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.

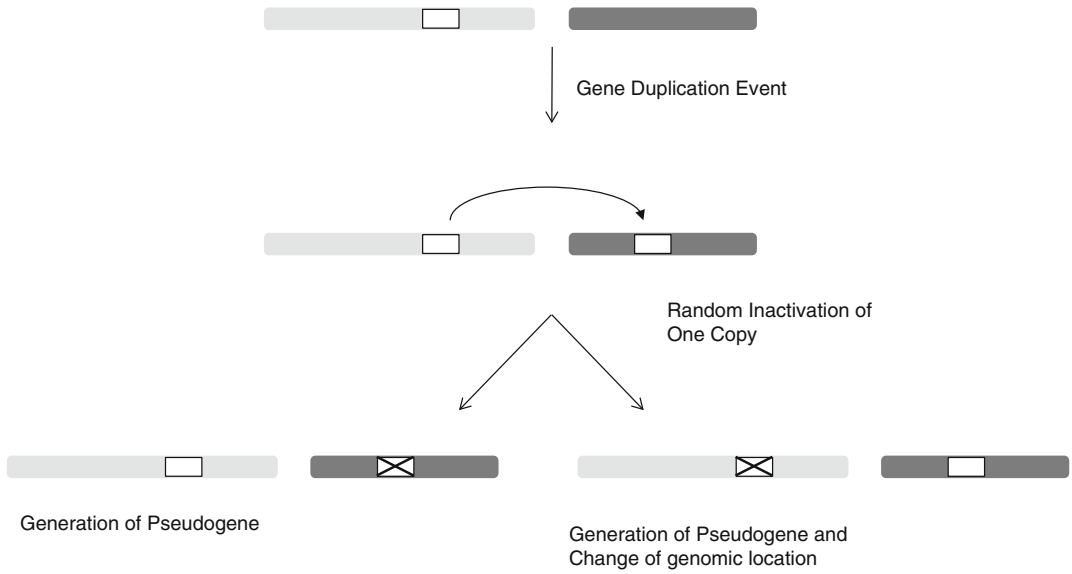

Fig. 2.1 Pseudogene generation by gene duplication and random inactivation. The creation of a novel pseudogene is initiated by a gene duplication event in which a sequence containing a functional gene (*white box*) is duplicated and inserted into a separate site in the genome (shown here as a duplication 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

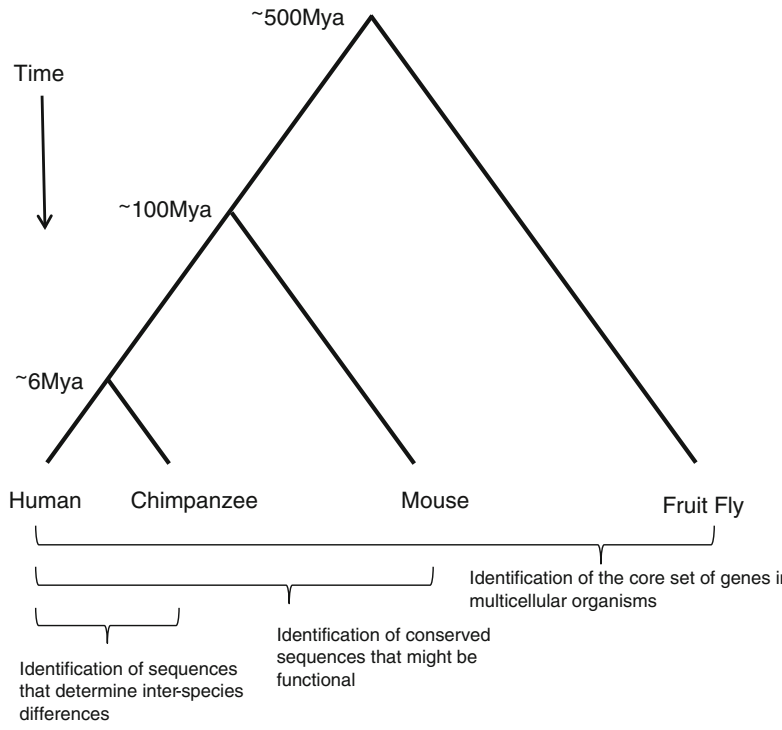

Fig. 2.2 Comparative genomics of species at different evolutionary distances. Genomic comparison of two species can yield different conclusions depending on the degree of genetic difference between them. The evolutionary tree shows the estimated time (in millions of years) from the divergence of human, 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;

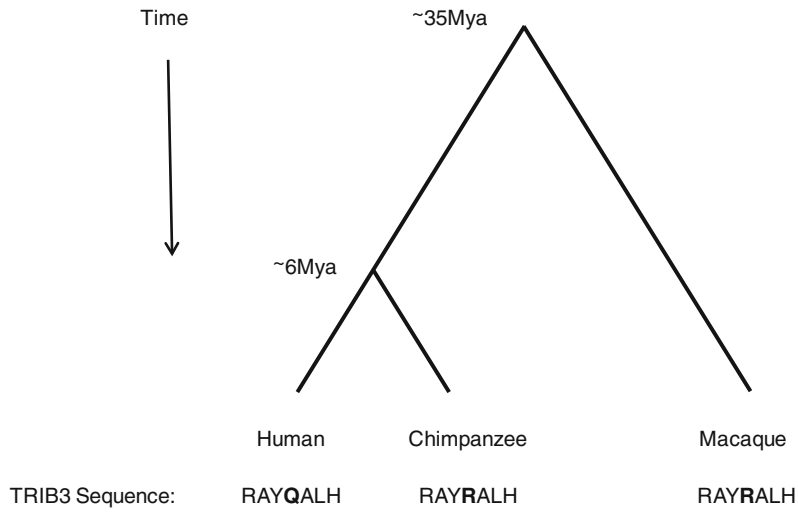

Fig. 2.3 Ternary analysis of closely related primate species. Evolutionary triangulation can identify the lineage in which a sequence variant evolved. The evolutionary tree shows the estimated time (in millions of years) from the divergence of human, chimp, and macaque from their common ancestor. As 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

14 populations from Europe, East Asia, sub-Saharan Africa, and the Americas, sequenced using combination of whole genome low coverage sequencing (2–6 x) and targeted deep sequencing (50–100 x) of the exome have been published recently (1000 Genomes Project Consortium 2012). The dataset provides a detailed view of variations across several populations. Individuals from different populations carry different profiles of rare and common variants, and low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. Most common variants (94 % with MAF ≥ 0.05) were previously known and their haplotype structures were also mapped through earlier projects (International HapMap Consortium 2007; 1000 Genomes Project Consortium 2010). In contrast, only 62 % of variants with the MAF range 0.005–0.05 and 13 % with MAF ≤ 0.005 had been described previously. A validated haplotype map of 38 million SNPs, 1.4 million short indels, and more than 14,000 larger deletions has been developed using this phase dataset. The Phase I data also show that at the most highly conserved coding sites, 85 % of the nonsynonymous variants and more than 90 % of stop-gain and splice disrupting variants are below 0.5 % in frequency, compared with 65 % of synonymous variants (1000 Genomes Project Consortium 2012).

2.3 Future Impact of the HGP: Where Are We Going?

2.3.1 Pharmacogenetics

Response to pharmacological interventions is variable and in most cases, difficult to predict. For instance, only about 66 % of individuals treated with beta blockers actually respond in the intended way with a reduction in blood pressure (Abbott 2003). In general, individuals can respond to drug treatment in one of three ways: favorably (i.e., as expected), unfavorably (i.e., adversely or with a blunted response), or not at all. Many factors, including age, ethnic background, gender, diet, interactions with other pharmaceuticals, and clearance function, influence the manner in which an individual will respond to a drug. In addition to these determinants, genetic factors are also known to impact the degree to which an individual will respond to a drug. The prediction of drug response based upon genetic variation has evolved into the field of pharmacogenetics. The closely related discipline of pharmacogenomics encompasses pharmacogenetics, but incorporates analysis of gene expression to understand genotype-drug interaction; thus, the main differences between the two disciplines lie mainly in the underlying technologies and the level at which a given gene is investigated. Because the focus of this chapter is the genome, we will address the intersection between genotype and drug response from a perspective favoring the field of pharmacogenetics.

Although the contribution of genetic variation on drug response has been recognized for decades, the availability of human genome reference sequence and a catalog of common genetic variation in the human genome has expanded the field tremendously (Collins et al. 2003). Indeed, it is in the field of pharmacogenetics that the clinical applicability of the HGP and HapMap resources has had the most impact. A number of genetic variants have been identified that at least partially predict drug response, including associations between *HLA-B* alleles and hypersensitivity to the anti-HIV therapeutic, abacavir (Hetherington et al. 2002). Among these patients, 46 % of individuals who had previously suffered an adverse immunological reaction to abacavir possessed the *HLA-B57* variant, compared to only 4 % of individuals who were not hypersensitive to the drug (Hetherington et al. 2002).

A specific haplotype within the vitamin K epoxide reductase gene (*VKORC1*) has been found to predict 21–25 % of required warfarin dose. When *VKORC1* haplotype is combined with genotypes in the cytochrome P450, subfamily IIC, polypeptide 9 gene (*CYP2C9*), 31 % of warfarin dose can be predicted (Rieder et al. 2005). This finding is particularly significant because warfarin, the most commonly prescribed anticoagulant, has a narrow therapeutic index and

requires careful and regular monitoring. Dosing above the required concentration produces potentially life-threatening side effects, while dosing below delays therapeutic benefit. The use of *VKORC1* and *CYP2C9* genotypes, combined with age, sex, body, and surface area, can predict up to 60 % of warfarin dose, thereby better ensuring achievement of optimal therapeutic dose (Rieder et al. 2005; Marsh and McLeod 2006).

Clopidogrel therapy improves cardiovascular outcomes in patients with acute coronary syndromes and following percutaneous coronary intervention by inhibiting adenosine diphosphate (ADP)-dependent platelet activation. However, nonresponsiveness to the drug is widely recognized and is related to recurrent ischemic events. The cytochrome P450 2C19 (*CYP2C19*) and *ABCB1* genotypes were found to be associated with platelet response to clopidogrel treatment and in the prediction of major cardiovascular events beyond stent thrombosis in coronary patients treated with clopidogrel (Shuldiner et al. 2009; Mega et al. 2010). Similarly, it has been shown that in patients with diabetes, vitamin E significantly increases HDL function in haptoglobin 2-2 but significantly decreases HDL function in haptoglobin 2-1. Thus, vitamin E therapy provides cardiovascular protection to individuals with the haptoglobin 2-2 genotype, but appears to increase cardiovascular risk in individuals with the haptoglobin 2-1 genotype. This pharmacogenetic interaction was paralleled by similar nonsignificant trends in HDL-associated lipid peroxides, glutathione peroxidase, and inflammatory cargo (Farbstein et al. 2011).

Pharmacogenetics is a rising concern in clinical oncology, because the therapeutic window of most anticancer drugs is narrow and patients with impaired ability to detoxify drugs will undergo life-threatening toxicities. In particular, genetic deregulations affecting genes coding for DPD, UGT1A1, TPMT, CDA, and CYP2D6 are now considered as critical issues for patients treated with 5-FU/capecitabine, irinotecan, mercaptopurine/azathioprine/thiopurine, gemcitabine/capecitabine/AraC, and tamoxifen, respectively (Evans 2004; Marques and Ikediobi 2010; Yang et al. 2011; O'Donnell and Ratain 2012). Examples like this serve to underscore the reality that the real clinical impact of pharmacogenetics will be in identifying those patients who are most likely to experience the desired therapeutic effect from the drug under consideration. For these individuals, quicker control of disease symptoms, reduced likelihood of adverse events, and better disease management will be provided by pharmacogenetics. Together, these factors will also impact public health by decreasing healthcare costs.

2.3.2 Nutrigenetics and Nutrigenomics

Nutrigenetics is the study of the relationship between genetic variation and metabolic, biochemical, or physiological response to foods. The related field of nutrigenomics comprises nutrient impact at the levels of gene expression, transcript stability, and posttranslational modifications (Young 2002; Ghosh et al. 2007). Completion of the HGP and availability of sequence variants have significantly fueled the development of these complementary disciplines; similar to the promise of pharmacogenetics, both nutrigenetics and nutrigenomics have the potential to influence the development of "personalized" nutrition by delineating dietary composition based upon specific genotype.

Several variants have been found to impact upon the metabolism of various dietary components (Ghosh et al. 2007; Raqib and Cravioto 2009). For example, individuals with phenylketonuria, an autosomal recessive disorder characterized by a deficiency in phenylalanine hydroxylase, are unable to metabolize phenylalanine and in the presence of foods high in this amino acid, such as meats, nuts, cheese, and the artificial sweetener aspartame, develop severe neurological disorders, including mental retardation. Simple avoidance of such foods prevents significant medical problems for patients with this genetic susceptibility. Likewise, variants in *HLA DQ2* and *DQ8* have been linked with gluten

in the development of celiac disease; more than 95 % of celiac patients are positive for either DQ2 or DQ8 (Sollid and Lie 2005). For individuals with these risk allelels, a gluten-free diet is recommended for disease management.

Considerable evidence also suggests that epigenetic abnormalities induced by diet are also among the most important factors affecting cancer risk. At least four distinct processes are involved with epigenetics: DNA methylation, histone modifications, microRNAs as well as other noncoding regulatory RNA, and chromatin modeling (Ross 2007). Some of the strongest data linking diet to epigenetic events come from studies with the agouti mouse model. Adding dietary factors (i.e., choline, betaine, or folic acid), which enhance methylation, to the maternal diet of pregnant agouti dams leads to a change in the phenotype of some of the offspring (Dolinoy 2008). Interestingly, adding genistein, which does not provide methyl groups, also leads to a change in the phenotype from a yellow to more agouti offspring (Dolinoy et al. 2006). Most importantly, these shifts in coat color are accompanied by a reduction in the risk of cancer, diabetes, and obesity. The shift in obesity in these animals is noteworthy because of the worldwide obesity epidemic. Such findings should serve as justification for additional attention to bioenergetic-epigenetic interrelationships, especially those that are modified by dietary factors.

Myzak and Dashwood (2006) have demonstrated that sulphoraphane, butyrate, and allyl sulfur are effective inhibitors of histone deacetylase (HDAC). HDAC inhibition was associated with global increases in histone acetylation, enhanced interactions of acetylated histones with the promoter regions of the *P21* and *BAX* genes, and elevated expression of p21Cip1/Waf1 and BAX proteins. Importantly, sulphoraphane has been reported to reduce HDAC activity in humans (Myzak et al. 2006). Future research likely needs to relate HDAC changes in humans to a change in cancer-related processes. Furthermore, since acetylation is only one method to regulate histone homeostasis (Ross 2007), greater attention needs to be given to how nutrition might influence the other types of histone modifications (Fenech et al. 2011).

In addition to the development of nutrient-related diseases, genetic variants can also interact with dietary components to produce subtle effects on metabolism. For example, a dose-dependent interaction between variants in the *APOA5* gene and dietary fat intake was found to increase risk for obesity in participants of the Framingham Heart Study (Corella et al. 2007). Similarly, individuals with the AA genotype at the G(-6)A marker in the angiotensinogen gene, which is associated with both higher circulating levels of angiotensinogen and elevated blood pressure, were more responsive to the effects of a diet high in fruits and vegetables and low in fat compared to individuals with the GG genotype (Svetkey et al. 2001). Other studies have found relationships between specific genetic variants and responsiveness to dietary components, and provide support for a role of dietary shifts in shaping human evolution. Perry et al. (2007) reported that individuals from populations with a typically high-starch diet (i.e., European Americans, Japanese, and Hadza hunter-gatherers) have more copies of the salivary amylase gene, which breaks down starch, compared to those from populations with a low-starch diet (i.e., Biaka, Mbuti, Datog pastoralists, and the Yakut). This finding is one of the first examples of positive selection on copy number variant, and further supports the idea that individuals may respond quite differently to the same diet given their respective genetic backgrounds.

2.4 Conclusions

The completion of HGP represents one of the momentous projects of modern scientific research. Delineation of the human genome sequence has consequently led to a greater understanding of human genetics and fueled the development of such diverse disciplines as comparative genomics, pharmacogenetics, and nutrigenomics. The fruits of the HGP directly contributed to the creation of the HapMap and the 1000 Genomes projects, which has since

provided the basis for WGA studies. Results from these investigations will be instrumental in the elucidation of the genetic variants that contribute to the development of complex diseases such as cancer, diabetes, autoimmune syndromes, and neurological disorders. Thus, the HGP has produced a significant impact upon a variety of different areas, and in completely unexpected ways.

References

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hur es ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65

Abbott A (2003) With your genes? Take one of these, three times a day. Nature 425:760–762

Adams MD, Celniker SE, Holt RA et al (2000) The genome sequence of Drosophila melanogaster. Science 287:2185–2195

Altshuler D, Daly M (2007) Guilt beyond a reasonable doubt. Nat Genet 39:813–815

Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84:775–790

Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) Recent segmental duplications in the human genome. Science 297:1003–1007

Bailey JA, Eichler EE (2006) Primate segmental duplications: crucibles of evolution, diversity and disease. Nat Rev Genet 7:552–564

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. Science 304:1321–1325

Bowcock AM (2007) Genomics: guilt by association. Nature 447:645–646

Chen FC, Li WH (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. Am J Hum Genet 68:444–456

Chen CT, Wang JC, Cohen BA (2007) The strength of selection on ultraconserved elements in the human genome. Am J Hum Genet 80:692–704

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154

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

Choudhuri S (2003) The path from nuclein to human genome: a brief history of DNA with a note on human genome sequencing and its impact on future research in biology. Bull Sci Technol Soc 23:360–367

Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci USA 104:19428–19433

Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, Ferriera S, Wang G, Zheng X, White TJ, Sninsky JJ, Adams MD, Cargill M (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. Science 302:1960–1963

Cohen J (2007) Genomics. DNA duplications and deletions help determine health. Science 317:1315–1317

Collins FS, Green ED, Guttmacher AE, Guyer MS, US National Human Genome Research Institute (2003) A vision for the future of genomics research. Nature 422:835–847

Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK (2006) A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 38:75–81

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J, Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, Hurles ME (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–712

Cook-Deegan RM (1989) The Alta summit, December 1984. Genomics 5:661–663

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 40:1199–1203

Corella D, Lai CQ, Demissie S, Cupples LA, Manning AK, Tucker KL, Ordovas JM (2007) APOA5 gene variation modulates the effects of dietary fat intake on body mass index and obesity risk in the Framingham Heart Study. J Mol Med (Berl) 85:119–128

Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). Science 302:1033–1035

Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL (2006) Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ Health Perspect 114:567–572

Dolinoy DC (2008) The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. Nutr Rev 66(Suppl 1):S7–S11

Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet 70:1490–1497

Elango N, Thomas JW, NISC Comparative Sequencing Program, Yi SV (2006) Variable molecular clocks in hominoids. Proc Natl Acad Sci USA 103:1370–1375

Enard W (2012) Functional primate genomics–leveraging the medical potential. J Mol Med 90:471–480

ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. Science 306:636–640

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. Nature 447:799–816

ENCODE Project Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol 9:e1001046. doi:10.1371/journal.pbio.1001046

ENCODE Project Consortium, Dunham I, Kundaje A et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74

Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui LC (2002) Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. Hum Mol Genet 11:1987–1995

Evans WE (2004) Pharmacogenetics of thiopurine S-methyltransferase and thiopurine therapy. Ther Drug Monit 26:186–191

Farbstein D, Blum S, Pollak M, Asaf R, Viener HL, Lache O, Asleh R, Miller-Lotan R, Barkay I, Star M, Schwartz A, Kalet-Littman S, Ozeri D, Vaya J, Tavori H, Vardi M, Laor A, Bucher SE, Anbinder Y, Moskovich D, Abbas N, Perry N, Levy Y, Levy AP (2011) Vitamin E therapy results in a reduction in HDL function in individuals with diabetes and the haptoglobin 2-1 genotype. Atherosclerosis 219:240–244

Fenech M, El-Sohemy A, Cahill L, Ferguson LR, French TA, Tai ES, Milner J, Koh WP, Xie L, Zucker M, Buckley M, Cosgrove L, Lockett T, Fung KY, Head R (2011) Nutrigenetics and nutrigenomics: viewpoints on the current status and applications in nutrition research and practice. J Nutr Nutr 4:69–89

Flicek P, Aken BL, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Koscielny G, Kulesha E, Lawson D, Longden I, Massingham T, McLaren W, Megy K, Overduin B, Pritchard B, Rios D, Ruffier M, Schuster M, Slater G, Smedley D, Spudich G, Tang YA, Trevanion S, Vilella A, Vogel J, White S, Wilder SP, Zadissa A, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Smith J, Searle SM (2010) Ensembl's 10th year. Nucleic Acids Res 38:D557–D562. doi:10.1093/nar/gkp972

Ghosh D, Skinner MA, Laing WA (2007) Pharmacogenomics and nutrigenomics: synergies and differences. Eur J Clin Nutr 61:567–574

Girirajan S, Campbell CD, Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45:203–226

Goldstein DB (2009) Common genetic variation and human traits. N Engl J Med 360:1696–1698

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Paabo S (2006) Analysis of one million base pairs of Neanderthal DNA. Nature 444:330–336

Gu Z, Rifkin SA, White KP, Li WH (2004) Duplicate genes increase gene expression diversity within and between species. Nat Genet 36:577–579

Gu X, Zhang Z, Huang W (2005) Rapid evolution of expression and regulatory divergences after yeast gene duplication. Proc Natl Acad Sci USA 102:707–712

Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, Castelo R, Eyras E, Ucla C, Gingeras TR, Harrow J, Hubbard T, Lewis SE, Reese MG (2006) EGASP: the human ENCODE genome annotation assessment project. Genome Biol 7(Suppl 1:S2):1–31

Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 (Suppl 1:S4):1–9

Hetherington S, Hughes AR, Mosteller M, Shortino D, Baker KL, Spreen W, Lai E, Davies K, Handley A, Dow DJ, Fling ME, Stocum M, Bowman C, Thurmond LM, Roses AD (2002) Genetic variations in HLA-B region and hypersensitivity reactions to abacavir. Lancet 359:1121–1122

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, Manolio TA (2012) A catalog of published genome-wide association studies. www.genome.gov/gwastudies. Accessed (2012)

Hirano T (2000) Chromosome cohesion, condensation, and separation. Annu Rev Biochem 69:115–144

Hittinger CT, Carroll SB (2007) Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681

Horvath JE, Willard HF (2007) Primate comparative genomics: lemur biology and evolution. Trends Genet TIG 23:173–182

Hua S, Kittler R, White KP (2009) Genomic antagonism between retinoic acid and estrogen signaling in breast cancer. Cell 137:1259–1271

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) Detection of large-scale variation in the human genome. Nat Genet 36:949–951

International HapMap Consortium (2003) The International HapMap project. Nature 426:789–796

International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320

International HapMap Consortium, Frazer KA, Ballinger DG, et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861

International HapMap 3 Consortium, Altshuler DM, Gibbs RA, et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945

Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316:1497–1502

Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155–156

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14:331–342

Kapranov P, Willingham AT, Gingeras TR (2007) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413–423

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tuzun E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64

Kim VN, Nam JW (2006) Genomics of microRNA. Trends Genet 22:165–173

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318:420–426

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234–236

Kryukov GV, Schmidt S, Sunyaev S (2005) Small fitness effect of mutations in highly conserved non-coding regions. Hum Mol Genet 14:2221–2229

Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lander ES (2011) Initial impact of the sequencing of the human genome. Nature 470:187–197

Li WH, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res 13:347–357

Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, Eichler EE (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79:275–290

Louis EJ (2007) Evolutionary genetics: making the most of redundancy. Nature 449:673–674

Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA, Chakravarti A, Patel PI (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66:219–232

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

Marques-Bonet T, Ryder OA, Eichler EE (2009) Sequencing primate genomes: what have we learned? Ann Rev Genomics Hum Genet 10:355–386

Marques SC, Ikediobi ON (2010) The clinical application of UGT1A1 pharmacogenetic testing: gene-environment interactions. Hum Genomics 4:238–249

Marsh S, McLeod, HL (2006) Pharmacogenomics: from bedside to clinical practice. Hum Mol Genet 15(Spec No 1):R89–R93

Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM, International HapMap Consortium (2006) Common deletion polymorphisms in the human genome. Nat Genet 38:86–92

Mega JL, Close SL, Wiviott SD, Shen L, Walker JR, Simon T, Antman EM, Braunwald E, Sabatine MS (2010) Genetic variants in ABCB1 and CYP2C19 and cardiovascular outcomes after treatment with clopidogrel and prasugrel in the TRITON-TIMI 38 trial: a pharmacogenetic analysis. Lancet 376:1312–1319

Meyer M, Kircher M, Gansauge MT, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paabo S (2012) A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226

Mighell AJ, Smith NR, Robinson PA, Markham AF (2000) Vertebrate pseudogenes. FEBS Lett 468:109–114

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628

Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K et al (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Mural RJ, Adams MD, Myers EW et al (2002) A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. Science 296:1661–1671

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ (2001) Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324

Myzak MC, Dashwood RH (2006) Histone deacetylases as targets for dietary cancer preventive agents: lessons learned with butyrate, diallyl disulfide, and sulforaphane. Curr Drug Targets 7:443–452

Myzak MC, Hardin K, Wang R, Dashwood RH, Ho E (2006) Sulforaphane inhibits histone deacetylase activity in BPH-1, LnCaP and PC-3 prostate epithelial cells. Carcinogenesis 27:811–819

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Paabo S, Pritchard JK, Rubin EM (2006) Sequencing and analysis of Neanderthal genomic DNA. Science 314:1113–1118

O'Donnell PH, Ratain MJ (2012) Germline pharmacogenomics in oncology: decoding the patient for targeting therapy. Mol Oncol 6:251–259

O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker C, Rieder MJ, Nickerson DA, Bernier R, Fisher SE, Shendure J, Eichler EE (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43:585–589

Oak Ridge National Laboratory http://www.ornl.gov. Accessed 28 Mar 2013

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, Nosrat B, Markowitz VM, Kyrpides NC (2012) The Genomes OnLine Database (GOLD) v. 4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 40:D571–D5799. doi:10.1093/nar/gkr1100

Patterson K (2011) 1000 genomes: a world of variation. Circ Res 108:534–536

Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444:499–502

Pennisi E (2007) Genetics. Working the (gene count) numbers: finally, a firm answer? Science 316:1113

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, Carter NP, Lee C, Stone AC (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39:1256–1260

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, Noonan K, Gribble S, Prigmore E, Donahoe PK, Smith RS, Park JH, Hurles ME, Carter NP, Lee C, Scherer SW, Feuk L (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. Nat Biotechnol 29:512–520

Poser I, Sarov M, Hutchins JR, Heriche JK, Toyoda Y, Pozniakovsky A, Weigl D, Nitzsche A, Hegemann B, Bird AW, Pelletier L, Kittler R, Hua S, Naumann R, Augsburg M, Sykora MM, Hofemeister H, Zhang Y, Nasmyth K, White KP, Dietzel S, Mechtler K, Durbin R, Stewart AF, Peters JM, Buchholz F, Hyman AA (2008) BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals. Nat Methods 5:409–415

Prabhakar S, Noonan JP, Paabo S, Rubin EM (2006) Accelerated evolution of conserved noncoding sequences in humans. Science 314:786

Raqib R, Cravioto A (2009) Nutrition, immunology, and genetics: future perspectives. Nutr Rev 67(Suppl 2): S227–S236

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EK, Dallaire S, Freeman JL, Gonzalez JR, Gratacos M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery L, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengol L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, Hurles ME (2006) Global variation in copy number in the human genome. Nature 444:444–454

Reich D, Green RE, Kircher M, Krause J, Patterson N, Durand EY, Viola B, Briggs AW, Stenzel U, Johnson PL, Maricic T, Good JM, Marques-Bonet T, Alkan C, Fu Q, Mallick S, Li H, Meyer M, Eichler EE, Stoneking M, Richards M, Talamo S, Shunkov MV, Derevianko AP, Hublin JJ, Kelso J, Slatkin M, Paabo S (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053–1060

Rhesus Macaque Genome Sequencing and Analysis Consortium, Gibbs RA, Rogers J et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234

Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, McLeod HL, Blough DK, Thummel KE, Veenstra DL, Rettie AE (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352:2285–2293

Roberts L (2001) The human genome. Controversial from the start. Science 291:1182–1188

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657

Ross SA (2007) Nutritional genomic approaches to cancer prevention research. Exp Oncol 29:250–256

Sasaki E, Suemizu H, Shimada A, Hanazawa K, Oiwa R, Kamioka M, Tomioka I, Sotomaru Y, Hirakawa R, Eto T, Shiozawa S, Maeda T, Ito M, Ito R, Kito C, Yagihashi C, Kawai K, Miyoshi H, Tanioka Y, Tamaoki N, Habu S, Okano H, Nomura T (2009) Generation of transgenic non-human primates with germline transmission. Nature 459:523–527

Sasidharan R, Gerstein M (2008) Genomics: protein fossils live on as RNA. Nature 453:729–731

Schaner P, Richards N, Wadhwa A, Aksentijevich I, Kastner D, Tucker P, Gumucio D (2001) Episodic evolution of pyrin in primates: human mutations recapitulate ancestral amino acid states. Nat Genet 27:318–321

Schmid M, Jensen TH (2010) Nuclear quality control of RNA polymerase II transcripts. Wiley Interdiscip Rev RNA 1:474–485

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, Wigler M (2004) Large-scale copy number polymorphism in the human genome. Science 305:525–528

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88

Shuldiner AR, O'Connell JR, Bliden KP, Gandhi A, Ryan K, Horenstein RB, Damcott CM, Pakyz R, Tantry US, Gibson Q, Pollin TI, Post W, Parsa A, Mitchell BD, Faraday N, Herzog W, Gurbel PA (2009) Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA 302:849–857

Sollid LM, Lie BA (2005) Celiac disease genetics: current concepts and practical applications. Clin Gastroenterol Hepatol 3:843–851

Steiper ME, Young NM, Sukarna TY (2004) Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. Proc Natl Acad Sci USA 101:17021–17026

Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, Tsalenko A, Sampas N, Bruhn L, Shendure J, 1000 Genomes Project, Eichler EE (2010) Diversity of human copy number variation and multicopy genes. Science 330:641–646

Svetkey LP, Moore TJ, Simons-Morton DG, Appel LJ, Bray GA, Sacks FM, Ard JD, Mortensen RM, Mitchell SR, Conlin PR, Kesari M, DASH Collaborative Research Group (2001) Angiotensinogen genotype and blood pressure response in the Dietary Approaches to Stop Hypertension (DASH) study. J Hypertens 19:1949–1956

Taylor JS, Raes J (2004) Duplication and divergence: the evolution of new genes and old ideas. Annu Rev Genet 38:615–643

Thomas JW, Touchman JW, Blakesley RW et al (2003) Comparative analyses of multi-species sequences from targeted genomic regions. Nature 424:788–793

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727–732

UCSC Genome Bioinformatics (2013) http://genome.ucsc.edu/ Accessed 28 Mar 2013

Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291:1304–1351

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4:e72

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale

identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63

Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, et al (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464:713–720

Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5:19–21

Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. Proc Natl Acad Sci USA 104:7145–7150

Yang CG, Ciccolini J, Blesius A, Dahan L, Bagarry-Liegey D, Brunet C, Varoquaux A, Frances N, Marouani H, Giovanni A, Ferri-Dessens RM, Chefrour M, Favre R, Duffaud F, Seitz JF, Zanaret M, Lacarelle B, Mercier C (2011) DPD-based adaptive dosing of 5-FU in patients with head and neck cancer: impact on treatment efficacy and toxicity. Cancer Chemothe Pharmacol 67:49–56

Young VR (2002) 2001 W.O. Atwater memorial lecture and the 2001 ASNS president's lecture: human nutrient requirements: the challenge of the post-genome era. J Nutr 132:621–629

Zhao XD, Han X, Chew JL, Liu J, Chiu KP, Choo A, Orlov YL, Sung WK, Shahab A, Kuznetsov VA, Bourque G, Oh S, Ruan Y, Ng HH, Wei CL (2007) Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell 1:286–298

Zheng D (2008) Asymmetric histone modifications between the original and derived loci of human segmental duplications. Genome Biol 9:R105