzing not only total whole cell RNAs but also those concentrated in the nucleus and other subcellular compartments, providing catalogs of potential microRNAs (miRNAs), snoRNA, promoter-associated short RNAs (PASRs) (Kapranov et al. 2007), and other short cellular RNAs. These analyses revealed that the human genome encodes a diverse array of transcripts. Additional transcript annotations include exonic regions and splice junctions, transcription start sites (TSSs), transcript 39 ends, spliced RNA length, locations of polyadenylation sites, and locations with direct evidence of protein expression (ENCODE Project Consortium 2011, 2012).

Transcriptional Regulatory Regions. Transcriptional regulatory regions include diverse functional elements such as promoters, enhancers, silencers, and insulators, which collectively modulate the magnitude, timing, and cell specificity of gene expression (Maston et al. 2006). The ENCODE Project is using multiple approaches to identify *cis*-regulatory regions, including localizing their characteristic chromatin signatures and identifying sites of occupancy of sequence-specific transcription factors. These approaches are being combined to create a comprehensive map of human *cis*-regulatory regions.

Chromatin Structure and Modification. Chromatin accessibility and histone modifications provide independent and complementary annotations of human regulatory DNA, and massively parallel, high-throughput DNA sequencing methods are being used by ENCODE to map these features on a genome-wide scale. Deoxyribonuclease I (DNaseI) hypersensitive sites (DHSs) and an expanding panel of histone modifications are also being mapped (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007). ENCODE chromatin annotation data such as chromatin accessibility, DNase I hypersensitive sites, and selected histone modifications are available through the UCSC browser (http://genome.ucsc.edu/).

Transcription Factor and RNA Polymerase Occupancy. Much of human gene regulation is determined by the binding of transcriptional regulatory proteins to their cognate sequence element in *cis*-regulatory region. To create an atlas of regulatory factor (i.e., transcription factors, RNA polymerase 2, both initiating and elongating, and RNA polymerase 3) binding, ENCODE is applying chromatin immunoprecipitation and DNA sequencing (ChIP-seq) technology, which enables genome-wide mapping of transcription factors occupancy pattern in vivo (Barski et al. 2007; Johnson et al. 2007; Robertson et al. 2007). Alternative technologies, such as epitope tagging of transcription factors in their native genomic context using recombineering (Poser et al. 2008; Hua et al. 2009), are also being explored.

ENCODE Additional Data. ENCODE is also generating additional data types to complement gene and regulatory region annotations and that includes data on DNA methylation, DNase I footprinting, long-range chromatin interaction, protein–RNA interaction, and genetic and structural variation in the cell types used in ENCODE production phase. The key features of the production phase include use of several cell types for the main data collections efforts and the use of these cell types by all project teams to maintain consistency. The cell types are organized into tiers to prioritize experimental investigations. These features are expected to enable better coordination of studies and interpretation of results.

2.2.1.2 Segmental Duplications

The HGP has also extended our understanding of segmental duplications (SDs). Eukaryotic organisms have evolved a complex, highly regulated

cellular machinery to insure the proper replication, condensation, and segregation of chromosomes during cell division (Hirano 2000). However, errors in the distribution of genetic material during cell division occasionally occur, leading to daughter cells that receive more or less than the usual complement of genomic DNA following cell division. If such an alteration in DNA copy number occurs in the germ cell lineage of a multicellular organism, then the progeny of that organism can inherit the change in DNA copy number. Over many generations, copy number changes that occur in a single individual can spread through a population, leading to a situation in which the copy number status of a chromosomal region can be considered a type of genetic polymorphism, typically referred to as a copy number polymorphism (CNP) or copy number variation (CNV) (Bailey et al. 2002; Sebat et al. 2004).

The human genome is enriched for SDs that vary extensively in copy number (Bailey et al. 2002; Iafrate et al. 2004; Redon et al. 2006; Kidd et al. 2008). There are about 25,000–30,000 SDs with ≥90 % sequence identity and ≥1 kb length have been identified in the human genome, which cover about 5–6 % of the total genome (Bailey et al. 2002). It has also been reported that SDs are highly enriched with genes and pseudogenes in the human genome (i.e., SDs comprise ~5 % of the genome and contain ~17.8 % of human genes and ~36.8 % of human pseudogenes) (Bailey et al. 2002; Zheng 2008).

When a SD contains a functional gene, the new sequence may contain a paralog performing the same function as the original gene or a new function. Duplicated pseudogenes are formed when the new sequence undergoes mutations that result in the loss of original function (Fig. 2.1). The process of SD such as retrotransposition events may also result in the loss of function (LOF) of the duplicated gene; such genes are referred as processed pseudogenes (Mighell et al. 2000; Harrison and Gerstein 2002). Processed pseudogenes usually lack promoter sequences, and hence are considered dead on arrival.

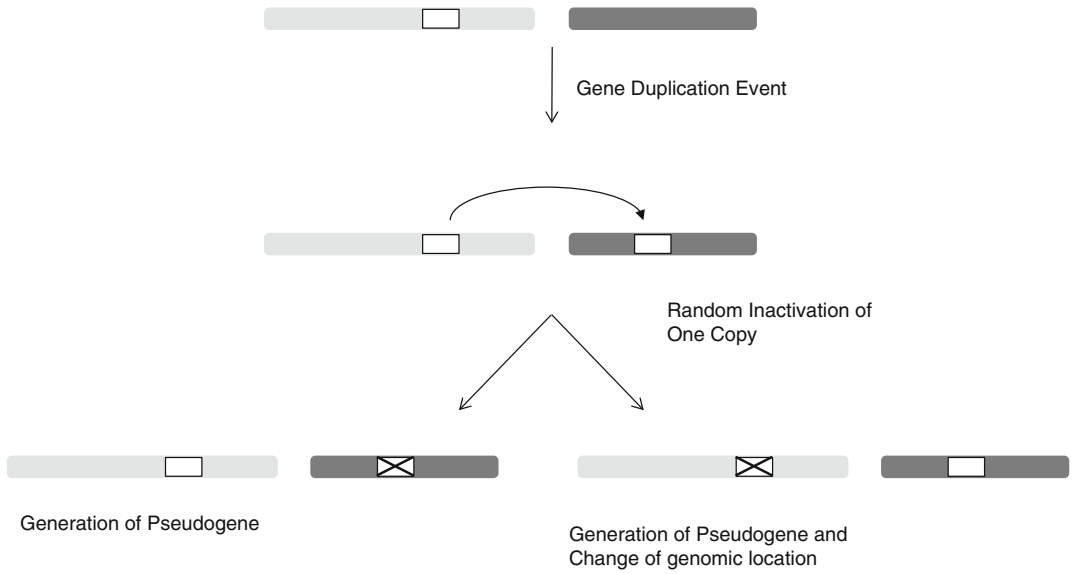

Fig. 2.1 Pseudogene generation by gene duplication and random inactivation. The creation of a novel pseudogene is initiated by a gene duplication event in which a sequence containing a functional gene (*white box*) is duplicated and inserted into a separate site in the genome (shown here as a duplication from one chromosome to another). In most cases of gene duplication, one of the two copies will be randomly silenced and inactivated by mutations, leading to the creation of a pseudogene (checked *white box*). Depending on which of the two copies is inactivated during this process, the genomic position of the original gene can change

Although pseudogenes are assumed to have lost the original coding functions of their parent genes due to the presence of disablements such as premature stop codons or frameshift mutations, recent studies indicate that they might have some regulatory roles (Sasidharan and Gerstein 2008). Automated methods of annotating genomic DNA sequences have identified more than 20,000 pseudogenes (International Human Genome Sequencing Consortium 2004).

Although studies have begun to define the important roles of SDs in generating novel genes through adaptive evolution, gene fusion, or exon exaptation (Lynch and Conery 2000; Taylor and Raes 2004; Bailey and Eichler 2006), it remains a mystery how duplicated copies have evolved from an initial state of complete redundancy (immediately after duplications) to a stable state where both copies are maintained by natural selection. Some glimpse into this important evolutionary process comes from the investigations of duplicated protein-coding genes or gene families showing that duplicated genes can evolve different expression patterns, leading to increased diversity and complexity of gene regulation, which in turn can facilitate an organism's adaptation to environmental change (Gu et al. 2004, 2005; Hittinger and Carroll 2007; Louis 2007). Furthermore, the studies of histone modification in human SDs have also demonstrated that parental and duplicated copies are not functionally identical even though they share ≥90 % identity in their primary sequences, suggesting that descendants in a new genomic environment are more likely the candidates for sequence degeneration or functional innovation (Zhao et al. 2007; Zheng 2008).

Despite recent technological advances in copy number detection, a global assessment of genetic variation of these regions has remained elusive. Commercial single nucleotide polymorphism (SNP) microarrays frequently bias against probe selection within these regions (Estivill et al. 2002; Locke et al. 2006; Cooper et al. 2008; Pinto et al. 2011). Array comparative genomic hybridization (array CGH) approaches have limited power to discern copy number differences, especially as the underlying number of duplicated genes increases and the difference in copy number with respect to a reference genome becomes vanishingly small (Locke et al. 2003; Sharp et al. 2005; Redon et al. 2006; Pinto et al. 2011). Even sequence-based strategies such as paired-end mapping (Tuzun et al. 2005; Korbel et al. 2007) frequently cannot unambiguously assign end sequences in duplicated regions, making it impossible to distinguish allelic and paralogous variation. Consequently, duplicated regions have been largely refractory to standard human genetic analyses (Conrad et al. 2010; Sudmant et al. 2010).

However, a great deal of interest has developed around the role of CNPs/CNVs in inherited diseases, since Lupski et al. (1991) showed for the first time, that a duplicated region on chromosome 17 caused an inherited form of Charcot–Marie–Tooth disease. Since that initial finding, numerous CNPs have been shown to be associated with several human diseases such as psoriasis, Crohn's disease, lupus, rheumatoid arthritis, Parkinson's, Alzheimer's, autism, neuroblastoma, obesity, coronary heart disease, and type 2 diabetes (Cohen 2007; Girirajan et al. 2011). While the number of such cases is still relatively small compared to the number of inherited diseases shown to be caused by point mutations in protein-coding sequences, the importance of CNPs/CNVs in human disease has become increasingly apparent over the past few years. It is now known that at least 15 % of human neurodevelopmental diseases are due to rare and large copy number changes that result in local dosage imbalance for dozens of genes (Giriraj et al. 2011). Other large CNVs, both inherited and de novo, have been implicated in the etiology of autism, schizophrenia, kidney dysfunction, and congenital heart disease. Surprisingly, studies of the general population suggest that although such alleles are rare, collectively they are quite common and under strong purifying selection. These features mean that a significant fraction of the human population carries an unbalanced genome. Such individuals may be sensitized for the effect of another variant that could potentially interact with these CNVs in a digenic manner. The co-occurrence of

multiple, rare CNVs has been used to explain the comorbidity and variable expressivity associated with particular variants in cases of severe developmental delay. There is circumstantial evidence that the full complement of both CNVs and SNPs may be important for understanding genetic diseases more broadly (O'Roak et al. 2011).

2.2.1.3 Comparative Genomics and Genome Evolution

Comparative genomics is the study of relationships among genome sequences of different species. Although a relatively young discipline, comparative genomics has been used to refine our understanding of a number of phenomena, including the evolutionary relationship between species, and the content and function of genomes. From an evolutionary perspective, the similarities and differences between genomic sequences can serve to infer phylogenetic relationships between species based upon molecular criteria in the same fashion that morphological and physiological criteria were used to distinguish species in the past. Identification of conserved regions may also help to elucidate functionally important sequences such as genes, regulatory sites, and structural elements.

Before the availability of whole genome assemblies, comparative genomic analyses were performed using a small number of homologous sequences that were individually isolated from different organisms and sequenced (Murphy et al. 2001). As crucial as these studies were for establishing broad phylogenetic relationships between and among species, the relatively small fraction of genomic sequence used for such analyses was a significant limitation. The recent explosion in the field of comparative genomics results directly from the efforts of numerous sequencing projects and the widespread availability of whole genome assemblies from a variety of different species. The Genomes Online Database (GOLD), which is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, documented 11,472 ongoing and completed genome projects by September 2011. These comprise 8,473 bacterial, 329 archaeal, and 2,204 eukaryal genomes. Additionally, 340 metagenomic projects are tracked with a total of 1,927 samples associated with them. GOLD also tracks well over 1,000 proprietary projects, currently not available to the public, whose metadata will be accessible once the principal investigators of these projects give consent for their public release. In terms of status, 1914 different organisms are completely sequenced and their final sequence has been released from GenBank. From those, 1,644 are bacterial, 117 are archaeal, and 153 are eukaryal. A constantly increasing number of sequencing projects are completed at the level of a draft genome and their final sequences are submitted in GenBank. These projects are identified as "Permanent Draft" genomes. There are currently 989 genomes at this stage (28 archaeal, 949 bacterial, and 12 eukaryal). As of September 2011, the total number of complete genomes is 2,907, which is the sum of the finished and the permanent draft genomes (Pagani et al. 2012).

With the availability of genomes representing multiple species, comprehensive comparisons have produced results that have been both informative and unexpected. Primarily, our understanding of the functional contents of the human genome has been substantially enhanced by comparisons with the genomes of other species. For example, comparison of the human genome with distantly related organisms (e.g., the fruit fly) has been critical for determining the core set of genes necessary for the development and function of multicellular eukaryotes. Similarly, comparison of genomes from humans and vertebrate species of intermediate evolutionary distance (e.g., the mouse) can identify both coding and noncoding sequences that are likely to be functional based on strong evolutionary conservation (Fig. 2.2). Finally, comparison of genomes from humans and closely related primates will help identify the small percentage of divergent sequence that is responsible for specifically human traits. The following paragraphs touch briefly on each of these kinds of comparisons.

The divergence of humans and fruit flies (*D. melanogaster*) from a common ancestor is

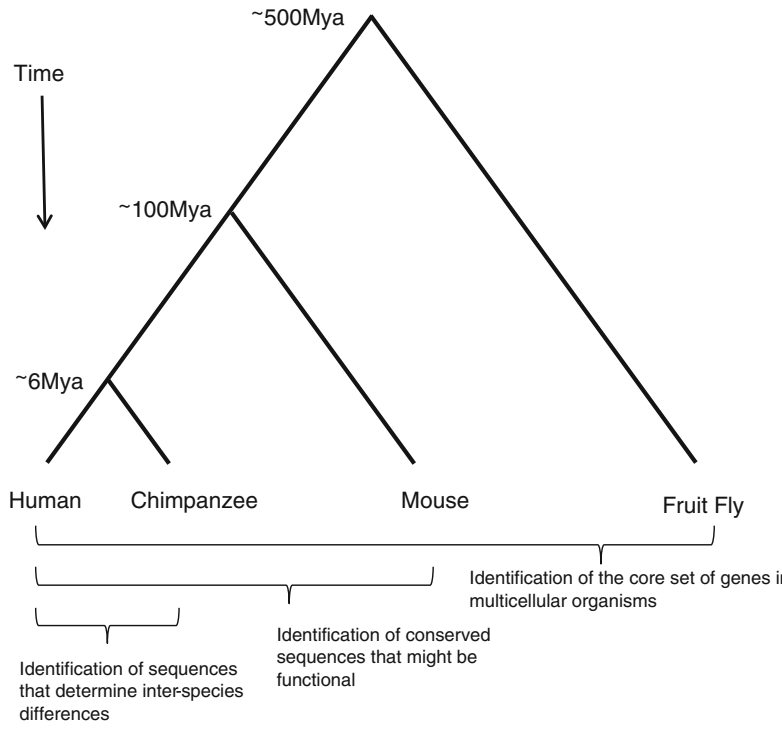

Fig. 2.2 Comparative genomics of species at different evolutionary distances. Genomic comparison of two species can yield different conclusions depending on the degree of genetic difference between them. The evolutionary tree shows the estimated time (in millions of years) from the divergence of human, chimp, mouse, and fruit fly from their common ancestor. The text at *bottom* indicates the information that can be inferred from comparing the human genome to that of a closely related species (chimp), a species of intermediate evolutionary distance (mouse), or a species of great evolutionary distance (fruit fly)

estimated to have occurred over half a billion years ago. The obvious morphological differences between the two species are reflected in the substantial differences at the level of the genome, with the most apparent discrepancies being genome size and gene content (Adams et al. 2000). The human genome spans ~3.1 billion base pairs compared to the 180 million base pairs comprising the *drosophila* genome (Adams et al. 2000), yet contains less than twice as many genes compared to the fly. This size–content disparity is generally consistent with the large expansion of nongenic sequence present in the human lineage, resulting mostly from simple repetitive elements, which are not present in the *drosophila* genome. Despite relatively comparable content, human genes undergo vastly greater amounts of alternative transcription and splicing events, which lead to a much greater diversity of protein products. For example, the ~20,000 genes comprising the human genome give rise to more than 100,000 proteins. Further comparison of protein-coding sequences from the genomes of both species reveals that many genes involved in basic cellular functions such as metabolism, DNA replication and repair, core transcriptional regulation, and cell cycle regulation are conserved. In contrast, human-specific gene expansions are observed for many different functional groups, several of which would be expected given the anatomical and physiological differences between the two species. In general, these expansions occur mainly in gene families involved in adaptive immunity (a vertebrate-specific process), neuronal function, hemostasis, and programmed cell death (Venter et al. 2001).

The first large-scale comparison of two mammalian genome assemblies was performed between human and mouse (*Mus musculus*), two species separated by 75–100 million years of evolution (Mouse Genome Sequencing Consortium 2002; Mural et al. 2002). The human and mouse genomes share ~80–90 % of the same genes, while the remaining unshared genes represent mostly species-specific expansions of functional groups including olfaction, immunology, reproduction, and detoxification (Mouse Genome Sequencing Consortium 2002). One of the most significant and unexpected findings of the human/mouse genome comparison was the large fraction of highly conserved sequences that are neither protein-encoding nor related to known genes (Mural et al. 2002). While ~5 % of the human genome is significantly conserved with that of the mouse (>70 % identity over 100 bp or more), only ~1.5 % of each genome was found to correspond to protein-coding sequence (Dermitzakis et al. 2003). This finding suggests that conserved nonprotein coding sequence is almost twice as abundant as conserved coding sequence. Further, the degree of conservation is estimated to be even greater for noncoding than coding sequences, implying a substantial degree of selective pressure on noncoding sequences (Dermitzakis et al. 2003). Recent comparisons of vertebrate genome assemblies from organisms as diverse as human, rat, mouse, dog, and chicken have provided additional support for this relationship by identifying hundreds of "ultra-conserved" elements, in which an extremely high level of conservation is present among sequences (>95 % over 200 bp or more), and with most of the conserved regions occurring outside of known genes (Bejerano et al. 2004). Although a substantial portion of this conserved sequence is posited to serve a regulatory function (Pennacchio et al. 2006; Prabhakar et al. 2006; Xie et al. 2007), and a very weak selection could also maintain the sequence conservation of ultraconserved elements in noncoding regions (Kryukov et al. 2005; Chen et al. 2007), the reason for this extremely high level of conservation in noncoding regions over millions of years remains unknown.

The completion of genomic assemblies from closely related primates has enabled focus on more recent events in the molecular evolution, molecular adaptation, and genome structure of *Homo sapiens* (Fig. 2.3). Currently, the genome sequences of 13 nonhuman primates are available and at least 11 are approved sequencing targets (Enard 2012). These genomic assemblies together with future sequencing will reveal basic insights into evolutionary processes of mutation, selection and recombination (Marques-Bonet et al. 2009), will be essential tools for primate model organisms (Sasaki et al. 2009), and will also be directly informative for medically relevant questions (Enard 2012). Among the first completed after human are chimpanzee (*Pan troglodytes*) (Chimpanzee Sequencing and Analysis Consortium 2005) and rhesus macaque (*Macaca mulatta*) (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007), which diverged from humans ~6 and 25 million years ago, respectively. Genome-wide comparative analyses of the human, macaque, and chimpanzee genomes have revealed some important features and general principles of primate genome evolution. The alignment of the majority of genomic sequence from closely related primates is relatively trivial (Ebersberger et al. 2002; Thomas et al. 2003) and shows a neutral pattern of single nucleotide variation consistent with the primate phylogeny, although the rate of single nucleotide variation has varied by a factor of threefold within different lineages (Li and Tanimura 1987; Steiper et al. 2004; Elango et al. 2006). Notably, the pattern of single nucleotide variation also varies as a function of chromosome structure and organization (Chimpanzee Sequencing and Analysis Consortium 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). On average, 10 % of the genomic sequence has proven more elusive in terms of orthologous alignment. This includes SDs, subtelomeric regions, pericentromeric regions, and lineage specific repeats.

Comparative sequence data highlight the value of genomic sequence from nonhuman primates to determine the ancestral and derived status of human alleles (Chen and Li 2001;

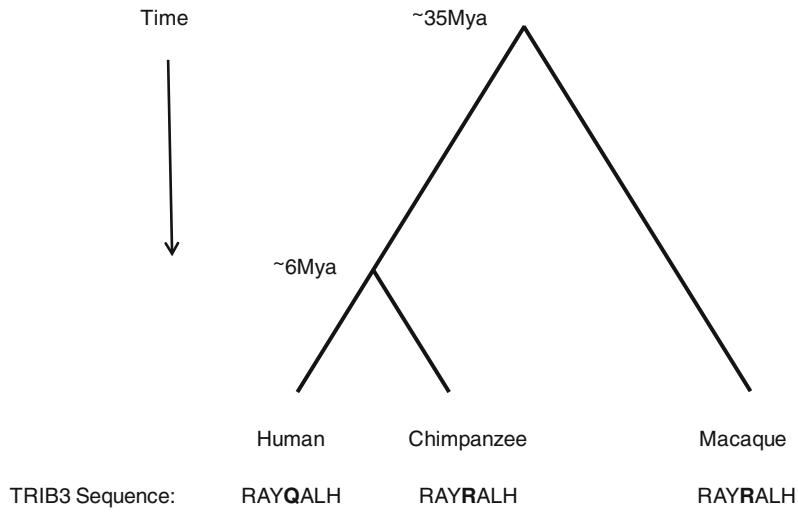

Fig. 2.3 Ternary analysis of closely related primate species. Evolutionary triangulation can identify the lineage in which a sequence variant evolved. The evolutionary tree shows the estimated time (in millions of years) from the divergence of human, chimp, and macaque from their common ancestor. As an example, the protein sequence shown at bottom is derived from a portion of the *TRIB3* gene from each species. Since the sequence variant in the *TRIB3* gene is common to chimp and macaque, it likely occurred in the human lineage within the last 6 million years. Interestingly, the ancestral *TRIB3* allele observed in the chimp and macaque is associated with insulin resistance when present in humans

Kaessmann et al. 2001). There have been some surprises. Phylogenetic analysis of resequenced regions among humans and the great apes reveal that as many as 18 % of genomic regions are inconsistent with the Homo-Pan clade, and, rather, support a Homo-Gorilla clade (Chen and Li 2001). This has been taken as evidence of lineage-sorting and/or an ancestral hominid population size greater than five times that of the effective human population size (n = 10,000). Another surprise has been the identification of ancestral allelic variants that now occur as disease alleles within the human population, i.e., phenylketonuria, macular dystrophy, and cystic fibrosis pyrin and familial Mediterranean fever (Schaner et al. 2001; Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). Such findings suggest that the functional and selective effects of mutations change over time, perhaps as a result of environmental changes or compensatory genetic mutations.

Despite the ease at which genomic sequences can be aligned among primate genomes, the number of genes that can be assigned to 1:1:1 orthologous group has changed only slightly with the first two nonhuman primate genomes sequenced. A three-way comparison involving chimp–human–mouse identified 7,645 orthologues (Clark et al. 2003) as compared to 10,376 by human–chimp–macaque (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007) over the total estimated 20,000 genes in the human genome, suggesting that a large fraction of human genes are yet to be subjected to orthologous comparisons and the pattern of selection operating on these genes is yet to be adequately interrogated. Among the primate order of mammals, comparative genomic studies have advanced more rapidly for taxa closely related to humans, chimpanzees, macaques, and baboons. As complete genome sequencing projects advance for other primate families, including the New

World monkeys (Cebidae) and strepsirhine primates (lemurs, lorises, aye-aye, pottos, and galagos), new insights are anticipated as, particularly for a lemur genome project, new information about primate adaptations and evolution can be anticipated (Horvath and Willard 2007).

However, identification of the most recent events in the speciation of *H. sapiens* will require comparative analyses between the genomes of humans and other members of the genus *Homo*. While genetic material for such species has been available for years, the reliable amplification and sequencing of DNA extracted from ancient bone samples has not been tenable until recently. Careful collection procedures, performed under exceedingly pristine conditions, have enabled 1.3x coverage from three Neanderthal individuals (Green et al. 2006; Noonan et al. 2006) and the 1.9x coverage from a small finger bone found in the Denisova cave in Siberia (Reich et al. 2010). These genomes are on average slightly more related to each other than to modern human genomes, but most genomic regions still fall within the variation of modern humans (Reich et al. 2010). Interestingly, those regions where this is not the case, i.e., where all modern humans are closely related to each other than to Denisovans or Neanderthals, are enriched for regions that have been positively selected after the population split some 270,000–440,000 years ago (Green et al. 2006). While a comprehensive comparison of human and Neanderthal DNA sequence has the potential to identify the relatively small number of genetic changes that occurred over the span of time in which *H. sapiens* evolved into a distinct species. Further data and the identification of additional fossils will lead to considerably better assemblies of these ancient genomes and 30x coverage data for Denisovans was recently made available (Meyer et al. 2012). Although it is unlikely that endogenous DNA sequences can be obtained from much older hominin fossils, the unexpected finding of Denisovans allows optimism that genomes from more hominins can be discovered and will improve our understanding of human evolution and even some aspects of human disease.

2.2.2 Genetic Studies of Complex Traits

Perhaps the greatest impact of the HGP has been on the manner in which researchers investigate the causes of complex human diseases. Unlike monogenic diseases, which arise due to a single genetic aberration, complex diseases result from a complicated interaction of multiple genetic and environmental determinants, none of which are amenable to identification and characterization using the traditional approaches to monogenic disease gene discovery. Completion of the HGP gave rise to the development of efforts and technology to characterize genetic variation on a genome-wide scale, including the genotyping of common variants, which has led directly to the application of whole genome association studies to identify common alleles which contribute to complex disease risk, or the very recent whole genome sequencing efforts to identify low-frequency and rare variants in diverse populations. Each of these areas is discussed in the following sections.

2.2.2.1 The International HapMap Project

The sequence data resulting from the HGP paved the way for the development of an effort lead by the International HapMap Consortium to characterize all common variation within the human genome (International HapMap Consortium 2005). The most common type of genetic variant is the SNP, which occurs with the presence of two or more different alleles at the same nucleotide position. In humans, polymorphisms occur at a rate of approximately one variant every kilobase (Wang et al. 1998; Lander et al. 2001), and the presence of 11 million SNP sites with a minimal minor allele frequency of 1 % that constitute ~90 % of the variation in the world's population has been estimated (Kruglyak and Nickerson 2001).

The HapMap Project, currently completed phase III, was officially launched in 2002 to create a public, genome-wide database of common

human sequence variation, providing information needed as a guide to genetic studies of clinical phenotypes and consists of collaborators from the United States, Canada, the United Kingdom, China, Nigeria, and Japan (International HapMap Consortium 2003).

The Phase I of the HapMap Project contains high-quality genotype data on more than 1 million SNPs, genotyped on 270 samples from 90 individuals (30 parent–parent–offspring trios) of European descent from Utah (CEU), 90 Yoruba individuals (30 trios) from Ibadan, Nigeria (YRI), 45 unrelated Japanese from Tokyo (JPT), and 45 unrelated Han Chinese from Beijing (CHB). Although the goal of Phase I was to genotype at least one common SNP (minor allele frequency ≥0.05) every 5 kb across the genome and SNP selection was agnostic to functional annotation, 11, 500 nonsynonymous SNPs are prioritized in choosing SNPs for each 5 kb region (International HapMap Consortium 2005).

The Phase I HapMap Project data had a central role in the development of methods for the design and analysis of Genome-Wide Association (GWA) studies. For example, the HapMap resource provides critical information regarding the extent of linkage disequilibrium among SNPs in each of the four distinct populations represented in the project. In this way, knowledge of a particular SNP allele at one site can predict specific alleles at nearby sites (allele combinations along a chromosome are known as haplotypes). Approximately, 50–75 % of all SNPs in the HapMap database are highly correlated with other genotyped markers and >90 % are associated with nearby SNPs at levels of statistical significance (International HapMap Consortium 2005). These advances, alongside the release of commercial platforms for performing economically viable genome-wide genotyping, have led to a new phase in human medical genetics.

Large-scale GWA studies have identified novel loci involved in multiple complex diseases (Altshuler and Daly 2007; Bowcock, 2007). In addition, the HapMap data have led to novel insights into the distribution and causes of recombination hotspots (International HapMap Consortium 2005, Myers et al. 2005), the prevalence of structural variation (Conrad et al. 2006; McCarroll et al. 2006), and the identity of genes that have experienced recent adaptive evolution (International HapMap Consortium 2005; Voight et al. 2006).

In Phase II of the HapMap project an additional 2.1 million SNPs were genotyped on the same individuals from Phase I. The resulting HapMap Phase I and II datasets (3.1 million SNPs) constitute ∼25–30 % of the 9–10 million estimated common SNPs (minor allele frequency ≥0.05) in the assembled human genome. The Phase II HapMap differs from the Phase I not only in SNP spacing, but also in minor allele frequency (MAF) distribution and patterns of linkage disequilibrium. Because the criteria for choosing additional SNPs did not include consideration of SNP spacing or preferential selection for high MAF, the SNPs added in Phase II are, on average, more clustered and have lower MAF than the Phase I SNPs. One notable consequence is that the Phase II HapMap includes a better representation of rare variation than the Phase I HapMap (International HapMap Consortium 2007). The HapMap dataset and other resources such as public catalog of variant sites (dbSNP) and databases of structural variants (SVs) have driven disease gene discovery in the first generation of GWA studies, wherein genotypes at several hundred thousand variant sites, combined with the knowledge of LD structure, allowed the vast majority of common variants (MAF ≥ 0.05) to be tested for association with disease (International HapMap Consortium 2007). Over 6–7 years, GWA studies have identified more than a thousand genomic regions associated with disease susceptibility and other common traits (Hindorff et al. 2012). Genome-wide collections of both common and rare SVs have similarly been tested for association with disease (Wellcome Trust Case Control Consortium 2010). Despite successes, these studies raise many questions, such as why the identified variants have low-associated risks and account for so little heritability (Goldstein 2009). Explanations for this apparent gap are being sought. It is possible that these studies were limited with respect to variant type, frequency, and population

diversity. Only common DNA variants (MAF ≥ 0.05) have been well studied, even though the contributions of rare variants, which were not captured by GWA studies; SVs, which were poorly captured, and other forms of genomic variation; or interactions between genes or between genes and environmental factors may be important (Manolio et al. 2009). Furthermore, despite their value in locating the vicinity of genomic variants that may be related to the susceptibility to disease, few of the SNPs identified in GWA studies have clear functional implications that are relevant to mechanisms of disease (Hindorff et al. 2009). Narrowing an implicated locus to a single variant with direct functional consequences has proven challenging. Together, these findings suggest that additional work will be necessary to achieve a deep understanding of the genetic contribution to human phenotypes and diseases (Manolio et al. 2009).

Once a region has been identified as harboring a risk locus, a detailed study of all genetic variants in the locus is required to discover the causal variant(s), to quantify their contribution to disease susceptibility, and to elucidate their roles in functional pathways. A much more complete catalog of human DNA variation is a prerequisite to fully understanding the role of common and low-frequency variants in human phenotypic variation. The efforts aimed at illuminating the gaps in the first generation of databases that contain mostly common variant sites were made. The HapMap project was expanded into Phase III to perform genome-wide SNP genotyping and CNP detection, as well as polymerase chain reaction (PCR) resequencing in selected genomic regions on a larger set of 1,184 samples from 11 populations (International HapMap3 Consortium 2010). Also during the same time another consortium project called "1,000 Genomes" aimed to discover additional genotypes and to provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations by next generation sequencing, was initiated (1000 Genomes Project Consortium 2010).

The HapMap Phase III. Despite great progress in identifying genetic variants influencing human diseases, most inherited risk remains unexplained. A more comprehensive strategy that fully examines the low-frequency and rare variants in populations of diverse ancestry is required to understand the genetic architecture of human diseases. Accordingly, the HapMap Phase I and II resources were expanded by genotyping 1.6 million SNPs and CNP detection in 1,184 samples from 11 populations. These included all Phase I and II samples, along with additional samples from the same four populations (i.e., samples from 165 individuals (trios) of European descent from Utah (CEU), 167 Yoruba individuals (trios) from Ibadan, Nigeria (YRI), 86 unrelated Japanese from Tokyo (JPT), and 84 unrelated Han Chinese from Beijing (CHB)), and an additional 682 samples from seven new populations (i.e., 83 individuals (trios) of African ancestry from southwestern USA (ASW); 85 unrelated Chinese individuals from metropolitan Denver, Colorado, USA (CHD); 88 unrelated Gujarati Indian individuals from Houston, Texas, USA (GIH); 90 unrelated Luhya individuals from Webuye, Kenya (LWK); 171 Maasai individuals (trios + unrelated) from Kinyawa, Kenya (MKK); 77 unrelated individuals of Mexican ancestry from Los Angeles, California, USA (MXL); and 88 unrelated Tuscan individuals from Italy (Toscani in Italia, TSI). The new populations were included to provide further variation data from each of the three continental regions, as well as data from some admixed populations. Unlike Phase I and II, a much larger sample size of 692 unrelated individuals from ten populations (i.e., ASW, CEU, CHB, CHD, GIH, JPT, LWK, MXL, TSI, and YRI) were sequenced for 100 kb each of the ten ENCODE regions (see International HapMap 3 Consortium 2010 publication for details) by direct PCR-Sanger capillary sequencing in the Phase III. This direct sequencing of the selected regions, unlike SNPs genotyped using microarray platforms, which are intentionally biased toward high frequency by the discovery and selection process, the SNPs discovered by sequencing provide a direct estimate of the underlying allele frequency spectrum in each population. As in previous phases, common (MAF ≥ 0.05) and low-

frequency (MAF = 0.005–0.05) variants account for the vast majority of the heterozygosity in each sample, but a large number of rare (MAF = 0.0005–0.005) and private (singletons and MAF < 0.0005) variants were also observed. Each population had 42–66 % of sites with a MAF < 0.05, compared to 10–13 % in the genotyping data; 37 % of SNPs with a MAF < 0.005 were observed in only one population. In total, 77 % of the discovered SNPs were new (that was, not in the SNP database (dbSNP) build 129) and 99 % of those had a MAF < 0.05 (International HapMap 3 Consortium 2010). The HapMap Phase III results underscored the need to characterize population-specific parameters, and for each stratum of allele frequency. As expected, lower frequency variation is less shared across populations, even closely related ones, highlighting the importance of sequencing and sampling widely to achieve a comprehensive understanding of human variation. With improvement in sequencing technology, whole genome sequencing is becoming increasingly accessible. This revolution will no doubt expand our ability to identify rare and private variations along with common variations to better understand the genetic architecture of human diseases.

2.2.2.2 The 1000 Genomes Project

Launched in 2008, the 1000 Genomes Project involving researchers from more than 75 institutions and companies in the United States, the United Kingdom, China, and Germany, set its sights on characterizing over 95 % of variants that have allele frequency of 1 %, or higher (MAF ≥ 0.01) in the five major population groups-West African, European, North American, and East and South Asian. The coding region of the genome was cataloged for variants of even lower allele frequencies (i.e., MAF ≥ 0.001) because coding regions will more often have variants with functional consequences, which may also have low allele frequency (1000 Genomes Project Consortium 2010; Patterson 2011).

The pilot phase of the project aimed at developing and comparing genome-wide sequencing strategies, sequenced three sets of samples at three different levels of sequencing coverage.

- Family trios: high coverage (average 42x) whole genome sequencing of two HapMap family trios (i.e., one YRI and one CEU).
- Low coverage: low coverage (2–6x) whole genome sequencing of 179 unrelated individuals from four HapMap populations (i.e., 59 from YRI, 60 from CEU, 30 from CHB, and 30 from JPT).
- Exon sequencing: targeted capture of the exons from nearly1,000 randomly selected protein-coding genes (total 1.4 Mb) followed by sequencing at high coverage (average > 50 x) in 697 individuals from 7 HapMap populations (i.e., YRI, LWK, CEU, TSI, CHB, JPT, and CHD).

The pilot project identified 15 million SNPs, 1 million short insertions and deletions of DNA, and 20,000 large SVs. Populations of African ancestry contributed the largest number of variants to the data, including the biggest portion of novel variants (1000 Genomes Project Consortium 2010). The pilot project data also showed that more than half of the genetic variants that were found were previously unknown. It has also been observed that an individual's genome contains many variants of functional consequence (10,000–11,000 nonsynonymous sites and 10,000–12,000 synonymous sites per genome that differs from reference). However, the number of variants with greater functional impact is much smaller (overall 340–400 premature stop codons, splice site disruptions, and frame shifts, affecting 250–300 genes per genome, as putative LOF variants). In addition, 50–100 of the variants had previously been associated with an inherited disease (1000 Genomes Project Consortium 2010).

The success of the pilot project paved the way for the production phase of the full 1000 Genomes Project, which aims to sequence 2,500 genomes from 27 populations worldwide. The data on genomes of 1,092 individuals from