Human genetics and genomics a decade after the release of the draft sequence of the human genome

*Nasheen Naidoo,[1] Yudi Pawitan,[2] Richie Soong,[3] David N. Cooper[4] and Chee-Seng Ku[1,3]**

[1]Centre for Molecular Epidemiology, Department of Epidemiology and Public Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[3]Cancer Science Institute of Singapore, National University of Singapore, Singapore
[4]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff, UK
**Correspondence to:* Tel: +65 (0)81388095; E-mail: g0700040@nus.edu.sg

Date received (in revised form): 15th May 2011

Abstract

Substantial progress has been made in human genetics and genomics research over the past ten years since the publication of the draft sequence of the human genome in 2001. Findings emanating directly from the Human Genome Project, together with those from follow-on studies, have had an enormous impact on our understanding of the architecture and function of the human genome. Major developments have been made in cataloguing genetic variation, the International HapMap Project, and with respect to advances in genotyping technologies. These developments are vital for the emergence of genome-wide association studies in the investigation of complex diseases and traits. In parallel, the advent of high-throughput sequencing technologies has ushered in the 'personal genome sequencing' era for both normal and cancer genomes, and made possible large-scale genome sequencing studies such as the 1000 Genomes Project and the International Cancer Genome Consortium. The high-throughput sequencing and sequence-capture technologies are also providing new opportunities to study Mendelian disorders through exome sequencing and whole-genome sequencing. This paper reviews these major developments in human genetics and genomics over the past decade.

Keywords: *Human Genome Project, International HapMap Project, 1000 Genomes Project, genome-wide association studies, single nucleotide polymorphisms, copy number variations, next-generation sequencing technologies, cancer genome sequencing, exome sequencing, complex disease, Mendelian disorders, personalised genomic medicine*

Introduction

Substantial progress has been made in human genetics and genomics research over the past 10 years since the publication of the draft sequence of the human genome.[1,2] The Human Genome Project (HGP) provided the basic raw DNA sequence that spawned a plethora of secondary studies which together greatly improved our knowledge of the architecture and function of the genome, yielding new insights with respect to (i) gene number and density, (ii) non-protein-coding RNA genes (or RNA genes), (iii) pervasive transcription, (iv) high copy number repeat sequences and (v) evolutionary conservation. These developments also have challenged the classical definition of the gene (see below).

In parallel, the design of studies investigating complex diseases and traits has gradually shifted from candidate-gene association and linkage studies to genome-wide association studies (GWASs). The

first proper GWAS study was published in 2005. This succeeded in identifying a common risk variant with a large effect size in the complement factor H (*CFH*) gene, which was associated with age-related macular degeneration.[3] By 2007, approximately 100 new GWASs had been published, relating to various complex diseases and traits.[4] There has, however, been some criticism of the inability of GWASs to identify many of the presumed disease-associated variants. Indeed, the validity of the common-disease common-variant (CD/CV) model has recently been challenged by virtue of the perceived 'missing heritability'.[5–7] This notwithstanding, the GWAS approach has dramatically changed the field of human disease genetics, from identifying mostly irreproducible disease associations in the pre-GWAS era to revealing thousands of statistically robust single nucleotide polymorphism (SNP) associations today.[8–11] The focus has also gradually shifted back to Mendelian disorders, with the advent of high-throughput sequence capture and sequencing technologies which have potentiated exome and whole-genome (re)sequencing (WGS).[12–16]

The rapid advances made in genotyping technologies over the past decade, from the arrival of the first 'whole-genome' SNP genotyping array (the Affymetrix GeneChip 10K [Affymetrix; Santa Clara, CA] in 2003) to current capacity able to genotype five million SNPs per array (Illumina Omni5.0 Beadchip [Illumina; San Diego, CA]),[17,18] have contributed substantially to GWASs (http://www.genome.gov/gwastudies). A total of 874 publications and 4,327 SNP associations with p-values $<1.0 \times 10^{-5}$ for approximately 500 complex diseases and traits had been included in the catalogue as of 13th May 2011.

The genotyping arrays have also contributed significantly to population genetics studies.[19–21] These arrays have been used to identify and characterise copy number variations (CNVs)[22,23] and regions of homozygosity (ROHs).[24,25] Research on CNVs and ROHs has also progressed rapidly since CNVs were first reported to be widespread in the human genome,[26,27] and ROHs have been found to be common in outbred populations.[28] In recognition of the progress achieved in the context of both GWASs and CNVs, 'human genetic variation' was considered the 'Breakthrough of The Year' in 2007 by *Science*.[4]

Advances have also been made in sequencing technologies, with the advent of the first next-generation sequencer in 2004 (Roche GS 20 System [Roche 454; Branford, CT]) and later, third-generation sequencing (TGS) technologies such as true single molecule sequencing (Helicos Biosciences, Cambridge, MA) and single molecule real-time sequencing (SMRT) (Pacific Biosciences Menlo Park, CA).[29–33] Developments of other more promising TGS or single-molecule sequencing technologies are on the horizon, such as nanopore sequencing and sequencing using transmission electron microscopy.[32,34–37] These developments have also marked the end of the era of the Sanger dideoxynucleotide or chain termination sequencing method, which has dominated the field since its introduction in 1977.[38]

The arrival of next-generation sequencing (NGS) technologies has also significantly changed the approaches applied in structural and functional genomics studies. Several microarray-based methods have been swiftly supplanted by sequencing-based approaches such as ChIP-Seq, RNA-Seq, Methyl-Seq and CNV-Seq (paired-end mapping [PEM] and depth-of-coverage approaches). Studies using these sequencing approaches have contributed significantly to both fields.[39–41] In addition to a variety of different applications in functional genomic studies, these sequencing technologies have also made it feasible, both technically and in terms of cost, to sequence a whole human genome within weeks, for tens rather than hundreds of thousands of US dollars.[42,43] Currently, the cost of WGS at several tenfold depth of sequencing coverage has been reduced to less than 5,000 US dollars.[44] The number of WGS studies for both normal and cancer genomes has grown rapidly over the past three years.[45] These studies have led to important discoveries in the context of both heritable genetic variation[42,43,46] and somatic mutations in cancer genomes.[47–49]

Such progress would not have been possible without the reference genome generated by the HGP. Also made possible by the high-throughput genotyping and sequencing technologies, several large-scale international projects have been launched, such as the International HapMap Project; the Encyclopedia of DNA Elements (ENCODE) Project, the 1000 Genomes Project, the International Cancer Genome Consortium, the National Institute of Health (NIH) Roadmap Epigenomics Program and the Human Microbiome Project. These projects have contributed substantially to our understanding and knowledge of human genetics and genomics.

This paper aims to review these major developments in human genetics and genomics over the past decade. Major developments and landmarks in human genetics and genomics are summarised in Table 1.

The HGP

Rapid progress has been made since the completion of the HGP, with the provision of a 'finished' reference DNA sequence for the human genome.[64] The project was initiated in 1990 and, upon its completion in 2003 it yielded important new insights into the architecture and function of the human genome. The sequencing of the HGP relied almost entirely upon the Sanger sequencing method.

The draft sequences of the HGP were imperfect because of the incomplete coverage of the euchromatic regions (euchromatin) — approximately 10 per cent of these regions were missing. In reality, the coverage was even less complete when the whole genome was considered (ie when the heterochromatic regions were included). Thus, in all, some 30 per cent of the genome was not initially covered. Furthermore, there was an extensive number of gaps between contigs, which rendered the genome sequence discontinuous.[1,2] The IHGSC subsequently published an improved version of the human genome sequence in 2004 and the HGP was then deemed to be 'complete'. This 'finished' version of the genome had achieved an almost complete coverage of all the euchromatic regions (ie approximately 99 per cent) and also significantly reduced the number of gaps between contigs to 341 from the initial hundreds of thousands.[64]

Significant further progress toward the total completion of the human genome sequence continued until 2006; the complete euchromatic sequences of all individual human chromosomes, including the annotation of genes and other features, have now been published (summarised in Table S1). Since November 2005, the National Center for Biotechnology Information (NCBI) Build 36 assembly of the human genome sequence has been available in public databases. The data comprise a reference assembly of the complete genome sequence plus the Celera WGS (Celera; Alameda, CA) and a number of alternative assemblies of individual haplotypic chromosomes or regions. The full list of assemblies in NCBI 36, as well as the genome sequences, is available through the following genome browsers:

- Ensembl (http://www.ensembl.org/)[114]
- UCSC (http://genome.ucsc.edu/)[115]
- NCBI (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606).[116]

Although both HGP and Celera Genomics had only sequenced the human haploid genome, the availability of the reference DNA sequence initiated a new era in the study of genetic variation and the functional characterisation of the human genome. The two global projects that subsequently ensued were the International HapMap Project and the ENCODE project.[63,65] The aim of the HapMap initiative was to validate several million SNPs that were identified during and after the completion of the HGP, and then to characterise the extent of their linkage disequilibrium (LD) patterns in populations of European, Asian and African ancestry. The ENCODE project was conceived to identify all the functional and regulatory elements in the human genome.

Table 1. Major developments and landmarks in human genetics and genomics, 1977 to date

Year	Development	References
1977	Sanger dideoxynucleotide/chain termination sequencing method developed	38
	Mammalian genes shown to contain introns	50
1978	First report of characterisation of gross gene deletions responsible for human inherited disease (α- and β-thalassaemia) by Southern blotting	51
1979	First single base-pair substitution causing a human inherited disease (β-thalassaemia) characterised by DNA sequencing	52
1980	Construction of a genetic linkage map in humans using restriction fragment length polymorphisms	53
1990	Initiation of the Human Genome Project (HGP)	54
1992	Second-generation linkage map of the human genome	55
1996	The Human Gene Mutation Database (HGMD), an attempt to collate known (published) gene lesions responsible for human inherited disease, established and made available at http://www.hgmd.org	56
	Genome-wide association studies (GWAS) approach for genetic studies of complex diseases first proposed	57
2001	Completion of draft DNA sequences of the human genome by the International Human Genome Sequencing Consortium (IHGSC) and Celera Genomics	1,2
	International SNP Map Working Group identifies 1.42 million SNPs in the human genome	58
	Genetic architecture of complex diseases subjected to intense debate	59,60
	Linkage disequilibrium (LD) patterns documented between SNPs in regions of the human genome	61,62
2003	Initiation of the International HapMap Project	63
	First whole-genome SNP genotyping array – Affymetrix GeneChip 10K	17
2004	IHGSC publishes the 'finished version' of the DNA sequence of the human genome	64
	Initiation of the ENCODE project	65
	Discovery of hundreds of copy number variations (CNVs) in the human genome	26,27
	Database of Genomic Variants (DGV) established to catalogue CNVs	27
	First new-generation sequencing (NGS) technology – Roche 454 GS 20 System	29,30
2005	Completion of the International HapMap Phase I Project	66
	First proper GWAS using a commercial whole-genome SNP genotyping array	3
2005-present	Rapid developments of whole-genome and custom SNP genotyping arrays and technologies	18
	Rapid developments of sequencing technologies	31,33
2006	Discovery of more than 1,000 regions of homozygosity >1 megabase (Mb) in the genomes of outbred populations	28
	First comprehensive map of CNVs in the HapMap populations	22
	An initial map of insertion and deletion variants in the human genome	67

Continued

Table 1. Continued

Year	Development	References
	Illumina sequencing platform commercially marketed	29,30
2007	The first human diploid genome (Craig Venter's genome) sequenced by the Sanger sequencing method	68
	Completion of the International HapMap Phase II Project and extension to Phase III	69
	Genome-wide detection and characterisation of positive selection in human populations	70
	Completion of the ENCODE project	71
	Explosion of GWAS publications ('Year of GWAS'), approximately 100 new GWASs	4
	'Human Genetic Variation' considered to be the 'Breakthrough of The Year' in 2007 by *Science*	4
	Sequence capture or enrichment methods and technologies developed	72–74
	Pervasive transcription documented	75
	Demonstration of paired-end mapping (PEM) to detect structural variation using NGS technologies	76
	Demonstration of ChIP-Seq to map transcription factor binding sites	77
	Demonstration of ChIP-Seq to interrogate histone modifications	78
	Life Technologies SOLiD sequencing platform commercially marketed	29,30
	A community resource project launched to sequence large-insert clones from many individuals, systematically discovering and resolving these complex variants at the DNA sequence level (The Human Genome Structural Variation Working Group)	79
2007–present	Microarray-based methods increasingly supplanted by sequencing-based approaches such as ChIP-Seq, RNA-Seq, Methyl-Seq and CNV-Seq	39,41,80,81
2008	First human diploid genome (James Watson's genome) sequenced by NGS technologies	46
	First whole cancer genome (acute myeloid leukaemia [AML]) sequenced	82
	Initiation of the 1000 Genomes Project	83
	Vast majority of human genes shown to undergo alternative splicing (RNA-Seq)	84,85
	Large scale mapping and sequencing of structural variation using a clone-based method	86
	Demonstration of depth-of-coverage approach to detect CNVs using NGS technologies	87
	First GWAS meta-analysis using imputation methods	88
	The issue of 'missing heritability' in GWASs recognised	89
2009	Feasibility of exome sequencing approach to identify a causal mutation for a Mendelian disorder first demonstrated	12
	Exome sequencing as a useful tool for diagnostic application demonstrated	90
	Third generation sequencing (TGS; single molecule sequencing) technology introduced — Heliscope Single Molecule Sequencer (Helicos Biosciences) commercially marketed	91

Continued

Table 1. Continued

Year	Development	References
	First human diploid genome sequenced by TGS technology	92
	Latest assembly of the human genome (Genome Reference Consortium, release GRCh37, February 2009), Genebuild published by Ensembl (database version 56.37a) includes 23,616 protein-coding genes, 6,407 putative RNA genes and 12,346 pseudogenes	http://www.ensembl.org/Homo_sapiens/Info/StatsTable
	Large intergenic non-coding RNAs (lincRNAs) found to represent a novel category of evolutionarily conserved RNAs	93,94
	Direct single molecule RNA sequencing without prior conversion of RNA to cDNA	95
	First human DNA methylomes at base resolution	96
	Comprehensive mapping of long-range chromatin interactions	97,98
2010	Number of disease-causing/disease-associated germline mutations collated in the Human Gene Mutation Database exceeds 100,000 in >3,700 different nuclear genes	99,100
	More than 17 million SNPs in the human genome catalogued in the SNP Database (dbSNP; http://www.ncbi.nlm.nih.gov/projects/SNP/)	101
	As of 2nd November 2010, DGV catalogued 66,741 CNVs, 953 inversions and 34,229 insertions and deletions (indels) (100 base pairs (bp) — 1 kilobase (kb) from 42 published studies	http://projects.tcag.ca/variation/
	1,048 microRNAs found in the human genome	miRBase, Release 16.0: September 2010, http://www.mirbase.org/
	Completion of the International HapMap Phase III Project	21
	Completion of pilot phase of the 1000 Genomes Project	102
	Second generation whole-genome SNP genotyping array (with SNP selection from the 1000 Genomes Project) launched	http://www.illumina.com/applications/gwas.ilmn
	Cost of whole-genome sequencing (at several tenfold of sequencing coverage depth) reduced to less than $5,000	44
	Metagenomic sequencing of human gut microbes accomplished using NGS technologies	103
	Exome sequencing study identifies causal mutations and genes for previously unexplained Mendelian disorders	13,14
	GWAS meta-analysis involving total sample size of >249,000	104
	Comprehensive mapping of CNVs using high-resolution tiling oligonucleotide microarrays (42 million probes)	105

Continued

Table 1. Continued

Year	Development	References
	Characterisation of 20 sequenced human genomes to evaluate the prospects for identifying rare functional variants	106
	Neanderthal genome sequenced	107
	The genome of an extinct Palaeo–Eskimo sequenced	108
	Exome sequencing of 200 individuals identifies an excess of low-frequency non-synonymous coding variants	109
	International Cancer Genome Consortium (ICGC) launched	110
	Largest GWAS of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls performed	111
2011	As of 13th May 2011, 874 publications and 4,327 SNPs documented in the National Human Genome Research Institute (NHGRI) 'A Catalog of Published Genome-wide Association Studies'	http://www.genome.gov/gwastudies/
	Comprehensive mapping of copy number variations based on whole-genome DNA sequencing data	112
	Developments of other TGS technologies, such as single-molecule real-time sequencing and nanopore sequencing, are on the horizon	32
	New addition to the NGS market — the Ion Torrent Personal Genome Machine (PGM), produced by Life Technologies (Carlsbad, CA)	http://www.iontorrent.com/
	Single-cell sequencing to infer tumour evolution	113

Architecture and function of the human genome

To coincide with the tenth anniversary of the release of the draft human genome sequences, the key findings from the HGP and their importance for the results of subsequent studies will now be recalled briefly. The findings emanating from the HGP and follow-on studies have had an enormous impact on the understanding of the architecture and function of the human genome.

Gene number and density

Initial annotation data indicated that the human genome encodes at least 20,000–25,000 protein-coding genes, with an indeterminate number of additional 'computationally derived genes' supported by somewhat weaker *in silico* evidence.[2,64] Many genes are now known to encode RNAs rather than proteins as their final products[117] but many still remain unannotated.[75] In the latest assembly of the human genome (Genome Reference Consortium, release GRCh37, February 2009), the Genebuild published by Ensembl (database version 56.37a) includes 23,616 protein-coding genes, 6,407 putative RNA genes and 12,346 pseudogenes (http://www.ensembl.org/Homo_sapiens/Info/StatsTable). The HUGO Human Gene Nomenclature Committee (http://www.genenames.org/index.html) has so far approved more than 28,000 human gene symbols, although some of these may yet turn out to correspond to functionally meaningless open reading frames.[118] It is nevertheless encouraging that at least 17,052 human genes have been shown to have orthologous counterparts in the mouse genome, suggesting that they do indeed correspond to real proteins.[119] The definition of what constitutes a gene is still fairly fluid, and hence, depending upon the precise definition adopted, it may be that many

additional human 'genes' still remain to be described and annotated.

To appreciate why definition is an issue here, one need only be aware of the many exceptions to genes being contiguous (as well as functionally and spatially distinct) entities, as classically envisaged. Thus, some genes are known to occur within the introns of other genes.[120–122] Some genes can overlap with each other either on the same or on different DNA strands,[123] resulting in the sharing of some of their coding and/or regulatory elements.[124,125] In addition, the vast majority of human genes are now known to undergo alternative splicing,[84] leading in some cases to quite different proteins being encoded by the same gene. For example, the human cyclin-dependent kinase inhibitor 2A gene (*CDKN2A*) (MIM# 600160) encodes an alternatively spliced variant ($p14^{ARF}$) which, through the inclusion of an alternative first exon, acquires an altered reading frame so as to specify a protein product that is structurally unrelated to the other p16 isoforms encoded by this gene.

Gene density varies between the human chromosomes and the gene distribution within chromosomes is also rather uneven. Strikingly gene-poor regions have been identified and are known as 'gene deserts'.[126] These are regions that are devoid of protein-coding genes over distances of several Mb but which may nevertheless contain regulatory sequences (Box 1).

RNA genes or non-protein-coding RNAs

A large proportion of the human transcriptome still remains to be annotated.[136] Although some of the overall transcriptional activity may simply be 'transcriptional noise',[137,138] at least a portion of it is likely to be associated with functional non-coding RNA genes, many of which are located in regions previously regarded as intergenic and/or non-coding.[71] Non-coding RNA genes are as widespread as they are diverse,[139] are transcribed from both strands of the genome and may well exceed protein-coding genes in terms of their number.[140,141]

Non-coding RNAs of known function include structural RNAs such as transfer RNAs, ribosomal RNAs and small nuclear RNAs, but also putative

Box 1. Gene deserts and their potential relevance to human inherited disease

A functional role(s) for gene deserts[127] has been supported by results from GWASs. Thus, multiple SNPs on chromosome 5p13.1 have been shown to be strongly associated with Crohn's disease, even though the region is located within a 1.2 Mb gene desert and the nearest annotated gene, that encoding prostaglandin E receptor EP4 (*PTGER4*), is about 270 kb away from the association signals.[128–131] Although the SNPs were consistently associated with the disease, their functional effect is not easy to infer because these SNPs could exert an effect either on the nearest gene or on other genes that are located further away. However, Libioulle *et al.* (2007)[128] integrated the GWAS results with gene expression data and found that the associated SNPs influenced the level of expression of *PTGER4*.

The majority of GWAS-SNPs are located in either intronic, intergenic or gene desert regions rather than within gene-coding or promoter sequences. These SNPs could nevertheless be of direct functional significance if their locations coincide with regulatory elements, either already known or yet to be characterised, such as enhancers, transcription factor binding sites and sequences encoding for microRNAs.[132]

The association of the SNP rs6983267 at 8q24 with colorectal and prostate cancer has been a mystery since its discovery because the risk allele is located in a gene desert >300 kb away from the nearest annotated gene, *MYC*. Recent studies have, however, found that the region containing the risk allele is a transcriptional enhancer that interacts with the *MYC* proto-oncogene.[133,134] In a similar vein, GWAS-SNPs in a 9p21-located gene desert (associated with coronary artery disease) have been found to impair the interferon-γ signalling response.[135]

regulatory RNAs (microRNAs, small interfering RNAs [siRNAs], piwi-interacting RNAs, transcription initiation RNAs [tiRNAs], transcription start site-associated RNAs [TSSa-RNAs], promoter upstream transcripts [PROMPTs], promoter-associated sRNAs [PASRs and PALRs] and longer non-coding RNAs such as XIST), which are involved in sequence-specific transcriptional and post-transcriptional modulation of gene expression.[142–148] Thus, more than 1,000 microRNA genes already have been identified in the human genome, with many more probably awaiting discovery (Box 2). In total, at least 1,500 non-coding RNA genes already have been annotated in the human genome reference sequence, with up to 5,000 more predicted by homology-based methods[117] (see Ensembl, database version 56.37a).

Indeed, large intergenic non-coding RNAs (lincRNAs) recently have been found to represent a novel category of evolutionarily conserved RNAs, with a diverse array of functions ranging from stem cell pluripotency to cellular proliferation;[93,94] lincRNAs appear to number at least 3,000 in the human genome.[155–158] Some lincRNAs guide chromatin-modifying complexes to specific genomic loci, to regulate gene expression.[94] LincRNAs also play an important role in the derivation of human-induced pluripotent stem cells.[156] Collectively, non-coding RNAs have been intensively studied over the past several years.[159,160]

Pervasive transcription: Transcripts of unknown function and unannotated transcripts

The ENCODE project, designed to analyse 30 Mb of DNA from 44 genomic regions to characterise the functional elements present, has identified complex patterns of regulation and 'pervasive transcription' of the human genome.[71] Although >90

Box 2. MicroRNAs

MicroRNA has been the most intensively studied non-coding RNA in the human genome. MicroRNA gene loci may be fairly numerous: already more than 15,000 microRNA gene loci have been identified in various species (miRBase, Release 16.0: September 2010; http://www.mirbase.org/), with 1,048 microRNAs being found in the human genome.

Biogenesis and function

The synthesis of microRNAs starts with the transcription of primary microRNAs by RNA polymerase II. The primary microRNAs will be processed further to become precursor microRNAs and then mature microRNAs. The mature microRNAs are short sequences of 18–25 nucleotides; they are incorporated into RNA-induced silencing complex (RISC) to exert their post-transcriptional regulatory roles through binding to the 3′ untranslated region (UTR) of target mRNAs. The binding of microRNA to target mRNAs can lead to two possible outcomes; either degradation or cleavage of the mRNAs or suppression of the translation of mRNAs into protein.[149]

Relevance to diseases

The importance of microRNAs as functional regulators increasingly has been interrogated by microarray and sequencing studies. Deregulation in the expression patterns of microRNAs was commonly associated with various cancers.[150–152] SNPs in the (i) sequences encoding microRNAs and (ii) 3′ UTR of mRNAs also have been found to be associated with various cancers.[153,154]

per cent of the human genome appears to be represented in nuclear primary transcripts, it has become clear that only 35–50 per cent of processed transcripts have so far been annotated as genes, implying that many genes may not yet have been recognised as such.[71,85,161,162] Thus, large numbers of hitherto unannotated transcripts may well yet turn out to be of functional significance.[161] Such transcripts have been collectively classified as transcripts of unknown function (TUFs) and are thought to include (i) antisense transcripts of protein-coding genes, (ii) isoforms of protein-coding genes and (iii) transcripts that either overlap introns of annotated gene transcripts (on the same strand) or which are derived entirely from intergenic regions. Although both the complexity and abundance of TUFs are remarkable, it should be realised that there is often no firm evidence for these transcripts being of functional significance. Indeed, unannotated non-polyadenylated transcripts originating from intergenic regions have been found to represent the bulk of the >90 per cent of the human genome that now appears to be transcribed.[161,163,164] Although the functional significance of pervasive transcription remains unclear, it is much more extensive than had previously been realised.[165]

In both humans and mice, up to 70 per cent of genomic loci exhibit evidence of transcription from the antisense strand, as well as the sense strand.[166–168] These naturally occurring antisense transcripts may modulate the level of expression of their associated sense transcripts (or otherwise influence their processing), thereby adding another level of complexity to the regulation of gene expression.[169,170] Although there is, as yet, no suggestion that the genomic sources of such antisense transcripts should be regarded as genes in their own right, their prevalence clearly renders our task of defining the gene that much more difficult.

High copy number repeat sequences

The HGP revealed that repeat sequences account for at least 50 per cent of the human genome sequence. These repeats may be classified as (i) transposon-derived repeats, (ii) partially retroposed copies of genes (referred to as processed pseudogenes), (iii) simple sequence repeats, (iv) blocks of tandemly repeated sequences at centromeres, telomeres and the short arms of acrocentric chromosomes and (v) segmental duplications (SDs) or low copy number repeats.

Segmental duplications

Both the number and the breadth of the distribution of SDs in the human genome (5 per cent) were surprising. SDs represent extensive inter- and intra-chromosomal duplications of genomic regions that contain genes as well as intergenic sequences.[1,2] She *et al.* extended the initial analyses of these low copy number repeats or SDs and initiated the characterisation of the duplicational landscape of the human genome.[171] SDs may be viewed as mutational hotspots, since they are prone to aberrant recombination events occurring between highly homologous paralogous SDs, and give rise to large deletions or duplications of the intervening sequences resulting in human genomic disorders.[172] Indeed, SDs have been shown to represent frequent sites of CNV between individuals, thereby contributing considerably to human genomic diversity.[173] The mechanism that generates CNVs in SDs is known as non-allelic homologous recombination.[174] These interspersed SDs confer susceptibility to recurrent microdeletions and microduplications upon approximately 10 per cent of the human genome through unequal crossing over. Furthermore, data have accumulated showing that specific recurrent rearrangements within these genomic hotspots are associated with both syndromic and non-syndromic diseases. Studies of common complex diseases have shown that these recurrent events play an important role in autism, schizophrenia and epilepsy.[175–177]

The above notwithstanding, the duplicated genomic regions have remained largely intractable, owing to difficulties in accurately resolving their structure, copy number and sequence content. New algorithms have been developed to map comprehensively next-generation sequence reads, allowing the prediction of absolute CNVs of duplicated

segments and genes. On average, 73–87 genes vary in copy number between any two individuals and these differences overwhelmingly correspond to segmental duplications.[178]

Pseudogenes

Whether processed or non-processed (duplicational), it has become clear that pseudogenes are almost as abundant as genes ('classical' or otherwise) in the human genome, with ~20 per cent of known pseudogenes being transcribed.[179–181] By means of a comparison of cytochrome P450 genes (*CYP*) from the mouse and human genomes, Nelson *et al.* (2004) demonstrated that the complete identification of all human pseudogene sequences is likely to be clinically important and proposed a naming procedure for *CYP* pseudogenes.[182]

It should, however, be appreciated that, although some pseudogenes may well be readily identifiable as lacking protein-coding potential by virtue of the interruption of their open-reading frames by premature stop codons or frameshift mutations, others will be less easily recognisable, especially if they are transcribed. The recent identification of short (≤300 bp) human pseudogenes generated via the retrotransposition of mRNAs,[183] however, suggests that pseudogenes may be even more common in the human genome than previously appreciated. Intriguingly, some of these pseudogenes are polymorphic, in that they have functional as well as non-functional alleles segregating in the extant human population.[184]

With the realisation that pseudogene-derived RNA transcripts may harbour functional elements,[181,185] the distinction between genes and pseudogenes has become somewhat blurred.[186] Indeed, some 'pseudogenes' appear to have a regulatory role,[187,188] providing additional examples of the potential functional significance of non-coding RNAs. At present it is unclear what proportion of the pseudogenes identified to date have either retained or acquired a function via their non-coding RNAs.

Transposable elements

Transposable elements, including Long INterspersed Elements (LINE-1), *Alu* and SINE–VNTR–Alu (SVA) elements (SVA is an unusual composite element derived from three other repeats: Short INterspersed Elements [SINE]-R, variable number tandem repeats [VNTR] and *Alu*), make up ~40 per cent of the human genome[189] and constitute a major source of inter-individual structural variability.[190] Some of these transposable elements have contributed gene-coding sequences to the human genome via 'exonisation'.[191] Other transposable elements have contributed functional non-coding sequence — for example, as regulatory elements,[192,193] microRNAs[194] or naturally occurring antisense transcripts.[195] Many more are likely to have functional significance, as suggested by their evolutionary conservation.[196,197]

Evolutionary conservation

Extensive evolutionary conservation of non-coding DNA sequences is evident in the human genome because only ~40 per cent of the evolutionarily constrained sequence occurs within protein-coding exons or their associated untranslated regions.[71] Studies of evolutionarily conserved non-coding sequences[198–201] have suggested that 5–20 per cent of the genome may be of functional importance, rather than just the ~2 per cent associated with the protein-coding portion.[202,203] Some non-coding regions (the genomic 'dark matter') contain 'ultraconserved elements' which not only exhibit enhancer function, but are also transcribed and often appear to have been subject to selection to the same extent as protein-coding regions.[204–206] Some non-coding regions contain CpG islands, which, although located far from the transcriptional initiation sites of genes, may nevertheless have some regulatory significance.[207] It should be appreciated, however, that the absence of evolutionary conservation does not necessarily denote lack of function. Indeed, human specific functional elements have been shown to be present within rapidly evolving non-coding sequences.[208,209]

Towards a new definition of the gene

It is clear from the above that precisely what constitutes a gene has become somewhat contentious. The unanticipated scale of the extent of

transcription in the genome, coupled with the widespread occurrence of overlapping genes and shared functional elements, hampers attempts to demarcate precisely and unambiguously where one gene ends and another one begins. As a consequence, the notion of the gene has become diffuse.[161,210] Indeed, as Kapranov *et al.*[211] opined, 'it is not unusual that a single base-pair can be part of an intricate network of multiple isoforms of overlapping sense and antisense transcripts, the majority of which are unannotated'. Gene regulatory elements that are often distant from the genes they regulate,[212] the existence of *trans*- as well as *cis*-regulatory elements[213] and the formation of non-co-linear transcripts through *trans*-splicing,[214] taken together with the abundance of non-coding RNA genes[215] and evolutionarily conserved non-coding regions,[199,201] have combined to challenge the classical notion of the gene.

On the basis of the findings of the ENCODE project, Gerstein *et al.*[210] proposed an updated definition of the gene as 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products'. An alternative definition of the gene as: 'A discrete genomic region whose transcription is regulated by one or more promoters and distal regulatory elements and which contains the information for the synthesis of functional proteins or non-coding RNAs, related by the sharing of a portion of genetic information at the level of the ultimate products (proteins or RNAs)' has been proposed by Pesole.[216] Irrespective of its precise definition, it is clear that the concept of the gene is inadequate to the task of building a lexicon of those functional genomic sequences that could harbour mutations causing human inherited disease. It is likely in the context of mutation detection, that we shall eventually have to consider the universe of functional genetic elements in the human genome as our hunting ground, rather than simply genes *per se*.

Development of the GWAS approach to complex diseases and traits

In this section, developments in cataloguing genetic variation (SNP and CNV), initiation and completion of the International HapMap Project, and advances in genotyping technologies are discussed. These developments are important prerequisites for the use of GWASs in the investigation of complex diseases and traits.

SNP discovery after the HGP

While the HGP was being completed, genetic variants, in particular SNPs, were also being discovered. By 2001, the International SNP Map Working Group had identified 1.42 million SNPs in the human genome.[58] Currently, more than 17 million SNPs in the human genome have been catalogued in the SNP Database (dbSNP; http://www.ncbi.nlm.nih.gov/projects/SNP/). It is, however, likely that at least some of the entries in the database are errors or artefacts rather than 'genuine' variants. A false-positive rate for the dbSNP of 15–17 per cent has been estimated.[101] Therefore, large-scale validation in population-based studies is necessary. The HapMap Project was conceived in 2003 with the aim of validating several million SNPs in order to obtain SNP and genotype frequency information, as well as to study their LD patterns in different populations.

SNPs are the most abundant type of genetic variation in the human genome. They occur at intervals of approximately one SNP to every kb of DNA sequence throughout the genome when the DNA sequences of any two unrelated individuals are compared. This is approximately equivalent to three million SNPs being carried by each individual genome. Therefore, the DNA sequences of any two unrelated genomes are estimated to be about 99.9 per cent identical; the 0.1 per cent comprises mainly SNPs, and these are believed to be responsible for many of the phenotypic differences noted among individuals in populations — for example, disease susceptibility, drug responses and physical traits such as height.[217]

The discovery of thousands of CNVs that collectively encompass hundreds of Mb of the genome[22,23,105] and the several hundred thousand short indels identified by WGS studies,[42,43] however, have cast doubt upon the initial estimate

of '99.9 per cent similarity' between any two genomes. Indeed, the DNA sequences of individuals within and between populations are genetically rather more diverse and varied than previously thought. This has been corroborated by a recent study demonstrating that the Craig Venter genome differs from the consensus reference sequence by approximately 1.2 per cent when indels and CNVs are considered, a further 0.1 per cent when SNPs are considered and ~0.3 per cent when inversions are considered — a grand total of ~1.6 per cent.[218]

Linkage disequilibrium and the International HapMap Project

Most SNPs are predicted to be neutral, without any functional effects. Owing to their abundance in the human genome, they may serve as useful genetic markers in GWASs, by comparison with other genetic variations, such as microsatellites, which in any case exhibit a mutation rate that is too high to be useful in this context. Early reports documented LD patterns between SNPs in parts of the human genome;[61,62,219] however, no large-scale effort had been undertaken to study the LD patterns in the whole genome until the initiation of the International HapMap Project. A total of more than three million SNPs were genotyped and validated in Phase I and Phase II of the project in four populations.[66,69] These populations were the US Utah population of Northern and Western European ancestry (CEU), Han Chinese from Beijing (CHB), Japanese from Tokyo (JPT) and the Yoruba from Ibadan, Nigeria (YRI).

One novel finding has been that 10–30 per cent of pairs of individuals within a population share at least one region of extended genetic identity arising from recent common ancestry. An additional discovery was that up to 1 per cent of all common variants are not tagged by SNPs, primarily because they are located within recombination hotspots.[69] Importantly, increased population differentiation with respect to non-synonymous SNPs was noted, by comparison with synonymous SNPs. These observations have also indicated systematic differences in the strength or efficacy of natural selection between populations from different geographical areas involving genes linked to the Lassa virus in West Africa, skin pigmentation in Europe and hair follicle development in Asia.[70]

The discovery of millions of SNPs has created a significant challenge in genotyping. It is neither technically feasible nor cost-effective to genotype all the SNPs in a GWAS, even with the latest genotyping technologies; however, the existence of LD significantly reduces the number of SNPs that need to be genotyped. The indirect association approach of GWASs is dependent on surrogate markers ('tag' SNPs) to locate disease variants through LD. As shown by the HapMap Project[69] and other published work,[220–222] approximately half a million SNPs are adequate to capture most of the SNPs that have been genotyped in the HapMap Phase I and II projects. However, the genome coverage of commercial genotyping arrays is population dependent (Box 3).

The HapMap project has created a useful and valuable resource for GWASs. In parallel, the public availability of the HapMap resource has driven the rapid development of genotyping arrays, in which the data are used to guide the selection of tag SNPs. Once the HapMap Phase I and II projects were completed, a number of genotyping arrays were designed and introduced onto the market.[223,224] The newer arrays (eg the Illumina Human 1M Beadchip and Affymetrix SNP Array 6.0) have significantly improved genome coverage and are also designed for CNV detection.[225] The HapMap Phase I and II projects led to the development of higher resolution genotyping arrays, which in turn were used in the HapMap Phase III project to investigate genetic variations (both SNPs and CNVs) in additional populations of diverse ancestry.[21]

The Phase III project, building on the success of the HapMap Phase I and II projects, included an additional seven populations and has recently been completed.[21] These additional populations involved people of African ancestry in the south-western USA (ASW), the Chinese community in Metropolitan Denver, CO (CHD), Gujarati Indians in Houston, TX (GIH), the Luhya in Webuye,

Box 3. Genome coverage

High genome coverage is important, since the underlying principle of this approach is the use of LD to detect disease variants. In SNP-scarce regions, *bona fide* disease variants could be missed because they are not in strong LD with any of the SNPs genotyped on the array. Genome coverage is an estimate of the proportion of SNPs (using the International HapMap data as a reference) that can be captured by the SNPs which are directly genotyped in an array with a preset r^2 threshold. Usually, a threshold of 0.8 is used to estimate genome coverage. These first-generation genotyping arrays used the International HapMap database for SNP selection and have poor coverage for SNPs with minor allele frequency (MAF) <5 per cent.[220–222]

Kenya (LWK), people of Mexican ancestry in Los Angeles, CA (MEX), the Maasai in Kinyawa, Kenya (MKK) and Tuscans in Italy (TSI). The ethos behind the HapMap Phase III project was that, in order to obtain a more complete understanding of human genetic variation, populations with a wider geographical/ancestral range needed to be studied. In total, the HapMap Phase III project genotyped approximately 1.6 million SNPs (using both the Illumina Human 1M Beadchip and Affymetrix SNP Array 6.0) in 1,184 individuals from 11 populations (four original and seven additional populations). The population-specific differences among low-frequency variants were characterised in addition to SNPs and common CNVs or copy number polymorphisms (CNPs). More importantly, it also demonstrated the feasibility of imputing newly discovered CNPs and SNPs, which are important for future GWASs and meta-analyses.[21]

Whole-genome SNP genotyping technologies

The paradigm shift from candidate-gene association and family linkage studies to GWASs has been attributed to several important developments, most notably the rapid advances in high-throughput SNP genotyping technologies, which have enabled researchers to interrogate up to one million SNPs simultaneously in a microarray.[18] GWASs employ an 'agnostic' approach in the search for unknown disease variants, and hence the ability to interrogate a large number of SNPs covering the entire human genome is a prerequisite for this study design. In parallel with the decreasing cost of genotyping, it has recently become technically feasible to genotype thousands of samples in GWASs. As a result, more than 800 GWASs have been published since 2005 (http://www.genome.gov/gwastudies/), of which almost all have used the commercially available whole-genome SNP genotyping arrays from Illumina or Affymetrix.

A series of whole-genome genotyping arrays have been introduced since 2005, such as the Affymetrix Human Mapping 100K 500K sets, and the Illumina HumanHap300 and HumanHap550 BeadChips.[223,224] These genotyping arrays provide different degrees of genome coverage in different populations; lower coverage was achieved in African populations because of the greater genetic diversity in these populations. For example, the Illumina HumanHap550 Beadchip, which contains approximately 550,000 tag SNPs selected from the HapMap Phase I and II projects, achieved genome coverage of 87 per cent and 83 per cent in CEU and CHB + JTP populations, respectively, but only 50 per cent in YRI.[220–222] Whole-genome genotyping arrays such as the Illumina Human 1M Beadchip and Affymetrix SNP Array 6.0 offer almost complete genome coverage (>90 per cent) for HapMap CEU and CHB + JPT populations (Box 3).

The more recent genotyping arrays, such as the Illumina Human 1M BeadChip and Affymetrix SNP Array 6.0, have enabled genotyping of up to one million SNPs and increased the sensitivity to detect CNVs because of higher marker density and more uniform marker distribution.[225] For example, the Affymetrix SNP Array 6.0 contains more than

1.8 million markers, half of which are SNPs, the remainder being non-polymorphic or copy number probes to enhance the power of detection of CNVs. Copy-number probes were deliberately selected so as to cover regions lacking SNPs or regions where SNPs are difficult to assay, such as repetitive sequences within segmental duplications.[226] In addition, markers were also chosen to target known copy number variable regions as reported in the Database of Genomic Variants (http://projects.tcag.ca/variation/). Employing such a design, these genotyping arrays have enabled researchers to discover novel CNVs, as well as to validate previously known CNVs. These more recent arrays were designed for the application of GWASs and CNV detection.

The first wave of GWASs utilised first-generation SNP genotyping arrays and focused mainly on common SNPs with MAF >5 per cent.[132] Thus, expanding the coverage to include less common or rarer SNPs (MAF 1–5 per cent) is essential for new discoveries to be made in future GWASs. This step is now technically feasible and practically achievable with the arrival of second-generation SNP genotyping arrays (Illumina HumanOmni2.5 and Omni5.0) in 2010; these are capable of genotyping 2.5 to 5.0 million SNPs (Illumina Whole-Genome Genotyping Product Roadmap; http://www.illumina.com/applications/gwas.ilmn). These arrays were designed to increase the coverage of SNPs down to a MAF of 1 per cent. In contrast to the first-generation arrays, the SNP selection in these latest genotyping arrays leverages the data from the 1000 Genomes Project.[102] However, the promise of second-generation genotyping arrays for new discoveries in GWASs is conditional upon the adequacy of the statistical power of the studies to identify the associations of rarer SNPs with complex traits. This suggests that larger sample sizes will be needed in future GWASs.

The era of GWASs

More than 4,000 SNPs have been reported to be associated with various human complex diseases and traits with varying degrees of replication and success (http://www.genome.gov/gwastudies/).

Despite some notable successes in revealing numerous novel SNPs and loci associated with complex phenotypes, the results from GWASs have been disappointing, in that all the GWAS-SNPs collectively account for only a small proportion of the heritability of complex phenotypes. This is due mainly to the small effect sizes of most GWAS-SNPs (odds ratio <1.5).[5,10,89] The small effect sizes of the GWAS-SNPs have also limited their applications in disease risk prediction.[227]

Although several diseases have been claimed to be investigated by GWASs and meta-analyses of sufficiently large sample sizes, most of their heritability still remains unaccounted for. This missing heritability has stimulated much discussion on future strategies for detecting the remaining genetic variants associated with complex phenotypes. The proposed strategies range from increasing the sample sizes by combining several GWASs through meta-analysis in order to attain a higher statistical power, to more complicated experiments such as epigenetic studies.[5,228] The methodologies for meta-analysis and for the merging of SNP genotype data from multiple GWASs employing different genotyping arrays are now well developed and rely upon newly developed genotype imputation methods.[229–231] By contrast, there are still many experimental and analytical uncertainties and challenges to be faced in the context of epigenetic studies of complex phenotypes.[232,233] Other approaches are summarised in Figure 1.

Figure 1 summarises a variety of approaches to the further identification of disease-associated variants: (1) GWASs of various complex diseases and traits ideally should be performed in different populations representing European, Asian and African ancestries, as most published studies have focused primarily on populations of European ancestry.[234,235] (2) Most GWASs have done fast-track replication by selecting the top few or top tens of SNPs with the most significant *p*-values in stage 1 and then proceeded to replicate them in stage 2 or stage 3 with larger sample sizes. Therefore, the next step should be to conduct a

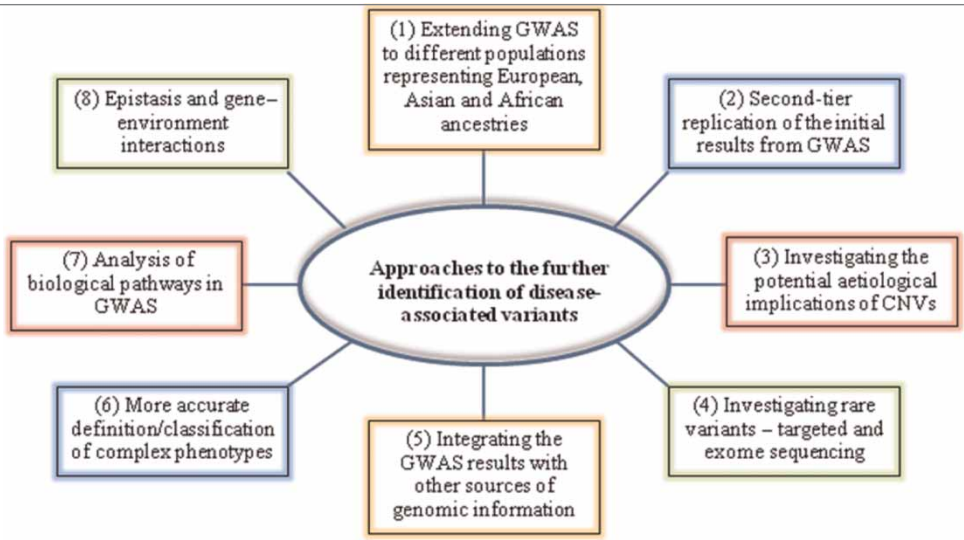

Figure 1. Summary of the approaches identifying disease-associated variants.

second tier of replication, where more SNPs from stage 1 are tested to assess their associations.[236] (3) The role of CNVs is increasingly recognised as being associated with complex diseases and traits; thus, it is important to investigate their associations with these complex phenotypes.[111] (4) Resequencing of the GWAS loci will be needed to uncover additional rarer variants. The success of this approach has been demonstrated in the discoveries of multiple rare variants for type 1 diabetes and hypertriglyceridaemia.[237,238] (5) Integrating GWAS results with other sources of genomic data, such as expression quantitative trait loci (eQTL) and ChIP-Seq, has led to the discovery of novel SNP associations.[239,240] (6) Subgroup analysis of disease phenotypes is a powerful approach to identifying genetic variants that are specific to certain subtypes. For instance, differences in SNP associations for oestrogen receptor-positive and -negative breast cancer have been shown.[241] (7) Pathway-based approaches have been developed using prior biological knowledge of gene function to facilitate more powerful analysis of GWAS datasets.[242] (8) Most studies have not taken epistasis and gene–environment interactions into account, which could account for a proportion of the missing heritability of complex phenotypes; however, challenges associated with studying these interactions should also be noted.[243,244]

Genetic architecture of complex diseases

The genetic architecture of complex diseases has been the subject of intense debate over the past decade[59,60] and has been polarised by the emergence of two opposing models: the CD/CV hypothesis and the multiple rare variant or common-disease rare-variant (CD/RV) hypothesis.[245] The CD/CV model formed the basis of the HapMap Project and largely influenced the development of commercial genotyping arrays with respect to SNP selection. Therefore, the published GWAS using the HapMap data mainly involved the interrogation of the association of common SNPs (MAF $>$ 5 per cent) with complex diseases and traits.

One of the reasons that the CD/CV model became favoured was because of the sequencing technologies available at that time. Sanger sequencing did not allow the survey of rare variants in the whole genome. By contrast, the convenient high-throughput genotyping platforms have enabled efficient interrogation of up to one million SNPs throughout the genome, which eventually indirectly leads to the capture of almost all the SNPs in the HapMap Project. Furthermore, it is more affordable to genotype (rather than to sequence) the entire genomes of several thousand cases and controls as part of an adequately powered association study.

Currently, the results from the GWASs focus on common SNPs and explain only a small fraction of the heritability of complex phenotypes.[5] The missing heritability has challenged the validity of the CD/CV hypothesis, and has also diverted research endeavours toward rare variants;[109,237,238,246,247] however, published data have revealed the contributions of both common and rare variants to complex phenotypes. The results from GWASs have strongly supported the involvement of common variants, especially common SNPs, in complex phenotypes.[132] Moreover, recent studies have shown that common SNPs can explain a greater proportion of the heritability than has been accounted for by recent GWASs. These SNPs, however, are often 'hidden' within the GWAS data, and will require larger sample sizes to be uncovered.[248,249]

The data supporting the roles of rare variants have also been accumulating from an increasing number of studies of less-common SNPs[109,237,238,246] and rare CNVs.[250–253] This suggests that the genetic architecture of complex phenotypes is likely to comprise both common and rare variants. The relative proportions of these variants remain to be determined and will remain unclear until all the genetic variants for most complex phenotypes are found; furthermore, the relative proportions are likely to vary between different complex phenotypes, with some phenotypes having a greater influence on the genetic susceptibility risk by common variants, whereas other phenotypes may be more affected by rare variants. Being able to predict the genetic architecture of complex phenotypes is critical, however, as it will determine the future strategies to be adopted in seeking disease variants.

Homozygosity mapping

Homozygosity mapping has been shown to be useful in the identification of disease susceptibility genes in complex diseases.[254,255] An ROH defines an uninterrupted stretch of a DNA sequence lacking heterozygosity in the diploid state (ie in the presence of both copies of the homologous DNA segment). Thus, all the genetic variants within the homologous DNA segments are represented by two identical alleles that contribute to the homozygosity.[28] Currently, there are no standardised criteria to define an ROH. Previous studies have focused on regions $\geq$1 Mb, however, and hence the true extent of homozygosity in the human genome could have been underestimated because shorter regions were not considered.[28,256,257] More recent studies have defined ROHs as having a minimum length of 500 kb,[258] the intention being to avoid underestimation of the number of such regions in the human genome.

Although long continuous ROHs were first documented a decade ago, until recently no large-scale population-based studies had been performed to assess the extent of ROHs in the human genome.[259] The recent advances in the genome-wide detection and characterisation of ROHs have been driven mainly by the availability of highly accurate SNP databases such as the HapMap project[28] and advanced genotyping technologies.[24,25] Genotyping a large number of SNPs on a microarray platform presents a powerful tool for detecting ROHs comprehensively across the whole genome, thereby enabling investigation of the number, length, location and distribution of the ROHs in the human genome in a more unbiased manner, as compared with microsatellite markers. It was not previously expected that the genomes of outbred populations would contain ROHs of several Mb in length until the early reports appeared in 2006/2007.[28,256,257]

Many novel causal genes or mutations underlying autosomal recessive disorders have been identified through homozygosity mapping. This approach is particularly useful for investigating these disorders in populations with a high prevalence of consanguinity, as is evident from the many recent studies that have identified causal mutations.[260–265]

The effects of consanguinity and recessive variants or heterozygosity levels on the risk of complex phenotypes (diseases and quantitative traits) are well established.[266–268] Higher levels of relative heterozygosity have been shown to be

associated with lower blood pressure and total and low-density lipoprotein (LDL) cholesterol by measuring genome-wide heterozygosity.[268] In addition to quantitative traits, inbreeding has also been found to be a significant positive predictor for a number of late-onset complex diseases, such as coronary heart disease, stroke, cancer and asthma.[266] These studies have strongly supported the hypothesis that the genetics of complex phenotypes include a component which corresponds to recessively acting variants. The importance of ROHs to complex phenotypes remains largely unexplored; however, several studies have shown significant differences in ROHs between cases and controls in genome-wide investigations for schizophrenia[269] and late-onset Alzheimer's disease.[270] Success was also achieved for complex quantitative traits such as height, where strong statistical evidence for an association of a particular ROH with height was obtained in a total sample size of >10,000. The height of individuals with this ROH was significantly higher (increased by 3.5 cm) than the individuals lacking the region.[258] Cataloguing ROHs in human genomes and investigating their associations with complex phenotypes by building on existing GWAS data should be fruitful areas for future research.

Beyond SNPs: CNVs

A new era of CNV discovery began when two separate studies, published concurrently in 2004, identified several hundred deletions and duplications in the human genome.[26,27] Such genetic abnormalities had actually been documented decades before, however, in clinical cytogenetics studies that found them to be a cause of various genomic or cytogenetic disorders.[271] The distinguishing feature of the recent studies was that these CNVs were found to be much more prevalent in the human genome than previously expected. These changes in copy number did not result in any clinical disorder or pathological phenotype and were found in the genomes of phenotypically normal individuals. As these submicroscopic (<5 Mb) deletions and duplications were below the detection limit of traditional cytogenetics tools such as fluorescence *in situ* hybridisation (FISH), these recent discoveries were credited to the use of whole-genome microarray technologies.[272]

Although these early whole-genome microarray studies discovered several hundred new CNVs, it was clear from the outset that that this would be a gross underestimate of the true total. These studies used 'low-resolution' microarrays such as representational oligonucleotide microarray analysis (ROMA) containing 85,000 probes with a resolution of approximately one probe per 35 kb[26] or the bacterial artificial chromosome–comparative genomic hybridisation (BAC-CGH) array with a resolution of approximately one probe per 1 Mb.[27] Further, these studies investigated a small sample size, which limited the efficiency of detection of less common CNVs. CNVs smaller than 50–100 kb would not have been detected because their size was below the resolution limit for these microarrays. Thus, both the sample size and the resolution of the microarray are critical factors that contribute to the discovery of less common and/or smaller CNVs.

The contribution of CNVs as a major source of genetic variation in human populations has become appreciated despite the limitations of the microarrays. The first comprehensive mapping of CNVs in 270 samples from the HapMap Phase I project identified a total of 1,447 copy number variable regions, covering 360 Mb. These regions contained hundreds of genes, disease loci, functional elements and segmental duplications.[22] The limitations of ROMA and the BAC-CGH arrays have been overcome in later studies by the use of higher-resolution microarrays and larger sample sizes comprising several hundred samples.[23,105,273–276] High-resolution tiling oligonucleotide microarrays, comprising 42 million probes, were used to generate a comprehensive map of 11,700 CNVs.[105] Yim *et al.*[275] screened CNVs in 3,578 healthy, unrelated Korean individuals, using the Affymetrix SNP Array 5.0.

Other types of chromosomal rearrangement, particularly inversions and balanced translocations, have received considerably less attention.[277–279] Inversions and translocations are also known as 'copy-neutral variations' or 'balanced chromosomal rearrangements', since they do not involve changes

in copy number. These copy-neutral variations have also been found to be associated with disease.[279] Collectively, these copy number and copy-neutral variations are broadly classified as 'structural variations'. As discussed, the genome-wide mapping and detection of CNVs in different populations has advanced considerably since 2004, being driven mainly by microarray technologies such as oligonucleotide-CGH and SNP microarrays. By contrast, the pace in identifying inversions and translocations in the human genome has been slower because more powerful and effective methods were not available until the advent of NGS technologies[76] (Boxes 4 and 5).

The discovery of a 20 kb deletion located immediately upstream of the immunity-related GTPase family M gene (*IRGM*) underlying Crohn's disease, and the identification of a 45 kb deletion that is in perfect LD with body mass index-associated SNPs near the neuronal growth regulator 1 gene (*NEGR1*),[287,288] together with other studies reporting evidence for LD of CNVs with GWAS-SNPs at $r^2 > 0.5$, suggest possible associations of CNVs with a variety of different human complex diseases and traits.[105] The genome-wide study performed by the Wellcome Trust Case Control Consortium (WTCCC) investigating the association between ~3,400 common CNVs and eight complex diseases in 19,000 samples did not yield any novel discoveries;[111] however, rare CNVs associated with various complex phenotypes have been identified in studies of schizophrenia,[250,289,290] epilepsy[251] and severe early-onset obesity.[252,253] The studies on schizophrenia found that rare structural variations that disrupt multiple genes in neurodevelopmental pathways are over-represented in cases, as compared with controls.[250,289]

High-throughput sequencing technologies and their impact on genomic studies

The advent of high-throughput sequencing technologies has initiated the 'personal genome sequencing' era for both normal and cancer genomes, and large-scale genome sequencing studies such as the 1000 Genomes Project and the International Cancer Genome Consortium. The high-throughput sequencing technologies also provide new opportunities to study Mendelian disorders through exome sequencing and WGS. Several international projects have also been launched to explore functional genomics.

High-throughput sequencing technologies

NGS technologies have only been on the market since 2004, but have now largely replaced Sanger sequencing technologies (owing to the ultra-high-throughput production capacity of NGS technologies, which is a thousand times greater than that of traditional sequencing). One of the major differences is the ability of next-generation sequencers to simultaneously sequence millions of DNA fragments; hence, they are also referred to as massively parallel sequencing technologies. This feature has considerably increased the number of nucleotides that can be sequenced per instrument run when compared with Sanger sequencing. The sequencing chemistry of NGS technologies, together with their ultra-high-throughput production capacity, has also reduced sequencing costs significantly, making large-scale or WGS studies much more affordable.[29–31] The sequencing technologies currently available can be broadly grouped into NGS technologies such as the Roche 454 Genome Sequencer FLX (GS FLX) System, Illumina Genome Analyzer (GA) and HiSeq and Life Technologies Supported Oligonucleotide Ligation Detection System (SOLiD), and TGS (or single-molecule sequencing) technologies such as the HeliScope Single Molecule Sequencer (Helicos Biosciences).[32]

One of the more laborious steps in WGS using the Sanger method was the *in vivo* amplification step using bacterial cloning. This has now been substituted by the *in vitro* amplification of millions of DNA fragments by NGS technologies using emulsion polymerase chain reaction (PCR) (Roche GS FLX and Life Technologies SOLiD) or bridge amplification on a solid surface (Illumina GA and HiSeq). The sequencing approach for NGS

Box 4. Characterising structural variation by means of sequencing

The discovery of copy-neutral variations has been attributed to the development of the PEM method and concurrent advances in NGS technologies. The PEM method has also contributed greatly to the discovery of CNVs in the human genome.[76,81,280] Further studies have also taken advantage of an important feature of sequencing data generated by NGS technologies, where several hundred million short sequence reads are produced per instrument run to detect CNVs based on the abundance or density of the sequence reads aligned to the reference genome. This approach is known as depth-of-coverage (DOC) and is similar to microarray-based methods, in that it is also unable to detect copy-neutral variations.[281]

PEM

In the PEM method, a library of DNA fragments with a fixed insert size is prepared and both ends of the DNA fragments are sequenced to generate 'paired-end sequences' (the sequences at both ends of the DNA fragments). This sequence information is then aligned against the reference genome. The underlying principle of PEM in detecting structural variations is reliance upon the discordance in insert size and orientation of the paired-end sequences being aligned to the reference genome to infer 'simple' deletion, insertion and inversion. Thus, when paired-end sequences aligned to the reference sequence display discordance from the expected insert size or distance, this is indicative of either a deletion or an insertion, whereas discordance in orientation suggests the presence of an inversion (ie paired-end sequences are incorrectly oriented by comparison with the reference genome). Hence, the paired-end sequences are usually classified as 'concordant pairs' or 'discordant pairs'; only the discordant pairs are informative for inferring structural variants. Other, more complex, rearrangements — such as 'everted duplications', 'linked insertions' and 'hanging insertions' — can also be detected.[282]

DOC

The DOC method utilises NGS data for CNV detection. This method is based on the DOC of the sequence reads to infer deletions and duplications. The DOC method is made possible by the production of several hundred million short sequence reads per instrument-run by NGS technologies. The principle underlying the DOC approach is based on the assumption that the sequencing process is uniform, so that the number of sequence reads mapping to a region follows a Poisson distribution. As such, the number of sequence reads should be proportional to the number of times that a particular region appears in the genome. Therefore, it is expected that a duplicated region will have more reads aligned with it, with the converse being true for deletions.[281,282] The assumption that the sequencing process is uniform may not be valid, however, because of the sequencing bias of the NGS technologies, which leads to certain regions of the genome being over- or under-sampled, resulting in spurious signals.[283] Despite their shortcomings, the PEM and DOC methods will continue to play a role in the discovery of structural variations until *de novo* genome assembly becomes more feasible.

Application in cancer studies

Both PEM and DOC have also proven useful in dissecting somatically acquired rearrangements in cancer genomes.[87,284] Sequencing of both ends of the DNA fragments derived from the genomes of two individuals with lung cancer was performed and 306 germline structural variations and 103 somatic rearrangements were identified to the single nucleotide level of resolution.[87]

Box 5. International effort to characterise structural variants using PEM

Proof-of-concept studies

The PEM method for detecting structural variants was first demonstrated by Tuzun *et al.* by mapping paired-end sequences data from a human fosmid DNA genomic library.[285] The average insert size of a fosmid library is approximately 40 kb. This study identified 297 structural variants (139 insertions, 102 deletions and 56 inversions); however, sequencing of fosmid clones by means of Sanger sequencing is laborious and costly.[285] These limitations have been overcome by NGS technologies which directly sequence the paired-end or mate-pair libraries without the need for cloning steps.[76] Both of these studies applied the PEM approach to investigate structural variants in the same sample (NA15510) from the International HapMap Project. Their library insert sizes differed, however, and this has enabled a comparison of the sensitivity between these studies. Korbel *et al.*[76] were able to confirm 41 per cent of all deletion and inversion events detected by fosmid paired-end sequencing. Moreover, they identified an additional 407 structural variants in NA15510 that previously had not been detected by fosmid paired-end sequencing. This further suggests that several libraries with different insert sizes are needed to increase the sensitivity of PEM.

Human Genome Structural Variation Working Group

In addition to individual studies, a large-scale effort is currently being undertaken by the Human Genome Structural Variation Working Group comprehensively to map structural variants in phenotypically normal individuals using the PEM approach.[79] More specifically, the objective is to characterise the pattern of human structural variants at the nucleotide sequence level from a collection of 48 individuals of European, Asian and African ancestry. This project plans to make fosmid clone libraries of approximately 40 kb insert size from the genomic DNA of 48 unrelated females. These samples were already genotyped by the HapMap Project. A larger insert size of approximately 150 kb prepared from BAC clone libraries will also be constructed from 14 unrelated HapMap males. This will aim to provide sequence information on structural variants that are too large to be included in the fosmid libraries, such as those associated with segmental duplications. As such, both the fosmid and BAC libraries will ensure the comprehensive capture of structural variants of varying sizes across the human genome.

Structural variation is biased toward complex duplicated and repetitive regions. Hence, developing clone libraries for a modest number of human genomes should serve as a valuable resource for characterising complex and difficult-to-assay regions of genome structural variation. Since the underlying clones can be retrieved, the complete sequence context of the discovered structural variant can also be obtained.[79] This is crucial for precise breakpoint delineation of structural variation, which is then important for understanding the mutational mechanisms responsible for human genome structural variation. A total of 1,695 structural variants were discovered with fosmid libraries derived from nine individuals. The study also showed that 50 per cent were seen in more than one individual and that nearly half lay outside regions of the genome previously described as structurally variant, indicating novel discoveries. More importantly, 525 new insertion sequences (that are not present in the human reference genome) were discovered and many of these were found to be variable in copy number between individuals.[86] This is important because it suggests that structural variants or CNVs could have gone undetected as part of the 'missing sequences' in the human reference genome. Complete sequencing of 261 structural variants provided insights into the different mutational processes that

have shaped the human genome. This study therefore provided the first high-resolution sequence map of human structural variation.[86] A subsequent study then expanded the Human Genome Structural Variation clone resource by including capillary end sequencing of 4.1 million additional fosmid clones from eight additional human genomes. The combined set includes 13.8 million clones derived from the genomes of six YRI, five Centre d'Etude du Polymorphisme Humain (CEPH) Europeans, three JPT, two CHB and one individual of unknown ancestry.[286] This study characterised the complete sequence of 1,054 large structural variants and analysed their breakpoint junctions to infer their potential mechanisms of origin. Three mechanisms were found to account for the bulk of germline structural variation: microhomology-mediated processes involving short (2–20 bp) stretches of sequence (28 per cent), non-allelic homologous recombination (22 per cent) and L1 retrotransposition (19 per cent).

technologies broadly can be divided into: (1) sequencing-by-synthesis mediated by DNA polymerase (ie pyrosequencing for Roche GS FLX and sequencing by reversible terminator chemistry for the Illumina sequencing platform); and (2) sequencing-by-synthesis mediated by DNA ligase for Life Technologies SOLiD.[29–31]

Whole-genome resequencing can now be accomplished relatively rapidly because of the availability of the HGP template for alignment of the billions of short sequence reads produced by next-generation sequencers. This is necessary because the NGS technologies are characterised by short sequence read lengths of approximately 50–125 bp for both Illumina and Life Technologies sequencing platforms.[29–31] This feature makes *de novo* sequencing, or the assembly of billions of short sequence reads into large contigs challenging — especially for large and complex genomes like the human genome.[291] A longer read length is key to obtaining larger contigs with fewer gaps between them during the assembly steps. Although the latest improvements in sequencing chemistry and systems allow the Roche GS FLX to achieve a sequence read length of 500 bp on average, this is still markedly lower than the 800 bp to 1 kb length achieved by Sanger sequencing (http://www.454.com/).[292] In addition to a short read length, NGS technologies have higher sequence error rates, although this gradually has been improving.[293]

A relatively new addition in the NGS market is the Ion Torrent Personal Genome Machine (PGM) produced by Life Technologies (http://www.iontorrent.com/). The earlier NGS technologies relied on emission of either fluorescent (Illumina and Life Technologies SOLiD sequencing platforms) or chemiluminescent (Roche GS FLX) light to detect and distinguish the nucleotides incorporated during sequencing. However, the Ion Torrent PGM uses proprietary semiconductor sensors to perform direct real-time measurement of the hydrogen ions released upon incorporation of nucleotides during sequencing. Several ion semiconductor sequencing chips will be available, with throughputs ranging from >10 Mb to >1 gigabase (Gb) per instrument run, but these are many-fold lower than the several hundred Gb of sequencing data generated by the latest Illumina HiSeq and Life Technologies SOLiD machines. The Ion Torrent PGM is therefore more suitable for smaller-scale targeted sequencing.

The first TGS instrument — the Heliscope Single Molecule Sequencer — is now commercially marketed by Helicos Biosciences. The Heliscope Single Molecule Sequencer or true single-molecule sequencing (tSMS) is vaguely classified as a TGS technology because it has features of both NGS and TGS technologies. It is considered to be a TGS platform because of its ability to perform single DNA molecule sequencing without the need for whole-genome amplification but the sequencing is still based on 'cyclic sequencing' (repeated cycles of sequencing) comprising several steps, such as flow of fluorescent-labelled nucleotides and reagents, nucleotides incorporation, washing and imaging

steps, in each cycle.[32] Therefore, one of the major distinctions between NGS and TGS is that TGS does not require whole-genome amplification steps.

Numerous other TGS technologies, such as SMRT sequencing, are on the horizon and will soon be marketed commercially,[294] whereas others — such as nanopore sequencing — may take several years to become a mature technology.[34,35] SMRT sequencing is performed by synthesising complementary strands of the single DNA molecules by DNA polymerase through incorporation of four different fluorescent colour-labelled nucleotides. The incorporation of each nucleotide into the synthesising DNA strands is monitored in real time by visualisation of 'pulses' of coloured light emitted from each zero-mode waveguide. Each waveguide corresponds to a single molecule of DNA fragment and the incorporation of nucleotides is distinguished by emission of four different colours of light. Similarly, nanopore sequencing requires no cyclic sequencing steps.[32] By comparison, companies such as Complete Genomics (Mountain View, CA) provide a sequencing service, rather than selling their sequencing machines to end-users. The sequencing platform achieves efficient imaging and low reagent consumption with combinatorial probe anchor ligation chemistry independently to assay each base from patterned nanoarrays of self-assembling DNA nanoballs.[44] As TGS is characterised by single DNA molecule sequencing, it has the potential further to increase the number of sequence reads or throughput per instrument run above their current capacity.

Whole-genome (re)sequencing

NGS and TGS technologies have now made possible the sequencing of the entire human genome within a few days. The first human WGS study using a next-generation sequencer was completed in 2008;[46] this marked the beginning of a new era in personalised genome sequencing. To date, more than 20 WGS studies have been completed using NGS and TGS technologies.[45] The number of genomes being sequenced is expected to increase in the coming years, as sequencing technologies and analytical and bioinformatics tools become more advanced and affordable.[295] The reference genome sequence from the HGP is needed for alignments of the large amount of sequence reads produced by the high-throughput sequencers. Clearly, these studies do not involve the *de novo* assembly of human genome sequences, but rather constitute genome resequencing studies.

The first human diploid genome sequence — Craig Venter's genome — appeared in 2007 and was sequenced using the Sanger sequencing method.[68] A year later, the genome of James Watson, who discovered the double-helical structure of the DNA molecule half a century ago, was also sequenced.[46] In contrast to Venter's genome, Watson's genome was sequenced using NGS technologies. A number of additional human genomes have now also been fully sequenced. For example, a single Caucasian/European;[92] a single African (ie NA18507 from the HapMap project, sequenced using two different NGS technologies);[42,296] two Koreans;[297,298] a single Han Chinese;[43] a single Japanese;[299] a single Irish individual[300] and a single Gujarati Indian[301] have been sequenced.

Two whole genomes of the indigenous hunter-gatherer peoples of southern Africa (Khoisan and Bantu) have also been sequenced, together with the protein-coding regions from an additional three hunter-gatherers from the Kalahari. This study has been important for understanding human diversity, as these genomes represent the oldest known lineage of modern humans. A better understanding of genomic differences between the hunter-gatherers and others may help to pinpoint genetic adaptations to an agricultural lifestyle.[302] In addition, the genome of an extinct Palaeo-Eskimo (~4,000-years old)[108] and a Neanderthal genome[107] have been sequenced. The sequencing work of most of these individual genomes was accomplished using NGS technologies.

These WGS studies have identified several hundred thousand new SNPs that had not been previously catalogued in the dbSNP database. For example, Bentley *et al.* (2008)[42] found about one million new SNPs in the African genome

(NA18507), and several hundred thousand new SNPs for other genomes. Most of the common SNPs in human populations have already been captured; thus, the new SNPs identified in these studies are probably representative of those from the lower-frequency spectrum. Data on population frequencies of the new SNPs are not available, since they were derived from individual genome-sequencing studies; however, these data should be available upon completion of the 1000 Genomes Project. In addition to SNPs, several hundred thousand short indels and several thousand structural variants have also been identified.

Schuster *et al.* characterised the extent of whole-genome and exome diversity among five individuals (two whole genomes and three exomes were sequenced) and identified 1.3 million novel DNA differences genome-wide.[302] Interestingly, in terms of nucleotide substitutions, the Bushmen would appear to be genetically more different from each other than Europeans and Asians are to each other. This is consistent with the view that the genetic diversity between African individuals is greater than between individuals from other ethnogeographic origins.[302] A total of 353,151 high-confidence SNPs were identified in the genome of the extinct Palaeo-Eskimo.[108] By comparing the high-confidence SNPs in this extinct human genome with contemporary populations to identify the populations most closely related to this individual, this study provided evidence for a migration from Siberia into the New World some 5,500 years ago. Comparisons of the Neanderthal genome with the genomes of five extant humans from different parts of the world identified a number of genomic regions that may have been affected by positive selection in ancestral modern humans, regions that include genes involved in metabolism and in cognitive and skeletal development.[107]

The WGS studies also identified a portion of the sequence reads that could not be mapped to the NCBI human reference genome, indicating that some sequences are 'missing' from the reference genome. For example, Wheeler *et al.* found that 1.5 million reads (approximately 1.4 per cent of the total sequence data) did not map to the reference genome.[46] These 'unmappable' sequence reads were then assembled into ~170,000 contigs spanning 48 Mb. Even after the removal of contigs that were <100 bp in size, there were still ~110,000 contigs spanning 29 Mb. This concurs with the estimated 25 Mb of euchromatic sequence that is absent from the reference genome. More recent studies using sequencing data have also identified new sequences that are absent in the human reference genome.[303,304]

1000 Genomes Project

The 1000 Genomes Project was initiated in 2008 with the aim of sequencing the genomes of at least 1,000 individuals from different populations around the world (http://www.1000genomes.org/). The main aim of this international collaborative project has been to provide a comprehensive map of human genetic variation for future disease association studies and population genetics. As with the HapMap project, the data from this project also will be made available publicly.

Owing to the ease of high-throughput genotyping technologies, SNPs have been widely used as genetic markers in GWASs to search for disease variants. Evidence has been accumulating to suggest that (common) SNPs alone are unlikely to account for all the heritable risk of complex disease, however.[5] Concurrently, the amount of data supporting associations of CNVs with complex diseases has been growing.[305] Similarly, the importance of rare variants in complex diseases is also increasingly being recognised.[306,307] This indicates that future disease association studies need to interrogate non-SNP and rare genetic variants, requiring a comprehensive catalogue of human genetic variants. Common SNPs have been well documented in the dbSNP, but rarer (or lower frequency) SNPs are still under-represented in the database and information on indels and structural variations is still incomplete.

The completion of the pilot phase of the 1000 Genomes Project identified approximately 15 million SNPs, one million short indels and 20,000 structural variations, most of which were previously

unreported.[102] In addition, the location, allele frequency and local haplotype structure of these genetic variants were described. The sequencing data also enabled characterisation of CNVs within heavily duplicated and near-identical regions.[308] Recently, a map of CNVs was constructed based on WGS data from 185 human genomes in the pilot phase of the project; this encompasses 22,025 deletions and 6,000 additional structural variations, including insertions and tandem duplications. More importantly, approximately half of the structural variations were mapped to single nucleotide resolution, thereby facilitating analysis of their origin and functional impact.[112] Precision in terms of the breakpoint delineation of structural variations is a prerequisite to obtain insights into their underlying mutational mechanisms.[286] The nucleotide resolution analysis of the breakpoints was hampered by the low resolution of the microarrays used in previous studies.

A recent study also identified approximately two million small indels, ranging from 1 bp to 10,000 bp in length, in the genomes of 79 humans. Interestingly, approximately half of these variants (ie 819,363 small indels) mapped to human genes. These small indels were frequently found in the coding exons of these genes, and several lines of evidence indicate that such variation is a major determinant of human biological diversity.[309] This study also found that many of the small indels had high levels of LD with both HapMap-SNPs and GWAS-SNPs, suggesting that a proportion of these indels have already been interrogated indirectly for their associations with complex phenotypes in GWASs through LD with the SNPs as surrogate markers. This also indicates that, in addition to SNPs and larger CNVs, small indel variation is likely to be a key factor underlying the genetics of human complex diseases and traits.

By comparison with WGS, which relies on a reference genome for aligning the sequence reads, *de novo* genome assembly will enable the more thorough and comprehensive detection of various genetic variations in the human genome ranging from single nucleotide variants and small indels, to large structural variations. Currently, *de novo* genome assembly is challenging and less practical because of the short sequence reads generated by NGS technologies, especially the Illumina and Life Technologies sequencing platforms. Recent studies have attempted to perform *de novo* human genome assembly using short sequence reads, with limited success.[291,310,311] One such study showed that *de novo* assemblies were 16.2 per cent shorter than the reference genome, with thousands of coding exons being completely absent.[312] *De novo* genome assembly and haplotype phasing will eventually become more feasible with longer sequence read lengths of up to tens of kb being generated by future sequencing technologies.[33]

Cancer genome sequencing and somatic mutations

Cancers differ from other complex diseases in several aspects. The involvement of somatic mutations in cancer initiation and progression, in addition to germline variations, is well recognised. Sporadic cancer is considered to be an 'acquired disease' caused by the accumulation of somatic mutations in the genome of the original cancer cell type over the lifespan of a patient. Direct sequencing of the cancer genome, and comparison with the genome sequence from constitutional DNA from the same individual as a reference, is required for the proper assessment of somatic mutations.[313,314]

Recent advances in the understanding of the somatic mutational profile of cancer genomes have been driven by NGS technologies, which have enabled numerous whole cancer genomes to be sequenced for the first time.[47–49] Nevertheless, many large-scale targeted resequencing studies of collections of cancer-relevant candidate genes, gene families or the RefSeq genes also have been performed previously using traditional PCR isolation and Sanger sequencing methods. The scale of these targeted studies previously has been limited by the lack of high-throughput sequence capture and sequencing methods.[315,316] By contrast, sequencing of the entire collection of exons in acute monocytic leukaemia was completed without PCR isolation and Sanger sequencing methods.[317]

Although somatic mutations have been found in many genes, only a few genes have been found to be frequently mutated across the tumour samples screened (ie mutated in a significant proportion of cancer samples). These genes have been referred to as 'mountains' — as opposed to the 'hills', which correspond to genes that are infrequently mutated or mutated at low frequency.[316,318–320] For example, the gene encoding V-erb-a erythroblastic leukaemia viral oncogene homolog 4 (avian) (*ERBB4*) was found to be the most highly mutated gene in melanoma and hence may be considered to be a 'mountain'; a considerable proportion of samples (19 per cent) were found to have somatic mutations in this gene, with some samples containing more than one mutation. The role of *ERBB4* was also supported by extensive functional studies showing that various missense mutations increased kinase activity and transformation ability, and the demonstration of reduced cell growth after knockdown of the gene in melanoma cells expressing mutant ERBB4.[320] Targeted cancer genome sequencing has demonstrated the potential to identify potential therapeutic targets for melanoma.

Despite cost constraints, the number of WGS studies performed on different cancers has been increasing[47–49] since the milestone first study that sequenced the cancer genome of an AML patient;[82] however, these WGS studies have generally sequenced only a few samples.[321–323] The ability of WGS to detect somatic mutations in abundance requires us to be able to identify the 'driver' mutations from among the myriad 'passenger' mutations. It has been predicted that approximately ten functional driver mutations are required to cause most cancers, yet up to tens of thousands of mutations may be identified in an analysis of a cancer genome.[313,314] Effective methods for identifying driver mutations in cancer genomes are not well developed, and the criteria for distinguishing driver mutations are not well defined. In addition, the set of driver mutations can be very different for different cancer types.

Although frequently mutated genes and recurrent mutations are of particular interest,[324] all of the current studies have interrogated only one or a few cancer genomes. Thus, these studies are unable to distinguish 'mountains' from 'hills', and recurrent mutations from other mutations that occur only once in the samples. Therefore, testing a subset of somatic mutations identified in the cancer genome in a larger number of cancer samples is required to identify this subset of genes or mutations[325] before the application of WGS in larger samples can be regarded as not only technically feasible, but also affordable (Box 6).

At present, somatic mutations in non-coding regions have received relatively scant attention and should be given more importance, since pervasive transcription beyond the protein-coding regions has now been demonstrated,[165,333,334] suggesting a regulatory role for the non-coding regions. These somatic mutations, and possibly driver mutations in the non-coding regions, can only be revealed by sequencing the whole cancer genome, as opposed to a targeted approach.

Revisiting Mendelian disorders

Mendelian or monogenic disorders make up approximately 7,000 known or suspected disorders and contribute significantly to the disease burden in society.[335–338] Over the past two decades, much progress has been made in identifying the causal mutations and candidate genes for Mendelian disorders through mainly traditional linkage studies.[339] Currently, causal mutations for >4,000 Mendelian disorders have been identified.[99] Indeed, a total of 112,864 different disease-causing and disease-associated mutations in 4,078 human genes are currently (as of May 2011) catalogued in the HGMD (http://www.hgmd.org/) (Box 7).

Although classical linkage studies have been the main tool for elucidating the genetics of Mendelian disorders, not all of these disorders are amenable to this study design. Homozygosity mapping is a more powerful and effective approach to studying recessive disorders in consanguineous families. For those disorders that are not amenable to these two conventional approaches, their causal mutations remain elusive. These disorders include: (a) extremely rare Mendelian disorders where only a small

Box 6. Challenges in cancer genome sequencing

Several major challenges at the forefront of cancer genome sequencing studies are outlined and discussed. The first relates to the collection of 'high-quality' samples of cancer cells or tissues for DNA extraction for sequencing.[48,49] Primary cancer tissues are usually contaminated by other normal cells that hamper our ability to detect somatic mutations in cancer genomes. The contamination with (or mixture of) DNA from non-cancerous cells is particularly problematic, and a higher depth of sequencing coverage will be required to detect somatic mutations in 'mixed DNA', increasing the cost of sequencing. For example, Ding *et al.* studied 188 primary lung adenocarcinoma samples, each containing a minimum of 70 per cent tumour cells independently determined by pathologists.[326] Single-cell sequencing is now emerging as a promising approach to resolving cancer tissue heterogeneity or mixed populations of cells, however, because it is potentially able to resolve genetic and/or cellular heterogeneity among the cancer cells. This single-cell sequencing approach was applied to investigate tumour population structure and evolution in two cases of human breast cancer. Analysis of 100 single cells from a polygenomic tumour revealed three distinct clonal subpopulations that probably represent sequential clonal expansions. Analysis of 100 single cells from a monogenomic primary tumour and its liver metastasis indicated that a single clonal expansion formed the primary tumour and seeded the metastasis.[113]

The second most important challenge is accurately to identify different types of somatic mutations in the cancer genome. NGS technologies are characterised by shorter sequence read lengths and higher sequencing error rates, by comparison with Sanger sequencing.[295] Data quality could be adversely affected if these sequencing errors are not properly filtered out.

Thirdly, the cost of whole-genome resequencing is still prohibitively expensive when it is to be applied to hundreds of samples. Furthermore, there are also significant bioinformatics and analytical challenges to processing and analysing huge amounts of sequencing data. These two constraints currently restrict whole-genome resequencing studies to studies of only a few cancer genomes. This in itself becomes a major barrier to identifying recurrent mutations (which are more likely to be functionally important) and driver mutations. Although the current approach to identifying recurrent mutations is to select a subset of somatic mutations detected in cancer genomes and then to test them in a larger study,[325] this approach cannot be used to screen for all mutations, resulting in many recurrent mutations remaining undetected. For example, a total of 64 mutations were detected in protein-coding genes, regulatory RNAs and highly conserved non-coding regions in the AML genome, but only four of these mutations were subsequently found in additional samples when tested for in more than 180 AML patients. By contrast, targeted resequencing in large sample sizes is able to identify recurrent mutations. This targeted approach focuses only on certain genes, however, and, as a consequence, those recurrent mutations located outside the targeted regions remain undetected. In addition to identifying recurrent mutations, a large sample size is also needed to distinguish 'mountains' from 'hills'.

Although the findings from targeted,[315,316,320,326,327] exome[317,328,329] and whole-genome resequencing[82,321–323,330–332] studies have increasingly provided new insights into cancer genomes, the greatest challenge for cancer genome sequencing lies in discerning driver mutations from the multitude of other (passenger) mutations. Effective methods for identifying driver mutations in cancer genomes are not yet well developed. In addition, driver mutations may differ between cancer types.

number of cases are available; (b) unrelated cases from different families; and (c) sporadic cases due to *de novo* mutations. Exome sequencing now offers new opportunities to study extremely rare disorders and sporadic cases caused by *de novo* mutations, such as Kabuki syndrome and Schinzel–Giedion syndrome.[14,340]

High-throughput sequence capture methods are able to isolate the universe of exons (the 'exome') in a more efficient and cost-effective way than traditional PCR-based methods. These methods are commercially marketed — for example, the NimbleGen Sequence Capture technology (NimbleGen, Madison, WI: http://www.nimblegen.com/) and the Agilent SureSelect Target Enrichment technology (Agilent; Santa Clara, CA: http://www.home.agilent.com). They allow researchers to target custom genomic regions of interest in the human genome of up to tens of Mb in length, and also enable isolation of the exome in a single experiment. This development, coupled with the high-throughput sequencing data produced by NGS technologies, ensures an adequate depth of sequencing coverage accurately to detect the genetic variations in the exome or targeted regions.[295,341,342]

Causal mutations have been identified for a number of previously unexplained rare disorders, such as Miller syndrome,[13] Sensenbrenner syndrome,[343] Perrault syndrome[344] and Fowler syndrome.[345] Exome sequencing is also a useful tool for diagnostic application and is anticipated to be used increasingly in molecular diagnosis.[90,346–348] The genetic diagnosis of congenital chloride diarrhoea in a patient with suspected Bartter syndrome was made through exome sequencing, which revealed a homozygous missense variant in the solute carrier family 26, member 3 gene (*SLC26A3*).[90] The position of this variant is completely conserved from invertebrates to humans. The diagnostic application was further illustrated by Lupski *et al.* through WGS of a proband with Charcot–Marie–Tooth disease.[16] One missense variant and one nonsense variant were detected in *SH3TC2*, and all affected individuals in the family of the proband were found to be compound heterozygotes for these variants.

Studying Mendelian disorders can, paradoxically, reveal genes for complex diseases and traits. For example, numerous GWAS-identified common SNPs which are associated with triglyceride, high-density lipoprotein (HDL) cholesterol and LDL cholesterol levels were also found in the candidate genes causing the monogenic form of these lipid metabolism disorders.[349,350] The discovery of causal

Box 7. Human Gene Mutation Database and the 'human mutome'

As the number of disease-causing or disease-associated germline mutations or variants increases, proper cataloguing is critically important. In this regard, the HGMD represents an attempt to collate all known (published) gene lesions responsible for human inherited disease.

Disease-causing or disease-associated germline mutations/variants collated in the HGMD now exceed 110,000 in >4,000 different nuclear genes. Newly described human gene mutations are currently being reported at a rate of ~10,000 per annum, with ~300 new 'inherited disease genes' being recognised every year. The HGMD has provided useful insights into the 'human mutome' (ie disease-causing or disease-associated germline mutations/variants in the entire human genome).[99,100] For a variety of reasons, however, this figure is likely to represent only a small proportion of the clinically relevant genetic variants present in the human genome. Those disease-causing or disease-associated variants that are located outside the gene-coding regions are likely to have been overlooked often as a direct consequence either of focusing exclusively on screening the protein-coding sequence or of the inherent limitations of the mutation detection techniques used. Such considerations are important for improving mutation screening strategies, as well as for facilitating the interpretation of findings from GWASs, exome sequencing and WGS.

mutations in the disease genes responsible for Mendelian disorders should help in acquiring an understanding of the underlying pathophysiology. For example, the identification of causal mutations in the gene encoding dihydroorotate dehydrogenase (*DHODH*) for Miller syndrome has provided new insights into the role of pyrimidine metabolism in craniofacial and limb development.[13] The potential discovery of new drug targets through study of the genetics of Mendelian disorders should also be emphasised. Thus, statins, the most commonly used drugs to lower cholesterol levels by inhibiting the enzyme 3-hydroxy-3-methyl-glutaryl-CoA (HMG-CoA) reductase, were discovered by studying familial hypercholesterolaemia.[351]

Currently, the return to Mendelian disorder research has been mainly due to the 'attraction' of the exome sequencing approach, coupled with the disappointment engendered by GWAS results that have served to explain only a small fraction of the heritability of complex diseases and traits. Nevertheless, studying complex diseases should not be abandoned, as GWASs have also revealed new biological insights, such as unravelling the autophagy and interleukin (IL)-23 receptor pathways for Crohn's disease.[352–354] The knowledge gained from studying Mendelian disorders and complex diseases will eventually complement each other and come together synergistically to enhance our understanding of genotype–phenotype relationships.

New efforts to identify further causal mutations underlying Mendelian disorders include a recent initiative by the National Human Genome Research Institute (USA) to establish 'A Center for Mendelian Disorders' whose mission will be to take on the sequencing of Mendelian disorders. This centre will be expected to explain the molecular basis of 40–50 disorders per year (NHGRI Large-Scale Sequencing Program May 2010, http://www.genome.gov/).

Sequencing-based approaches to the study of functional genomics

The NGS technologies, since their introduction in 2004, have been increasingly applied in studies of protein–DNA interactions and histone modifications (ChIP-Seq), transcriptomic profiling of mRNAs and non-coding RNAs (RNA-Seq), and bisulphite sequencing of DNA methylation (Methyl-Seq).[39–41]

ChIP-Seq

Previous studies of protein–DNA interactions — such as the identification of transcription factor binding sites — have relied on several low-throughput methods and have been focused on a few specific genomic regions. In the era of microarrays, the genome-wide studies of protein–DNA interactions and histone modifications were performed using a method known as ChIP-chip.[355] Undeniably, microarray development has enabled interrogation on a genome-wide scale but the detection of the immunoprecipitated DNA sequences is still dependent upon the availability of probes to capture them. Although the development of high-density tiling arrays,[356] where oligonucleotide probes are placed in high density throughout the whole genome, has improved the sensitivity of the ChIP-chip, the cost for such tiling arrays is expensive, especially for large genomes like the human genome.[357] By contrast, for ChIP-Seq, the immunoprecipitated DNA sequences are not hybridised on microarrays (thereby avoiding the problems inherent in probe hybridisation experiments) but instead are directly sequenced to detect their presence and measure their abundance. This allows detection of all the DNA fragments or sequences that are immunoprecipitated without any bias in relation to probe selection.[357] This is a key advantage of ChIP-Seq over microarrays.

ChIP-Seq or chromatin immunoprecipitation with the paired-end ditag sequencing (ChIP-PET) methods have led to major advances in the genome-wide mapping of binding sites for transcription factors (eg p53 transcription factor binding sites),[358] and for DNA binding proteins such as neurone restrictive silencer factor (NRSF) and signal transducer and activator of transcription (STAT1).[77,359] Studies of histone modifications have also been revolutionised by means of ChIP-Seq methodology;[78] this has expanded our

knowledge of how this epigenetic mechanism regulates gene expression in the human genome. ChIP-Seq has made an important contribution to the studies of protein–DNA interactions and histone modifications.[360,361]

RNA-Seq

Studies of gene expression are important because they constitute immediate molecular traits that are directly affected by variation in DNA sequences and epigenetics. The term 'gene expression' usually refers to the expression of protein-coding genes. Previous studies were focused on mRNA expression, as mRNAs serve as the templates for protein synthesis; however, this perception changed after the completion of the pilot phase of the ENCODE project. This project and other studies revealed 'pervasive transcription' in the human genome.[165,333,334] Previously it had been thought that only the protein-coding regions or sequences (ie genes) would undergo transcription followed by translation; however, accumulating data are compatible with the view that transcription also occurs in non-protein-coding regions, indicating the importance of studying non-coding RNAs.

The advent of NGS technologies has spawned new approaches to exploring the transcriptome (eg RNA-Seq).[362,363] This method allows the study of the expression of mRNAs and non-coding RNAs, and is also able to detect and identify new transcripts (coding and non-coding) that have not been formally annotated. The applications of sequencing-based approaches in transcriptomic studies have included genome annotation and the discovery of new transcripts,[364] the investigation of the alternative splicing patterns,[84,365] detection of gene fusions in cancer[366] and allele-specific expression analysis,[367] as well as the discovery and measurement of non-coding RNA expression.

Methyl-Seq

Substantial progress has also been achieved in the context of DNA methylation analysis with the advent of NGS technologies allowing the determination of the DNA methylome at a single-base resolution.[96,368–371] The 'gold standard' for the detection of DNA methylation (or cytosine methylation) is sodium bisulphite conversion of DNA followed by sequencing. The sodium bisulphite treatment will convert the unmethylated cytosine to uracil (subsequently read as thymine during sequencing), whereas methylated cytosine remains unchanged. One of the limitations of this method, however, is that it cannot distinguish between 5-methycytosine and 5-hydroxymethylcytosine. The importance of studying 5-hydroxymethylcytosine for its biological roles will become clearer when more powerful methods to distinguish them become available.[372] The SMRT sequencer produced by Pacific Biosciences holds out great promise directly to sequence (and distinguish) 5-methycytosine and 5-hydroxymethylcytosine.[373] Nanopore sequencing technologies have also demonstrated the ability to directly detect methylated cytosines.[35] The revolution in sequencing approaches to exploring functional genomics in the human genome has also led to the initiation of several international projects (Box 8).

Personalised genomic medicine

The translation of genomic information to the clinical setting has shown great promise. In the field of pharmacogenetics, the US Food and Drug Administration (FDA) has approved genotyping tests for the screening of genetic variants in candidate genes that influence the responses and adverse effects of several commonly used anticancer drugs (eg the genes encoding thiopurine S-methyltransferase [*TPMT*] and UDP-glucuronosyltransferase *1A1* [*UGT1A1*] for thiopurine drugs and irrinotecan, respectively). Pharmacogenetic information is important to guide the optimal dose prescription.[384] Similarly, the FDA has also approved genotyping tests for two genes (*CYP2C9* and the vitamin K epoxide reductase complex, subunit 1 gene [*VKORC1*]) in the prescription of warfarin, a drug of low therapeutic index.[385]

The over-expression status of human epidermal growth factor receptor 2 (HER-2) receptors in breast cancer patients is clinically informative in

Box 8. International projects that are exploring functional genomics

The advent of NGS and TGS will facilitate the undertaking of several international projects (http://commonfund.nih.gov/). These large-scale projects would not have been technically feasible without NGS and TGS technologies, which have potentiated sequencing-based approaches in studying functional genomics. These projects will contribute significantly to functional genomics.

The NIH Roadmap Epigenomics Program

The NIH Roadmap Epigenomics Program aims to generate new research tools, technologies, datasets and infrastructure to accelerate our understanding of the role of epigenetics.[374] This will improve our understanding of instances of transcriptional regulation that are not dependent on the DNA sequence. This will be important in understanding diseases attributed to epigenetic aberrations involving DNA methylation or histone modifications.[375] For example, many cancers are commonly associated with epigenetic aberrations.[376]

The Genotype–Tissue Expression (GTeX) Project

Transcriptional regulation is modulated not only by epigenetics, but also by genetic variation in the DNA sequence. Therefore, the GTeX Project aims to study human gene expression and regulation in multiple tissues, providing valuable insights into the mechanisms of gene regulation and, in the future, its disease-relevant aberrations. Genetic variation between individuals will be examined for a correlation with differences in gene expression level. Major advances have been made in studies of eQTL through the use of high-throughput genotyping and sequencing technologies.[377–381] For example, Montgomery *et al.* sequenced the mRNA fraction of the transcriptome in 60 HapMap individuals of European descent and integrated the data with SNP information from the HapMap Phase III project, an undertaking which led to discoveries of novel eQTLs and sequence variants responsible for alternative splicing.[380]

The Human Microbiome Project

The Human Microbiome Project aims to characterise the microbial communities found at several different sites in the human body, such as oral cavities, skin, gastrointestinal tract and the urogenital tract. This project is important in providing insights into the roles of these microbes in human health and disease.[382] The first metagenomic sequencing of gut microbes was accomplished using NGS technologies.[103] A human gut microbial gene catalogue was established by characterisation of 3.3 million non-redundant microbial genes derived from faecal samples from 124 European individuals. This research is important in gaining better understanding of the influence of gut microbes on human health and disease.[103]

The International Cancer Genome Consortium

New developments have also occurred in cancer genomics, where the International Cancer Genome Consortium aims to obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumour types and subtypes.[110] This is in accordance with the notion of integrative analyses incorporating multiple sources of genomics data.[383] This project will be important in dissecting the somatic genetic heterogeneity, a general hallmark of cancer, through studying various tumour types and subtypes.

deciding whether a given patient would benefit from trastuzumab treatment. Similarly, the deletion of *CYP2D6* predicts whether a patient would benefit from tamoxifen treatment, as this prodrug requires bioactivation into its active metabolite, 4-hydroxytamoxifen, which is catalysed by the CYP2D6 enzyme. Thus, breast cancer patients who would not benefit from trastuzumab and tamoxifen treatments should be prescribed alternative drugs, such as aromatase inhibitors. In terms of prognosis, breast cancer prognostic gene expression arrays such as MammaPrint and Oncotype DX are informative and relevant to clinical management, as they help to determine which patients should receive adjuvant therapy after surgery.[386–388] These examples highlight the potential clinical utility of genomic information in prescribing and optimising treatments.

Genomics information has also been used to develop molecular-targeted cancer therapies. The discovery of the breakpoint cluster region–c-abl oncogene 1 nav-receptor tyrosine kinase (*BCR–ABL*) genomic translocation ultimately led to the development of a molecular-targeted drug as a treatment for chronic myeloid leukaemia (CML), namely imatinib — a tyrosine kinase inhibitor targeting the tyrosine kinase domain of the fusion protein.[389] The identification of somatic mutations in the epidermal growth factor receptor (EGFR) in non-small-cell lung cancer led to the development of gefitinib. Further, somatic mutations in EGFR have also been found to be informative in predicting sensitivity to gefitinib and in explaining inter-ethnic variability in drug responses.[390] Advances in epigenetics have led to drug developments such as inhibitors of DNA methylation (DNMTs); indeed, 5-azacytidine and 5-aza-2′-deoxycytidine have been approved in the treatment of AMLs and myelodysplastic syndromes by the US FDA.[391,392] These show that genomic discoveries can be directly translated into clinical applications.

Given the advances in the field, more discoveries will eventually translate into clinical applications and management of patients. For example, GWASs have led to several promising discoveries, such as the identification of genetic variants in *IL28B* that influence the spontaneous clearance of hepatitis C virus and affect the individual response to chronic hepatitis C of interferon-α plus ribavirin therapy.[393,394] Similarly, cancer genome sequencing has identified promising somatic mutations in candidate genes (eg the isocitrate dehydrogenase 1 gene [*IDH1*]) as potential targets for drug interventions. Recurrent mutations in *IDH1* have been found in 12 per cent of glioblastoma multiforme patients.[318] The importance of this gene is not confined to glioblastoma multiforme, as mutations in *IDH1* were also found in 16 per cent of AML patients.[325]

In the era of GWASs and WGS, the great challenge lies in data interpretation and how genomic information can be used to discover new drugs or molecular biomarkers for clinical applications that will eventually translate into patient benefit. The ultimate goal of these studies is to improve the clinical management of patients and to bring about personalised medicine[395,396] through the development of new therapeutic agents tailored to the individual, based upon their genetic information. Although progress made towards achieving these goals has been promising, many challenges in the translational phase remain. Hence, it is still unclear how long it will take for personalised genomic medicine to become an everyday reality.

Summary

The analysis of the sequence of the human genome has had a major impact on biomedical research over the past few years. The HGP has made possible a multitude of genome-wide scale analyses and has thus provided a wealth of information about the architecture of the human genome. In many ways, the HGP has paved the way for what is coming to be called individualised or personalised genome medicine. The development of new (genotyping and sequencing) technologies for improved, less cost-intensive and more precise genome sequencing and assembly has been driven by the overwhelming success of the HGP.

In summary, the advances discussed in this review would not have been possible without the reference genome sequence produced a decade ago

Box 9. Bioinformatics — computational and analytical tools — in the NGS era

Bioinformatics — and computational and analytical tools — play a key role in the NGS era, an era in which huge amounts of sequencing data are being generated. Parallel developments in bioinformatics tools have contributed greatly to recent advances in the field of human structural and functional genomics where NGS technologies have been applied. A detailed discussion of the development of these analytical tools and methodological pipelines is beyond the scope of this paper. However, bioinformatics, computational and analytical tools have been developed for a variety of applications at different stages of the analysis of data generated by both structural and functional genomics studies. Exemplars are given below.

Base calling, alignment, mapping and assembly

1. Base-calling for NGS platforms.[397]
2. Survey of sequence alignment algorithms for NGS.[398]
3. Evaluation of NGS software in mapping and assembly.[399]
4. *De novo* assembly of short sequence reads.[291]
5. Assembly algorithms for NGS data.[400]

Structural genomics (discovery of genetic variations)

6. Computational methods for discovering structural variation with NGS.[282]
7. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes.[401]
8. A framework for variation discovery and genotyping using NGS DNA data.[402]

Functional genomics

9. Introduction to the analysis of high-throughput-sequencing based epigenome data.[403]
10. Computation for ChIP-seq and RNA-seq studies.[404]
11. Bioinformatics approaches for genomics and post-genomics applications of NGS.[405]

Association studies

12. Association studies for NGS.[406]

by the HGP. These advances have greatly improved our understanding of human genetic diversity, disease genetics and functional genomics. The development of powerful analytical and bioinformatics tools is crucially important in the era of genome sequencing (Box 9). The ongoing large-scale international projects will further contribute to the fields of human genetics, as well as human genomics, transcriptomics, epigenomics and metagenomics upon their completion. These projects will provide vital resources for future studies. Continued progress over the next ten years will bring us closer to the final goal of personalised genomic medicine.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C. *et al.* (2001), 'Initial sequencing and analysis of the human genome', *Nature* Vol. 409, pp. 860–921.
2. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W. *et al.* (2001), 'The sequence of the human genome', *Science* Vol. 291, pp. 1304–1351.
3. Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y. *et al.* (2005), 'Complement factor H polymorphism in age-related macular degeneration', *Science* Vol. 308, pp. 385–389.
4. Pennisi, E. (2007), 'Breakthrough of the year. Human genetic variation', *Science* Vol. 318, pp. 1842–1843.
5. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B. *et al.* (2009), 'Finding the missing heritability of complex diseases', *Nature* Vol. 461, pp. 747–753.
6. Clarke, A.J. and Cooper, D.N. (2010), 'GWAS: Heritability missing in action?', *Eur. J. Hum. Genet.* Vol. 18, pp. 859–861.
7. Lander, E.S. (2011), 'Initial impact of the sequencing of the human genome', *Nature* Vol. 470, pp. 187–197.
8. Altshuler, D., Daly, M.J. and Lander, E.S. (2008), 'Genetic mapping in human disease', *Science* Vol. 322, pp. 881–888.
9. Donnelly, P. (2008), 'Progress and challenges in genome-wide association studies in humans', *Nature* Vol. 456, pp. 728–731.
10. Manolio, T.A. and Collins, F.S. (2009), 'The HapMap and genome-wide association studies in diagnosis and therapy', *Annu. Rev. Med.* Vol. 60, pp. 443–456.
11. Feero, W.G., Guttmacher, A.E. and Collins, F.S. (2010), 'Genomic medicine — An updated primer', *N. Engl. J. Med.* Vol. 362, pp. 2001–2011.
12. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D. *et al.* (2009), 'Targeted capture and massively parallel sequencing of 12 human exomes', *Nature* Vol. 461, pp. 272–276.
13. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W. *et al.* (2010), 'Exome sequencing identifies the cause of a Mendelian disorder', *Nat. Genet.* Vol. 42, pp. 30–35.
14. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C. *et al.* (2010), 'Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome', *Nat. Genet.* Vol. 42, pp. 790–793.
15. Rios, J., Stein, E., Shendure, J., Hobbs, H.H. *et al.* (2010), 'Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia', *Hum. Mol. Genet.* Vol. 19, pp. 4313–4318.
16. Lupski, J.R., Reid, J.G., Gonzaga-Jauregui, C., Rio Deiros, D. *et al.* (2010), 'Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy', *N. Engl. J. Med.* Vol. 362, pp. 1181–1191.
17. Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M. *et al.* (2003), 'Large-scale genotyping of complex DNA', *Nat. Biotechnol.* Vol. 21, pp. 1233–1237.
18. Ragoussis, J. (2009), 'Genotyping technologies for genetic research', *Annu. Rev. Genomics Hum. Genet.* Vol. 10, pp. 117–133.
19. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R. *et al.* (2008), 'Genotype, haplotype and copy-number variation in worldwide human populations', *Nature* Vol. 451, pp. 998–1003.
20. Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J. *et al.* (2009), 'Mapping human genetic diversity in Asia', *Science* Vol. 326, pp. 1541–1545.
21. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E. *et al.* (2010), 'Integrating common and rare genetic variation in diverse human populations', *Nature* Vol. 467, pp. 52–58.
22. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L. *et al.* (2006), 'Global variation in copy number in the human genome', *Nature* Vol. 444, pp. 444–454.
23. McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S. *et al.* (2008), 'Integrated detection and population-genetic analysis of SNPs and copy number variation', *Nat. Genet.* Vol. 40, pp. 1166–1174.
24. McQuillan, R., Leutenegger, A.L., Abdel-Rahman, R., Franklin, C.S. *et al.* (2008), 'Runs of homozygosity in European populations', *Am. J. Hum. Genet.* Vol. 83, pp. 359–372.
25. Nothnagel, M., Lu, T.T., Kayser, M. and Krawczak, M. (2010), 'Genomic and geographic distribution of SNP-defined runs of homozygosity in Europeans', *Hum. Mol. Genet.* Vol. 19, pp. 2927–2935.
26. Sebat, J., Lakshmi, B., Troge, J., Alexander, J. *et al.* (2004), 'Large-scale copy number polymorphism in the human genome', *Science* Vol. 305, pp. 525–528.
27. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L. *et al.* (2004), 'Detection of large-scale variation in the human genome', *Nat. Genet.* Vol. 36, pp. 949–951.
28. Gibson, J., Morton, N.E. and Collins, A. (2006), 'Extended tracts of homozygosity in outbred human populations', *Hum. Mol. Genet.* Vol. 15, pp. 789–795.
29. Shendure, J. and Ji, H. (2008), 'Next-generation DNA sequencing', *Nat. Biotechnol.* Vol. 26, pp. 1135–1145.
30. Mardis, E.R. (2008), 'Next-generation DNA sequencing methods', *Annu. Rev. Genomics Hum. Genet.* Vol. 9, pp. 387–402.
31. Metzker, M.L. (2010), 'Sequencing technologies — The next generation', *Nat. Rev. Genet.* Vol. 11, pp. 31–46.
32. Schadt, E.E., Turner, S. and Kasarskis, A. (2010), 'A window into third-generation sequencing', *Hum. Mol. Genet.* Vol. 19, pp. R227–R240.
33. Mardis, E.R. (2011), 'A decade's perspective on DNA sequencing technology', *Nature* Vol. 470, pp. 198–203.
34. Branton, D., Deamer, D.W., Marziali, A., Bayley, H. *et al.* (2008), 'The potential and challenges of nanopore sequencing', *Nat. Biotechnol.* Vol. 26, pp. 1146–1153.
35. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A. *et al.* (2009), 'Continuous base identification for single-molecule nanopore DNA sequencing', *Nat. Nanotechnol.* Vol. 4, pp. 265–270.
36. Derrington, I.M., Butler, T.Z., Collins, M.D., Manrao, E. *et al.* (2010), 'Nanopore DNA sequencing with MspA', *Proc. Natl. Acad. Sci. USA* Vol. 107, pp. 16060–16065.
37. Treffer, R. and Deckert, V. (2010), 'Recent advances in single-molecule sequencing', *Curr. Opin. Biotechnol.* Vol. 21, pp. 4–11.
38. Sanger, F., Nicklen, S. and Coulson, A.R. (1977), 'DNA sequencing with chain-terminating inhibitors', *Proc. Natl. Acad. Sci. USA* Vol. 74, pp. 5463–5467.
39. Mardis, E.R. (2008), 'The impact of next-generation sequencing technology on genetics', *Trends Genet.* Vol. 24, pp. 133–141.
40. Morozova, O. and Marra, M.A. (2008), 'Applications of next-generation sequencing technologies in functional genomics', *Genomics* Vol. 92, pp. 255–264.
41. Werner, T. (2010), 'Next generation sequencing in functional genomics', *Brief. Bioinform.* Vol. 11, pp. 499–511.
42. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P. *et al.* (2008), 'Accurate whole human genome sequencing using reversible terminator chemistry', *Nature* Vol. 456, pp. 53–59.
43. Wang, J., Wang, W., Li, R., Li, Y. *et al.* (2008), 'The diploid genome sequence of an Asian individual', *Nature* Vol. 456, pp. 60–65.
44. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., *et al.* (2010), 'Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays', *Science* Vol. 327, pp. 78–81.
45. Venter, J.C. (2010), 'Multiple personal genomes await', *Nature* Vol. 464, pp. 676–677.
46. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y. *et al.* (2008), 'The complete genome of an individual by massively parallel DNA sequencing', *Nature* Vol. 452, pp. 872–876.
47. Mardis, E.R. and Wilson, R.K. (2009), 'Cancer genome sequencing: A review', *Hum. Mol. Genet.* Vol. 18, pp. R163–R168.
48. Meyerson, M., Gabriel, S. and Getz, G. (2010), 'Advances in understanding cancer genomes through second-generation sequencing', *Nat. Rev. Genet.* Vol. 11, pp. 685–696.
49. Robison, K. (2010), 'Application of second-generation sequencing to cancer genomics', *Brief Bioinform.* Vol. 11, pp. 524–534.
50. Jeffreys, A.J. and Flavell, R.A. (1977), 'The rabbit beta-globin gene contains a large large insert in the coding sequence', *Cell* Vol. 12, pp. 1097–1108.

51. Orkin, S.H., Alter, B.P., Altay, C., Mahoney, M.J. *et al.* (1978), 'Application of endonuclease mapping to the analysis and prenatal diagnosis of thalassemias caused by globin-gene deletion', *N. Engl. J. Med.* Vol. 299, pp. 166–172.
52. Chang, J.C. and Kan, Y.W. (1979), 'Beta 0 thalassemia, a nonsense mutation in man', *Proc. Natl. Acad. Sci. USA* Vol. 76, pp. 2886–2889.
53. Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980), 'Construction of a genetic linkage map in man using restriction fragment length polymorphisms', *Am. J. Hum. Genet.* Vol. 32, pp. 314–331.
54. Watson, J.D. (1990), 'The human genome project: Past, present, and future', *Science* Vol. 248, pp. 44–49.
55. Weissenbach, J., Gyapay, G., Dib, C., Vignal, A. *et al.* (1992), 'A second-generation linkage map of the human genome', *Nature* Vol. 359, pp. 794–801.
56. Cooper, D.N., Ball, E.V. and Krawczak, M. (1998), 'The human gene mutation database', *Nucleic Acids Res.* Vol. 26, pp. 285–287.
57. Risch, N. and Merikangas, K. (1996), 'The future of genetic studies of complex human diseases', *Science* Vol. 273, pp. 1516–1517.
58. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M. *et al.* (2001), 'A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms', *Nature* Vol. 409, pp. 928–933.
59. Pritchard, J.K. (2001), 'Are rare variants responsible for susceptibility to complex diseases?', *Am. J. Hum. Genet.* Vol. 69, pp. 124–137.
60. Reich, D.E. and Lander, E.S. (2001), 'On the allelic spectrum of human disease', *Trends Genet.* Vol. 17, pp. 502–510.
61. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. *et al.* (2001), 'High-resolution haplotype structure in the human genome', *Nat. Genet.* Vol. 29, pp. 229–232.
62. Reich, D.E., Cargill, M., Bolk, S., Ireland, J. *et al.* (2001), 'Linkage disequilibrium in the human genome', *Nature* Vol. 411, pp. 199–204.
63. International HapMap Consortium (2003), 'The International HapMap Project', *Nature* Vol. 426, pp. 789–796.
64. International Human Genome Sequencing Consortium (2004), 'Finishing the euchromatic sequence of the human genome', *Nature* Vol. 431, pp. 931–945.
65. ENCODE Project Consortium (2004), 'The ENCODE (ENCyclopedia Of DNA Elements) Project', *Science* Vol. 306, pp. 636–640.
66. International HapMap Consortium (2005), 'A haplotype map of the human genome', *Nature* Vol. 437, pp. 1299–1320.
67. Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A. *et al.* (2006), 'An initial map of insertion and deletion (INDEL) variation in the human genome', *Genome Res.* Vol. 16, pp. 1182–1190.
68. Levy, S., Sutton, G., Ng, P.C., Feuk, L. *et al.* (2007), 'The diploid genome sequence of an individual human', *PLoS Biol.* Vol. 5, p. e254.
69. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A. *et al.* (2007), 'A second generation human haplotype map of over 3.1 million SNPs', *Nature* Vol. 449, pp. 851–861.
70. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J. *et al.* (2007), 'Genome-wide detection and characterization of positive selection in human populations', *Nature* Vol. 449, pp. 913–918.
71. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R. *et al.* (2007), 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature* Vol. 447, pp. 799–816.
72. Hodges, E., Xuan, Z., Balija, V., Kramer, M. *et al.* (2007), 'Genome-wide in situ exon capture for selective resequencing', *Nat. Genet.* Vol. 39, pp. 1522–1527.
73. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J. *et al.* (2007), 'Microarray-based genomic selection for high-throughput resequencing', *Nat. Methods* Vol. 4, pp. 907–909.
74. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L. *et al.* (2007), 'Direct selection of human genomic loci by microarray hybridization', *Nat. Methods* Vol. 4, pp. 903–905.
75. Kapranov, P., Cheng, J., Dike, S., Nix, D.A. *et al.* (2007), 'RNA maps reveal new RNA classes and a possible function for pervasive transcription', *Science* Vol. 316, pp. 1484–1488.
76. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B. *et al.* (2007), 'Paired-end mapping reveals extensive structural variation in the human genome', *Science* Vol. 318, pp. 420–426.
77. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007), 'Genome-wide mapping of in vivo protein-DNA interactions', *Science* Vol. 316, pp. 1497–1502.
78. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y. *et al.* (2007), 'High-resolution profiling of histone methylations in the human genome', *Cell* Vol. 129, pp. 823–837.
79. Eichler, E.E., Nickerson, D.A., Altshuler, D., Bowcock, A.M. *et al.* (2007), 'Completing the map of human genetic variation', *Nature* Vol. 447, pp. 161–165.
80. Hirst, M. and Marra, M.A. (2010), 'Next generation sequencing based approaches to epigenomics', *Brief. Funct. Genomics* Vol. 9, pp. 455–465.
81. Xi, R., Kim, T.M. and Park, P.J. (2010), 'Detecting structural variations in the human genome using next generation sequencing', *Brief. Funct. Genomics* Vol. 9, pp. 405–415.
82. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B. *et al.* (2008), 'DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome', *Nature* Vol. 456, pp. 66–72.
83. Qiu, J. and Hayden, E.C. (2008), 'Genomics sizes up', *Nature* Vol. 451, p. 234.
84. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. *et al.* (2008), 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nat. Genet.* Vol. 40, pp. 1413–1415.
85. Sultan, M., Schulz, M.H., Richard, H., Magen, A. *et al.* (2008), 'A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome', *Science* Vol. 321, pp. 956–960.
86. Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S. *et al.* (2008), 'Mapping and sequencing of structural variation from eight human genomes', *Nature* Vol. 453, pp. 56–64.
87. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S. *et al.* (2008), 'Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing', *Nat. Genet.* Vol. 40, pp. 722–729.
88. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F. *et al.* (2008), 'Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes', *Nat. Genet.* Vol. 40, pp. 638–645.
89. Maher, B. (2008), 'Personal genomes: The case of the missing heritability', *Nature* Vol. 456, pp. 18–21.
90. Choi, M., Scholl, U.I., Ji, W., Liu, T. *et al.* (2009), 'Genetic diagnosis by whole exome capture and massively parallel DNA sequencing', *Proc. Natl. Acad. Sci. USA* Vol. 106, pp. 19096–19101.
91. Bowers, J., Mitchell, J., Beer, E., Buzby, P.R. *et al.* (2009), 'Virtual terminator nucleotides for next-generation DNA sequencing', *Nat. Methods* Vol. 6, pp. 593–595.
92. Pushkarev, D., Neff, N.F. and Quake, S.R. (2009), 'Single-molecule sequencing of an individual human genome', *Nat. Biotechnol.* Vol. 27, pp. 847–850.
93. Guttman, M., Amit, I., Garber, M., French, C. *et al.* (2009), 'Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals', *Nature* Vol. 458, pp. 223–227.
94. Khalil, A.M., Guttman, M., Huarte, M., Garber, M. *et al.* (2009), 'Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression', *Proc. Natl. Acad. Sci. USA* Vol. 106, pp. 11667–11672.
95. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G. *et al.* (2009), 'Direct RNA sequencing', *Nature* Vol. 461, pp. 814–818.
96. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D. *et al.* (2009), 'Human DNA methylomes at base resolution show widespread epigenomic differences', *Nature* Vol. 462, pp. 315–322.
97. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J. *et al.* (2009), 'An oestrogen-receptor-alpha-bound human chromatin interactome', *Nature* Vol. 462, pp. 58–64.
98. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M. *et al.* (2009), 'Comprehensive mapping of long-range interactions

reveals folding principles of the human genome', *Science* Vol. 326, pp. 289–293.
99. Cooper, D.N., Chen, J.M., Ball, E.V., Howells, K. *et al.* (2010), 'Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics', *Hum. Mutat.* Vol. 31, pp. 631–655.
100. Chen, J.M., Ferec, C. and Cooper, D.N. (2010), 'Revealing the human mutome', *Clin. Genet.* Vol. 78, pp. 310–320.
101. Day, I.N. (2010), 'dbSNP in the detail and copy number complexities', *Hum. Mutat.* Vol. 31, pp. 2–4.
102. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A. *et al.* (2010), 'A map of human genome variation from population-scale sequencing', *Nature* Vol. 467, pp. 1061–1073.
103. Qin, J., Li, R., Raes, J., Arumugam, M. *et al.* (2010), 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature* Vol. 464, pp. 59–65.
104. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L. *et al.* (2010), 'Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index', *Nat. Genet.* Vol. 42, pp. 937–948.
105. Conrad, D.F., Pinto, D., Redon, R., Feuk, L. *et al.* (2010), 'Origins and functional impact of copy number variation in the human genome', *Nature* Vol. 464, pp. 704–712.
106. Pelak, K., Shianna, K.V., Ge, D., Maia, J.M. *et al.* (2010), 'The characterization of twenty sequenced human genomes', *PLoS Genet.* Vol. 6, p. e1001111.
107. Green, R.E., Krause, J., Briggs, A.W., Maricic, T. *et al.* (2010), 'A draft sequence of the Neandertal genome', *Science* Vol. 328, pp. 710–722.
108. Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J.S. *et al.* (2010), 'Ancient human genome sequence of an extinct Palaeo-Eskimo', *Nature* Vol. 463, pp. 757–762.
109. Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E. *et al.* (2010), 'Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants', *Nat. Genet.* Vol. 42, pp. 969–972.
110. Hudson, T.J., Anderson, W., Artez, A., Barker, A.D. *et al.* (2010), 'International network of cancer genome projects', *Nature* Vol. 464, pp. 993–998.
111. Craddock, N., Hurles, M.E., Cardin, N., Pearson, R.D. *et al.* (2010), 'Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls', *Nature* Vol. 464, pp. 713–720.
112. Mills, R.E., Walter, K., Stewart, C., Handsaker, R.E. *et al.* (2011), 'Mapping copy number variation by population-scale genome sequencing', *Nature* Vol. 470, pp. 59–65.
113. Navin, N., Kendall, J., Troge, J., Andrews, P. *et al.* (2011), 'Tumour evolution inferred by single-cell sequencing', *Nature* Vol. 472, pp. 90–94.
114. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B. *et al.* (2007), 'Ensembl 2007', *Nucleic Acids Res.* Vol. 35, pp. D610–617.
115. Kent, W.J., Hsu, F., Karolchik, D., Kuhn, R.M. *et al.* (2005), 'Exploring relationships and mining data with the UCSC Gene Sorter', *Genome Res.* Vol. 15, pp. 737–741.
116. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H. *et al.* (2008), 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Res.* Vol. 36, pp. D13–D21.
117. Griffiths-Jones, S. (2007), 'Annotating noncoding RNA genes', *Annu. Rev. Genomics Hum. Genet.* Vol. 8, pp. 279–298.
118. Clamp, M., Fry, B., Kamal, M., Xie, X. *et al.* (2007), 'Distinguishing protein-coding and noncoding genes in the human genome', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 19428–19433.
119. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C. *et al.* (2009), 'The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes', *Genome Res.* Vol. 19, pp. 1316–1323.
120. Herzog, H., Darby, K., Hort, Y.J. and Shine, J. (1996), 'Intron 17 of the human retinoblastoma susceptibility gene encodes an actively transcribed G protein-coupled receptor gene', *Genome Res.* Vol. 6, pp. 858–861.
121. Vuoristo, J.T., Berrettini, W.H. and Ala-Kokko, L. (2001), 'C18orf2, a novel, highly conserved intronless gene within intron 5 of the GNAL gene on chromosome 18p11', *Cytogenet. Cell Genet.* Vol. 93, pp. 19–22.
122. Yu, P., Ma, D. and Xu, M. (2005), 'Nested genes in the human genome', *Genomics* Vol. 86, pp. 414–422.
123. Denoeud, F., Kapranov, P., Ucla, C., Frankish, A. *et al.* (2007), 'Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions', *Genome Res.* Vol. 17, pp. 746–759.
124. van Bokhoven, H., Rawson, R.B., Merkx, G.F., Cremers, F.P. *et al.* (1996), 'cDNA cloning and chromosomal localization of the genes encoding the alpha- and beta-subunits of human Rab geranylgeranyl transferase: The 3′ end of the alpha-subunit gene overlaps with the transglutaminase 1 gene promoter', *Genomics* Vol. 38, pp. 133–140.
125. Yang, M.Q. and Elnitski, L.L. (2008), 'Diversity of core promoter elements comprising human bidirectional promoters', *BMC Genomics* Vol. 9 (Suppl. 2), p. S3.
126. Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C. *et al.* (2005), 'Evolution and functional classification of vertebrate gene deserts', *Genome Res.* Vol. 15, pp. 137–145.
127. Taylor, J. (2005), 'Clues to function in gene deserts', *Trends Biotechnol.* Vol. 23, pp. 269–271.
128. Libioulle, C., Louis, E., Hansoul, S., Sandor, C. *et al.* (2007), 'Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4', *PLoS Genet.* Vol. 3, p. e58.
129. Parkes, M., Barrett, J.C., Prescott, N.J., Tremelling, M. *et al.* (2007), 'Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility', *Nat. Genet.* Vol. 39, pp. 830–832.
130. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H. *et al.* (2008), 'Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease', *Nat. Genet.* Vol. 40, pp. 955–962.
131. Franke, A., McGovern, D.P., Barrett, J.C., Wang, K. *et al.* (2010), 'Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci', *Nat. Genet.* Vol. 42, pp. 1118–1125.
132. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M. *et al.* (2009), 'Potential etiologic and functional implications of genome-wide association loci for human diseases and traits', *Proc. Natl. Acad. Sci. USA* Vol. 106, pp. 9362–9367.
133. Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P. *et al.* (2009), 'The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer', *Nat. Genet.* Vol. 41, pp. 882–884.
134. Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O. *et al.* (2009), 'The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling', *Nat. Genet.* Vol. 41, pp. 885–890.
135. Harismendy, O., Notani, D., Song, X., Rahim, N.G. *et al.* (2011), '9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response', *Nature* Vol. 470, pp. 264–268.
136. Peters, B.A., St Croix, B., Sjoblom, T., Cummins, J.M. *et al.* (2007), 'Large-scale identification of novel transcripts in the human genome', *Genome Res.* Vol. 17, pp. 287–292.
137. Louro, R., Smirnova, A.S. and Verjovski-Almeida, S. (2009), 'Long intronic noncoding RNA transcription: Expression noise or expression choice?', *Genomics* Vol. 93, pp. 291–298.
138. Ponjavic, J., Ponting, C.P. and Lunter, G. (2007), 'Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs', *Genome Res.* Vol. 17, pp. 556–565.
139. Borel, C., Gagnebin, M., Gehrig, C., Kriventseva, E.V. *et al.* (2008), 'Mapping of small RNAs in the human ENCODE regions', *Am. J. Hum. Genet.* Vol. 82, pp. 971–981.
140. Affymetrix ENCODE Transcriptome Project, Cold Spring Harbor Laboratory ENCODE Transcriptome Project (2009), 'Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs', *Nature* Vol. 457, pp. 1028–1032.

141. Washietl, S., Hofacker, I.L., Lukasser, M., Huttenhofer, A. *et al.* (2005), 'Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome', *Nat. Biotechnol.* Vol. 23, pp. 1383–1390.
142. Collins, L.J. and Penny, D. (2009), 'The RNA infrastructure: dark matter of the eukaryotic cell?', *Trends Genet.* Vol. 25, pp. 120–128.
143. Kawaji, H. and Hayashizaki, Y. (2008), 'Exploration of small RNAs', *PLoS Genet.* Vol. 4, p. e22.
144. Mattick, J.S. (2009), 'The genetic signatures of noncoding RNAs', *PLoS Genet.* Vol. 5, p. e1000459.
145. Mercer, T.R., Dinger, M.E. and Mattick, J.S. (2009), 'Long non-coding RNAs: insights into functions', *Nat. Rev. Genet.* Vol. 10, pp. 155–159.
146. Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S. *et al.* (2008), 'RNA exosome depletion reveals transcription upstream of active human promoters', *Science* Vol. 322, pp. 1851–1854.
147. Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W. *et al.* (2008), 'Divergent transcription from active promoters', *Science* Vol. 322, pp. 1849–1851.
148. Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C. *et al.* (2009), 'Tiny RNAs associated with transcription start sites in animals', *Nat. Genet.* Vol. 41, pp. 572–578.
149. He, L. and Hannon, G.J. (2004), 'MicroRNAs: Small RNAs with a big role in gene regulation', *Nat. Rev. Genet.* Vol. 5, pp. 522–531.
150. Calin, G.A. and Croce, C.M. (2006), 'MicroRNA signatures in human cancers', *Nat. Rev. Cancer* Vol. 6, pp. 857–866.
151. Nevins, J.R. and Potti, A. (2007), 'Mining gene expression profiles: Expression signatures as cancer phenotypes', *Nat. Rev. Genet.* Vol. 8, pp. 601–609.
152. Farazi, T.A., Spitzer, J.I., Morozov, P. and Tuschl, T. (2011), 'miRNAs in human cancer', *J. Pathol.* Vol. 223, pp. 102–115.
153. Nicoloso, M.S., Sun, H., Spizzo, R., Kim, H. *et al.* (2010), 'Single-nucleotide polymorphisms inside microRNA target sites influence tumor susceptibility', *Cancer Res.* Vol. 70, pp. 2789–2798.
154. Ryan, B.M., Robles, A.I. and Harris, C.C. (2010), 'Genetic variation in microRNA networks: The implications for cancer research', *Nat. Rev. Cancer* Vol. 10, pp. 389–402.
155. Huarte, M., Guttman, M., Feldser, D., Garber, M. *et al.* (2010), 'A large intergenic noncoding RNA induced by pp53 mediates global gene repression in the pp53 response', *Cell* Vol. 142, pp. 409–419.
156. Loewer, S., Cabili, M.N., Guttman, M., Loh, Y.H. *et al.* (2010), 'Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells', *Nat. Genet.* Vol. 42, pp. 1113–1117.
157. Huarte, M. and Rinn, J.L. (2010), 'Large non-coding RNAs: Missing links in cancer?', *Hum. Mol. Genet.* Vol. 19, pp. R152–R161.
158. Orom, U.A. and Shiekhattar, R. (2011), 'Long non-coding RNAs and enhancers', *Curr. Opin. Genet. Dev.* Vol. 21, pp. 194–198.
159. Mattick, J.S. and Makunin, I.V. (2006), 'Non-coding RNA', *Hum. Mol. Genet.* Vol. 15 (Spec. No. 1), pp. R17–R29.
160. Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M. *et al.* (2010), 'Non-coding RNAs: regulators of disease', *J. Pathol.* Vol. 220, pp. 126–139.
161. Gingeras, T.R. (2007), 'Origin of phenotypes: Genes and transcripts', *Genome Res.* Vol. 17, pp. 682–690.
162. Rozowsky, J.S., Newburger, D., Sayward, F., Wu, J. *et al.* (2007), 'The DART classification of unannotated transcription within the ENCODE regions: Associating transcription with known and novel loci', *Genome Res.* Vol. 17, pp. 732–745.
163. Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S. *et al.* (2002), 'Large-scale transcriptional activity in chromosomes 21 and 22', *Science* Vol. 296, pp. 916–919.
164. Kapranov, P., Willingham, A.T. and Gingeras, T.R. (2007), 'Genome-wide transcription and the implications for genomic organization', *Nat. Rev. Genet.* Vol. 8, pp. 413–423.
165. Dinger, M.E., Amaral, P.P., Mercer, T.R. and Mattick, J.S. (2009), 'Pervasive transcription of the eukaryotic genome: Functional indices and conceptual implications', *Brief. Funct. Genomic Proteomic* Vol. 8, pp. 407–423.
166. Grinchuk, O.V., Jenjaroenpun, P., Orlov, Y.L., Zhou, J. *et al.* (2010), 'Integrative analysis of the human cis-antisense gene pairs, miRNAs and their transcription regulation patterns', *Nucleic Acids Res.* Vol. 38, pp. 534–547.
167. Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K. *et al.* (2005), 'Antisense transcription in the mammalian transcriptome', *Science* Vol. 309, pp. 1564–1566.
168. Werner, A., Carlile, M. and Swan, D. (2009), 'What do natural antisense transcripts regulate?', *RNA Biol.* Vol. 6, pp. 43–48.
169. Faghihi, M.A. and Wahlestedt, C. (2009), 'Regulatory roles of natural antisense transcripts', *Nat. Rev. Mol. Cell Biol.* Vol. 10, pp. 637–643.
170. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. *et al.* (2008), 'The antisense transcriptomes of human cells', *Science* Vol. 322, pp. 1855–1857.
171. She, X., Jiang, Z., Clark, R.A., Liu, G. *et al.* (2004), 'Shotgun sequence assembly and recent segmental duplications within the human genome', *Nature* Vol. 431, pp. 927–930.
172. Shaw, C.J. and Lupski, J.R. (2004), 'Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease', *Hum. Mol. Genet.* Vol. 13 (Spec. No. 1), pp. R57–R64.
173. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z. *et al.* (2005), 'Segmental duplications and copy-number variation in the human genome', *Am. J. Hum. Genet.* Vol. 77, pp. 78–88.
174. Hastings, P.J., Lupski, J.R., Rosenberg, S.M. and Ira, G. (2009), 'Mechanisms of change in gene copy number', *Nat. Rev. Genet.* Vol. 10, pp. 551–564.
175. Sharp, A.J., Hansen, S., Selzer, R.R., Cheng, Z. *et al.* (2006), 'Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome', *Nat. Genet.* Vol. 38, pp. 1038–1042.
176. Mefford, H.C. and Eichler, E.E. (2009), 'Duplication hotspots, rare genomic disorders, and common disease', *Curr. Opin. Genet. Dev.* Vol. 19, pp. 196–204.
177. Girirajan, S. and Eichler, E.E. (2010), 'Phenotypic variability and genetic susceptibility to genomic disorders', *Hum. Mol. Genet.* Vol. 19, pp. R176–R187.
178. Alkan, C., Kidd, J.M., Marques-Bonet, T., Aksay, G. *et al.* (2009), 'Personalized copy number and segmental duplication maps using next-generation sequencing', *Nat. Genet.* Vol. 41, pp. 1061–1067.
179. Harrison, P.M., Zheng, D., Zhang, Z., Carriero, N. *et al.* (2005), 'Transcribed processed pseudogenes in the human genome: An intermediate form of expressed retrosequence lacking protein-coding ability', *Nucleic Acids Res.* Vol. 33, pp. 2374–2383.
180. Sakai, H., Koyanagi, K.O., Imanishi, T., Itoh, T. *et al.* (2007), 'Frequent emergence and functional resurrection of processed pseudogenes in the human and mouse genomes', *Gene* Vol. 389, pp. 196–203.
181. Zheng, D., Frankish, A., Baertsch, R., Kapranov, P. *et al.* (2007), 'Pseudogenes in the ENCODE regions: Consensus annotation, analysis of transcription, and evolution', *Genome Res.* Vol. 17, pp. 839–851.
182. Nelson, D.R., Zeldin, D.C., Hoffman, S.M., Maltais, L.J. *et al.* (2004), 'Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants', *Pharmacogenetics* Vol. 14, pp. 1–18.
183. Terai, G., Yoshizawa, A., Okida, H., Asai, K. *et al.* (2010), 'Discovery of short pseudogenes derived from messenger RNAs', *Nucleic Acids Res.* Vol. 38, pp. 1163–1171.
184. Zhang, Z.D., Frankish, A., Hunt, T., Harrow, J. *et al.* (2010), 'Identification and analysis of unitary pseudogenes: Historic and contemporary gene losses in humans and other primates', *Genome Biol.* Vol. 11, p. R26.
185. Khachane, A.N. and Harrison, P.M. (2009), 'Assessing the genomic evidence for conserved transcribed pseudogenes under selection', *BMC Genomics* Vol. 10, p. 435.

186. Zheng, D. and Gerstein, M.B. (2007), 'The ambiguous boundary between genes and pseudogenes: The dead rise up, or do they?', *Trends Genet*. Vol. 23, pp. 219–224.
187. Hirotsune, S., Yoshida, N., Chen, A., Garrett, L. *et al.* (2003), 'An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene', *Nature* Vol. 423, pp. 91–96.
188. Svensson, O., Arvestad, L. and Lagergren, J. (2006), 'Genome-wide survey for biologically functional pseudogenes', *PLoS Comput. Biol.* Vol. 2, p. e46.
189. Mills, R.E., Bennett, E.A., Iskow, R.C. and Devine, S.E. (2007), 'Which transposable elements are active in the human genome?', *Trends Genet*. Vol. 23, pp. 183–191.
190. Xing, J., Zhang, Y., Han, K., Salem, A.H. *et al.* (2009), 'Mobile elements create structural variation: Analysis of a complete human genome', *Genome Res.* Vol. 19, pp. 1516–1526.
191. Lin, L., Jiang, P., Shen, S., Sato, S. *et al.* (2009), 'Large-scale analysis of exonized mammalian-wide interspersed repeats in primate genomes', *Hum. Mol. Genet*. Vol. 18, pp. 2204–2214.
192. Jordan, I.K., Rogozin, I.B., Glazko, G.V. and Koonin, E.V. (2003), 'Origin of a substantial fraction of human regulatory sequences from transposable elements', *Trends Genet*. Vol. 19, pp. 68–72.
193. Thornburg, B.G., Gotea, V. and Makalowski, W. (2006), 'Transposable elements as a significant source of transcription regulating signals', *Gene* Vol. 365, pp. 104–110.
194. Piriyapongsa, J., Marino-Ramirez, L. and Jordan, I.K. (2007), 'Origin and evolution of human microRNAs from transposable elements', *Genetics* Vol. 176, pp. 1323–1337.
195. Conley, A.B., Miller, W.J. and Jordan, I.K. (2008), 'Human cis natural antisense transcripts initiated by transposable elements', *Trends Genet*. Vol. 24, pp. 53–56.
196. Lowe, C.B., Bejerano, G. and Haussler, D. (2007), 'Thousands of human mobile element fragments undergo strong purifying selection near developmental genes', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 8005–8010.
197. Nishihara, H., Smit, A.F. and Okada, N. (2006), 'Functional noncoding sequences derived from SINEs in the mammalian genome', *Genome Res.* Vol. 16, pp. 864–874.
198. Asthana, S., Noble, W.S., Kryukov, G., Grant, C.E. *et al.* (2007), 'Widely distributed noncoding purifying selection in the human genome', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 12410–12415.
199. Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J. *et al.* (2006), 'Conserved noncoding sequences are selectively constrained and not mutation cold spots', *Nat. Genet*. Vol. 38, pp. 223–227.
200. Parker, S.C., Hansen, L., Abaan, H.O., Tullius, T.D. *et al.* (2009), 'Local DNA topography correlates with functional noncoding regions of the human genome', *Science* Vol. 324, pp. 389–392.
201. Ponting, C.P. and Lunter, G. (2006), 'Signatures of adaptive evolution within human non-coding sequence', *Hum. Mol. Genet*. Vol. 15 (Spec. No. 2), pp. R170–R175.
202. Eory, L., Halligan, D.L. and Keightley, P.D. (2010), 'Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes', *Mol. Biol. Evol*. Vol. 27, pp. 177–192.
203. Pheasant, M. and Mattick, J.S. (2007), 'Raising the estimate of functional human sequences', *Genome Res.* Vol. 17, pp. 1245–1253.
204. Katzman, S., Kern, A.D., Bejerano, G., Fewell, G. *et al.* (2007), 'Human genome ultraconserved elements are ultraselected', *Science* Vol. 317, p. 915.
205. Licastro, D., Gennarino, V.A., Petrera, F., Sanges, R. *et al.* (2010), 'Promiscuity of enhancer, coding and non-coding transcription functions in ultraconserved elements', *BMC Genomics* Vol. 11, p. 151.
206. McLean, C. and Bejerano, G. (2008), 'Dispensability of mammalian DNA', *Genome Res.* Vol. 18, pp. 1743–1751.
207. Medvedeva, Y.A., Fridman, M.V., Oparina, N.J., Malko, D.B. *et al.* (2010), 'Intergenic, gene terminal, and intragenic CpG islands in the human genome', *BMC Genomics* Vol. 11, p. 48.
208. Bird, C.P., Stranger, B.E., Liu, M., Thomas, D.J. *et al.* (2007), 'Fast-evolving noncoding sequences in the human genome', *Genome Biol*. Vol. 8, p. R118.
209. Prabhakar, S., Noonan, J.P., Paabo, S. and Rubin, E.M. (2006), 'Accelerated evolution of conserved noncoding sequences in humans', *Science* Vol. 314, p. 786.
210. Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D. *et al.* (2007), 'What is a gene, post-ENCODE? History and updated definition', *Genome Res.* Vol. 17, pp. 669–681.
211. Kapranov, P., Drenkow, J., Cheng, J., Long, J. *et al.* (2005), 'Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays', *Genome Res.* Vol. 15, pp. 987–997.
212. Kleinjan, D.A. and Lettice, L.A. (2008), 'Long-range gene control and genetic disease', *Adv. Genet*. Vol. 61, pp. 339–388.
213. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L. *et al.* (2004), 'Genetic analysis of genome-wide variation in human gene expression', *Nature* Vol. 430, pp. 743–747.
214. Gingeras, T.R. (2009), 'Implications of chimaeric non-co-linear transcripts', *Nature* Vol. 461, pp. 206–211.
215. Zhang, C. (2008), 'MicroRNomics: A newly emerging approach for disease biology', *Physiol. Genomics* Vol. 33, pp. 139–147.
216. Pesole, G. (2008), 'What is a gene? An updated operational definition', *Gene* Vol. 417, pp. 1–4.
217. Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J. (2009), 'Human genetic variation and its contribution to complex traits', *Nat. Rev. Genet*. Vol. 10, pp. 241–251.
218. Pang, A.W., MacDonald, J.R., Pinto, D., Wei, J. *et al.* (2010), 'Towards a comprehensive structural variation map of an individual human genome', *Genome Biol*. Vol. 11, p. R52.
219. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M. *et al.* (2002), 'The structure of haplotype blocks in the human genome', *Science* Vol. 296, pp. 2225–2229.
220. Barrett, J.C. and Cardon, L.R. (2006), 'Evaluating coverage of genome-wide association studies', *Nat. Genet*. Vol. 38, pp. 659–662.
221. Eberle, M.A., Ng, P.C., Kuhn, K., Zhou, L. *et al.* (2007), 'Power to detect risk alleles using genome-wide tag SNP panels', *PLoS Genet*. Vol. 3, pp. 1827–1837.
222. Li, M., Li, C. and Guan, W. (2008), 'Evaluation of coverage variation of SNP chips for genome-wide association studies', *Eur. J. Hum. Genet*. Vol. 16, pp. 635–643.
223. Matsuzaki, H., Loi, H., Dong, S., Tsai, Y.Y. *et al.* (2004), 'Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array', *Genome Res.* Vol. 14, pp. 414–425.
224. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. *et al.* (2005), 'A genome-wide scalable SNP genotyping assay using microarray technology', *Nat. Genet*. Vol. 37, pp. 549–554.
225. Cooper, G.M., Zerr, T., Kidd, J.M., Eichler, E.E. *et al.* (2008), 'Systematic assessment of copy number variant detection via genome-wide SNP genotyping', *Nat. Genet*. Vol. 40, pp. 1199–1203.
226. Shen, F., Huang, J., Fitch, K.R., Truong, V.B. *et al.* (2008), 'Improved detection of global copy number variation using high density, non-polymorphic oligonucleotide probes', *BMC Genet*. Vol. 9, p. 27.
227. Kraft, P. and Hunter, D.J. (2009), 'Genetic risk prediction — Are we there yet?', *N. Engl. J. Med*. Vol. 360, pp. 1701–1703.
228. Eichler, E.E., Flint, J., Gibson, G., Kong, A. *et al.* (2010), 'Missing heritability and strategies for finding the underlying causes of complex disease', *Nat. Rev. Genet*. Vol. 11, pp. 446–450.
229. Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P. *et al.* (2010), 'Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis', *Nat. Genet*. Vol. 42, pp. 579–589.
230. Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C. *et al.* (2010), 'Biological, clinical and population relevance of 95 loci for blood lipids', *Nature* Vol. 466, pp. 707–713.
231. Marchini, J. and Howie, B. (2010), 'Genotype imputation for genome-wide association studies', *Nat. Rev. Genet*. Vol. 11, pp. 499–511.
232. Petronis, A. (2010), 'Epigenetics as a unifying principle in the aetiology of complex traits and diseases', *Nature* Vol. 465, pp. 721–727.
233. Maunakea, A.K., Chepelev, I. and Zhao, K. (2010), 'Epigenome mapping in normal and disease states', *Circ. Res.* Vol. 107, pp. 327–339.

234. Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A. *et al.* (2010), 'Genome-wide association studies in diverse populations', *Nat. Rev. Genet*. Vol. 11, pp. 356–366.
235. Need, A.C. and Goldstein, D.B. (2009), 'Next generation disparities in human genomics: Concerns and remedies', *Trends Genet*. Vol. 25, pp. 489–494.
236. Eeles, R.A., Kote-Jarai, Z., Al Olama, A.A., Giles, G.G. *et al.* (2009), 'Identification of seven new prostate cancer susceptibility loci through a genome-wide association study', *Nat. Genet*. Vol. 41, pp. 1116–1121.
237. Nejentsev, S., Walker, N., Riches, D., Egholm, M. *et al.* (2009), 'Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes', *Science* Vol. 324, pp. 387–389.
238. Johansen, C.T., Wang, J., Lanktree, M.B., Cao, H. *et al.* (2010), 'Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia', *Nat. Genet*. Vol. 42, pp. 684–687.
239. Fransen, K., Visschedijk, M.C., van Sommeren, S., Fu, J.Y. *et al.* (2010), 'Analysis of SNPs with an effect on gene expression identifies UBE2L3 and BCL3 as potential new risk genes for Crohn's disease', *Hum. Mol. Genet*. Vol. 19, pp. 3482–3488.
240. Ramagopalan, S.V., Heger, A., Berlanga, A.J., Maugeri, N.J. *et al.* (2010), 'A ChIP-seq defined genome-wide map of vitamin D receptor binding: Associations with disease and evolution', *Genome Res.* Vol. 20, pp. 1352–1360.
241. Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K. *et al.* (2008), 'Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics', *PLoS Genet*. Vol. 4, p. e1000054.
242. Wang, K., Li, M. and Hakonarson, H. (2010), 'Analysing biological pathways in genome-wide association studies', *Nat. Rev. Genet*. Vol. 11, pp. 843–854.
243. Cordell, H.J. (2009), 'Detecting gene–gene interactions that underlie human diseases', *Nat. Rev. Genet*. Vol. 10, pp. 392–404.
244. Thomas, D. (2010), 'Gene-environment-wide association studies: emerging approaches', *Nat. Rev. Genet*. Vol. 11, pp. 259–272.
245. Schork, N.J., Murray, S.S., Frazer, K.A. and Topol, E.J. (2009), 'Common vs. rare allele hypotheses for complex diseases', *Curr. Opin. Genet. Dev.* Vol. 19, pp. 212–219.
246. Bowes, J., Lawrence, R., Eyre, S., Panoutsopoulou, K. *et al.* (2010), 'Rare variation at the TNFAIP3 locus and susceptibility to rheumatoid arthritis', *Hum. Genet*. Vol. 128, pp. 627–633.
247. Bansal, V., Libiger, O., Torkamani, A. and Schork, N.J. (2010), 'Statistical analysis strategies for association studies involving rare variants', *Nat. Rev. Genet*. Vol. 11, pp. 773–785.
248. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S. *et al.* (2010), 'Common SNPs explain a large proportion of the heritability for human height', *Nat. Genet*. Vol. 42, pp. 565–569.
249. Park, J.H., Wacholder, S., Gail, M.H., Peters, U. *et al.* (2010), 'Estimation of effect size distribution from genome-wide association studies and implications for future discoveries', *Nat. Genet*. Vol. 42, pp. 570–575.
250. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M. *et al.* (2008), 'Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia', *Science* Vol. 320, pp. 539–543.
251. Mefford, H.C., Muhle, H., Ostertag, P., von Spiczak, S. *et al.* (2010), 'Genome-wide copy number variation in epilepsy: Novel susceptibility loci in idiopathic generalized and focal epilepsies', *PLoS Genet*. Vol. 6, p. e1000962.
252. Walters, R.G., Jacquemont, S., Valsesia, A., de Smith, A.J. *et al.* (2010), 'A new highly penetrant form of obesity due to deletions on chromosome 16p11.2', *Nature* Vol. 463, pp. 671–675.
253. Bochukova, E.G., Huang, N., Keogh, J., Henning, E. *et al.* (2010), 'Large, rare chromosomal deletions associated with severe early-onset obesity', *Nature* Vol. 463, pp. 666–670.
254. Miyazawa, H., Kato, M., Awata, T., Kohda, M. *et al.* (2007), 'Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients', *Am. J. Hum. Genet*. Vol. 80, pp. 1090–1102.
255. Jiang, H., Orr, A., Guernsey, D.L., Robitaille, J. *et al.* (2009), 'Application of homozygosity haplotype analysis to genetic mapping with high-density SNP genotype data', *PLoS One* Vol. 4, p. e5280.
256. Li, L.H., Ho, S.F., Chen, C.H., Wei, C.Y. *et al.* (2006), 'Long contiguous stretches of homozygosity in the human genome', *Hum. Mutat*. Vol. 27, pp. 1115–1121.
257. Simon-Sanchez, J., Scholz, S., Fung, H.C., Matarin, M. *et al.* (2007), 'Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals', *Hum. Mol. Genet*. Vol. 16, pp. 1–14.
258. Yang, T.L., Guo, Y., Zhang, L.S., Tian, Q. *et al.* (2010), 'Runs of homozygosity identify a recessive locus 12q21.31 for human adult height', *J. Clin. Endocrinol. Metab.* Vol. 95, pp. 3777–3782.
259. Broman, K.W. and Weber, J.L. (1999), 'Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain', *Am. J. Hum. Genet*. Vol. 65, pp. 1493–1500.
260. Harville, H.M., Held, S., Diaz-Font, A., Davis, E.E. *et al.* (2010), 'Identification of 11 novel mutations in eight BBS genes by high-resolution homozygosity mapping', *J. Med. Genet*. Vol. 47, pp. 262–267.
261. Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M.K. *et al.* (2010), 'Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82', *Am. J. Hum. Genet*. Vol. 87, pp. 90–94.
262. Pang, J., Zhang, S., Yang, P., Hawkins-Lee, B. *et al.* (2010), 'Loss-of-function mutations in HPSE2 cause the autosomal recessive urofacial syndrome', *Am. J. Hum. Genet*. Vol. 86, pp. 957–962.
263. Lapunzina, P., Aglan, M., Temtamy, S., Caparros-Martin, J.A. *et al.* (2010), 'Identification of a frameshift mutation in Osterix in a patient with recessive osteogenesis imperfecta', *Am. J. Hum. Genet*. Vol. 87, pp. 110–114.
264. Nicolas, E., Poitelon, Y., Chouery, E., Salem, N. *et al.* (2010), 'CAMOS, a nonprogressive, autosomal recessive, congenital cerebellar ataxia, is caused by a mutant zinc-finger protein, ZNF592', *Eur. J. Hum. Genet*. Vol. 18, pp. 1107–1113.
265. Collin, R.W., Safieh, C., Littink, K.W., Shalev, S.A. *et al.* (2010), 'Mutations in C2ORF71 cause autosomal-recessive retinitis pigmentosa', *Am. J. Hum. Genet*. Vol. 86, pp. 783–788.
266. Rudan, I., Rudan, D., Campbell, H., Carothers, A. *et al.* (2003), 'Inbreeding and risk of late onset complex disease', *J. Med. Genet*. Vol. 40, pp. 925–932.
267. Rudan, I., Campbell, H., Carothers, A.D., Hastie, N.D. *et al.* (2006), 'Contribution of consanguinuity to polygenic and multifactorial diseases', *Nat. Genet*. Vol. 38, pp. 1224–1225.
268. Campbell, H., Carothers, A.D., Rudan, I., Hayward, C. *et al.* (2007), 'Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits', *Hum. Mol. Genet*. Vol. 16, pp. 233–241.
269. Lencz, T., Lambert, C., DeRosse, P., Burdick, K.E. *et al.* (2007), 'Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia', *Proc. Natl. Acad. Sci. USA* Vol. 104, pp. 19942–19947.
270. Nalls, M.A., Guerreiro, R.J., Simon-Sanchez, J., Bras, J.T. *et al.* (2009), 'Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease', *Neurogenetics* Vol. 10, pp. 183–190.
271. Lee, C., Iafrate, A.J. and Brothman, A.R. (2007), 'Copy number variations and clinical cytogenetic diagnosis of constitutional disorders', *Nat. Genet*. Vol. 39, pp. S48–S54.
272. Carter, N.P. (2007), 'Methods and strategies for analyzing copy number variation using DNA microarrays', *Nat. Genet*. Vol. 39, pp. S16–S21.
273. Matsuzaki, H., Wang, P.H., Hu, J., Rava, R. *et al.* (2009), 'High resolution discovery and confirmation of copy number variants in 90 Yoruba Nigerians', *Genome Biol*. Vol. 10, p. R125.
274. Park, H., Kim, J.I., Ju, Y.S., Gokcumen, O. *et al.* (2010), 'Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing', *Nat. Genet*. Vol. 42, pp. 400–405.

275. Yim, S.H., Kim, T.M., Hu, H.J., Kim, J.H. *et al.* (2010), 'Copy number variations in East-Asian population and their evolutionary and functional implications', *Hum. Mol. Genet.* Vol. 19, pp. 1001–1008.
276. Ku, C.S., Pawitan, Y., Sim, X., Ong, R.T. *et al.* (2010), 'Genomic copy number variations in three Southeast Asian populations', *Hum. Mutat.* Vol. 31, pp. 851–857.
277. Feuk, L., Carson, A.R. and Scherer, S.W. (2006), 'Structural variation in the human genome', *Nat. Rev. Genet.* Vol. 7, p. 85–97.
278. Feuk, L. (2010), 'Inversion variants in the human genome: Role in disease and genome architecture', *Genome Med.* Vol. 2, p. 11.
279. Stankiewicz, P. and Lupski, J.R. (2010), 'Structural variation in the human genome and its role in disease', *Annu. Rev. Med.* Vol. 61, pp. 437–455.
280. Alkan, C., Coe, B.P. and Eichler, E.E. (2011), 'Genome structural variation discovery and genotyping', *Nat. Rev. Genet.* Vol. 12, pp. 363–376.
281. Yoon, S., Xuan, Z., Makarov, V., Ye, K. *et al.* (2009), 'Sensitive and accurate detection of copy number variants using read depth of coverage', *Genome Res.* Vol. 19, pp. 1586–1592.
282. Medvedev, P., Stanciu, M. and Brudno, M. (2009), 'Computational methods for discovering structural variation with next-generation sequencing', *Nat. Methods* Vol. 6, pp. S13–S20.
283. Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X. *et al.* (2009), 'Evaluation of next generation sequencing platforms for population targeted sequencing studies', *Genome Biol.* Vol. 10, p. R32.
284. Stephens, P.J., McBride, D.J., Lin, M.L., Varela, I. *et al.* (2009), 'Complex landscapes of somatic rearrangement in human breast cancer genomes', *Nature* Vol. 462, pp. 1005–1010.
285. Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R. *et al.* (2005), 'Fine-scale structural variation of the human genome', *Nat. Genet.* Vol. 37, pp. 727–732.
286. Kidd, J.M., Graves, T., Newman, T.L., Fulton, R. *et al.* (2010), 'A human genome structural variation sequencing resource reveals insights into mutational mechanisms', *Cell* Vol. 143, pp. 837–847.
287. McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D. *et al.* (2008), 'Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease', *Nat. Genet.* Vol. 40, pp. 1107–1112.
288. Willer, C.J., Speliotes, E.K., Loos, R.J., Li, S. *et al.* (2009), 'Six new loci associated with body mass index highlight a neuronal influence on body weight regulation', *Nat. Genet.* Vol. 41, pp. 25–34.
289. International Schizophrenia Consortium (2008), 'Rare chromosomal deletions and duplications increase risk of schizophrenia', *Nature* Vol. 455, pp. 237–241.
290. Mulle, J.G., Dodd, A.F., McGrath, J.A., Wolyniec, P.S. *et al.* (2010), 'Microdeletions of 3q29 confer high risk for schizophrenia', *Am. J. Hum. Genet.* Vol. 87, pp. 229–236.
291. Paszkiewicz, K. and Studholme, D.J. (2010), 'De novo assembly of short sequence reads', *Brief. Bioinform.* Vol. 11, pp. 457–472.
292. Rothberg, J.M. and Leamon, J.H. (2008), 'The development and impact of 454 sequencing', *Nat. Biotechnol.* Vol. 26, pp. 1117–1124.
293. Li, Y. and Wang, J. (2009), 'Faster human genome sequencing', *Nat. Biotechnol.* Vol. 27, pp. 820–821.
294. Eid, J., Fehr, A., Gray, J., Luong, K. *et al.* (2009), 'Real-time DNA sequencing from single polymerase molecules', *Science* Vol. 323, pp. 133–138.
295. Koboldt, D.C., Ding, L., Mardis, E.R. and Wilson, R.K. (2010), 'Challenges of sequencing human genomes', *Brief. Bioinform.* Vol. 11, pp. 484–498.
296. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F. *et al.* (2009), 'Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding', *Genome Res.* Vol. 19, pp. 1527–1541.
297. Ahn, S.M., Kim, T.H., Lee, S., Kim, D. *et al.* (2009), 'The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group', *Genome Res.* Vol. 19, pp. 1622–1629.
298. Kim, J.I., Ju, Y.S., Park, H., Kim, S. *et al.* (2009), 'A highly annotated whole-genome sequence of a Korean individual', *Nature* Vol. 460, pp. 1011–1015.
299. Fujimoto, A., Nakagawa, H., Hosono, N., Nakano, K. *et al.* (2010), 'Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing', *Nat. Genet.* Vol. 42, pp. 931–936.
300. Tong, P., Prendergast, J.G., Lohan, A.J., Farrington, S.M. *et al.* (2010), 'Sequencing and analysis of an Irish human genome', *Genome Biol.* Vol. 11, p. R91.
301. Kitzman, J.O., Mackenzie, A.P., Adey, A., Hiatt, J.B. *et al.* (2011), 'Haplotype-resolved genome sequencing of a Gujarati Indian individual', *Nat. Biotechnol.* Vol. 29, pp. 59–63.
302. Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P. *et al.* (2010), 'Complete Khoisan and Bantu genomes from southern Africa', *Nature* Vol. 463, pp. 943–947.
303. Kidd, J.M., Sampas, N., Antonacci, F., Graves, T. *et al.* (2010), 'Characterization of missing human genome sequences and copy-number polymorphic insertions', *Nat. Methods* Vol. 7, pp. 365–371.
304. Li, R., Li, Y., Zheng, H., Luo, R. *et al.* (2010), 'Building the sequence map of the human pan-genome', *Nat. Biotechnol.* Vol. 28, pp. 57–63.
305. Wain, L.V., Armour, J.A. and Tobin, M.D. (2009), 'Genomic copy number variation, human health, and disease', *Lancet* Vol. 374, pp. 340–350.
306. Bodmer, W. and Bonilla, C. (2008), 'Common and rare variants in multifactorial susceptibility to common diseases', *Nat. Genet.* Vol. 40, pp. 695–701.
307. Bodmer, W. and Tomlinson, I. (2010), 'Rare genetic variants and the risk of cancer', *Curr. Opin. Genet. Dev.* Vol. 20, pp. 262–267.
308. Sudmant, P.H., Kitzman, J.O., Antonacci, F., Alkan, C. *et al.* (2010), 'Diversity of human copy number variation and multicopy genes', *Science* Vol. 330, pp. 641–646.
309. Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U. *et al.* (2011), 'Natural genetic variation caused by small insertions and deletions in the human genome', *Genome Res.* Vol. 21, pp. 830–839.
310. Li, R., Zhu, H., Ruan, J., Qian, W. *et al.* (2010), 'De novo assembly of human genomes with massively parallel short read sequencing', *Genome Res.* Vol. 20, pp. 265–272.
311. Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010), 'State of the art de novo assembly of human genomes from massively parallel sequencing data', *Hum. Genomics* Vol. 4, pp. 271–277.
312. Alkan, C., Sajjadian, S. and Eichler, E.E. (2011), 'Limitations of next-generation genome sequence assembly', *Nat. Methods* Vol. 8, pp. 61–65.
313. Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009), 'The cancer genome', *Nature* Vol. 458, pp. 719–724.
314. Stratton, M.R. (2011), 'Exploring the genomes of cancer cells: Progress and promise', *Science* Vol. 331, pp. 1553–1558.
315. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L. *et al.* (2007), 'Patterns of somatic mutation in human cancer genomes', *Nature* Vol. 446, pp. 153–158.
316. Wood, L.D., Parsons, D.W., Jones, S., Lin, J. *et al.* (2007), 'The genomic landscapes of human breast and colorectal cancers', *Science* Vol. 318, pp. 1108–1113.
317. Yan, X.J., Xu, J., Gu, Z.H., Pan, C.M. *et al.* (2011), 'Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia', *Nat. Genet.* Vol. 43, pp. 309–315.
318. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C. *et al.* (2008), 'An integrated genomic analysis of human glioblastoma multiforme', *Science* Vol. 321, pp. 1807–1812.
319. Jones, S., Zhang, X., Parsons, D.W., Lin, J.C. *et al.* (2008), 'Core signaling pathways in human pancreatic cancers revealed by global genomic analyses', *Science* Vol. 321, pp. 1801–1806.
320. Prickett, T.D., Agrawal, N.S., Wei, X., Yates, K.E. *et al.* (2009), 'Analysis of the tyrosine kinome in melanoma reveals recurrent mutations in ERBB4', *Nat. Genet.* Vol. 41, pp. 1127–1132.

321. Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J. *et al.* (2010), 'A comprehensive catalogue of somatic mutations from a human cancer genome', *Nature* Vol. 463, pp. 191–196.
322. Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J. *et al.* (2010), 'A small-cell lung cancer genome with complex signatures of tobacco exposure', *Nature* Vol. 463, pp. 184–190.
323. Lee, W., Jiang, Z., Liu, J., Haverty, P.M. *et al.* (2010), 'The mutation spectrum revealed by paired genome sequences from a lung cancer patient', *Nature* Vol. 465, pp. 473–477.
324. Ivanov, D., Hamby, S.E., Stenson, P.D., Phillips, A.D. *et al.* (2011), 'Comparative analysis of germline and somatic microlesion mutational spectra in 17 human tumor suppressor genes', *Hum. Mutat.* Vol. 32, pp. 620–632.
325. Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E. *et al.* (2009), 'Recurring mutations found by sequencing an acute myeloid leukemia genome', *N. Engl. J. Med.* Vol. 361, pp. 1058–1066.
326. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R. *et al.* (2008), 'Somatic mutations affect key pathways in lung adenocarcinoma', *Nature* Vol. 455, pp. 1069–1075.
327. Dalgliesh, G.L., Furge, K., Greenman, C., Chen, L. *et al.* (2010), 'Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes', *Nature* Vol. 463, pp. 360–363.
328. Wei, X., Walia, V., Lin, J.C., Teer, J.K. *et al.* (2011), 'Exome sequencing identifies GRIN2A as frequently mutated in melanoma', *Nat. Genet.* Vol. 43, pp. 442–446.
329. Varela, I., Tarpey, P., Raine, K., Huang, D. *et al.* (2011), 'Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma', *Nature* Vol. 469, pp. 539–542.
330. Totoki, Y., Tatsuno, K., Yamamoto, S., Arai, Y. *et al.* (2011), 'High-resolution characterization of a hepatocellular carcinoma genome', *Nat. Genet.* Vol. 43, pp. 464–469.
331. Chapman, M.A., Lawrence, M.S., Keats, J.J., Cibulskis, K. *et al.* (2011), 'Initial genome sequencing and analysis of multiple myeloma', *Nature* Vol. 471, pp. 467–472.
332. Link, D.C., Schuettpelz, L.G., Shen, D., Wang, J. *et al.* (2011), 'Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML', *JAMA* Vol. 305, pp. 1568–1576.
333. Jacquier, A. (2009), 'The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs', *Nat. Rev. Genet.* Vol. 10, pp. 833–844.
334. Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M. *et al.* (2010), 'Annotating non-coding regions of the genome', *Nat. Rev. Genet.* Vol. 11, pp. 559–571.
335. Ropers, H.H. (2007), 'New perspectives for the elucidation of genetic disorders', *Am. J. Hum. Genet.* Vol. 81, pp. 199–207.
336. Ropers, H.H. (2010), 'Single gene disorders come into focus — Again', *Dialogues Clin. Neurosci.* Vol. 12, pp. 95–102.
337. Antonarakis, S.E. and Beckmann, J.S. (2006), 'Mendelian disorders deserve more attention', *Nat. Rev. Genet.* Vol. 7, pp. 277–282.
338. Antonarakis, S.E., Chakravarti, A., Cohen, J.C. and Hardy, J. (2010), 'Mendelian disorders and multifactorial traits: The big divide or one for all?', *Nat. Rev. Genet.* Vol. 11, pp. 380–384.
339. Botstein, D. and Risch, N. (2003), 'Discovering genotypes underlying human phenotypes: Past successes for Mendelian disease, future approaches for complex disease', *Nat. Genet.* Vol. 33 (Suppl.), pp. 228–237.
340. Hoischen, A., van Bon, B.W., Gilissen, C., Arts, P. *et al.* (2010), 'De novo mutations of SETBP1 cause Schinzel-Giedion syndrome', *Nat. Genet.* Vol. 42, pp. 483–485.
341. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I. *et al.* (2010), 'Target-enrichment strategies for next-generation sequencing', *Nat. Methods* Vol. 7, pp. 111–118.
342. Turner, E.H., Ng, S.B., Nickerson, D.A. and Shendure, J. (2009), 'Methods for genomic partitioning', *Annu. Rev. Genomics Hum. Genet.* Vol. 10, pp. 263–284.
343. Gilissen, C., Arts, H.H., Hoischen, A., Spruijt, L. *et al.* (2010), 'Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome', *Am. J. Hum. Genet.* Vol. 87, pp. 418–423.
344. Pierce, S.B., Walsh, T., Chisholm, K.M., Lee, M.K. *et al.* (2010), 'Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome', *Am. J. Hum. Genet.* Vol. 87, pp. 282–288.
345. Lalonde, E., Albrecht, S., Ha, K.C., Jacob, K. *et al.* (2010), 'Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing', *Hum. Mutat.* Vol. 31, pp. 918–923.
346. Bonnefond, A., Durand, E., Sand, O., De Graeve, F. *et al.* (2010), 'Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome', *PLoS One* Vol. 5, p. e13630.
347. Worthey, E.A., Mayer, A.N., Syverson, G.D., Helbling, D. *et al.* (2011), 'Making a definitive diagnosis: Successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease', *Genet. Med.* Vol. 13, pp. 255–262.
348. Montenegro, G., Powell, E., Huang, J., Speziani, F. *et al.* (2011), 'Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family', *Ann. Neurol.* Vol. 69, pp. 464–470.
349. Kathiresan, S., Musunuru, K. and Orho-Melander, M. (2008), 'Defining the spectrum of alleles that contribute to blood lipid concentrations in humans', *Curr. Opin. Lipidol.* Vol. 19, pp. 122–127.
350. Hegele, R.A. (2009), 'Plasma lipoproteins: Genetic influences and clinical implications', *Nat. Rev. Genet.* Vol. 10, pp. 109–121.
351. Brinkman, R.R., Dube, M.P., Rouleau, G.A., Orr, A.C. *et al.* (2006), 'Human monogenic disorders — A source of novel drug targets', *Nat. Rev. Genet.* Vol. 7, pp. 249–260.
352. Mathew, C.G. (2008), 'New links to the pathogenesis of Crohn's disease provided by genome-wide association scans', *Nat. Rev. Genet.* Vol. 9, pp. 9–14.
353. Cho, J.H. (2008), 'The genetics and immunopathogenesis of inflammatory bowel disease', *Nat. Rev. Immunol.* Vol. 8, pp. 458–466.
354. Hirschhorn, J.N. (2009), 'Genomewide association studies — Illuminating biologic pathways', *N. Engl. J. Med.* Vol. 360, pp. 1699–1701.
355. Hanlon, S.E. and Lieb, J.D. (2004), 'Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays', *Curr. Opin. Genet. Dev.* Vol. 14, pp. 697–705.
356. Mockler, T.C., Chan, S., Sundaresan, A., Chen, H. *et al.* (2005), 'Applications of DNA tiling arrays for whole-genome analysis', *Genomics* Vol. 85, pp. 1–15.
357. Park, P.J. (2009), 'ChIP-seq: Advantages and challenges of a maturing technology', *Nat. Rev. Genet.* Vol. 10, pp. 669–680.
358. Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P. *et al.* (2006), 'A global map of pp53 transcription-factor binding sites in the human genome', *Cell* Vol. 124, pp. 207–219.
359. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M. *et al.* (2007), 'Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing', *Nat. Methods* Vol. 4, pp. 651–657.
360. Farnham, P.J. (2009), 'Insights from genomic profiling of transcription factors', *Nat. Rev. Genet.* Vol. 10, pp. 605–616.
361. Zhou, V.W., Goren, A. and Bernstein, B.E. (2011), 'Charting histone modifications and the functional organization of mammalian genomes', *Nat. Rev. Genet.* Vol. 12, pp. 7–18.
362. Wang, Z., Gerstein, M. and Snyder, M. (2009), 'RNA-Seq: A revolutionary tool for transcriptomics', *Nat. Rev. Genet.* Vol. 10, pp. 57–63.
363. Ozsolak, F. and Milos, P.M. (2011), 'RNA sequencing: advances, challenges and opportunities', *Nat. Rev. Genet.* Vol. 12, pp. 87–98.
364. Denoeud, F., Aury, J.M., Da Silva, C., Noel, B. *et al.* (2008), 'Annotating genomes with massive-scale RNA sequencing', *Genome Biol.* Vol. 9, p. R175.
365. Pickrell, J.K., Pai, A.A., Gilad, Y. and Pritchard, J.K. (2010), 'Noisy splicing drives mRNA isoform diversity in human cells', *PLoS Genet.* Vol. 6, p. e1001236.

366. Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S. *et al.* (2009), 'Transcriptome sequencing to detect gene fusions in cancer', *Nature* Vol. 458, pp. 97–101.
367. Heap, G.A., Yang, J.H., Downes, K., Healy, B.C. *et al.* (2010), 'Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing', *Hum. Mol. Genet.* Vol. 19, pp. 122–134.
368. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M. *et al.* (2008), 'Genome-scale DNA methylation maps of pluripotent and differentiated cells', *Nature* Vol. 454, pp. 766–770.
369. Li, Y., Zhu, J., Tian, G., Li, N. *et al.* (2010), 'The DNA methylome of human peripheral blood mononuclear cells', *PLoS Biol.* Vol. 8, p. e1000533.
370. Harris, R.A., Wang, T., Coarfa, C., Nagarajan, R.P. *et al.* (2010), 'Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications', *Nat. Biotechnol.* Vol. 28, pp. 1097–1105.
371. Lister, R. and Ecker, J.R. (2009), 'Finding the fifth base: Genome-wide sequencing of cytosine methylation', *Genome Res.* Vol. 19, pp. 959–966.
372. Dahl, C., Gronbaek, K. and Guldberg, P. (2011), 'Advances in DNA methylation: 5-hydroxymethylcytosine revisited', *Clin. Chim. Acta* Vol. 412, pp. 831–836.
373. Flusberg, B.A., Webster, D.R., Lee, J.H., Travers, K.J. *et al.* (2010), 'Direct detection of DNA methylation during single-molecule, real-time sequencing', *Nat. Methods* Vol. 7, pp. 461–465.
374. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B. *et al.* (2010), 'The NIH Roadmap Epigenomics Mapping Consortium', *Nat. Biotechnol.* Vol. 28, pp. 1045–1048.
375. Portela, A. and Esteller, M. (2010), 'Epigenetic modifications and human disease', *Nat. Biotechnol.* Vol. 28, pp. 1057–1068.
376. Esteller, M. (2007), 'Cancer epigenomics: DNA methylomes and histone-modification maps', *Nat. Rev. Genet.* Vol. 8, pp. 286–298.
377. Gilad, Y., Rifkin, S.A. and Pritchard, J.K. (2008), 'Revealing the architecture of gene regulation: The promise of eQTL studies', *Trends Genet.* Vol. 24, pp. 408–415.
378. Veyrieras, J.B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T. *et al.* (2008), 'High-resolution mapping of expression-QTLs yields insight into human gene regulation', *PLoS Genet.* Vol. 4, p. e1000214.
379. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F. *et al.* (2010), 'Understanding mechanisms underlying human gene expression variation with RNA sequencing', *Nature* Vol. 464, pp. 768–772.
380. Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P. *et al.* (2010), 'Transcriptome genetics using second generation sequencing in a Caucasian population', *Nature* Vol. 464, pp. 773–777.
381. Montgomery, S.B. and Dermitzakis, E.T. (2011), 'From expression QTLs to personalized transcriptomics', *Nat. Rev. Genet.* Vol. 12, pp. 277–282.
382. Peterson, J., Garges, S., Giovanni, M., McInnes, P. *et al.* (2009), 'The NIH Human Microbiome Project', *Genome Res.* Vol. 19, pp. 2317–2323.
383. Hawkins, R.D., Hon, G.C. and Ren, B. (2010), 'Next-generation genomics: An integrative approach', *Nat. Rev. Genet.* Vol. 11, pp. 476–486.
384. Coate, L., Cuffe, S., Horgan, A., Hung, R.J. *et al.* (2010), 'Germline genetic variation, cancer outcome, and pharmacogenetics', *J. Clin. Oncol.* Vol. 28, pp. 4029–4037.
385. Tan, G.M., Wu, E., Lam, Y.Y. and Yan, B.P. (2010), 'Role of warfarin pharmacogenetic testing in clinical practice', *Pharmacogenomics* Vol. 11, pp. 439–448.
386. Hoskins, J.M., Carey, L.A. and McLeod, H.L. (2009), 'CYP2D6 and tamoxifen: DNA matters in breast cancer', *Nat. Rev. Cancer* Vol. 9, pp. 576–586.
387. Kim, C. and Paik, S. (2010), 'Gene-expression-based prognostic assays for breast cancer', *Nat. Rev. Clin. Oncol.* Vol. 7, pp. 340–347.
388. Hartman, M., Loy, E.Y., Ku, C.S. and Chia, K.S. (2010), 'Molecular epidemiology and its current clinical use in cancer management', *Lancet Oncol.* Vol. 11, pp. 383–390.
389. Mauro, M.J., O'Dwyer, M., Heinrich, M.C. and Druker, B.J. (2002), 'STI571: A paradigm of new agents for cancer therapeutics', *J. Clin. Oncol.* Vol. 20, pp. 325–334.
390. Paez, J.G., Janne, P.A., Lee, J.C., Tracy, S. *et al.* (2004), 'EGFR mutations in lung cancer: Correlation with clinical response to gefitinib therapy', *Science* Vol. 304, pp. 1497–1500.
391. Pinto, A. and Zagonel, V. (1993), '5-aza-2′-deoxycytidine (decitabine) and 5-azacytidine in the treatment of acute myeloid leukemias and myelodysplastic syndromes: Past, present and future trends', *Leukemia* Vol. 7 (Suppl. 1), pp. 51–60.
392. Kelly, T.K., De Carvalho, D.D. and Jones, P.A. (2010), 'Epigenetic modifications as therapeutic targets', *Nat. Biotechnol.* Vol. 28, pp. 1069–1078.
393. Thomas, D.L., Thio, C.L., Martin, M.P., Qi, Y. *et al.* (2009), 'Genetic variation in IL28B and spontaneous clearance of hepatitis C virus', *Nature* Vol. 461, pp. 798–801.
394. Suppiah, V., Moldovan, M., Ahlenstiel, G., Berg, T. *et al.* (2009), 'IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy', *Nat. Genet.* Vol. 41, pp. 1100–1104.
395. Guttmacher, A.E., McGuire, A.L., Ponder, B. and Stefansson, K. (2010), 'Personalized genomic information: preparing for the future of genetic medicine', *Nat. Rev. Genet.* Vol. 11, pp. 161–165.
396. Altman, R.B., Kroemer, H.K., McCarty, C.A., Ratain, M.J. *et al.* (2011), 'Pharmacogenomics: Will the promise be fulfilled?', *Nat. Rev. Genet.* Vol. 12, pp. 69–73.
397. Ledergerber, C. and Dessimoz, C. (2011), 'Base-calling for next-generation sequencing platforms', *Brief. Bioinform*, In press.
398. Li, H. and Homer, N. (2010), 'A survey of sequence alignment algorithms for next-generation sequencing', *Brief. Bioinform.* Vol. 11, pp. 473–483.
399. Bao, S., Jiang, R., Kwan, W., Wang, B. *et al.* (2011), 'Evaluation of next-generation sequencing software in mapping and assembly', *J. Hum. Genet.*, In press.
400. Miller, J.R., Koren, S. and Sutton, G. (2010), 'Assembly algorithms for next-generation sequencing data', *Genomics* Vol. 95, pp. 315–327.
401. Hormozdiari, F., Alkan, C., Eichler, E.E. and Sahinalp, S.C. (2009), 'Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes', *Genome Res.* Vol. 19, pp. 1270–1278.
402. Depristo, M.A., Banks, E., Poplin, R., Garimella, K.V. *et al.* (2011), 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nat. Genet.* Vol. 43, pp. 491–498.
403. Huss, M. (2010), 'Introduction into the analysis of high-throughput-sequencing based epigenome data', *Brief. Bioinform.* Vol. 11, pp. 512–523.
404. Pepke, S., Wold, B. and Mortazavi, A. (2009), 'Computation for ChIP-seq and RNA-seq studies', *Nat. Methods* Vol. 6, pp. S22–S32.
405. Horner, D.S., Pavesi, G., Castrignano, T., De Meo, P.D. *et al.* (2010), 'Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing', *Brief. Bioinform.* Vol. 11, pp. 181–197.
406. Luo, L., Boerwinkle, E. and Xiong, M. (2011), 'Association studies for next-generation sequencing', *Genome Res.* Vol. 21, pp. 1099–1108.

Table S1. Special features of human autosomes 1–22 and the sex chromosomes, including respective lengths, gene number and density

Chromosome	Chromosome length (bp)[a]	Number of known protein-coding genes per chromosome[a]	Gene density (genes/ Mb)	Special features	Reference
1	247,249,719	2,189	8.85	Largest human chromosome. Rich in disease genes. Huge (~30 Mb) pericentromeric heterochromatic region at 1q12 spans ~5% of the length of the chromosome. Contains clusters of amylase genes (1p21), U1 snRNA genes (1q12-q22) and 5S RNA genes (1q) as well as multiple (~250) tRNA genes	1
2	242,951,149	1,328	5.47	Chromosome 2 (along with chromosome 4) exhibits the lowest recombination rate of all the autosomes. Contains at 2q13 an ancient telomere–telomere fusion junction at the position where two ape chromosomes once fused to give rise to this human chromosome	2
3	199,501,827	1,112	5.57	Lowest rate of segmental duplication of all human chromosomes. Contains several olfactory receptor gene clusters	3
4	191,273,063	797	4.17	Chromosome 4 (along with chromosome 2) exhibits the lowest recombination rate of all the autosomes. Highest percentage of LINE elements among all chromosomes	2
5	180,857,866	903	4.99	Rich in intra-chromosomal duplications. Contains interleukin and protocadherin gene clusters on 5q31	4
6	170,899,992	1,133	6.62	Harbours the major histocompatibility complex and the largest tRNA gene cluster in the human genome. Contains at least three imprinted genes	5
7	158,821,424	1,023	6.44	Contains the highest number of intra-chromosomal duplications among all human chromosomes. Contains at least six imprinted genes	6, 7
8	146,274,826	747	5.11	Contains a fast-evolving 15 Mb region on distal 8p with genes related to the innate immunity and nervous systems that appear to have evolved under positive selection	8
9	140,273,252	929	6.62	Structurally highly polymorphic. Contains the large (~14 Mb) block of pericentromeric heterochromatin. Contains large numbers of intra- and inter-chromosomal segmental duplications, as well as the largest interferon gene cluster in the human genome (9p22)	9
10	135,374,737	834	6.16	Region of extensive segmental duplication located on 10q11	10

Continued

Table S1. Continued

Chromosome	Chromosome length (bp)[a]	Number of known protein-coding genes per chromosome[a]	Gene density (genes/ Mb)	Special features	Reference
11	134,452,384	1,385	10.30	Rich in both genes and disease genes. Contains 40% of all olfactory receptor gene clusters. Contains at least nine imprinted genes	11
12	132,349,534	1,080	8.16	Chromosome 12 has a unique history of evolutionary rearrangements that occurred in the rodent and primate lineages. Contains clusters of proline-rich protein and type II keratin genes at 12q13	12
13	114,142,980	361	3.16	Low gene density in general; contains a central 38 Mb segment where the gene density drops to only 3.1 genes per Mb. This acrocentric chromosome contains ribosomal RNA genes at 13p12 and at least one imprinted gene	13
14	106,368,585	669	6.29	This acrocentric chromosome contains ribosomal RNA genes at 14p12. Contains two 1 Mb regions of crucial importance to the immune system (T cell receptor and immunoglobulin heavy chain genes). Contains serpin gene cluster at 14q32.1 and several regions with imprinted genes	14
15	100,338,915	641	6.39	This acrocentric chromosome contains ribosomal RNA genes at 15p12. Two large clusters of clinically important segmental duplications are located in the proximal and distal regions of 15q. Contains a number of imprinted genes	15
16	88,827,254	925	10.41	Relatively high gene density. Contains a large number of segmental duplications	16
17	78,774,742	1,236	15.69	High gene density. Has undergone extensive intra-chromosomal rearrangement, many of which were probably mediated by segmental duplications. High G + C content of 45% (genome average: 41%)	17
18	76,117,153	295	3.88	Low gene density overall. Contains serpin gene cluster at 18q21.3	18

Continued

Table S1. Continued

Chromosome	Chromosome length (bp)[a]	Number of known protein-coding genes per chromosome[a]	Gene density (genes/ Mb)	Special features	Reference
19	63,811,651	1,443	22.61	Highest gene density of all human chromosomes. One quarter of the genes on chromosome 19 belong to tandemly arranged gene families, encompassing 25% of the length of the chromosome. High G + C content of 48–49% (genome average: 41%). Repetitive sequences constitute 53–57% of the chromosome, as compared with a genome average of 40–44%. Contains clusters of olfactory receptor genes and cytochrome P450 genes, and multiple clusters of zinc finger genes, and at least two imprinted genes	19
20	62,435,964	617	9.88	Smallest metacentric autosome. Rich in both genes and disease genes. Contains type 2 cystatin gene cluster and at least two imprinted genes	20
21	46,944,323	284	6.05	Smallest human chromosome with fewer genes than any other autosome. This acrocentric chromosome contains ribosomal RNA genes at 21p12	21
22	49,691,432	519	10.44	This acrocentric chromosome contains ribosomal RNA genes at 22p12. Relatively high gene density. Clusters of segmental duplications at 22q11.2 are associated with several genomic disorders	22
X	154,913,754	891	5.75	Contains the pseudoautosomal regions, PAR1 and PAR2, at the tips of the short and long arms, respectively. These regions are essential for normal male meiosis and recombination. PAR1 undergoes an obligate crossover with the Y chromosome, thereby giving this region the highest recombination rate in the human genome, at least in males. One X chromosome is subject to inactivation in females. Highly enriched in interspersed repeats and has a low G + C content of 39% (genome average: 41%)	23

Continued

Table S1. Continued

Chromosome	Chromosome length (bp)[a]	Number of known protein-coding genes per chromosome[a]	Gene density (genes/ Mb)	Special features	Reference
Y	57,772,954	80	1.38	Lowest gene density of all human chromosomes (contains only 82 known genes). Contains the male-specific region which is a mosaic of heterochromatin and euchromatic X-transposed, X-degenerate and ampliconic sequences that make up 30% of the euchromatin. PAR1 undergoes an obligate crossover with the X chromosome. The virtual absence of homologous recombination between the X and the Y chromosomes has led to a gradual degeneration of Y chromosomal genes over evolutionary time. However, the absence of recombination, at least within the extensive non-recombining region of the Y chromosome, has also favoured the evolutionary accumulation of transposable elements on the Y chromosome	24

[a]Chromosome lengths and the numbers of genes per chromosome are according to the Ensembl database, version 47.36. The chromosome length corresponds to the length of each chromosome that has been sequenced so far. The number of known protein-coding genes represents a conservative estimate of the likely total number, comprising genes which have been fully annotated. An earlier version of this table was published by Kehrer-Sawatzki and Cooper.[25]

[1]Gregory, S.G., Barlow, K.F, McLay, K.E., Kaul, R. *et al.* (2006), 'The DNA sequence and biological annotation of human chromosome 1', *Nature* Vol. 441, pp. 315–321.
[2]Hillier, L.W., Graves, T.A., Fulton, R.S., Fulton, L.A. *et al.* (2005), 'Generation and annotation of the DNA sequences of human chromosomes 2 and 4', *Nature* Vol. 434, pp. 724–731.
[3]Muzny, D.M., Scherer, S.E., Kaul, R., Wang, J. *et al.* (2006), 'The DNA sequence, annotation and analysis of human chromosome 3', *Nature* Vol. 440, pp. 1194–1198.
[4]Schmutz, J., Martin, J., Terry, A., Couronne, O. *et al.* (2004), 'The DNA sequence and comparative analysis of human chromosome 5', *Nature* Vol. 431, pp. 268–274.
[5]Mungall, A.J., Palmer, S.A., Sims, S.K., Edwards, C.A. *et al.* (2003), 'The DNA sequence and analysis of human chromosome 6', *Nature* Vol. 425, pp. 805–811.
[6]Hillier, L.W., Fulton, R.S., Fulton, L.A., Graves, T.A. *et al.* (2003), 'The DNA sequence of human chromosome 7', *Nature* Vol. 424, pp. 157–164.
[7]Scherer, S.W., Cheung, J., MacDonald, J.R., Osborne, L.R. *et al.* (2003), 'Human chromosome 7: DNA sequence and biology', *Science* Vol. 300, pp. 767–772.
[8]Nusbaum, C., Mikkelsen, T.S., Zody, M.C., Asakawa, S. *et al.* (2006), 'DNA sequence and analysis of human chromosome 8', *Nature* Vol. 439, pp. 331–335.
[9]Humphray, S.J., Oliver, K., Hunt, A.R., Plumb, R.W. *et al.* (2004), 'DNA sequence and analysis of human chromosome 9', *Nature* Vol. 429, pp. 369–374.
[10]Deloukas, P., Earthrowl, M.E., Grafham, D.V., Rubenfield, M. *et al.* (2004), 'The DNA sequence and comparative analysis of human chromosome 10', *Nature* Vol. 429, pp. 375–381.
[11]Taylor, T.D., Noguchi, H., Totoki, Y., Toyoda, A. *et al.* (2006), 'Human chromosome 11 DNA sequence and analysis including novel gene identification', *Nature* Vol. 440, pp. 497–500.
[12]Scherer, S.E., Muzny, D.M., Buhay, C.J., Chen, R. *et al.* (2006), 'The finished DNA sequence of human chromosome 12', *Nature* Vol. 440, pp. 346–351.
[13]Dunham, A., Matthews, L.H., Burton, J., Ashurst, J.L. *et al.* (2004), 'The DNA sequence and analysis of human chromosome 13', *Nature* Vol. 428, pp. 522–528.
[14]Heilig, R., Eckenberg, R., Petit, J.L., Fonknechten, N. *et al.* (2003), 'The DNA sequence and analysis of human chromosome 14', *Nature* Vol. 421, pp. 601–607.
[15]Zody, M.C., Garber, M., Sharpe, T., Young, S.K. *et al.* (2006), 'Analysis of the DNA sequence and duplication history of human chromosome 15', *Nature* Vol. 440, pp. 671–675.
[16]Martin, J., Han, C., Gordon, L.A., Terry, A. *et al.* (2004), 'The sequence and analysis of duplication-rich human chromosome 16', *Nature* Vol. 432, pp. 988–994.
[17]Zody, M.C., Garber, M., Adams, D.J., Sharpe, T. *et al.* (2006), 'DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage', *Nature* Vol. 440, pp. 1045–1049.
[18]Nusbaum, C., Zody, M.C., Borowsky, M.L., Kamal, M. *et al.* (2005), 'DNA sequence and analysis of human chromosome 18', *Nature* Vol. 437, pp. 551–555.
[19]Grimwood, J., Gordon, L.A., Olsen, A., Terry, A. *et al.* (2004), 'The DNA sequence and biology of human chromosome 19', *Nature* Vol. 428, pp. 529–535.
[20]Deloukas, P., Matthews, L.H., Ashurst, J., Burton, J. *et al.* (2001), 'The DNA sequence and comparative analysis of human chromosome 20', *Nature* Vol. 414, pp. 865–871.
[21]Hattori, M., Fujiyama, A., Taylor, T.D., Watanabe, H. *et al.* (2000), 'The DNA sequence of human chromosome 21', *Nature* Vol. 405, pp. 311–319.
[22]Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S. *et al.* (1999), 'The DNA sequence of human chromosome 22', *Nature* Vol. 402, pp. 489–495.
[23]Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S. *et al.* (2005), 'The DNA sequence of the human X chromosome', *Nature* Vol. 434, pp. 325–337.
[24]Skaletsky, H., Kuroda–Kawaguchi, T., Minx, P.J., Cordum, H.S. *et al.* (2003), 'The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes', *Nature* Vol. 423, pp. 825–837.
[25]Kehrer–Sawatzki, H. and Cooper, D.N. (2008), 'Sequencing the human genome: novel insights into its structure and function', in: *Encyclopedia of Life Sciences (ELS)*, John Wiley & Sons Ltd, Chichester.

