REVIEW

doi:10.1038/nature09792

# Initial impact of the sequencing of the human genome

Eric S. Lander[1]

**The sequence of the human genome has dramatically accelerated biomedical research. Here I explore its impact, in the decade since its publication, on our understanding of the biological functions encoded in the genome, on the biological basis of inherited diseases and cancer, and on the evolution and history of the human species. I also discuss the road ahead in fulfilling the promise of genomics for medicine.**

On 15 February 2001, a decade ago this week, *Nature* published a 62-page paper entitled 'Initial sequencing and analysis of the human genome', reporting a first global look at the contents of the human genetic code. The paper[1] marked a milestone in the international Human Genome Project (HGP), a discovery programme conceived in the mid-1980s and launched in 1990. The same week, *Science* published a paper[2] from the company Celera Genomics, reporting a draft human sequence based on their own prodigious data, as well as data from the public HGP.

The human genome has had a certain tendency to incite passion and excess: from early jeremiads that the HGP would strangle research by consuming the NIH budget (it never rose to more than 1.5%); to frenzied coverage of a late-breaking genome race between public and private protagonists; to a White House announcement of the draft human sequence in June 2000, 8 months before scientific papers had actually been written, peer-reviewed and published; to breathless promises from Wall Street and the press about the imminence of genetic 'crystal balls' and genome-based panaceas; to a front-page news story on the tenth anniversary of the announcement that chided genome scientists for not yet having cured most diseases.

The goal of this review is to step back and assess the fruits of the HGP from a scientific standpoint, addressing three questions: what have we learned about the human genome itself over the past decade? How has the human sequence propelled our understanding of human biology, medicine, evolution and history? What is the road ahead?

The past decade has shown the power of genomic maps and catalogues for biomedical research. By providing a comprehensive scaffold, the human sequence has made it possible for scientists to assemble often fragmentary information into landscapes of biological structure and function: maps of evolutionary conservation, gene transcription, chromatin structure, methylation patterns, genetic variation, recombinational distance, linkage disequilibrium, association to inherited diseases, genetic alterations in cancer, selective sweeps during human history and three-dimensional organization in the nucleus. By providing a framework to cross-reference information across species, it has connected the biology of model systems to the physiology of the human. Furthermore, by providing comprehensive catalogues of genomic information, it has enabled genes and proteins to be recognized based on unique 'tags'—allowing, for example, RNA transcripts to be assayed with arrays of oligonucleotide probes and proteins by detection of short peptide fragments in a mass spectrometer. In turn, these measurements have been used to construct 'cellular signatures' characteristic of specific cell types, states and responses, and catalogues of the contents of organelles such as the mitochondria. The intensity of interest can be seen in the 2.5 million queries per week on the major genome data servers and in the flowering of a rich field of computational biology.

The greatest impact of genomics has been the ability to investigate biological phenomena in a comprehensive, unbiased, hypothesis-free manner. In basic biology, it has reshaped our view of genome physiology, including the roles of protein-coding genes, non-coding RNAs and regulatory sequences.

In medicine, genomics has provided the first systematic approaches to discover the genes and cellular pathways underlying disease. Whereas candidate gene studies yielded slow progress, comprehensive approaches have resulted in the identification of ~2,850 genes underlying rare Mendelian diseases, ~1,100 loci affecting common polygenic disorders and ~150 new recurrent targets of somatic mutation in cancer. These discoveries are propelling research throughout academia and industry.

The following sections contain only a small number of citations due to space limitations; a more extensive bibliography tied to each section can be found as Supplementary Information.

## Genome sequencing

### The view from 2000

Genome sequencing was a daunting task in late 2000. The catalogue of organisms with published genome sequences was small: thirty-eight bacteria, one fungus (*Saccharomyces cerevisiae*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*) and one plant (*Arabidopsis thaliana*), all with relatively small and simple genomes.

The human genome was much more challenging, being roughly an order of magnitude larger than the total of all previous genomes and filled with repetitive sequences. The human draft sequence reported in 2001, although a landmark, was still highly imperfect. It covered only ~90% of the euchromatic genome, was interrupted by ~250,000 gaps, and contained many errors in the nucleotide sequence[1].

### Finishing the human

After the draft sequence, the HGP consortium quickly turned to producing a high-quality reference sequence, through lapidary attention to individual clones spanning the genome. In 2004, it published a near-complete sequence that was a vast improvement[3]: it contained ~99.7% of the euchromatic genome, was interrupted by only ~300 gaps, and harboured only one nucleotide error per 100,000 bases. The gaps consisted of ~28 megabases (Mb) of euchromatic sequence, mostly involving repetitive regions that could not be reliably cloned or assembled, and ~200 Mb of heterochromatic sequence, including the large centromeres and the short arms of acrocentric chromosomes.

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.

This 'finished' sequence allowed accurate inference of gene structure and detection of polymorphisms and mutation across the genome. Importantly, the clone-based approach also provided access to regions with recent segmental duplications, which are poorly represented in whole-genome shotgun sequencing. These segmental-duplication-rich regions have extraordinarily high rates of copy-number variation within and between species, and they mediate large-scale chromosomal rearrangement. In addition, they are nurseries for rapidly evolving gene families under positive selection, and they account for a disproportionate share of disease burden, particularly for neurological and neuropsychiatric disorders.

### Expanding the bestiary

Additional vertebrate genomes followed rapidly, to help interpret the human genome through comparative analysis and to enable experimental studies in model systems. Key genomes included mouse[4], dog[5], rat[6], chimpanzee[7] and cow, as well as a marsupial[8], monotreme and bird. Remarkably, partial genome sequences have even been obtained from several extinct species, notably the woolly mammoth and our closest relative, Neanderthal[9]. The current catalogue includes ~250 eukaryotes (totalling ~120 gigabases (Gb)) and ~4,000 bacteria and viruses (~5 Gb). Sequencing has extended to microbial communities, including samples from the mid-ocean and environmental remediation sites and human samples from gut and skin[10,11].

### Massively parallel sequencing

The HGP used essentially the same sequencing method introduced by Sanger in 1977: electrophoretic separation of mixtures of randomly terminated extension products, although dramatically improved with fluorescently labelled terminators and automated laser detectors. The past half-decade has seen a tectonic shift in sequencing technology, based on *in situ* sequencing in which two-dimensional optical imaging is used to monitor sequential addition of nucleotides to spatially arrayed DNA templates. Whereas electrophoretic methods could deploy ~$10^2$ parallel channels, optical imaging can now follow ~$10^9$ templates. The per-base cost of DNA sequencing has plummeted by ~100,000-fold over the past decade, far outpacing Moore's law of technological advance in the semiconductor industry. The current generation of machines can read ~250 billion bases in a week, compared to ~25,000 in 1990 and ~5 million in 2000.

A drawback of the new technology is that the sequence reads are much shorter than the ~700 bases routinely provided by electrophoretic methods. Because it is challenging to assemble a genome sequence *de novo* from such short reads, most applications have focused on placing reads onto the scaffold of an existing genome sequence to count their density or to look for differences from a reference sequence.

### Applications

An early application of massively parallel sequencing was to create 'epigenomic maps', showing the locations of specific DNA modifications, chromatin modifications and protein-binding events across the human genome. Chromatin modification and protein binding can be mapped by chromatin immunoprecipitation-sequencing (ChIP-Seq)[12,13], and the sites of DNA methylation can be found by sequencing DNA in which the methylated cytosines have been chemically modified (Methyl-Seq)[14].

As the technology has improved, the focus has turned to re-sequencing human samples to study inherited variation or somatic mutations. One can re-sequence the whole genome[15] to varying degrees of coverage or use hybridization-capture techniques[16] to re-sequence a targeted subset, such as the protein-coding sequences (referred to as the 'exome').

Sequencing is also being extensively applied to RNA transcripts (RNA-Seq), to count their abundance, identify novel splice forms or spot mutations[17]. A harder challenge is reconstructing a transcriptome *de novo*, but good algorithms have recently been developed[18,19].

The hardest challenge is *de novo* assembly of entire genomes, but even here there has been recent progress in achieving long-range connectivity. For the human, initial efforts yielded scaffolds of modest size (~500 kb), and recent algorithms[20] have approached the typical range for capillary-based sequencing (11.5-Mb scaffolds, containing ~90% of the genome). Encouraged by this progress, a scientific consortium has begun laying a plan to sequence 10,000 vertebrate genomes—roughly one from every genus.

### The road ahead

The ultimate goal is for sequencing to become so simple and inexpensive that it can be routinely deployed as a general-purpose tool throughout biomedicine. Medical applications will eventually include characterizing patients' germline genomes (to detect strongly predictive mutations for presymptomatic counselling where treatments exist, to search for causes of disease of unknown aetiology, and to detect heterozygous carriers for prenatal counselling); cancer genomes (by identifying somatic mutations to compare tumour and normal DNA); immune repertoires (by reading the patterns of B-cell and T-cell receptors to infer disease exposures and monitor responses to vaccines); and microbiomes (by associating patterns of microbial communities with diseases processes). Research applications will include characterizing genomes, epigenomes and transcriptomes of humans and other species, as well as using sequencing as a proxy to probe diverse molecular interactions.

To fulfil this potential, the cost of whole-genome sequencing will need eventually to approach a few hundred US dollars. With new approaches under development and market-based competition, these goals may be feasible within the next decade.

## Genome anatomy and physiology

### The view from 2000

Our knowledge of the contents of the human genome in 2000 was surprisingly limited. The estimated count of protein-coding genes fluctuated wildly. Protein-coding information was thought to far outweigh regulatory information, with the latter consisting largely of a few promoters and enhancers per gene. The role of non-coding RNAs was largely confined to a few classical cellular processes. And, the transposable elements were largely regarded as genomic parasites.

A decade later, we know that all of these statements are false. The genome is far more complex than imagined, but ultimately more comprehensible because the new insights help us to imagine how the genome could evolve and function.

### Protein–coding genes

Since the early 1970s, the total number of genes (the vast majority assumed to be protein-coding) had been variously estimated at anywhere from ~35,000 to well over 100,000, based on genetic load arguments, hybridization experiments, the average size of genes, the number of CpG islands and shotgun sequencing of expressed sequence tags. The HGP paper suggested a total of 30,000–40,000 protein-coding genes, but the estimate involved considerable guesswork owing to the imperfections of the draft sequence and the inherent difficulty of gene identification.

Today, the human genome is known to contain only ~21,000 distinct protein-coding genes[21]. Generating a reliable gene catalogue required eliminating the many open reading frames (ORFs) that occur at random in transcripts, while retaining those that encode bona fide proteins. The key insight was to identify those ORFs with the evolutionary signatures of bona fide protein-coding genes (such as amino-acid-preserving substitutions and reading-frame-preserving deletions) and prove that most ORFs without such conservation are not newly arising protein-coding genes. Recent RNA-Seq projects have confirmed the gene catalogue, while illuminating alternative splicing, which seems to occur at >90% of protein-coding genes and results in many more proteins than genes.

The proteome is now known to be similar across placental mammals, with about two-thirds of protein-coding genes having 1:1 orthologues across species and most of the rest belonging to gene families that undergo regular duplication and divergence—the invention of fundamentally new proteins is rare.

## Conserved non-coding elements

The most surprising discovery about the human genome was that the majority of the functional sequence does not encode proteins. These features had been missed by decades of molecular biology, because scientists had no clue where to look.

Comparison of the human and mouse genomes showed a substantial excess of conserved sequence, relative to the neutral rate in ancestral repeat elements[4]. The excess implied that at least 6% of the human genome was under purifying selection over the past 100 million years and thus biologically functional. Protein-coding sequences, which comprise only ~1.5% of the genome, are thus dwarfed by functional conserved non-coding elements (CNEs). Subsequent comparison with the rat and dog genomes confirmed these findings[5,6].

Although the initial analysis provided a bulk estimate of the amount of conserved sequence, it could only pinpoint the most highly conserved elements. Among them are nearly 500 ultraconserved elements (200 bases or more perfectly conserved across human, mouse and rat), most of which neither overlap protein-coding exons nor show evidence of being transcribed[22]. On the basis of statistical measures of constraint, tens of thousands of additional highly conserved non-coding elements (HCNEs) were identified[5,23]. In many cases the evolutionary origins of these HCNEs could be traced back to the common ancestor of human and fish. HCNEs preferentially reside in the gene deserts that often flank genes with key functions in embryonic development[5,22,23]. Large-scale screens of these sequences in transgenic mice revealed that they are highly enriched in tissue-specific transcriptional enhancers active during embryonic development[24], revealing a stunning complexity of the gene regulatory architecture active in early development[24] (Fig. 1).

Sequencing additional genomes has gradually increased our power to pinpoint the less stringently conserved CNEs. Recent comparison with 29 mammalian genomes has identified millions of additional conserved elements, comprising about two-thirds of the total conserved sequence.

Evolutionary analyses of CNEs have also enabled the discovery of distinct types of functional elements, including regulatory motifs present in the promoters and untranslated regions of co-regulated genes, insulators that constrain domains of gene expression, and families of conserved secondary structures in RNAs. Nonetheless, the function of most CNEs remains to be discovered.

## Nature of evolutionary innovation

Although some CNEs show deep conservation across vertebrate evolution, most evolve and turn over at a faster pace than protein-coding sequences. At least 20% of the CNEs conserved among placental mammals are absent in marsupial mammals, compared to only ~1% for protein-coding sequences[8]. These elements arose in the period between our common ancestor with marsupial mammals (~180 million years (Myr) ago) and our common ancestor with placental mammals (~90 Myr ago), or else were ancestral elements lost in the marsupial lineage. The proportion of CNEs having detectable conservation with birds is much lower (~30% detectable, ~310 Myr ago) and with fish is near zero (~450 Myr ago). The more rapid change of CNEs provides experimental support for the notion[25] that evolution of species depends more on innovation in regulatory sequences than changes in proteins.

Extrapolation from the marsupial–placental comparison indicates that at least ~20% of functional non-coding elements in human should be absent in mouse. Interestingly, ChIP-Seq studies report even greater differences in the localization of transcription factor binding sites between mammals. However, physical binding may not imply biological function: the evolutionary and biochemical data still need to be reconciled.

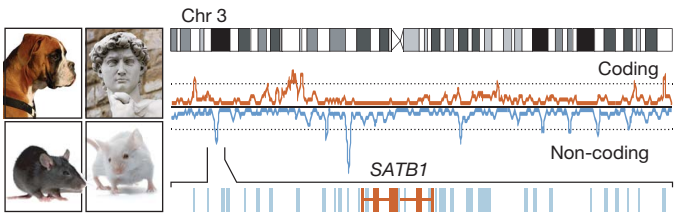


**Figure 1 | Evolutionary conservation maps.** Comparison among the human, mouse, rat and dog genomes helps identify functional elements in the genome. The figure shows the density of protein-coding sequences (red) and the most highly conserved non-coding sequences (blue) along chromosome 3. Highly conserved non-coding sequences are enriched in gene-poor regions, each of which contained a gene involved in early development (such as *SATB1*, shown). Images courtesy of iStock Photo.

## Transposons as drivers of evolutionary innovation

The HGP paper included a detailed analysis of transposon-derived sequences, but largely viewed transposons as a burden on the genome.

Comparative genomics, however, began to change this picture. The first hint was a handful of families of CNEs that had clearly been derived from transposons[26]. Comparison of placental and marsupial genomes then revealed that at least 15% of the CNEs that arose during the period from 180 Myr ago to 90 Myr ago were derived from transposon sequences[8]; the true total is likely to be considerably larger, because the flanking transposon-derived sequences will have degenerated in many cases. In retrospect, the advantage seems obvious. First, most transposons contain sequences that interact with the host transcriptional machinery, and therefore provide a useful substrate for evolution of novel regulatory elements. Second, a regulatory control that evolved at one locus could give rise to coordinated regulation across the genome by being picked up by a transposon, scattered around the genome and retained in advantageous locations. Over evolutionary timescales, transposons may earn their keep.

## Small non-coding RNAs

The HGP paper analysed all known classes of functional human non-protein-coding RNAs (ncRNAs), which consisted largely of those supporting protein translation (ribosomal, transfer and small nucleolar RNAs) and transcript splicing (small nuclear RNAs).

In late 2000, vertebrates were found to harbour an important new type of ncRNA first discovered in *C. elegans*. Called microRNAs (miRNAs), these products bind target mRNAs and decrease their stability[27]. Today, the human genome is known to encode ~100 evolutionarily conserved families of miRNAs. Genomic analysis proved critical in identifying the target mRNAs: evolutionarily conserved 7-base sequences in 3′ untranslated regions complementary to bases 2–8 of conserved miRNAs[28]. A typical conserved miRNA has ~200 target mRNAs with conserved binding sites. A few dozen miRNAs have been shown to have key regulatory roles, such as in cancer and development. Many of the others may help to fine-tune gene expression, although some may be too subtle to have detectable phenotypes in laboratory experiments. Recently, a new class of small RNAs, called PIWI-interacting RNAs, has been discovered that functions through a similar molecular machinery—they act to silence transposons in the germline.

## Ubiquitous transcription

In 2000, transcription was thought to be largely confined to regions containing protein-coding genes. Only a handful of non-classical large functional ncRNAs was known, such as telomerase RNA, 7SL signal recognition RNA, Xist and H19, and these were regarded as quirky exceptions. Pioneering studies of the human sequence soon began to provide hints that additional large RNA molecules might exist. Hybridization of RNA to microarrays of genomic sequence suggested that more than 10% of the genome was represented in mature transcripts, with most lying outside protein-coding exons[29], and random cDNA sequencing turned up many transcripts that could not be linked to protein-coding genes[30] (Fig. 2). With increasingly sensitive assays, it was concluded by 2007 that virtually every nucleotide in the euchromatic genome was likely to be represented in primary (unspliced) transcripts in at least some cell type at some time[31]. Many of these transcripts, however, have extremely low expression levels and show little evolutionary

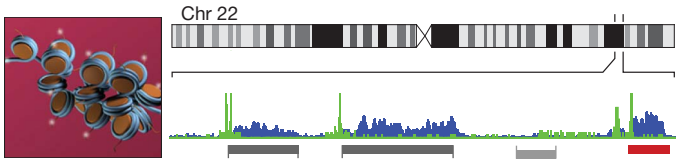


**Figure 2 | Chromatin state maps.** The genomic sites of chromatin modifications or protein binding can be mapped, using chromatin immunoprecipitation (ChIP) and massively parallel sequencing. The figure highlights chromatin marks associated with the active promoters (green) and actively transcribed regions (blue), in a region on chromosome 22. The four features shown correspond to two active protein-coding (dark grey), one inactive protein-coding (light grey) and one long intergenic non-coding RNA (maroon). Image courtesy of B. Wong (ClearScience).

conservation; these may represent 'transcriptional noise' (that is, reproducible, tissue-specific transcription from loci with randomly occurring weak regulatory signals). Exactly how much of the ubiquitous transcription is biologically functional remains controversial.

### Large intergenic non-coding RNAs

Epigenomic maps facilitated the discovery of a large class of thousands of genes encoding evolutionarily conserved (and thus clearly functional) transcripts, now called large intergenic non-coding RNAs (lincRNAs)[32]. The genes were pinpointed because they carry the distinctive chromatin patterns of actively transcribed genes but lack any apparent protein-coding capacity[32]. On the basis of their expression patterns, they have diverse roles in processes such as cell-cycle regulation, immune responses, brain processes and gametogenesis. A substantial fraction binds chromatin-modifying proteins and may modulate gene expression, for example, in the HOX complex[33] and in the p53-response pathway. Although their mechanism of action remains to be elucidated, lincRNAs may act analogously to telomerase RNA by serving as 'flexible scaffolds'[34] that bring together protein complexes to elicit a specific function.

lincRNAs are not the end of the ncRNA story. RNA-Seq studies have begun to define catalogues of antisense RNAs that overlap protein-coding genes[18]. Unlike lincRNAs, these transcripts show little evolutionary conservation (beyond the coding region)[18] and may function by base-pairing with the overlapping transcript or simply by causing chromatin changes through the act of transcription.

### Epigenomic maps

Recognizing the distinctive functional domains in the genome of a cell is a key challenge, both for genome scientists and for the cell itself. With thousands of genome-wide epigenomic maps, it is now clear that functionally active domains are associated with specific patterns of epigenomic marks[12,13,31,35] (Fig. 2). For example, active promoters show DNase hypersensitivity, histone acetylation and histone 3 lysine 4 trimethylation; transcribed regions are marked by histone 3 lysine 4 trimethylation; and enhancers show binding of the p300 acetyltransferase. Other features are seen at exons, insulators and imprinting control regions. The binding sites of transcription factors can also be read out, given an antibody with adequate specificity.

Moreover, it is possible to study dynamic behaviour and developmental potential by comparing epigenomic maps from related cellular states. For example, bivalent chromatin domains (both histone 3 lysine 4 trimethylation and histone 3 lysine 27 trimethylation) mark genes that are poised to play key parts in subsequent lineage decisions[36]. Epigenomic maps can also reveal genes that serve as obstacles to cellular reprogramming, and DNA methylation maps are helping identify aberrant functions in cancer[37]. Ultimately, hundreds of thousands of epigenomic marks will be layered atop the genome sequence to provide an exquisite description of genomic physiology in a cell type.

### Three-dimensional structure of the genome

Whereas general features of chromosomal packaging had been worked out through classical techniques such as X-ray diffraction, little was known about *in vivo* physical contacts between genomic loci more than a few kilobases apart.

The (one-dimensional) genome sequence enabled technologies for mapping the genome in three dimensions. Chromosome confirmation capture (3C) could test whether two loci are nearby in the nucleus, based on proximity-based ligation followed by locus-specific polymerase chain reaction[38]. It revealed, for example, that β-globin's locus control region forms an 'active hub' involving physical contact between genomic elements separated by 100 kb or more.

New approaches, such as a method called Hi-C, extend 3C to examine all physical contacts in an unbiased genome-wide fashion[39]. It has revealed that the genome is organized into two compartments, corresponding to open and closed chromatin, and, at megabase scale, exhibits folding properties consistent with an elegant structure called a fractal globule.

### The road ahead

The ultimate goal is to understand all of the functional elements encoded in the human genome. Over the next decade, there are two key challenges. The first will be to create comprehensive catalogues across a wide range of cell types and conditions of (1) all protein-coding and non-coding transcripts; (2) all long-range genomic interactions; (3) all epigenomic modifications; and (4) all interactions among proteins, RNA and DNA. Some efforts, such as the ENCODE and Epigenomics Roadmap projects, are already underway[31]. Among other things, these catalogues should help researchers to infer the biological functions of elements; for example, by correlating the chromatin states of enhancers with the transcriptional activity of nearby genes across cell types and conditions. These goals should be feasible with massively parallel sequencing and assay miniaturization, although they will require powerful ways to purify specific cell types *in vivo*, and the fourth goal will require a concerted effort to generate specific affinity reagents that recognize the thousands of proteins that interact with nucleic acids.

The second and harder challenge is to learn the underlying grammar of regulatory interactions; that is, how genomic elements such as promoters and enhancers act as 'processors' that integrate diverse signals. Large-scale observational data will not be enough. We will need to engage in large-scale design, using synthetic biology to create, test and iteratively refine regulatory elements. Only when we can write regulatory elements *de novo* will we truly understand how they work.

## Genomic variation

### The view from 2000

Since the early 1980s, humans were known to carry a heterozygous site roughly every 1,300 bases. Genetic maps containing a few thousand markers, adequate for rudimentary linkage mapping of Mendelian diseases, were constructed in the late 1980s and early 1990s. Systematic methods to discover and catalogue single nucleotide polymorphisms (SNPs) were developed in the late 1990s and resulted in the report of 1.42 million genetic variants in a companion to the HGP paper[40]. Still, the list was far from complete. Moreover, there was no way to actually assay the genotypes of these SNPs in human samples.

Today, the vast majority of human variants with frequency >5% have been discovered and 95% of heterozygous SNPs in an individual are represented in current databases. Moreover, geneticists can readily assay millions of SNPs in an individual.

### Linkage disequilibrium, HapMaps and SNP chips

Two critical advances propelled progress in the study of genomic variation: one conceptual and one technical. The first was the discovery of the haplotype structure of the human genome[41]; that is, that genetic variants in a region are tightly correlated in structures called haplotypes, reflecting linkage disequilibrium and separated by hotspots of recombination. Linkage disequilibrium was a classical concept, but its genome-wide structure had never been characterized in any organism. Humans turned out to have a surprisingly simple structure, reflecting recent expansion from a

small founding population. Tight correlations seen in a few dozen regions[41] implied that a limited set of ~500,000–1,000,000 SNPs could capture ~90% of the genetic variation in the population. The International Haplotype Map (HapMap) Project soon defined these patterns across the entire genome, by genotyping ~3 million SNPs[42]. The second advance was the development of genotyping arrays (often called SNP chips), which can now assay up to ~2 million variants simultaneously.

### Copy-number polymorphisms

Large-scale genomic aberrations (deletions and duplications) were long known to occur in cancers and congenital disorders. Biologists using DNA microarrays to study these events made a surprising observation: even normal, apparently healthy individuals showed copy-number polymorphism (CNP) in many genomic segments[43]. A typical person carries ~100 heterozygous CNPs covering ~3 Mb; the figure is vastly lower than initial estimates but still considerable[44]. Most are ancient variants that are tightly correlated with SNPs, which has enabled the association of CNVs with phenotypes using proxy SNPs. An intriguing minority of CNVs, however, seems to arise from recent and *de novo* mutations and may have important roles in psychiatric disorders[45–47].

### The road ahead

The ultimate goal is to create a reference catalogue of all genetic variants common enough to be encountered recurrently in populations, so that they can be examined for association with phenotypes and interpreted in clinical settings. Efforts towards this goal are already well underway. The 1000 Genomes Project[48] (which plans to study many more than one thousand genomes) aims to find essentially all variants with frequency >1% across the genome and >0.1% in protein-coding regions.

## Medicine: Mendelian and chromosomal disorders

### The view from 2000

At the time when the HGP was launched, fewer than 100 disease genes had been identified, because finding them largely relied on guesswork about the underlying biochemistry. Genetic linkage mapping in affected families offered a general solution in principle, but it was slow and tedious. With the genetic and physical maps created in the first stages of the HGP, the list of identified disease genes quickly began to grow. With the human sequence, it has exploded. A decade after the HGP, more than 2,850 Mendelian disease genes have been identified.

### Mendelian diseases

Given enough affected families, one can genetically map a monogenic disease to a chromosomal region and compare patient and reference sequences to search for the causative gene and mutation. With massively parallel sequencing, the simplest way to interrogate a region is often by whole-exome or whole-genome re-sequencing. Increasingly, investigators are attempting to eliminate the step of genetic mapping. The task is not entirely straightforward: even when all common variants can be filtered out based on the fruits of the 1000 Genomes Project, a typical person will still have ~150 rare coding variants affecting ~1% of their genes (as well as 100-fold more rare non-coding variants). For recessive diseases caused by protein-coding mutations, it may just be possible to discover a disease gene based on a single patient by looking for two mutations in the same gene. In general, though, pinpointing the right gene will require accumulating evidence from multiple patients.

### Clinical applications

DNA sequencing is being increasingly used in the clinic, including applying whole-exome sequencing to assign patients with an unclear diagnosis to a known disease[49]. Genomics is also becoming a routine tool in cytogenetics laboratories, where DNA microarrays have greatly increased the sensitivity to detect clinically significant chromosomal imbalances, advancing diagnostic evaluation of children with idiopathic developmental delay, major intellectual disability, autism and birth defects.

### The road ahead

New sequencing technologies should propel tremendous progress in elucidating the ~1,800 uncloned disorders in the current catalogue, and in recognizing undescribed Mendelian disorders in patients with unexplained congenital conditions. By comparing patients to parents using highly accurate sequencing, it will even be possible to spot newly arising mutations responsible for dominant lethal disorders. As whole-genome sequencing costs fall, it may become routinely used by couples before conception, and by paediatricians to explain idiopathic conditions in children.

Because Mendelian recessive disorders often involve the complete absence of a protein, they are typically difficult to treat, except where enzyme replacement is feasible. Genomics has produced some remarkable recent exceptions, such as the mechanism-based recognition that Marfan syndrome might be treated with an existing inhibitor of TGF-β[50]. A critical challenge will be to find systematic ways to treat Mendelian disorders by gene-based therapies, or by developing small-molecule therapeutics.

## Medicine: common diseases and traits

### The view from 2000

In contrast to rare Mendelian diseases, extensive family-based linkage analysis in the 1990s was largely unsuccessful in uncovering the basis of common diseases that afflict most of the population. These diseases are polygenic, and there were no systematic methods for identifying underlying genes. As of 2000, only about a dozen genetic variants (outside the HLA locus) had been reproducibly associated with common disorders.

A decade later, more than 1,100 loci affecting more than 165 diseases and traits have been associated with common traits and diseases, nearly all since 2007.

### Common disease

To study common diseases, geneticists conceived principles for genetic mapping using populations rather than families. A first systematic example was genome-wide association study (GWAS), which involves testing a comprehensive catalogue of common genetic variants in cases and controls from a population to find those variants associated with a disease[51,52]. Rare Mendelian diseases are almost always caused by a spectrum of rare mutations, because selection acts strongly against these alleles. By contrast, the 'common disease–common variant' (CD/CV) hypothesis[53] posited that common genetic variants (polymorphisms, classically defined as allele frequency >1%) could have a role in the aetiology of common diseases. By testing all common variants, one could pinpoint key genes and shed light on underlying mechanisms.

The CD/CV hypothesis rested on the following premise: because the vast majority (~99%) of genetic variance in the population is due to common variants, the susceptibility alleles for a trait will include many common variants except if the alleles have had a large deleterious effect on reproductive fitness over long periods. For common diseases or traits, many susceptibility alleles may have been only mildly deleterious, neutral or even advantageous. Examples may include diseases of late onset, diseases resulting from recent changes in living conditions such as diabetes and heart disease, morphological traits, and alleles with pleiotropic effects that result in balancing selection. Notably, humans are a favourable case for genetic mapping by association studies, because the small historical population size means that the force of selection is weaker and the allelic spectrum is simpler[53].

With the development of catalogues of common variants, haplotype maps, genotyping arrays and rigorous statistical methods[54], the CD/CV hypothesis was finally put to the test beginning in late 2006; it has been richly confirmed by an explosion of discoveries.

### Revealing disease pathways

Three key results have emerged from these studies: (1) most traits can be influenced by a large number of loci; (2) the vast majority of the common variants at these loci have a moderate effect, increasing risk by 10–50% (similar to effects of many environmental risk factors); and (3) the loci

include most of the genes found by linkage analysis, but reveal many more genes not previously implicated. Some early commentators argued that such discoveries were not useful for understanding disease, because loci with moderate effects were too hard to study and could not have important therapeutic consequences. The results, however, have proved otherwise.

By discovering large collections of genes that can modulate a phenotype, GWAS has begun to reveal underlying cellular pathways and, in some cases, already pointed to new therapeutic approaches. In effect, GWAS is the human analogue to mutagenesis experiments in animal models: they provide a systematic, unbiased way to identify genes and pathways underlying a biological process to allow subsequent physiological studies. A recent study of the genetic control of lipid levels illustrates many of the points (Box 1). Some other important examples are given in the following paragraphs.

**Adult macular degeneration.** GWAS uncovered the aetiology of this leading cause of blindness affecting millions of elderly patients, by finding strong associations with five loci with common variants of large effect, including multiple genes in the alternative complement system. The results pointed to a failure to inhibit specific inflammatory responses, spurring new therapeutic approaches.

**BOX 1**

## Genetics of lipids

A recent GWAS of plasma lipid levels, a major risk factor for myocardial infarction, demonstrates the power of the approach. A study of >100,000 individuals of European ancestry identified 95 loci associated to at least one of three major lipids: low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides[66]. Moreover, most also showed association in African and Asian cohorts, indicating their generality.

Although the variants have only moderate effects, their combined impact can be considerable. Together, the loci explain ~25% of the genetic variance for LDL and HDL levels. Furthermore, the top quartile of individuals with triglyceride-associated variants has a 44-fold higher risk of hypertriglyceridaemia than the bottom quartile.

Notably, the 95 loci with common variants include nearly all of the 18 genes previously implicated in rare Mendelian lipid disorders, indicating that GWAS will often help to pinpoint genes harbouring rare variants. Of note, at loci where both common and rare variants were studied, the former explain much more of the heritability than the latter.

The study underscores that loci with only moderate effects in GWAS may have major therapeutic implications. The *HMGCR* locus has a common variant at 40% frequency that changes LDL by a modest 2.8 mg $dl^{-1}$ and no known rare mutations of large effect, presumably because they would be lethal. Yet, the encoded protein is the target of statins, drugs taken by tens of millions of patients that can significantly reduce both LDL levels and myocardial infarction risk.

A number of the new loci identified have already been confirmed to affect lipid biology, through rapid transgenic animal studies and human clinical studies. The sortilin gene (*SORT1*), for example, contains a common variant that creates a novel transcription-factor-binding site that alters hepatic expression in humans, and transgenic studies in mouse show that it alters plasma LDL levels[100]. The gene thus defines a previously unknown regulatory pathway for LDL.

Separate studies have identified a pair of common nonsense variants in the *PCSK9* gene in African-Americans (2.6% combined frequency) that markedly reduces both LDL levels and risk of myocardial infarction[69]. The finding that human homozygotes for this presumably null allele are healthy indicates that a drug against the encoded protein should be safe and effective, provoking considerable activity by pharmaceutical companies.

**Crohn's disease.** Studies have so far identified 71 risk loci[55] for this severe inflammatory bowel disease (Fig. 3). The genes have together revealed previously unknown roles for such processes as innate immunity, autophagy and interleukin-23 receptor signalling. Specific mutations identified in patients have been confirmed to be pathogenic in cellular and animal models, providing new strategies for therapeutic development.

**Control of fetal haemoglobin.** Fetal haemoglobin (HbF) levels vary among individuals and higher levels can ameliorate symptoms of sickle cell anaemia and β-thalassaemia, but hopes for treating these diseases by increasing HbF had been stymied by ignorance of the mechanism that downregulates HbF expression. GWAS revealed three loci that modulate erythroid development, which together explain >25% of the genetic variance in HbF levels and are associated with reduced severity in sickle cell anaemia and β-thalassaemias[56]. Although the common SNP in the *BCL11A* gene has only a modest effect, strong perturbation of the gene in cells results in half of the cell's haemoglobin being HbF.

**Type 2 diabetes.** Studies have identified 39 loci so far in this disorder, which affects 300 million people worldwide[57]. Notably, genes previously implicated in glucose regulation based on biochemical studies do not seem to be associated with type 2 diabetes, but are associated with fasting glucose levels. The pathophysiology thus probably involves different molecular mechanisms. Many of the genes also point to insulin secretion rather than insulin resistance as the primary cause.

**Autoimmune diseases.** Studies have found ~100 loci related to autoimmune diseases such as type 1 diabetes, rheumatoid arthritis and coeliac disease. These studies point to many loci that have roles across multiple autoimmune diseases and probably involve fundamental regulatory pathways, as well as many that are disease specific.

**Height.** Adult height is a classic polygenic trait with high heritability. A recent analysis of 180,000 individuals identified 180 loci[58], with many showing multiple distinct alleles. With the large number of loci and new analytical tools, various biological pathways have been implicated, such as TGF-β signalling.

**Kidney disease.** GWAS for two common renal disorders revealed variants in *APOL1* that are common in African chromosomes but absent in European chromosomes, which account for much of the increased risk of kidney disease in African-Americans. The variants seem to have reached high frequency in Africa because they protect carriers from the deadliest subtype of African sleeping sickness.

**Psychiatric disorders.** Studies of psychiatric diseases have been relatively limited in scope, and much less is known than for other diseases. Genotyping studies have identified both common variants (in bipolar disorder and schizophrenia[59,60]) and rare deletions (in autism[46,61] and schizophrenia[62]). Each class of variants so far accounts for a few per cent of the genetic variance, but analyses indicate that both classes will have a major role in a highly polygenic aetiology[45,47,60,63]. Given our near-total ignorance of the underlying cellular pathways, larger genetic studies are essential.

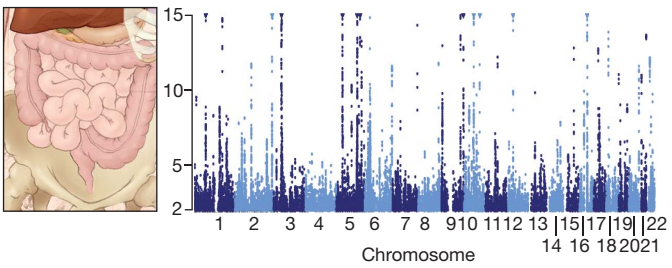


**Figure 3 | Disease association maps.** Geneticists can now test the association between a common disease and millions of individual genetic variants. The figure shows a 'Manhattan plot' from a study of Crohn's disease, a form of inflammatory bowel disease. For each variant across the genome, the height reflects its correlation with disease (measure by $\log_{10}$(significance)). The Manhattan plot reveals 71 'skyscrapers', corresponding to regions associated with Crohn's disease. Image courtesy of B. Wong (ClearScience).

**Pharmacogenetics.** Association studies have revealed genetic factors underlying hypersensitivity to the antiretroviral drug abacavir, drug-induced myopathies associated with cholesterol-lowering drugs, cardiovascular risks in patients receiving the antiplatelet drug clopidogrel and variations in metabolic clearance of the anticoagulant warfarin.

**eQTLs.** GWAS have also proved powerful for studying gene regulation, by using cellular gene expression as a phenotype in its own right. Following early studies associating variation in transcript levels with nearby common variants[64,65], studies have identified thousands of quantitative trait loci affecting gene expression (eQTLs), both in *cis* and *trans*.

Notwithstanding the moderate effect sizes in disease studies, investigators have been able to definitely implicate specific mutations in humans, by combining eQTL studies and transgenic animal models[66]. Although many affect coding sequences, a larger proportion seems to lie in non-coding regions and may affect gene regulation, consistent with the abundance of functional non-coding sequences in the genome. A stunning example is the region around the gene encoding the cell-cycle regulator CDKN2A/B, in which distinct non-coding regions are associated with a dozen different diseases, including diabetes, heart disease and various cancers. Some of the causal regulatory variants have been discovered, but most remain mysterious. These findings call for improved methods to understand the function of non-coding regions and seem likely to reveal new mechanisms of gene regulation.

## Genetic architecture of disease

Despite this success, the results have provoked some handwringing in the scientific community and beyond because initial studies often explained only a small proportion of the heritability (defined as the additive genetic variance). The so-called 'missing heritability' has provoked renewed interest in the 'genetic architecture' of human disease and traits—a topic that was the subject of much debate early in the twentieth century. The explanation will surely involve multiple contributions.

First, it is becoming clear that the heritability due to common variants is greater than initially appreciated. With larger GWAS, the heritability explained has continued to grow, reaching 20–25% for various diseases and traits (Table 1). Moreover, current estimates substantially understate the actual role of common variants for two reasons. One reason is that current GWAS miss many common variants of lower frequency (1–10%), because existing genotyping arrays often lack useful proxies. Many disease-related alleles probably fall into this frequency class, which is enriched for variants under mildly deleterious selection. New genotyping arrays based on the 1000 Genomes Project should be able to capture these common variants. Another reason is that GWAS also miss many common variants of smaller effect, due to limited sample size and stringent statistical thresholds imposed to ensure reproducibility. Recent efforts have sought to infer the contribution of loci that fall just short of statistical significance[67]. Beyond, there are surely many more common variants with still smaller effects (the standing variation expected under Darwinian evolution): although their individual contributions may be too small to ever detect with feasible sample sizes, they may collectively explain a significant fraction of heritability. Elegant indirect analyses indicate that common variants must account for >55% of heritability for height and >33% for schizophrenia[60,68].

**Table 1 | GWAS for common diseases and traits**

| Phenotype | Number of GWAS loci | Proportion of heritability explained (%)* |
|---|---|---|
| Type 1 diabetes | 41 | ~60 |
| Fetal haemoglobin levels | 3 | ~50 |
| Macular degeneration | 3 | ~50 |
| Type 2 diabetes | 39 | 20–25 |
| Crohn's disease | 71 | 20–25 |
| LDL and HDL levels | 95 | 20–25 |
| Height | 180 | ~12 |

HDL, high-density lipoprotein; LDL, low-density lipoprotein.
* Fraction of heritability explained is calculated by dividing the phenotypic variance explained by variants at loci identified by GWAS by the total heritability as inferred from epidemiological parameters. (See Supplementary Bibliography.)

Second, rare variants of larger effect will also play an important part in common diseases, although their role has barely been explored. Studying them requires sequencing protein-coding (or other) regions in genes to identify those in which the aggregate frequency of rare variants is higher in cases than controls. So far, studies have focused on candidate genes implicated through Mendelian diseases, mouse mutants, biochemical pathways or GWAS studies. Early studies reported findings at *MC4R* in extreme, early-onset obesity and *PPARG* in insulin resistance (both of which also have common variants affecting the traits). Recent important examples include rare variants in six candidate genes affecting lipid levels[69] and three candidate genes affecting blood pressure[70]. A study of a GWAS region associated with HbF levels implicated the *MYB* gene, explaining additional heritability beyond the common variants[71].

Whether rare variants will reveal many new genes must await systematic whole-exome sequencing. Given the background rate of rare variants (~1% per gene), many thousands of samples will be needed to achieve statistical significance. Similarly, the total heritability due to rare variants is still unclear. Although the inferred effect sizes are larger, the overall contribution to the heritability may still be small due to their low frequency. For example, only ~1/400 of the heritability is explained by rare variants at each of the three loci affecting blood pressure. Whatever their contribution to heritability, rare variants of large effect will be valuable by enabling direct physiological studies of pathways in human carriers.

Finally, some of the missing heritability may simply be an illusion. Heritability is estimated by applying formulae for inferring additive genetic effects from epidemiological data. The estimates may be inflated because the methods are not very effective at excluding the (nonlinear) contributions of genetic interactions or gene-by-environment interactions, which are likely to be significant.

## Biological mechanisms versus risk prediction

It is important to distinguish between two distinct goals. The primary goal of human genetics is to transform the treatment of common disease through an understanding of the underlying molecular pathways. Knowledge of these pathways can lead to therapies with broad utility, often applicable to patients regardless of their genotype. The past decade has seen remarkable progress towards identifying disease genes and pathways, with greater advances ahead.

Some seek a secondary goal: to provide patients with personalized risk prediction. Although partial risk prediction will be feasible and medically useful in some cases, there are likely to be fundamental limits on precise prediction due to the complex architecture of common traits, including common variants of tiny effect, rare variants that cannot be fully enumerated and complex epistatic interactions, as well as many non-genetic factors.

## The road ahead

The discovery of more than 1,100 loci within a few years is an excellent start, but just a start. Over the next decade, we need genetic studies of tens of thousands of patients for most common diseases, with appropriate combinations of GWAS and sequencing. In turn, intensive functional studies will be required to characterize the genes and pathways, and to construct animal models that mimic human physiology. Importantly, complete explanation of a disease is not required for progress.

# Medicine: cancer

## The view from 2000

With the establishment by the early 1980s that cancer results from somatic mutations, Dulbecco declared in an influential article in 1986 that sequencing the human genome was a critical priority, saying "We have two options: either to try to discover the genes important in malignancy by a piecemeal approach, or to sequence the whole genome of a selected animal species"[72]. By 2000, ~80 cancer genes involved in solid tumours had been discovered, most through viral oncogenes and transformation assays and the remainder through positional cloning of inherited cancer

syndromes and somatically deleted regions in cancer. (Many additional genes were found in blood cancers, where translocations could be readily visualized and cloned.)

A decade later, cancer gene discovery is being driven by systematic genome-wide efforts, involving a powerful combination of DNA sequencing, copy-number analysis (using genotyping arrays developed for GWAS), gene expression analysis and RNA interference. The number of cancer genes in solid tumours has nearly tripled to ~230, revealing new biological mechanisms and important therapeutic leads (see ref. 73 and http://www.sanger.ac.uk/genetics/CGP/cosmic). More generally, systematic efforts are beginning to parse the vast heterogeneity of cancer into more homogeneous groupings based on mechanism.

## Early sequencing efforts

Given the limited capacity of capillary-based sequencing, initial sequencing efforts focused on targeted gene sets, such as kinases in signalling pathways underlying cell growth; they soon hit pay dirt. (1) *BRAF* mutations were discovered in >50% of melanomas[74]. Pharmaceutical companies raced to develop inhibitors of RAF and MEK, a gene downstream of *RAF*. Initial results were disappointing, but recent clinical trials (using more potent inhibitors and studying tumours carrying relevant mutations) have shown response rates exceeding 80%. (2) *PIK3CA* mutations were discovered in >25% of colorectal cancers[75]; pharmaceutical programmes are at an earlier stage. (3) *EGFR* mutations were discovered in 10–15% of lung cancers and predicted which patients would respond to gefitinib and erlotinib, drugs that had had only patchy efficacy[76]. Such treatment soon became the standard-of-care for patients with the relevant mutation and has been shown to extend life.

An early exome-wide sequencing study of glioblastoma found a new class of cancer gene involved in basic cellular metabolism: a recurrent mutation in *IDH1* alters the active site of the enzyme isocitrate dehydrogenase[77], causing it to aberrantly generate an 'oncometabolite', 2-hydroxygluterate. Pharmaceutical companies are already working towards the development of inhibitors of the neo-enzyme.

## Microarray-based studies

Genotyping arrays allowed genome-wide, high-resolution analysis of amplifications and deletions. A recent genomic study of >3,000 tumours across 26 cancer types catalogued more than 150 recurrent focal copy-number alterations[78]; only one-quarter contain known cancer genes, indicating that many more remain to be discovered. Studies of specific events have identified many new cancer genes. Amplifications revealed an entirely new class consisting of transcription factors necessary for lineage-specific survival (*MITF* in melanoma, *NKX2.1* in lung cancer and *SOX2* in oesophageal cancer)[79]. Recurrent deletions in paediatric acute lymphoblastic leukaemias were found in *PAX5*, *IKZF1* and other regulators of lymphocyte differentiation.

Gene-expression arrays similarly revealed new mutational targets, including functional translocations involving one of several ETS transcription factors in >50% of prostate tumours[80], disproving the long-held belief that translocations have a major role in blood cancers, but not epithelial solid tumours. Translocations involving *ALK* have since been found in some lung cancers, and new pharmaceuticals directed against *ALK* are already showing impressive results in clinical trials[81].

Genome-wide expression analysis has also had a central role in classifying cancers based on their molecular properties[82,83], rather than anatomic sites. Studies have revealed distinctive subtypes and shed light on metastatic potential[84]. Expression signatures are already used in the clinic to predict which breast cancer patients will benefit most from adjuvant chemotherapy after surgery. More generally, gene expression analysis has become a routine aspect of both basic and clinically oriented discovery research. Expression signatures have also been proposed as a lingua franca for connecting the molecular mechanisms of drugs, genes and diseases[85], sometimes revealing new ways to use old drugs[85].

Genomic studies have been integrated with genome-wide RNA-interference-based screens to identify genes that are both genomically amplified and essential for cell viability in specific cancer cell lines. Examples include *IKBKE* in breast cancer, *CDK8* in colorectal cancer and the nuclear export protein XPO4 in hepatocellular cancer[86]. Genomically characterized cell lines can also be screened to identify 'synthetic lethals'—that is, genes essential only in the presence of particular recurrent cancer mutations—such as *PLK1*, *STK33* and *TBK1*, required in the setting of KRAS mutations.

## Genome-wide sequencing

With massively parallel sequencing, attention has now focused on comprehensive exome-wide and genome-wide sequencing in large numbers of samples (Fig. 4). In acute myelogenous leukaemia and clear-cell ovarian cancer, recurrent mutations in *DNMT3A* and *ARID1A*, respectively, point to epigenomic dysregulation[87,88]. Studies of tumours from ~40 multiple myeloma patients have uncovered frequent mutations in genes not previously known to have a role in cancer (such as *DIS3* and *FAM46C*), implicating novel pathways such as protein translation and homeostasis, as well as NF-κB activation[89]. Whole-genome sequencing of prostate cancers has shed light on the origins of tumour rearrangements.

## The road ahead

The ultimate goal is to markedly decrease death from cancer. To guide therapeutics, we must develop over the next decade a comprehensive catalogue of all genes that are significant targets of somatic alteration in all human cancer types, all animal models and all cancer cell lines.

The patterns of mutated genes will: (1) define direct drug targets in some cases; (2) identify cellular pathways and synthetic lethal interactors to target in others; (3) direct the creation of animal models; (4) allow chemical screening against cancer cells with defined molecular mechanisms; and (5) guide the design of human clinical trials. We must also chart all ways in which tumours develop resistance to specific interventions in patients. Ideally, we should obtain this information from pre-clinical studies, so that we can plan countermeasures, even as we develop a drug. Effective cancer treatment will surely require combination therapies, like those used against HIV, that target multiple vulnerabilities to markedly diminish the chance of resistance.

Projects such as The Cancer Genome Atlas and the International Cancer Genome Consortium plan to study ~500 tumours per cancer type, but these goals will need to be dramatically expanded. As genomics permeates clinical practice, we should create a mechanism by which all cancer patients can choose to contribute their genomic and clinical data to an open collaborative project to accelerate biomedical progress.

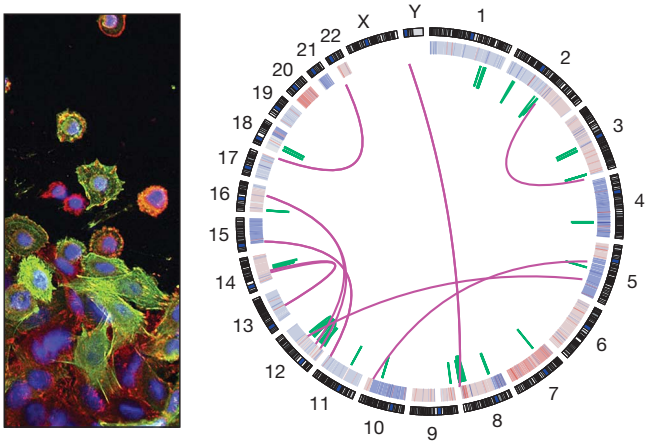


**Figure 4 | Cancer genome maps.** Whole-genome sequencing has provided powerful new views of cancers. The left panel shows an image of colon cancer (Wellcome Trust). The right panel shows the genome of a colon cancer sample (Broad Institute), including interchromosomal translocations (purple), intrachromosomal translocations (green) and amplifications and deletions (red and blue, on the inner ring). Individual nucleotide mutations are not shown.

## Human history

### The view from 2000

Well before 2000, studies of genetic variation at handfuls of loci such as mitochondrial DNA and blood groups across worldwide populations had given rise to an intellectual synthesis according to which modern humans arose in Africa, dispersed from the continent in a single migration event 50,000–100,000 years ago, replaced resident archaic human forms, and gave rise to modern populations largely through successive population splits without major mixture events[90]. It was difficult, however, to reconstruct the details of these events from the limited data. In addition, little was known about the role that positive selection may have played in shaping the biology of human populations as they migrated and expanded.

A decade later, genomic data have radically reshaped our understanding of the peopling of the globe, yielding a vastly richer picture of population mixing and natural selection. These studies have been made possible by the growth in catalogues and maps of genetic variation among human populations, as well as differences with our closest relatives, such as Neanderthal and chimpanzee.

### Population mixture in human history

Shifting the focus from stories based on individual loci (such as 'mitochondrial Eve') to large collections of genetic markers has provided powerful new insights. It is now clear that the migration of humans out of Africa was more complicated than previously thought, and that human history involved not just successive population splits, but also frequent mixing.

There is now strong genetic evidence, for example, showing that south Asians are the product of an ancient mixture. European populations have mixed so extensively with their neighbours that their genes mirror geography, rather than reflecting the paths of human migrations or language families.

Most strikingly, genome analysis showed that anatomically modern humans mixed with the Neanderthals. Europeans and Asians (but not Africans) have all inherited 1–4% of their genome from Neanderthals, indicating gene flow in the Middle East on the way out of Africa[9].

### Positive selection in the last 10,000 years

By studying dense collections of genetic markers, it has become possible to spot the signs left of recent positive natural selection—even without knowing the specific trait under selection. One such signature is a high-frequency, long-range haplotype, which results when an advantageous genetic variant rapidly sweeps through a population and carries along neighbouring variants[91]. Analysis of data from the HapMap has revealed at least 300 genomic regions that have been under positive selection during the past ~5,000–30,000 years. New methods have narrowed these signals to small regions, often with only a single gene, and sometimes implicated specific candidate variants[92] (Fig. 5).

Although much work will be needed to decipher each gene's unique story, the implicated genes are already beginning to point to specific processes and pathways. Many encode proteins related to skin pigmentation, infectious agents, metabolism and sensory perception[92]. In Europe, powerful selection around the dawn of agriculture favoured a regulatory variant that causes lifelong expression of lactase, the enzyme required to digest milk; a similar mutation was independently selected in cattle-herding groups in East Africa. In West Africa, strong selection for a gene encoding a receptor for the Lassa fever virus may indicate a resistance allele. Comparisons of nearby populations living under different conditions can be particularly informative. Tibetans, a population living at 14,000 feet, and Han Chinese are closely related, but show striking population differentiation at a locus encoding a protein involved in sensing oxygen levels[93].

### Positive selection in human speciation

A favourite question of philosophers and scientists alike is 'what makes us uniquely human?' The human and chimpanzee genome sequences made it possible to give a definitive, if unsatisfying answer: we can now enumerate the ~40 million genetic differences[7]. Unfortunately, the vast majority of them represent random drift.

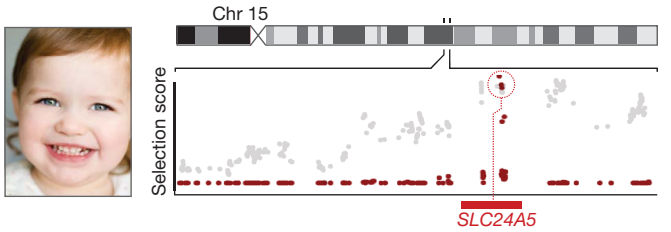


**Figure 5 | Positive selection maps.** Genetic variation patterns across the genome can help to reveal regions that have undergone strong positive selection during human history. The figure shows a region of 1 million bases on chromosome 15 that has been under strong selection in European populations. The initial analysis (grey) provides diffuse localization, whereas fine-structure mapping (red) pinpoints the signal to a gene (*SLC24A5*) known to affect skin colour. Image courtesy of iStock Photo.

The best clues as to which changes drove human evolution have come from methods to detect ancient positive selection. At several hundred loci termed 'human accelerated regions' (HARs), the rate of nucleotide change since the divergence from the common ancestor of human and chimpanzee is exceptionally high relative to the rate of change over the previous 100 million years[94]. HAR1 is part of a non-coding RNA that is expressed in a region of the brain that has undergone marked expansion in humans. HAR2 includes a transcriptional enhancer that may have contributed towards the evolution of the opposable thumb in humans. Similarly, the *FOXP2* gene has undergone accelerated amino-acid substitution along the human lineage. Because null alleles of *FOXP2* affect language processing, it has been suggested that these changes may be related to our acquisition of language. Of course, large signals of accelerated evolution represent only a piece of the puzzle: many critical changes surely involved only isolated nucleotide changes. Identifying these changes will require combining insights from both genotypic and phenotypic differences. Several dozen candidates have been suggested.

### The road ahead

The ultimate goal is to use genomic information to reconstruct as much as possible about the salient events of human history. This includes a complete accounting of the structure of the ancestral human population in Africa; the subsequent population dispersals and their relationship to landmarks such as the spread of agriculture; gene flow with archaic hominins both before and after the Out-of-Africa migration; and the impact of positive natural selection in recent and ancient times.

Over the next decade, we should assemble large-scale genomic databases from all modern human populations, hominin fossils and great ape relatives. New laboratory techniques will be needed to infer and test the functional role of human variations. Advances in statistical methodology will be needed, including better ways to date events and exploit haplotype information to infer common ancestry.

## New frontiers

Twenty-five years ago, biologists debated the value of sequencing the human genome. Today, young scientists struggle to imagine the nature of research in the antediluvian era, before the flood of genomic data. Genomics has changed the practice of biology in fundamental ways. It has revealed the power of comprehensive views and hypothesis-free exploration to yield biological insights and medical discoveries; the value of scientific communities setting bold goals and applying teamwork to accomplish them; the essential role of mathematics and computation in biomedical research; the importance of scale, process and efficiency; the synergy between large-scale capabilities and individual creativity; and the enormous benefits of rapid and free data sharing.

Yet, this is only a step towards transforming human health. We must now extend these principles to new frontiers. Our goal should be to

dramatically accelerate biomedical progress by systematically removing barriers to translational research and unleashing the creativity of a new generation.

Within genomics, we must complete biomedicine's 'periodic table'[51]. This will take at least another decade, through systematic efforts to define the components such as those described above. Connecting these components fully with disease will require something like an international 'One Million Genomes Project', analysing well-annotated patient samples from many disorders.

We must also apply systematic approaches more broadly. Some specific goals are given in the following paragraphs.

### Modular cell biology

Just as it was once inconceivable to possess complete catalogues of cellular components, it seems quixotic today to seek a comprehensive picture of cellular circuitry. Yet, it is time to turn systematic attention to this next level of organization. Cellular circuitry is not infinitely complex. It is organized around a limited repertoire of modules[95], whose reuse is fundamental to evolvability. These modules include protein complexes, *cis*-regulatory circuits, metabolic pathways, and signalling networks, each involving tightly coupled cores with hierarchies of condition-dependent interactions[96]. In yeast, we can already begin to glimpse the basic modular organization that controls environmental responses. In mammals, the picture will be far more complex but the number of fundamental modules is likely to be tractable—perhaps a few thousand. The goal of a complete catalogue of cell modules will involve many challenges. Conceptually, we will need diverse ways to infer and test candidate modules (for example, correlations in gene expression, protein modification and evolutionary retention and by systematic perturbation). Technically, we will need powerful platforms—'cell observatories', so to speak—including systematic reagents for modulating components and interactions; new instruments for single-cell measurements; analytical methods to derive mechanistic models; and access to many cell types[96].

### Cell programming

We must learn to program cells with the same facility with which we program computers. The past decade has set the stage. Yamanaka's stunning discovery that adult cells can be re-programmed into pluripotent cells has inspired screens for particular gene cocktails to induce other transformations. Independently, a cadre of creative young synthetic biologists have begun to invent new cellular circuits. The key challenges ahead include developing a general recipe to trans-differentiate any cell type into any other, and general combinatorial tools that make it is easy to create circuits activated only in a specific cellular state. Cell programming will draw inspiration from native cellular modules and in turn provide tools to study them.

### Chemical biology and therapeutic science

Accelerating treatments for disease will require a renaissance in therapeutic science. The pharmaceutical industry will remain the locus of drug development, but it rightly focuses on proven methods and commercial markets. Academia must become a hotbed for heterodox approaches that combine creativity and scale, exploit genomic approaches and targets, and empower a new generation of scientists. Key goals include developing large libraries of small molecules whose chemical properties favour selectivity, potency and rapid optimization[97]; powerful phenotypic assays using genomic signatures and other general approaches that make it possible to find modulators for any cellular process[98,99]; and systematic methods that can rapidly and reliably find the protein targets and mechanisms of action of 'hits'. Ultimately, we should aim for a comprehensive arsenal of small-molecule modulators for all cellular targets and processes to probe physiology and test therapeutic hypotheses.

Medical revolutions require many decades to achieve their full promise. Genomics has only just begun to permeate biomedical research: advances must proceed through fundamental tools, basic discoveries, medical studies, candidate interventions, clinical trials, regulatory approval and widespread adoption. We must be scrupulous not to promise the public a pharmacopoeia of quick pay-offs. At the same time, we should remain unabashed about the ultimate impact of genomic medicine, which will be to transform the health of our children and our children's children.


1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
3. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431,** 931–945 (2004). **The draft sequences reported in refs 1 and 2 provided the first comprehensive look at ~90% of the human genome; the finished sequence in ref. 3 increased the completeness to >99% and the accuracy to >99.999%, providing a solid foundation for biomedicine.**
4. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002). **Comparison with the mouse genome led to the discovery that the vast majority of functional sequence in the human genome does not encode protein.**
5. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438,** 803–819 (2005).
6. Rat Genome Sequencing Project Consoritum. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
7. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
8. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447,** 167–177 (2007).
9. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
10. Grice, E. A. *et al.* Topographical and temporal diversity of the human skin microbiome. *Science* **324,** 1190–1192 (2009).
11. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312,** 1355–1359 (2006).
12. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448,** 553–560 (2007).
13. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).
14. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
15. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456,** 53–59 (2008).
16. Okou, D. T. *et al.* Microarray-based genomic selection for high-throughput resequencing. *Nature Methods* **4,** 907–909 (2007).
17. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5,** 621–628 (2008).
18. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnol.* **28,** 503–510 (2010).
19. Yassour, M. *et al. Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl Acad. Sci. USA* **106,** 3264–3269 (2009).
20. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*. doi:10.1073/pnas.1017351108 (27 December, 2010).
21. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA* **104,** 19428–19433 (2007).
22. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304,** 1321–1325 (2004).
23. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
24. Pennacchio, L. A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).
25. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
26. Bejerano, G. *et al.* A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441,** 87–90 (2006).
27. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature* **408,** 86–89 (2000).
28. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115,** 787–798 (2003).
29. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).
30. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309,** 1559–1563 (2005).
31. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447,** 799–816 (2007).
32. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).
33. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129,** 1311–1323 (2007).
34. Zappulla, D. C. & Cech, T. R. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. *Cold Spring Harb. Symp. Quant. Biol.* **71,** 217–224 (2006).

35. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39,** 311–318 (2007).
36. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
37. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128,** 683–692 (2007).
38. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295,** 1306–1311 (2002).
39. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–293 (2009).
40. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).
41. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29,** 229–232 (2001).
42. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
**References 40–42 laid the foundation for genetic studies of common disease, which have so far identified more than 1,100 loci associated with diseases.**
43. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305,** 525–528 (2004).
44. McCarroll, S. A. Copy number variation and human genome maps. *Nature Genet.* **42,** 365–366 (2010).
45. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455,** 237–241 (2008).
46. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358,** 667–675 (2008).
47. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320,** 539–543 (2008).
**References 45–47 revealed an important role of rare genetic deletions in psychiatric diseases.**
48. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
49. Bilgüvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467,** 207–210 (2010).
50. Habashi, J. P. *et al.* Losartan, an AT1 antagonist, prevents aortic aneurysm in a mouse model of Marfan syndrome. *Science* **312,** 117–121 (2006).
51. Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).
52. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273,** 1516–1517 (1996).
53. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17,** 502–510 (2001).
54. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).
55. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genet.* **42,** 1118–1125 (2010).
56. Uda, M. *et al.* Genome-wide association study shows *BCL11A* associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia. *Proc. Natl Acad. Sci. USA* **105,** 1620–1625 (2008).
57. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genet.* **42,** 579–589 (2010).
58. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).
59. Stefansson, H. *et al.* Common variants conferring risk of schizophrenia. *Nature* **460,** 744–747 (2009).
60. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460,** 748–752 (2009).
61. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).
62. Sebat, J., Levy, D. L. & McCarthy, S. E. Rare structural variants in schizophrenia: one disorder, multiple mutations; one mutation, multiple disorders. *Trends Genet.* **25,** 528–535 (2009).
63. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455,** 232–236 (2008).
64. Cowles, C. R., Hirschhorn, J. N., Altshuler, D. & Lander, E. S. Detection of regulatory variation in mouse genes. *Nature Genet.* **32,** 432–437 (2002).
65. Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297,** 1143 (2002).
66. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).
67. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42,** 570–575 (2010).
68. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42,** 565–569 (2010).
69. Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in *PCSK9*. *Nature Genet.* **37,** 161–165 (2005).
70. Ji, W. *et al.* Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genet.* **40,** 592–599 (2008).
71. Galarneau, G. *et al.* Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. *Nature Genet.* **42,** 1049–1051 (2010).
72. Dulbecco, R. A turning point in cancer research: sequencing the human genome. *Science* **231,** 1055–1056 (1986).
73. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4,** 177–183 (2004).
74. Davies, H. *et al.* Mutations of the *BRAF* gene in human cancer. *Nature* **417,** 949–954 (2002).
**The discovery of *BRAF* mutations in melanoma has led to new drugs for melanoma with high response rates.**
75. Samuels, Y. *et al.* High frequency of mutations of the *PIK3CA* gene in human cancers. *Science* **304,** 554 (2004).
76. Paez, J. G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304,** 1497–1500 (2004).
77. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321,** 1807–1812 (2008).
78. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).
79. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436,** 117–122 (2005).
80. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310,** 644–648 (2005).
81. Kwak, E. L. *et al.* Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N. Engl. J. Med.* **363,** 1693–1703 (2010).
82. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96,** 9212–9217 (1999).
83. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
84. van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536 (2002).
85. Lamb, J. *et al.* The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313,** 1929–1935 (2006).
86. Boehm, J. S. *et al.* Integrative genomic approaches identify *IKBKE* as a breast cancer oncogene. *Cell* **129,** 1065–1079 (2007).
87. Ley, T. J. *et al. DNMT3A* mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).
88. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330,** 228–231 (2010).
89. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* doi:10.1038/nature09837 (in the press).
90. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and geography of Human Genes* 518, 541 (Princeton Univ. Press, 1994).
91. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).
92. Grossman, S. R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327,** 883–886 (2010).
93. Yi, X. *et al.* Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* **329,** 75–78 (2010).
94. Pollard, K. S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443,** 167–172 (2006).
95. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402,** C47–C52 (1999).
96. Davidson, E. H. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution* (Academic, 2006).
97. Clemons, P. A. *et al.* Small molecules of different synthetic and natural origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proc. Natl Acad. Sci. USA* **107,** 18787–18792 (2010).
98. Peck, D. *et al.* A method for high-throughput gene expression signature analysis. *Genome Biol.* **7,** R61 (2006).
99. Yarrow, J. C., Feng, Y., Perlman, Z. E., Kirchhausen, T. & Mitchison, T. J. Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Throughput Screen.* **6,** 279–286 (2003).
100. Musunuru, K. *et al.* From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2010).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** This review reflects collective ideas, insightful conversations and contributions shared by many colleagues at the Broad Institute and elsewhere. In particular, I wish to express my gratitude to D. Altshuler, J. Baldwin, B. Bernstein, B. Birren, C. Burge, F. Collins, M. Daly, M. DePristo, E. Eichler, A. Futreal, L. Garraway, T. Golub, E. Green, C. Gunter, M. Guyer, M. Guttman, D. Haussler, E. Hechter, J. Hirschhorn, D. Hung, D. Jaffe, S. Kathiresan, L. Kruglyak, E. Lieberman, R. Lifton, K. Lindblad-Toh, S. McCarroll, A. Meissner, T. Mikkelsen, R. Myers, R. Nicol, C. Nusbaum, L. Pennacchio, R. Plenge, A. Regev, D. Reich, J. Rinn, P. Sabeti, V. Sankaran, S. Schreiber, P. Sklar, M. Stratton, H. Varmus, P. Visscher, A. Wolf and O. Zuk. I also thank B. Wong and L. Gaffney for assistance with figures.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence should be addressed to E.S.L. (lander@broadinstitute.org).