# 11 The Human Genome

## THE PROJECT

The flagship project of the age was, of course, the Human Genome Project (HGP). We devote a chapter to it, primarily in the light of its controversies and controversial findings, the interpretation of which bears heavily on the understanding of genetic programming and the use of genomic information in healthcare.

The HGP was first mooted at a conference organized at the University of California Santa Cruz in 1985 by the biophysical molecular biologist Robert Sinsheimer,[1] attended by David Botstein, John Sulston, Bob Waterston, Leroy Hood, Walter Gilbert and George Church, among others. It was formally proposed in 1986 by the cancer virologist Renato Dulbecco[2] at a meeting in Santa Fe attended by, among others, Sinsheimer, Watson and Charles DeLisi from the US Department of Energy (DOE).[3–5] The HGP was then recommended for funding in 1987 by a subcommittee of the Office of Health and Environmental Research of the DOE (including Sinsheimer, Dulbecco and Hood),[6] supported by many luminaries of the time, albeit with reservations.[7] The first human genome sequencing conference was held in 1989 at Wolf Trap Farm near Washington.[4]

The project captured the imagination and ambition of the US government – the biomedical equivalent of the Apollo Space Program – which provided most of the funding through the National Institutes of Health (NIH) and the DOE, with a large contribution from the Wellcome Trust in the UK and support from the governments of Japan, France, Germany and China – the 'public' project. There was also a parallel project undertaken by a private company, Celera Genomics Corporation, headed by the *bête noir* of the human genetic establishment, Craig Venter. The public project was officially launched in 1990, but there was a lot of civil and not-so-civil toing-and-froing before it got seriously underway.

There are three aspects worth recalling. The first is the debate about whether to sequence just mRNAs (cDNAs, as an extension of Venter's 1995 study) or to sequence the entire genome. Why spend all that money of sequencing acres of junk?[8] Moreover, the view of "a surprisingly vocal group" was that the project (in any case) was a waste of money that would be better allocated to other areas of research or to healthcare, exacerbated by a fear of, or antagonism to, 'big science'.[7,8]

For example:

> It is doubtful that much of the resulting information will provide insights into human diseases or fundamental biological processes … (repeated sequences and introns) serve mainly to space exons or represent junk DNA. Obtaining the sequence of these genomic regions is, in my view, simply a waste of money and effort … Genome projects should be severely curtailed or, better still, abandoned.[9]

And Brenner, astride the fence: "If something like 98% of the genome is junk, then the best strategy would be to find the important 2%, and sequence it first".[10]

By contrast from Sinsheimer:

> There is currently a facile assumption that only 1 or 2 or 5 percent of the genome is 'of interest.' I am not convinced we know that. Surely, in an evolutionary sense, much more will be of interest. Knowledge of the variability among the genomes of individuals will surely shed light on variations in physiology and susceptibility to disease, as well as on questions of human origin.[11]

Others, Watson[a] in particular (who was made initial director of the project, and whose genome was the second to be sequenced[12]), agreed and maintained that the human genome could not be understood unless it was sequenced in its entirety, including its non-coding elements, whatever their extent and form might be.[13]

The second aspect was the speculation at the time about the numbers of 'genes' in the human genome,

[a] Watson's recollections may be found at https://wellcomecollection.org/works/m4kr8fz5.

DOI: 10.1201/9781003109242-11

which had declined from early and seemingly ludicrous estimates of millions (based on genome size and bacterial-like gene density; Chapter 5) to somewhere in the range of 30,000–150,000.[14–19] As always the underlying assumption was that, apart from those specifying infrastructural RNAs involved in mRNA splicing and translation, and a few others, the gene complement would be mainly protein-coding.

The third was the effort to assemble a coordinated international consortium to undertake the project, mainly at three meetings in Bermuda, co-chaired by the NIH, DOE and the UK's Wellcome Trust, in 1996, 1997 and 1998. They set out the so-called 'Bermuda principles',[20] which held that the human genome sequence data should be made public immediately, promulgated by the Wellcome Trust (and its senior investigators, notably John Sulston), which had no external stakeholders to satisfy, and the NIH, which likely realized that US interests would benefit most because of their capacity for fast adoption. This initially made life difficult for those from other countries, notably Germany, France and Japan, whose governments wanted to capture commercial value from their investment, but eventually the main players prevailed.[20–22]

The Bermuda conferences also discussed which group(s) would take responsibility for, and have provenance over, the sequencing of specific chromosomes, or parts thereof, based on their historical work on mapping of genetic disorders, and the resources they had developed along the way – based on the clone and map then sequence strategy proposed by Gilbert (see [8]). Venter announced that he would just sequence the whole lot, shotgun-style, and assemble the genome from overlapping 'contigs', which was met with a mixed reaction.[b]

The competition between the 'public' and 'private' genome sequencing initiatives had two interesting consequences: it spurred the funding agencies, notably the Wellcome Trust, to increase their investment in the project;[c] reciprocally Celera used the flow of data releases from the public project to accelerate the development of its draft of the human genome sequence.[23]

In any event, as is so often the case, competition was a good thing, and the project was completed ahead of time and under budget, with an estimated total cost around $USD 3 billion. Rapprochement was achieved and in 2001 'first drafts' of the sequence (totaling ~2.9 gigabases, or ~90%) of a composite genome amalgamated from a number of anonymous individuals by the public consortium and of Craig Venter's genome by Celera were published contemporaneously in *Nature*[24] and *Science*,[25] respectively. These publications were accompanied by fanfare announcements on both sides of the Atlantic by then US President Clinton (flanked by Venter and Francis Collins, who coordinated the public project as the then director of the National Human Genome Research Institute) and UK Prime Minister Blair. A more complete sequence was published in 2004 (Figure 11.1).[26]

**FIGURE 11.1** Industrial scale sequencing for the human genome. (Reproduced from the Human Genome Sequencing Consortium[24] with permission of Springer Nature.)

Analysis of the assembled sequences showed that just ~1% of the genome is protein-coding, with ~2% of the total represented in mRNAs (including the 5' and 3'UTRs that control mRNA localization, translation and turnover),[d] whereas 24% is intronic and 74% is 'intergenic' DNA. The genome was found to contain fewer protein-coding genes than expected, the initial counts by the two camps being 30,000–40,000[24] and 26,588 with "an additional approximately 12,000 computationally derived genes with mouse matches or other weak supporting evidence".[25]

[b] Personal recollection of JSM.

[c] https://www.sanger.ac.uk/news_item/1998-05-13-wellcome-trust-announces-major-increase-in-human-genome-sequencing/.

[d] More recent estimates indicate that only 0.77% of the human genome contains protein-coding information, and that exons in mature mRNAs occupy 1.74% of the genome.[27]

Even these surprisingly low estimates also turned out to be inflated, likely biased by prior expectations. The actual number of human protein-coding genes has since been revised downward to ~20,000,[27–29] although increasingly offset by growing numbers of genes found to express small and large non-protein-coding RNAs[30] (Chapters 12 and 13).

## ASSESSMENT OF FUNCTIONALITY

The subsequent publication and comparative analysis of the mouse genome sequence in 2002 (almost half of which can be aligned to the human genome, with 99% protein orthology[e]) included the estimate that only ~5% of the sequences in mammalian genomes has been 'conserved' during evolution, and by imputation is functional.[34]

The estimate assumed that ancient 'repeats' (i.e., transposon-derived sequences) that have persisted in both genomes since their divergence over 100 million years ago are non-functional and can be used to determine the rate and distribution of 'neutral' evolution of unconstrained sequences over time. Applying this estimate to the remainder of the alignable sequences showed that 95% had diverged to similar extent, with only 5% diverging more slowly, under evolutionary pressure for preservation of particular sequences, termed 'purifying selection',[34] despite dramatic variations in the 'neutral' substitution rates across the genome (Figure 11.2).[35–38]

The conclusion that most of the human genome is not under evolutionary selection and is therefore not functional was widely accepted. It supported the orthodox view and has remained a central plank of the argument of that most of the genome is junk,[39,40] and is therefore important to address.

There are several logical problems with the analysis upon which this conclusion relies. First, it is entirely circular: the assumption that ancient transposon-derived sequences that are orthologous in both genomes are non-functional was used to justify the conclusion that most of the rest of the genome is also non-functional. If the assumption is correct (although there was no evidence to support it), the conclusion is reasonable. If the assumption is wrong, then the conclusion is also wrong.[41]

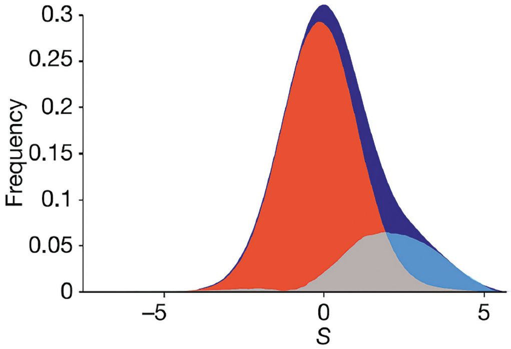

FIGURE 11.2 Comparison of the distribution of sequence divergence between the alignable fraction of the human and mouse genomes (dark blue), decomposed into a mixture of two scaled component distributions: neutrally evolving recognizable common ancient repeats (red) and sequences imputed to be under selection after subtraction of the red distribution from the blue distribution (light blue and gray), corresponding to approximately 5% of the total, which contains most of the orthologous protein-coding sequences (estimated to be about 1.5%). The remainder is assumed to be conserved regulatory elements. Note that if the red curve comprises only the recognizable highly conserved end of the original distribution, the presumed neutral rate of sequence divergence will have been underestimated (the red distribution will be shifted left), and the proportion of the genome imputed to be under selection will be higher. (Reproduced from Waterston et al.[34] with permission of Springer Nature.)

Indeed, this was a questionable assumption given that the reference sequences have been retained independently in the mouse and human genomes for over 100 million years, especially in view of the considerable evidence of the biological functions of TEs, the known cases of which, however, were regarded as exceptions rather than examples of a general phenomenon. This increasingly appears to be incorrect (Chapter 10).

Second, even if the assumption that most ancient retrotransposon-derived sequences that date back to the common ancestor are non-functional is correct, the analysis had an inherent flaw: many of the 'ancient repeats' used in the comparison are barely recognizable as being orthologous, because their sequences have drifted apart, which means there may be, and likely are, an unknown number that have diverged further, to the point of being unrecognizable.[42–45] Indeed the mouse genome analysis stated:

> The ability … to detect (common ancestral) repeats was found to fall off rapidly for

[e] Most protein-coding genes are conserved in vertebrates, with a large proportion of proteins shared between human, birds and fish,[31,32] many in all metazoans[33] (Chapter 10).

> divergence levels above about 37%. If we simulate the events … the proportion of the genome that would still be recognizable as ancestral repeats falls to only 6%.[34]

Consequently, the rate of (supposed) neutral evolution in mammalian genomes, and therefore the extent of their functionality, was underestimated to an unknown extent, and even a small increase in the true neutral evolution rate results in a large increase in the proportion of the human genome that is under 'purifying' selection.[41]

Third, while sequence conservation imputes function – highly structured RNAs like rRNAs and proteins are constrained by their physicochemical structure-function relationships – lack of sequence conservation imputes nothing.[46] Not only do non-conserved, lineage-specific sequences underlie evolutionary novelties, regulatory sequences (including gene promoters and some enhancers) can and indeed do evolve quickly,[36,47–50] like language,[f] under different sequence-function constraints and positive selection for adaptive radiation.[40,41] A high proportion of non-coding RNAs and other regulatory sequences, including many that have been functionally validated, show little sequence conservation, use TE-derived modular elements and are lineage-restricted, sometimes with only short conserved sequence and structural 'motifs' embedded in large RNA molecules (Chapters 13 and 16).

Subsequent studies showed that there are at least seven different rate classes of sequence evolution in the human genome,[51] at least 18% of the human genome is conserved at the level of predicted RNA structure,[52] there is strong negative selection across both coding and non-coding sequences,[53] and the vast majority of sequence variations influencing complex traits and diseases occurs in the non-coding regions of the genome (see below).

A pairwise comparison of eight mammalian species concluded that "there is a high rate of turnover of functional non-coding elements in the mammalian genome, so measures of functional constraint based on human-mouse comparisons may seriously underestimate the true value",[36] later reaffirmed by analyzing broader genomic datasets in avian lineages.[54] This challenges the use of primary sequence conservation and *a priori* dismissal of repetitive sequences in the assessment of genome functionality.

## THE MAJORITY OF THE GENOME IS ACTIVE

The possibility that the widely held belief that the vast majority of the human genome is non-functional may be wrong soon became evident in other ways.

The large-scale transcriptome sequencing projects that followed the genome projects revealed that most of the mammalian genome is differentially transcribed, producing an extraordinarily complex interlacing suite of coding and non-protein-coding RNAs, the latter exhibiting exquisitely precise expression patterns (Chapter 13).

The subsequent ENCODE ('Encyclopedia of DNA Elements') project, a large international study which aimed to identify functional elements in the human genome, encompassing RNA expression, the distribution of chromatin modifications (Chapter 14), transcription factor binding sites, DNase hypersensitive (exposed) regions, promoters, etc.[g] in different cell types (Figure 11.3), concluded, in its 2007 'pilot' publication covering 1% of the genome, that most of the studied regions exhibited (these) biochemical indices of function.[55,56]

This figure, however, was at odds with the estimate in the same paper (reiterating that from the earlier human-mouse genome comparison) that only "5% of the bases in the genome can be confidently identified as being under evolutionary constraint in mammals" … and therefore that "Surprisingly, many functional elements are seemingly unconstrained across mammalian evolution".[55]

After some internal (pre-submission) debate among the authors, a decision was made not to canvas the alternative possibility that the estimate of the extent of 'conservation' of the genome- and that only clearly conserved sequences are functional-might be incorrect.[36,47,48] Instead, the incongruity was rationalized as the existence of "a large pool of neutral elements that are biochemically active but provide no specific benefit to the organism".[55] This somewhat contradictory statement became a key talking point and led to a wager publicized in *Nature* as to whether more or less than 20% of the human genome

[f] The evolution of language is a useful comparison. The English word 'brother' and the French word 'frère' have no obvious homology, but not only do both have meaning, they have the same meaning and are derived from a common antecedent, having diverged under loose sequence-function constraints (sender-receiver recognition, as in regulatory circuits).[40]

[g] Unfortunately, the project did not include an examination of the incidence or distribution of alternative DNA structures.

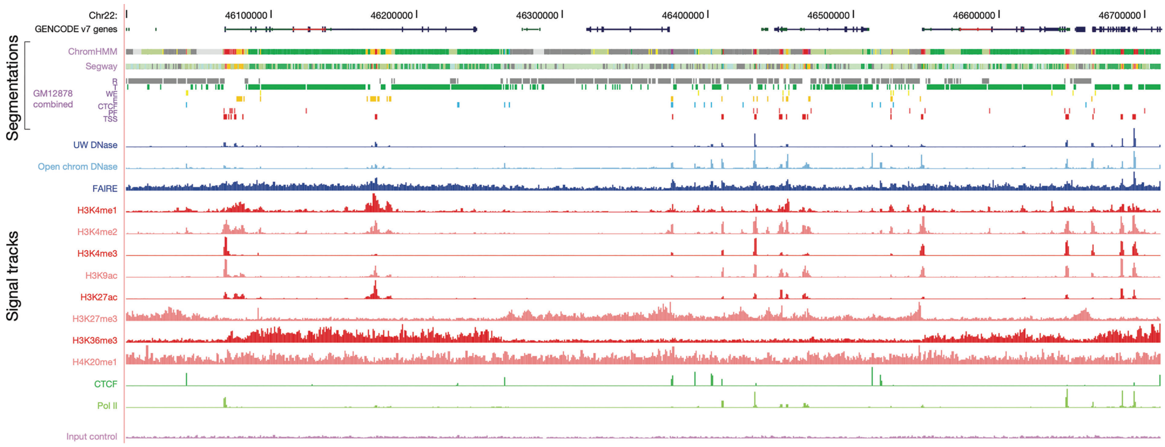

**FIGURE 11.3** Representative compilation of ENCODE genomic features cataloged on part of human chromosome 22 in the GM12878 lymphoblastoid cell line. Annotated (protein-coding) genes and their exon-intron structures are shown at top. Chromosome segmentation refers to blocks sharing similar features. Other tracks show predicted enhancers (E, Chapter 14), transcription start sites (TSS), RNA polymerase II binding sites (Pol II), open chromatin (DNase accessibility), nucleosome depleted sequences (FAIRE[56]) and the positions of nucleosomes marked with various histone modifications (Chapter 14). (Reproduced from Dunham et al.[60] with permission of Springer Nature.)

is functional,[57] which at the time of writing had still not been settled.

The more comprehensive 2012 genome-wide ENCODE paper[h] confirmed that at least 80% of the human genome "participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type", and addressed the conservation conundrum by stating that

> an appreciable proportion of the unconstrained elements are lineage-specific elements required for organismal function … and the remainder are probably 'neutral' elements that are not currently under selection but may still affect cellular or larger scale phenotypes without an effect on fitness.[60]

This paper spawned another round of controversy,[61,62] with some apoplectic at the suggestion that a large fraction of the genome may be functional, invoking the C-value paradox, mutational load and circular conservation arguments (Chapter 7), while rejecting any suggestion that dynamic transcription or differential chromatin modifications in non-protein-coding regions might be valid indices of genetic function, including in a species- and clade-restricted fashion.[39,63,64] It is clear that some of the antagonism was related to the invocation of junk in genomes as a line of argument against proponents of intelligent design,[40] who seize and misuse scientific ideas and observations to try to justify non-scientific, untestable beliefs.

## DAMAGED GENES

Naturally, the human genome was a major focus of genetic mapping by medical geneticists to identify genes responsible for serious inherited 'Mendelian' metabolic, physiological, developmental and/or cognitive disorders.[i] These diseases are the result of "catastrophic component damage",[68] i.e., disruptive mutations (mainly) in protein-coding sequences, which are generally lethal or severely disabling in the homozygous state, and many deleterious in the heterozygous state.

[h] The data included identification of ~2.9 million DNase hypersensitivity sites, ~580,000 of which could be connected to promoters,[58] and evidence that over 75% of the genome is differentially transcribed,[59] identifying many more transcripts than just those encoding 20,000 proteins and their splice variants (Chapter 13).

[i] Most such mutations are recessive, meaning that two damaged copies are required for the disorder to manifest, and reciprocally that there may be high frequency of heterozygous carriers, especially for mutations that may, like cystic fibrosis (1 in 25 carrier frequency in Caucasian populations)[65] and sickle cell anemia (common in tropical and subtropical regions), have provided protection in the heterozygous state against tuberculosis[66] and malaria,[67] respectively, a positive evolutionary trade-off.

However, of course, controlled breeding was not possible to construct conventional genetic maps of the human genome to locate and identity damaged genes and, in any case, the number of known genes with trackable allelic variants that segregated in large families was limited. A different approach was needed.

The solution, proposed by Ellen Solomon and Walter Bodmer in 1979[69] and again by David Botstein, Ray White, Mark Skolnick and Ron Davis in 1980,[70] was to take advantage of single nucleotide polymorphisms ('SNPs') in genomes that resulted in gain or loss of restriction endonuclease sites and a resulting change in the size of the corresponding fragments. Such restriction fragment length polymorphisms, or RFLPs[j] could be tracked as surrogate genetic markers by hybridization of Southern blots with cloned sequence probes, and linked to the inheritance of a condition in extended families.[72]

The search was assisted by the construction of chromosome-specific cloned libraries by flow cytometry sorting of metaphase chromosomes in the early 1980s by Kay Davies, Bryan Young, Rob Krumlauf and colleagues,[73,74] by the use of somatic cell hybrids developed by Bodmer and colleagues and Stephen Goss and Henry Harris in the 1970s,[75,76] and 'radiation hybrid' mapping developed in 1990 by David Cox, Richard Myers and colleagues, whereby individual human chromosomes or parts thereof can be separated and maintained in mouse cell lines.[77–79]

These approaches were made feasible by high penetrance disorders, which are easy to trace in affected pedigrees, especially if dominant or located on the X-chromosome, i.e., commonly exposed in males. On the other hand, the difficulty in identifying the causative gene was increased by the relatively large genomic regions identified by genetic mapping, a needle in a haystack problem.

One of the complications was that meiotic recombination rates across the human genome (and indeed across mammalian genomes in general) are not uniform, but rather occur at hotspots,[k] between which there is little recombinational exchange,[81] referred to as 'linkage disequilibrium'. The recombination-poor regions between hotspots are termed 'haplotype blocks',[82] which parse the genome into 'HapMaps'[83,84] that subsequently formed the analytical platform for population-scale mapping of genetic variations influencing complex traits and multifactorial diseases[85] (see below).

The identification of damaged genes by cloning and mapping approaches was, at the time, a *tour-de-force*, achieved by high-resolution mapping of the chromosomes carrying affected genes, and searching for markers (i.e., sequence variants) that are co-inherited with the condition, aided by homozygosity mapping, since most damaged genes are recessive.[86] Coarse mapping to haplotype blocks was relatively easy, but fine mapping to locate the affected gene within the region, especially in the absence of obvious candidates, relied on rare recombinational or deletion events in particular families, which were hard to find.

Eventually the hard grind paid off, culminating in the identification in 1986 by Tony Monaco, Lou Kunkel and colleagues of the protein-coding gene on the X-chromosome that is damaged in Duchenne's Muscular Dystrophy (*dystrophin*).[87–89] Dystrophin is required for the maintenance of muscle integrity and, as noted previously, is one of the largest genes and proteins[l] in vertebrates.[90,91] In 1989, Lap-Chee Tsui and colleagues identified the gene responsible for cystic fibrosis on chromosome 7, which encodes a chloride ion transporter ('Cystic Fibrosis Transmembrane Regulator' or 'CFTR')[92,93] and explained the symptoms of the disease, including salty sweat, in both cases providing targets for diagnosis and gene therapy.[65,93–99]

In 1991, a CGG trinucleotide repeat expansion was identified in the 5'UTR of the *FMR1* gene (which encodes a synaptic protein) in Fragile X Syndrome, the most common form of inherited intellectual disability[100,101] (which is also associated with autism).[102] Similar repeat expansions were subsequently identified in other genes causing X-linked or autosomal dominant neurological disorders such as Kennedy's Disease, Myotonic Dystrophy, Huntington's Disease and Spinocerebellar Ataxia,[103–113] which were initially thought to result in defective proteins or translation (since many lie in the introns or UTRs) but may also be RNA toxicity disorders, an increasingly prominent theme in neurodegenerative diseases (Chapter 16) (Figure 11.4).

[j] Sequence polymorphisms in intronic sequences that altered restriction sites were later patented as a means of diagnosing tightly linked genetic disorders,[71] a surprising decision by patent offices in view of the long history of linkage mapping.

[k] Human genetic data suggests that 60% of recombination events happen in 6% of the genome.[80]

[l] Human dystrophin is composed of 79 exons (encoding 3,684 amino acids) that account for 0.6% of its 2.4Mb sequence.[90]

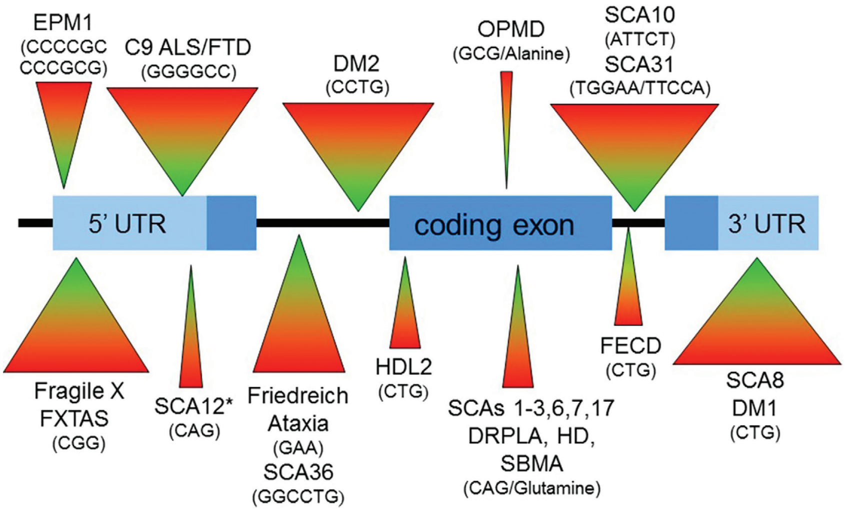

**FIGURE 11.4** Schematic of gene showing repeat expansions that cause neurologic diseases. The differing sizes of the associated triangles roughly reflect the range of repeat expansion sizes in each disease. The SCA12* repeat is in an intron. (Reproduced from Paulson[113] with permission of Elsevier.)

Others followed, and it became easier as the technology improved.

## A PLETHORA OF 'RARE DISEASES'

About 3%–5% of all children are born with a serious physical or intellectual disability due to a mutation in a protein-coding gene or a chromosomal abnormality,[114] which are also major causes of miscarriage.[115,116] While some genetic disorders, like cystic fibrosis and thalassemia, are relatively common, most are individually rare, due to damage to any one of thousands of protein-coding genes. Collectively, however, they account for a high proportion of all infant deaths and pediatric hospital admissions, as well as a lifetime burden on survivors, their families and health systems.[m]

Such damaged genes, because of their low allele frequency in the population and mostly recessive nature, often lie silent in family histories, as the incidence of homozygosity is low.[n] The high collective frequency of defective alleles,[121–123] however, means that at least 1 in 10 couples are at serious risk of bearing a disabled child with every pregnancy, due to the 1 in 4 chance of each transmitting to their child a damaged gene that they unknowingly have in common. There is also a surprisingly high frequency of new ('de novo') mutations that result in intellectual disability.[124,125]

The identification of damaged genes in individuals suffering severe disabilities is now done not by genetic mapping (impossible due to their rarity) but by whole 'exome' or genome sequencing, comparing their genome (and usually those of their parents, called a 'trio') with a reference, to identify mainly variations in protein-coding sequences (or proximal non-coding regions, such as splicing signals) that introduce a frame shift or stop codon that result in a truncated protein, or codon changes that disrupt protein structure and function.[o]

Exome sequencing[p] has been favored by many clinical geneticists and others because of the

[m] Protein-coding mutations, whose presence may not be evident in early life, account for up to 50% of pediatric hospital conditions and 10%–20% of all hospital admissions, as well morbidity and premature death in later life.[117–119] Examples of such (damaged) genes are those causing familial hypercholesterolemia and cardiac defects, which result in catastrophic heart failure in otherwise healthy adults. Many adults carry protein-coding mutations that have yet to become pathogenic.[120]

[n] Higher in communities that have consanguineous, i.e., mostly first cousin, marriages, which increases the odds of homozygosity of mutated genes in the children.

[o] There are many more possible amino acid changes (allelic variants) with more subtle effects.

[p] Exome sequencing is accomplished by oligonucleotide-based hybridization capture of known protein-coding sequences, thereby removing ~99% of the genome prior to sequencing.

emphasis on protein-coding mutations and because it is cheaper and easier to analyze than whole genome sequencing. Its diagnostic yield is, however, lower than whole genome sequencing because it is limited to annotated exons in annotated genes, has technical biases, and is generally unable to detect other types of damage such as translocations and copy number variations, which can be found using other means.[126–128]

Nonetheless the process is becoming increasingly efficient, supported by databases that have cataloged thousands of genetic disorders, and sophisticated software that can sift through millions of individual sequence variations. It is also being aided by increasing detection of 'expressed regions' in transcriptome studies, leading to better annotation of both coding and non-coding genes.[127,129]

Currently, several genetic conditions are polled at birth by the so-called 'Guthrie' heel prick blood test, which uses biochemical and genetic tests to screen for genetic disorders that can be treated by early intervention. The prototype, and good example, is the test for phenylketonuria, a rare recessive disorder whereby infants cannot metabolize the aromatic amino acid phenylalanine, leading to mental retardation, which can be avoided by dietary modification.[130] In the near future, it is likely that the Guthrie test will be replaced, conditional on parental consent, with whole genome sequencing,[q] which will provide a much more comprehensive view of incipient genetic problems and allow early intervention to prevent or mitigate their effects.

Moreover, the surprise finding that a significant proportion (~6%) of the DNA circulating in the blood of pregnant women comes from the fetus[131,132] has led to the rapid rise of non-invasive prenatal testing (NIPT), which can detect chromosomal trisomies (such as trisomy 21, Down's Syndrome) more accurately and with no threat to the embryo (unlike the preexisting amniocentesis and chorionic villi sampling tests).[133] This has also contributed to the progressive demise of medical cytogeneticists, whose other main activity is detecting chromosomal translocations in cancer and balanced translocations in reproductive failure, also soon to be replaced by genome sequencing.

The identification of mutations causing serious disorders will inevitably become fully automated, as computers match the spectrum of patient sequence variation and clinical features with recorded cases, reducing and eventually obviating the need for *ad hoc* sleuthing by clinical geneticists in hospital laboratories.[134] Computerization will also allow such information, and recommended evidence-based actions based on the latest publications and national guidelines, to be delivered to the desktop of health professionals, including general practitioners on the front line.

## COMPLEX TRAITS AND DISORDERS

Human genetic analyses have progressively moved from protein-centric (such as the classic blood-group and HLA allele frequencies) to variable microsatellite loci[135–137] and more recently to genome-wide approaches involving very large numbers of individuals.[138,139] Human genomes vary by ~0.1%, i.e., unrelated individuals have 4–5 million sequence differences, although the total number of differences that occur among humans is many times greater, with no absolute differences yet found between populations, although allele frequencies vary.[140–142] Studies comparing the congruence or difference between identical and non-identical twins showed that genetic factors play a substantial part in susceptibility to almost all human traits and disorders, including to infectious diseases.[143,144]

However, the identification of genetic loci contributing to complex traits and diseases is not amenable to the approaches used in mapping severe genetic deficiencies because each causal locus often only makes a small contribution to overall heritability.[145,146] The problem was to a significant extent solved by the development of haplotype maps and oligonucleotide arrays (originally developed by Patrick Brown and colleagues for transcriptome analysis[147]) that poll common sequence variants.[141] SNP arrays are cheap to produce and enabled large population-scale surveys, termed 'genome-wide association studies' (GWAS), which compare the distribution of sentinel SNPs (and by imputation other variants that co-segregate in the same haplotype block) to identify variants that are statistically over- or under-represented with respect to the trait or disease under study.[146] The statistical probabilities are then graphed across the genome to produce so-called 'Manhattan' plots, with

[q] Whole genome sequencing can also allow detection of mutations in non-coding regulatory RNAs, such that occur in some cases of phenylketonuria – see Chapter 13.

a P-value of $\sim 10^{-8}$ commonly used as a significance threshold.[148,149]

The first GWAS was conducted in 2002 on 94 Japanese individuals who had suffered myocardial infarction and 658 controls using (protein-coding) gene-centric SNPs, identifying, among others, an intronic SNP that enhanced the transcriptional level of the lymphotoxin-alpha gene, confirmed in a more focused analysis of over 1,000 affected individuals and controls.[150] This was followed by a study in 2005 on 96 individuals suffering macular degeneration with 50 controls, which identified two significantly associated SNPs in an intron of a gene encoding a blood complementation factor.[151] Two years later the Wellcome Trust Case Control Consortium published a multilateral GWAS involving 14,000 cases of seven common diseases (~2,000 individuals for each of coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension) with 3,000 shared controls.[152] Other studies at the time led to the discovery of many variants in non-coding regions regulating the developmental expression of human fetal hemoglobin.[153–155]

Since then, study sizes have grown to millions of individuals and have encompassed over 1,000 different conditions and traits, usually using 'biobank' samples integrated by international consortia.[156]

Despite the still-present difficulties in teasing out the interplay of genetic variations and heterogeneous social-environmental factors in complex traits, the phenotypes examined include psychological traits such as temperament,[157] neuropsychiatric disorders (such as autism[158,159]), schizophrenia and bipolar disorder,[160–162] ADHD (attention deficit hyperactivity disorder), panic disorder and depression,[163–167] vertigo,[168] neurodegenerative diseases such as Alzheimer's and Parkinson's Disease,[161,169–171] as well as various types of cancer,[172] immunological disorders (such as ankylosing spondylitis, ectopic dermatitis, asthma and inflammatory bowel disease),[173–176] hypertension,[177] height and body mass index,[178] bone density and osteoporosis,[179] alcoholism and other drug dependences,[180–184] caffeine consumption,[185] handedness,[186] insomnia,[187] aging,[188] and even cognitive performance,[189] intelligence[190–193] and correlated (and environmentally contingent) educational attainment (Figure 11.5).[194]

Two general findings emerged from this fleet of GWAS, apart from the identification of tens of thousands of SNPs/haplotype blocks associated with various conditions and traits.

The first is that GWAS does not appear to identify, quantitatively, all of the genetic contribution to complex traits, traditionally determined by pedigree estimates and twin studies, although these are

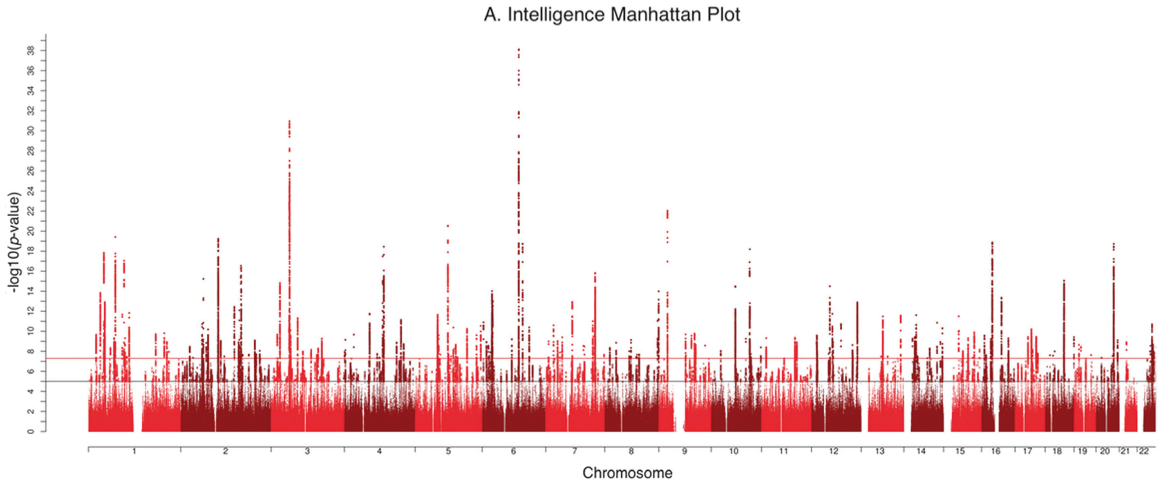

**FIGURE 11.5** Combined Manhattan plot of two large genome-wide association studies of education and intelligence. (Reproduced from Hill et al.[193] under Creative Commons Attribution 4.0 International License.) The red line indicates threshold for genome-wide significance and the black line the threshold for suggestive associations. The data suggest that genes involved in neurogenesis, myelination, expressed in the synapse and involved in the regulation of the nervous system play a role in the variation in intelligence.

limited by confounding environmental and methodological factors.[143,195–197] The emblematic example is height, a deceptively simple trait that is known to be highly (80%–90%) genetically determined after controlling for environmental variables such as nutrition,[198] implicit in twin studies, but where only ~25% of the variance could be accounted by GWAS-identified loci, of which there at least 180.[178,199,200] More extensive studies identified several thousand "near-independent" DNA markers and rare variants that appear to account for 60%–70% of the genetic contributions to height,[178,201] possibly overestimated due to uncorrected stratification.[197] Another confounding factor is 'hidden epistasis' (i.e., synergistic interactions between loci involving regulatory networks).[195,202,203] There is similar complexity of polygenic contributions to other traits such as urate, insulin-like growth factor 1 and testosterone levels.[204]

In addition to the plethora of environmental factors and life histories that can interact with genotypes differently, haplotype analysis masks 'private' mutations and tandem repeat variations that have occurred in individual lineages since the divergence of the common versions of haplotype blocks, which occurred hundreds of generations ago.[205] For example, whole genome sequencing of families found that

> Variation in height in our sample arises from a combination of a small number of QTLs[r] with large effects - which are not tagging previously identified common variants, and so cannot be imputed from them - and a large number of common variants with small effects.[206]

Tandem repeat variations make a significant contribution to autism,[207–209] as also do rare mutations that are only a few generations old.[210] There is also an unknown contribution of transgenerational epigenetic inheritance (Chapter 17), which cannot be polled by DNA variants.

The second general finding from GWAS is that (unsurprisingly) the vast majority of genetic variations associated with complex traits and diseases, including cancer predisposition,[172] occur outside of protein-coding sequences, in intronic and intergenic sequences.[211–219]

Although some pleiotropic SNPs (affecting multiple traits) occur in coding sequences, UTRs and promoters, loci containing multiple-trait associated variants cover the majority of the genome.[220] A high proportion of the imputed loci exhibit the signatures of being (real) genes, including promoters characterized by DNase hypersensitivity, typical chromatin modification signatures and transcription.[68,213,221–225] Variation between individuals also occurs by recombination between endogenous proviral sequences,[226] as well as in tandem repeat sequences.[227]

There is enrichment of variations in enhancers (Chapters 14 and 16) that are active in disease-relevant cell types associated with developmental abnormalities, cancers, Alzheimer's Disease, schizophrenia, autoimmune diseases including diabetes, rheumatoid arthritis and multiple sclerosis, and cardiovascular disorders. Functional analyses are uncovering increasing numbers of causal variations, many linked to non-coding RNAs transcribed from these loci.[160,228–238] Indeed, while most haplotype blocks identified as being associated with complex traits and diseases in GWAS studies are devoid of protein-coding genes ('gene deserts'),[216,239,240] most produce multi-exonic non-protein-coding RNAs,[223–225,241–243] at least some of which comprise or are candidates for the molecular basis of trait association[234,235,244–251] (Chapter 13).

Identifying the relevant variations within haplotype blocks among the many differences between individuals is a huge challenge but will likely be achieved by analysis of large datasets of genome sequences, as recently in the case of autism.[252] These analyses will be informed by model organism[s] studies,[254] RNA expression and predicted structural variants, DNA and histone modifications in affected tissues (epigenome-wide association scans/studies or EWAS) and transcription factor binding profiles, to link genomic variants with molecular and phenotypic indices.[53,255–262]

## THE TRANSFORMATION OF MEDICAL RESEARCH AND HEALTHCARE

Frustration has often been expressed at the delay in the delivery of health benefits from the HGP, many based on promises and expectations that arguably had their roots in a century-old 'genes for'

[r] QTL = Quantitative trait loci.

[s] High-throughput genetic screening of *C. elegans* orthologs of human obesity-candidate genes reported in GWAS identified 17 protein-coding loci that are causally linked to obesity across phylogeny.[253]

mentality, which was boosted by successes in the study of monogenic disorders but does not reflect the complexity of most human diseases and traits. Nonetheless, by 2011 it was estimated that there had been a 140-fold economic return on investment by the US government in the HGP,[263] and, fueled by the major scientific advances that the genome sequences have made possible, there is renewed interest in harnessing the information in whole genome sequences for healthcare at individual and population scales.

The $1,000 human genome sequence cost barrier was breached in 2014 and is likely to decline further. New competing approaches and technologies are emerging and reaching the market, such as long-read sequencing and the combination of chromosome conformation capture and deep sequencing for chromosome-length assembly of large genomes.[264] Other technologies using solid-state devices and high-resolution microscopy may not be far away, with the $100 human genome sequence in sight.

The declining cost of sequencing prompted the establishment of projects to explore human genetic diversity and the etiology of cancer, beginning with the 1,000 Genomes[140,141] and the International Cancer Genomes Consortium projects,[265,266] followed quickly by the UK 100,000 Genomes Project, the first to apply population-scale genomic sequencing to the diagnosis of genetic disorders and cancer,[267] completed in 2018. Larger projects are underway, with the UK announcing a minimum of 1 million genomes (and an "ambition" of 5 million genomes) to be sequenced by the UK Biobank and the National Health Service[268] and 1 million genomes to be sequenced by the US 'All of US' program, along with accompanying clinical and lifestyle data,[269] with similar projects under way in China and many other places. In fact, with the accumulation of such studies (with high-coverage WGS of very large number of individuals with diverse ancestral and admixed backgrounds, deep phenotyping, longitudinal assessment and improved imputation methods[270]), the focus is shifting to the dissection of the contribution of rare non-coding variants to human phenotypic variation,[215,271,272] including polygenic risk variant calling for complex diseases[273–275] (an approach that is still controversial[276]) and pharmacogenomic indices to guide drug selection and dose.[277]

Sequencing of tumor DNA is also revolutionizing the understanding and treatment of cancer,[278] showing that cancers that arise in different tissues are caused by a similar spectrum of mutations.[279,280] Most of the main 'driver' mutations occur in protein-coding genes, such as TP53, whereas there are many other non-coding variants that also contribute,[t] including previously undetected 'weak drivers' with aggregated effects on cancer phenotypes.[266,283,284] Increasing numbers of the protein mutations can be treated with targeted drugs[285,286] and, in the case of tumors with high mutational load, with immunotherapies, which are proving extraordinarily successful in increasing survival.[287–289]

Genetic screening is allowing the identification of cancer-susceptibility genes and monogenic diseases at population level, finding a large number of previously unsuspected carriers and individuals with latent disease risk.[290–292] Within a decade or so it is likely, depending on the pace of reduction in sequencing and storage/analysis costs, that genomic analysis will become routine in the early detection, identification, treatment and prevention or mitigation of genetically linked disorders and risks, including those usually manifested later in life, such as familial hypercholesterolemia, arthritis and cancer. Although the evidence framework and underlying databases are still evolving, identification of underlying mutations is already leading to improvements in outcomes, through either the selection of targeted drugs or the likely response to immunotherapies, leading to substantial increases in life expectancy and, sometimes, to permanent remission.[285,293]

Twenty years after the publication of the human genome drafts, there is virtually complete coverage of entire chromosomes[294,295] and a full picture of the diversity of human genomes (as 'the genome' is a misnomer) is emerging,[296] including the secrets of the highly repetitive and heterochromatic regions and the functional impact of their variations.

The acquisition of human whole genome sequences at scale will continue to illuminate human biology and transform medical research, drug discovery and healthcare over the coming decades. Millions of genomes, accompanied by billions of data points from clinical records, self-phenotyping and smart sensors (which record real-time physiological and environmental parameters), will create a multidimensional information ecology that can be mined for new genotype-phenotype correlations using machine learning and other

[t] Some non-coding mutations in regulatory regions, such as – prominently – in the promoter of the *TERT* telomerase component, also have driver malignant effects.[281,282]

methods of artificial intelligence, consequently refining patient stratification and treatment.[u] Once the infrastructure is in place to analyze and report the consequences of genomic variants to clinicians (and patients), medicine will change from the art of crisis management to the science of good health, and radically improve the quality, efficiency and sustainability of healthcare, arguably the most important and fastest growing industry in the world.

## FURTHER READING

Davies K. (2010) *$1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine* (Free Press, New York).

Tam V., et al. (2019) Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 20: 467–84.

Watson J.D. (1990) The human genome project: past, present, and future. *Science* 248: 44.

Watson J.D. and Cook-Deegan R.M. (1991) Origins of the human genome project. *The FASEB Journal* 5: 8–11.

[u] Machine learning on transcriptome and genomic data in relation to cell differentiation stage will also transform understanding of the genes and genetic variants controlling development.