The Human Genome Project—An Overview

David R. Bentley

The Sanger Centre, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Abstract: The human genome sequence will underpin human biology and medicine in the next century, providing a single, essential reference to all genetic information. The international program to determine the complete DNA sequence (3,000 million bases) is well underway. As of January 2000, 50% of the sequence is available in the public domain. A comprehensive working draft is expected this year, and the entire sequence is projected to be finished in 2003. DNA sequencing is carried out on mapped, overlapping bacterial clones of 150–200 kb. The working draft comprises assembled unfinished sequence and is released immediately in the public domain. The draft sequence of each clone is then completed, by closing any remaining gaps and resolving any ambiguities, before the entire sequence is checked, annotated, and submitted to the public databases. The sequence of each clone is finished to an accuracy of >99.99%. The availability of a reference sequence of the genome provides the basis for studying the nature of sequence variation, particularly single nucleotide polymorphisms (SNPs), in human populations. SNP typing is a powerful tool for genetic analysis, and will enable us to uncover the association of loci at specific sites in the genome with many disease traits. SNPs occur at a frequency of ~1 SNP/kb throughout the genome when the sequence of any two individuals is compared. Programs to detect and map SNPs in the human genome are underway with the aim of establishing a SNP map of the genome during the next two years. The human genome sequence will provide a complete description of all the genes. Annotation of the sequence with the gene structures is achieved by a combination of computational analysis (predictive and homology-based) and experimental confirmation by cDNA sequencing. Detecting homologies between newly defined gene products and proteins of known function helps to postulate biochemical functions for them, which can then be tested. Establishing the association of specific genes with disease phenotypes by mutation screening, particularly for monogenic disorders, provides further assistance in defining the functions of some gene products, as well as helping to establish the cause of the disease. As our knowledge of gene sequences and sequence variation in populations increases, we will pinpoint more and more of the genes and proteins that are important in common, complex diseases. A more detailed understanding of the function of the human genome will be achieved as we identify sequences that control gene expression. Given the availability of gene sequences, the expression status of genes in particular tissues can be monitored in parallel. By comparing corresponding genomic sequences in different species (for example: man, mouse, chicken, and zebrafish), regions that have been highly conserved during evolution can be identified, many of which reflect conserved functions such as gene regulation.

Correspondence to: David R. Bentley

These approaches promise to greatly accelerate our interpretation of the human genome sequence.

Key words: human genome; DNA sequence; genetics; genes; polymorphism

1. INTERNATIONAL HUMAN GENOME PROJECT

The goal of the International Human Genome Project is to decode the entire 3,000 million bases (Mb) of DNA sequence, and to interpret the information stored within it in order to provide a new foundation for understanding human genetics and biology. To obtain the most value from this endeavor, it is important to make the complete sequence freely available in the public domain.[1] Interpretation will also be considerably enhanced by comparing the human genome sequence with the complete genome sequences of other organisms as they become available. For example, the 15 Mb genome sequence of baker's yeast (*Saccharomyces cerevisiae*) was completed in 1996,[2] and contains all the genes that encode this example of a functional eukaryotic cell; the 97 Mb genome sequence of the free-living soil nematode worm, *Caenorhabditis elegans*, was completed in 1998[3] and is the first multicellular animal for which a complete gene set is available. Finding matches between genes in these organisms and human DNA sequence is already proving invaluable in human gene classification.

The pioneering work that led to completion of these model genomes provided many important technical insights in genome analysis that have since been applied to the human genome. A common strategy was developed in which the entire genomic DNA of the organism was broken into random fragments and cloned into bacteria (using vectors such as bacteriophage lambda, 15–20 kb; or cosmids, 30–45 kb). For bigger genomes, larger fragments (0.1–1.5 Mb) were also cloned directly into yeast as yeast artificial chromosomes.[4] A physical map of the cloned fragments was then constructed, in which the clones were arranged in overlapping sets to form a continuum of clones, or contig. Minimally overlapping clones were chosen for determination of the complete DNA sequence representing the region. It is a tribute to the success of this strategy that the clone map of the 97Mb nematode has only two remaining internal gaps.

An early milestone in tackling the much larger human genome was the construction of a genome-wide genetic map. Polymorphic sequences composed of variable numbers of the dinucleotide repeat CA (microsatellites) were used as genetic markers to establish a linkage map. 2,000 of the 6,000 markers typed were ordered at high odds ($>$1000:1), thus providing the first framework of unique landmarks across the genome.[5] The use of overlapping yeast artificial chromosomes, and, later, radiation hybrid (RH) mapping[6] resulted in construction of maps of much higher densities of landmarks for each of the human chromosomes (e.g., Refs. 7–9). A feature of these maps was that all genes, genetic markers, and other unique markers were integrated on a single scale. An important product of this stage of the project was a comprehensive gene map,[10] constructed by RH mapping using 30,181 gene-based markers derived from expressed sequence tags (ESTs), which had been clustered into nonredundant sets, or Unigenes (Ref. 11; see www.ncbi.nlm.nih.gov/UniGene).

The strategy for constructing maps of overlapping bacterial clones for sequencing is to screen large insert (150–200 kb) P1, or bacterial artificial chromosome libraries using the available landmarks from the detailed chromosome framework maps (described in detail in Ref. 12). Pairwise comparison of restriction fingerprints is also used to determine the degree of overlap between clones and, thus, to aid in the assembly of the bacterial clone map. Contigs are extended and joined by walking, using new landmarks generated from a sequence derived from the end of a cloned insert, where required. Additional mapping information on the order, orientation, and gap size between contigs can be established using clones as probes in fluorescent *in situ* hybridization (FISH) analysis of interphase nuclei or extended DNA fibers (described in Ref. 13; see figure 1 and also front cover illustration of fiber FISH analysis of overlapping chromosome 22 clones).

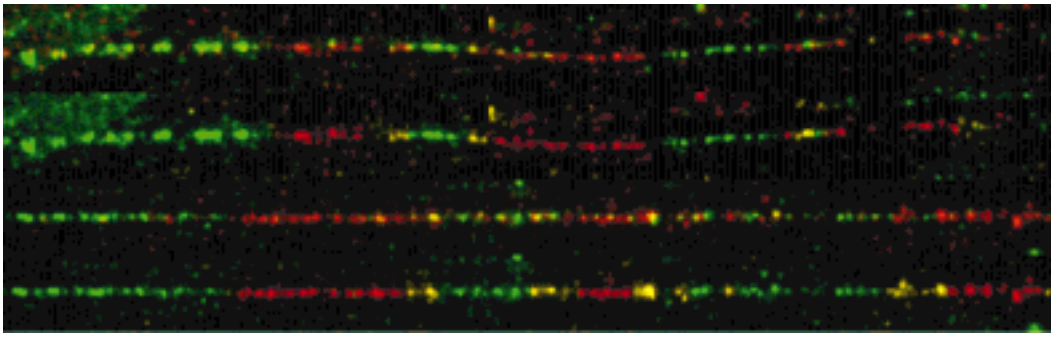

Figure 1. Fiber FISH analysis of overlapping clones on chromosome 22. A minimally overlapping set of bacterial clones (in this case cosmids) have been selected from the physical map for sequencing. Alternate clones were labeled to fluoresce either red or green, mixed and hybridized to extended fibers of human DNA. Overlaps between clones are apparent as indicated by the yellow signals (reproduced with permission from N. Carter)

Determination of the DNA sequence itself is carried out on each selected 150–200 kb bacterial clone. In an initial "shotgun" phase, 2–3,000 sequence reads are determined at random and assembled automatically, and the assembled unfinished sequence is released in the public domain. This is followed by a directed "finishing" phase in which all gaps are closed, ambiguities are resolved, and the entire sequence is checked before submission to the public databases (see Ref. 12 and references therein). The sequence of each clone is finished to an accuracy of >99.99%, leaving no gaps. As of January 2000, there is over 1,500 Mb of human DNA sequence available, of which 480 Mb is finished (see www.ncbi.nlm.nih.gov; and Genome MOT at www.ebi.ac.uk).

The release of the human genome sequence in mapped segments of 100–200 kb means that all other genetic information can be immediately integrated into the map using the public sequence as a reference. This includes all the isolated gene fragments (ESTs, cDNAs) and polymorphic markers (microsatellites), and also all genomic sequence that is released in the public domain as a result of other large-scale genome sequencing efforts, notably whole-genome shotgun sequencing.[14] It is a unifying feature of DNA sequence itself that all efforts that become publicly available can be combined to provide the best possible picture of the human genome for posterity.

2. SINGLE NUCLEOTIDE POLYMORPHISMS

Single nucleotide polymorphisms (SNPs) will provide the genetic markers for the next era in human genetics. Particular advantages of using SNPs over other types of genetic markers include the relative ease of automating SNP typing assays for cheap, robust, large-scale genotyping, and the abundance of SNPs throughout the genome. Comparison of the DNA sequence between two individuals reveals, on average, one nucleotide difference every 1,000 bases.[15] This frequency increases as more samples are included in the comparison.[16,17] The majority of changes are single base substitutions, termed single nucleotide polymorphisms (SNPs), although deletions and insertions of one or more bases are also observed. Heritable sequence variations arise as a result of a copying error, or damage and misrepair of DNA that results in alteration of the genomic sequence in germ cells. As a result, the changes are then passed on in subsequent generations, and the two or more alternatives at a particular position, or locus, become established in the population. Further diversity arises as a result of recombination between homologous segments of chromosomes during meiosis.[18]

A number of experimental approaches have been used to detect SNPs. As the genome sequence is determined from overlapping BACs or PACs, sequence differences can be detected directly by

comparison of the two genomic sequences in the region of overlap whenever the two overlapping clones originate from different genotypes.[15] Methods used to target SNP discovery in specific regions (for example, in genes of interest) include direct sequencing of PCR-amplified segments of the genome from multiple individuals, and sequence differences are detected directly by analysis of the sequence output.[16,17,19] Alternative techniques have also been used to scan selected sequences for variations, such as denaturing high-pressure liquid chromatography (dHPLC), single-stranded conformational polymorphism, and sequencing by hybridization to oligonucleotides attached to a solid phase (i.e., microarrays or gene chips).

In contrast to the targeted approaches, random strategies are aimed at accumulating a large number of SNPs randomly distributed throughout the genome. Large-scale "shotgun" sequencing (i.e., of clones picked at random) of libraries of either individual chromosomes, or from total genomic DNA, can be used to obtain sequence from one or more individuals. The resulting sequences are aligned to the available human genome reference sequence, and high-confidence sequence differences are recorded as candidate SNPs. This approach has been tested in a pilot study using a library made from flow-sorted chromosome 22 DNA, and results in the detection of SNPs at a frequency of 1SNP/850 bases (Mullikin et al., unpublished results). A further modification of this strategy involves shotgun sequencing of a library made from a specific size fraction of gel-purified restriction fragments derived from a mixture of unrelated genomic DNA samples. Limiting the shotgun sequencing to a small, defined subset of genomic fragments results in sequence sampling of a reduced representation, or subset, of the genome. As the sequencing proceeds, multiple examples of each fragment are sampled, and can be aligned to each other. Because the starting DNA for the library is derived from a mixture of different individual genotypes, the aligned reads represent sequences of different genetic origin, and differences between them (including SNPs) can be detected. This approach, termed the reduced representation shotgun (RRS) strategy, does not require the reference human genome sequence in order to detect candidate SNPs. The RRS strategy has been adopted as the initial approach to identify SNPs by a recently formed consortium of pharmaceutical companies, the Wellcome Trust and international genome centers [The SNP Consortium (TSC); see http://snp.cshl.org].

As a result of the combined efforts already in progress in the public domain, it is likely that >400,000 candidate SNPs will be identified within the next two years (the stated goals of the TSC and the NIH-funded programs to date are to identify 300,000 and 100,000 SNPs, respectively). SNPs are being deposited in public databases, notably dbSNP (at www.ncbi.nlm.nih.gov), which contains over 26,000 SNPs as of January 2000.

SNPs will provide a powerful tool for genetic analysis, enabling us to uncover the association of loci at specific sites in the genome with many disease traits (see Ref. 20 for review of analysis of complex disease traits). DNA sample collections, selected for a particular phenotypic attribute such as type I diabetes, hypercholesterolaemia, or many other common and complex conditions, can be typed with a genome-wide set of SNPs to detect an association between a region of the genome (defined by one or more of the SNPs) and the condition under study. Thereafter, the genes in the region can be investigated further, initially by searching for sequence variants within each gene. This may enable identification of the gene product that is directly involved, which in turn may be a new target for therapeutic intervention.

SNP typing will also be of great use in identifying functional variation in the response of individuals to different drugs, and this area of pharmacogenomics promises to revolutionize the precision in predicting side-effects that are specific to the genetic make-up of each individual.[21] This will open up the way to tailor the prescription of existing drugs to the individual, as well as driving the development of new drugs.

A variety of assay systems are currently under development, which promise to make high-throughput genotyping a reality in the next few years, although none has yet been employed on a large scale. Foremost examples include the use of synthetic oligonucleotide microarrays, which are hybridized to labeled amplified DNA from an individual.[22] The presence of different alleles of each

SNP is detected by measuring the intensity of fluorescence bound at each allele-specific oligonucleotide. To date, chips containing assays for up to 2,000 SNPs have been developed (www.affymetrix.com). Other assays in use include single base primer extension ("minisequencing") either on gels or microarrays,[23] mass spectroscopy on tagged beads, and solution assays in which allele-specific oligonucleotides are cleaved or joined at the position of the SNP allele, resulting in activation of a fluorescent reporter system. Examples of the latter include the oligonucleotide ligation assay, the Taqman and others (reviewed in Ref. 24); and this list is by no means exhaustive.

3. GENE ANNOTATION

The human genome sequence will ultimately provide a complete description of all the genes. Annotation of the gene structures is achieved by a combination of computational analysis (predictive and homology-based) and experimental confirmation by partial cDNA sequencing. Initially, the finished sequence of each clone is masked to exclude known repeats, and then searched for matches with gene or protein sequence in the public databases using BLAST.[25] In parallel, the unmasked sequence is analyzed by prediction programs (e.g., Genscan,[26] Fgenes,[27] and others reviewed in Ref. 28) to identify candidate exons, and other features such as potential CpG islands,[29] which may indicate the presence of a gene. The results of the analyses are combined and displayed in a database. Following visual inspection, a conservative set of gene features is selected. Further experimental analysis is then carried out to confirm and extend the results as follows. Primers are designed to the selected gene features and tested for their ability to amplify the corresponding segments from cDNA libraries for sequencing. Alignment of the cDNA sequence to the genomic sequence results in definition of the gene structure, including intron-exon boundaries. Further extension of the annotation at the 5′ and 3′ ends, if necessary, can be done using further rounds of cDNA library screening,[30] and also using 5′ or 3′ RACE (random cDNA extension),[31] which involves primer extension on cDNA synthesized directly from cellular RNA. The extension products are sequenced and the data is incorporated into the annotated gene structure. The resulting annotated sequence is submitted to the public databases.

4. GENES, DISEASE, AND PROTEIN FUNCTION

Establishing the association of a specific gene (in a mutated form) with a disease phenotype assists in defining the function of the gene product, and also provides us with an understanding of the biochemical basis of the disease. Most disease-gene associations described to date have been for monogenic disorders, where identification of the causative gene involves a search based on knowledge of the biological function of the gene product, or by positional cloning (reviewed in Ref. 32). In the latter approach, the position of a disease locus is initially defined by establishing linkage of the phenotype to particular genetic markers, or by detection of a cytogenetic rearrangement, which is associated with the disease. Candidate genes within the critical region are then identified. The DNA of patients that exhibit the disease phenotype is screened for the presence of mutations within suitable candidate genes. The presence of mutations that alter the expression or function of the gene product in a manner which is consistent with the disease phenotype provides conclusive evidence of the association between the disease and the causative gene.

Notable examples of disease genes discovered by positional cloning include cystic fibrosis, Duchenne muscular dystrophy,[33] Bruton's disease,[34] and many others (reviewed in Ref. 35). Positional cloning projects are increasingly making use of genomic sequence analysis (as described above) to provide knowledge of the candidate genes. A recent example is the identification of the gene involved in X-linked lymphoproliferative disease (XLP), a disease that manifests as a severe immune deficiency in young boys, which is elicited by exposure to Epstein–Barr virus. The XLP gene was localized to within a 3 Mb critical region on the long arm of the X-chromosome, on the ba-

sis that this region was deleted in one patient, who was, therefore, presumed to be missing the XLP gene. An extensive search within the annotated genomic sequence of the region revealed two candidate genes that were located within the deletion, one of which (the SH2D1A gene) was found to contain mutations in multiple independent XLP patients.[36]

Genomic sequence information can also help to elucidate the function of a gene product. The sequence of the SH2D1A gene product exhibited a high degree of homology with the peptide sequence of the SH2 (src-homology type 2) domains first discovered in the transforming *src* gene of Rous sarcoma virus. SH2 domains have since been found in many protein kinases that participate in cell signaling pathways.[37] In contrast to most proteins that contain SH2 domains, the SH2D1A gene product appears to contain no catalytic domain. Direct biochemical studies led to the discovery that SH2D1A interacts with the intracellular domain of the signal lymphocyte activating molecule (SLAM), apparently competing with the binding of another SH2 domain protein SHP2.[38] The SLAM-SHP2 pathway is involved in signaling between B and T cells. The putative function of SH2D1A is, therefore, to down-regulate the immune response to EBV infection by competitive inhibition of the SHP2-SLAM signaling pathway. The absence of SH2D1A activity in XLP patients results in failure to down-regulate the lymphoproliferative response, causing the observed phenotype.

5. COMPARATIVE SEQUENCING AND REGULATORY SEQUENCES

Comparison of genomic sequences among different species can be used to identify sequences that have been conserved during evolution. Striking similarities are found in protein coding regions, even of genomes that are extremely divergent (for example, 37% of proteins in budding yeast share homologues in man; see Ref. 3). Comparisons between genome sequences that are more closely related (i.e., those that have diverged more recently in evolution) indicate similarities not only in protein coding sequences, but also marked similarities in flanking regions, which are important in other conserved functions such as transcriptional regulation. This is illustrated in a recent example of comparative sequencing of the stem-cell leukemia (SCL-1) gene, which encodes a helix-loop-helix transcription factor that is the earliest known regulator in haematopoiesis. Alignment of human, mouse, chicken, and puffer fish SCL-1 gene sequences resulted in identification of a range of conserved elements in the 5′ flanking region and within introns.[39] In particular, the alignment of human and chicken gene sequences provided precise definition of regulatory sequence motifs in the promoter region, whereas the human and mouse sequences were too similar to highlight specific motifs. This study illustrates two important points. First, comparison of genome sequences will greatly accelerate the process of identifying biologically important sequence motifs as well as protein coding sequences. Second, multiple genome sequences (of species with different degrees of evolutionary divergence) will all contribute important information—a simple comparison of man and mouse alone will not be enough.

6. WHAT WILL THE GENOME DELIVER?

We are on the verge of a revolution in biology. The first draft of the human genome sequence will become available during 2000, and the reference version is scheduled for completion to a high degree of accuracy by 2003. The complete sequence of the human genome will draw together all the available genetic information, which will be positioned on a single metric to an accuracy of a single base pair. The human sequence should not be viewed in isolation. A growing number of whole genome sequences of other organisms will help us to obtain detailed, accurate annotation of the important biological sequence information that is stored within every genome. We will accrue complete

gene sets, and also gain a full picture of the nature and diversity of proteins. It will be possible to understand how specific sequences regulate the expression of genes in the genome. Most important, genome sequences will provide the foundation for a new era of experimental and computational biology, providing the essential resources for future study.

REFERENCES

1. Bentley DR. Genomic sequence information should be released immediately and freely in the public domain. Science 1996;274:533–534.
2. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
3. The Caenorhabditis elegans Sequencing Consortium. Genome sequence of the Nematode Caenorhabditis elegans. A platform for investigating biology. Science 1998;282:2010–2018.
4. Burke DT, Carle GF, Olson, MV. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 1987;236:806–812.
5. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, et al. A comprehensive genetic map of the human genome based on 5,264 microsatellites [see comments]. Nature 1996;380:152–154.
6. Walter MA, Spillett DJ, Thomas P, Weissenbach J, Goodfellow PN. A method for constructing radiation hybrid maps of whole genomes. Nature Genet 1994;7:22–28.
7. Foote S, Vollrath D, Hilton A, Page DC. The human Y chromosome: overlapping DNA clones spanning the euchromatic region. Science 1992;258:60–66.
8. Hudson TJ, Stein LD, Gerety SS, Ma J, Castle AB, Silva J, et al. An STS-based map of the human genome. Science 1995;270:1945–1954.
9. Collins JE, Cole CG, Smink LJ, Garrett CL, Leversha MA, Soderlund CA, et al. A high-density YAC contig map of human chromosome 22. Nature 1995;377:367–379.
10. Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez Tome P, et al. A physical map of 30,000 human genes. Science 1998;282:744–746.
11. Boguski MS, Schuler GD. Establishing a human transcript map. Nature Genet 1995;10:369–371.
12. The Sanger Centre and The Washington University Genome Sequencing Center. Toward a complete human genome sequence. Genome Res 1998;8:1097–1108.
13. Leversha M. Fluorescence in situ hybridisation. In: Dear PH, editor. Genome mapping — A practical approach. Oxford: IRL; 1997. p 199–225.
14. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M. Shotgun sequencing of the human genome. Science 1998;280:1540–1542.
15. Taillon–Miller P, Gu Z, Li Q, Hillier L, Kwok PY. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. Genome Res 1998;8:748–754.
16. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
17. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, et al. DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. Nature Genet1998;19:233–240.
18. Chakravarti A. It's raining SNPs, hallelujah? Nature Genet 1998;19:216–217.
19. Kwok PY, Deng Q, Zakeri H, Taylor SL, Nickerson DA. Increasing the information content of STS-based genome maps: identifying polymorphisms in mapped STSs. Genomics 1996;31:123–126.
20. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037–48.
21. Housman D, Ledley FD. Why pharmacogenomics? Why now? Nature Biotech 1998;16:492–493.
22. Lipshutz RJ, Fodor SPA, Gingeras TR, Lockhart, DJ. High density synthetic oligonucleotide arrays. Nature Genet 1999;21:20–24.
23. Pastinen T, Kurg A, Metspalu A, Peltonen L, Syvanen AC. Minisequencing: a specific tool for DNA analysis and diagnostics on oligonucleotide arrays. Genome Res 1997;7:606–614.
24. Landegren U, Nilsson M, Kwok PY. Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. Genome Res 1998;8:769–776.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.

26. Kulp D, Haussler D, Reese MG, Eeckman FH. A generalized hidden Markov model for the recognition of human genes in DNA. In: Proc Fourth Int Conf Intel Sys Molecular Bio; 1996. p 134–142.
27. Solovyev VV, Salamov AA, Lawrence CB. Identification of human gene structure using linear discriminant functions and dynamic programming. In: Proc Third Int Conf Intel Sys Molecular Bio; 1995. p 367–375.
28. Burset M, Guigo R. Evaluation of gene structure prediction programs. Genomics 1996;34:353–367.
29. Antequera F, Bird A. Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci USA 1993;90:11995–11999.
30. Huang SH, Yang AY, Holcenberg J. Amplification of gene ends from gene libraries by polymerase chain reaction with single-sided specificity. In: White BA, editor. Methods in molecular biology, PCR protocols: Current methods and applications. Totowa, NJ: Humana; 1993. p 357–363.
31. Frohman MA, Dush MK, Martin GR.. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci USA 1988;85:8998–9002.
32. Collins FS. Positional cloning: let's not call it reverse anymore. Nature Genet 1992;1:3–6.
33. Monaco AP, Neve RL, Colletti Feener C, Bertelson CJ, Kurnit DM, Kunkel LM. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 1986;323:646–650.
34. Vetrie D, Vorechovsky I, Sideras P, Holland J, Davies A, Flinter F, et al. The gene involved in X-linked agammaglobulinaemia is a member of the src family of protein-tyrosine kinases. Nature 1993;361:226–233.
35. Collins FS. Positional cloning moves from perditional to traditional. Nature Genet 1995;9:347–350.
36. Coffey AJ, Brooksbank RA, Brandau O, Oohashi T, Howell GR, Bye J M, et al. Host response to EBV infection in X-linked lymphoproliferative disease results from mutations in an SH2-domain encoding gene. Nature Genet 1998;20:129–135.
37. Koch CA, Anderson D, Moran MF, Ellis C, Pawson T. SH2 and SH3 domains: elements that control interactions of cytoplasmic signaling proteins. Science 1991;252:668–674.
38. Sayos J, Wu C, Morra M, Wang N, Zhang X, Allen D, et al. The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. Nature 1998;395:462–469.
39. Gottgens B, Barton LM, Gilbert GJR, Bench AJ, Sanchez MJ, Gering M, et al. Functional genomic analysis of vertebrate SCL loci identifies conserved enhancers. Nature Biotech, to appear.

David R. Bentley *is the Head of Human Genetics at the Sanger Centre, Hinxton, Cambridge, UK. Research at the Sanger Centre is directed to the understanding of genomes, particularly through DNA sequencing and analysis.*