
CHAPTER 7

Genetic and Physical Map Resources—an Integrated View

MICHAEL R. BARNES

GlaxoSmithKline Pharmaceuticals, Harlow, Essex, UK

- 7.1 Introduction
 - 7.1.1 What is a genome map?
- 7.2 Genetic maps
 - 7.2.1 Human genetic maps
 - 7.2.2 The Genethon genetic linkage map
 - 7.2.3 The Marshfield genetic linkage map
 - 7.2.4 TSC SNP linkage map
 - 7.2.5 SNP-based haplotype and linkage disequilibrium (LD) maps
- 7.3 Physical maps
 - 7.3.1 Cytogenetic maps
 - 7.3.2 Fluorescence *in situ* hybridization (FISH) mapping
 - 7.3.3 Radiation Hybrid (RH) mapping
 - 7.3.4 Human RH-mapping panels
- 7.4 Physical contig maps
 - 7.4.1 Yeast Artificial Chromosome (YAC) maps
 - 7.4.2 Bacterial Artificial Chromosome (BAC) maps
- 7.5 The role of physical and genetic maps in draft sequence curation
 - 7.5.1 Electronic PCR (e-PCR)
- 7.6 The human genome sequence—the ultimate physical map?
- 7.7 QC of genomic DNA—resolution of marker order and gap sizes
- 7.8 Tools and databases for map analysis and integration
 - 7.8.1 Entrez Map View (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search)
 - 7.8.1.1 Searching and browsing Map View
 - 7.8.2 The Genome Database (GDB) (www.gdb.org)
 - 7.8.3 The Unified Database for human genome mapping (UDB) (<http://bioinformatics.weizmann.ac.il/udb/>)
- 7.9 Conclusions
- References

7.1 INTRODUCTION

Not so many years ago, maps of the human genome were restricted to a handful of very low resolution diallelic RFLP marker maps of specific loci. Physical mapping following linkage analysis required a laborious laboratory-based process of contig construction using yeast and bacterial artificial chromosome (YAC and BAC) clones or cosmids. This involved consecutive rounds of library screening and clone characterization to identify overlaps between clones and build contigs. In recent years, as the human genome sequence nears completion, practical approaches to the characterization of genomic loci have changed quite dramatically. Today the process which took many months or even years can be completed in an afternoon using web-based resources. These tools might lead us to believe that the human genome sequence is the only map we need to know, but it actually represents just one dimension of a multifaceted map. Other maps including genetic, cytogenetic and radiation hybrid maps, represent different aspects of the structure, content and behaviour of chromosomes. These properties really need to be integrated with sequence-based maps to fully understand the properties and genomic landmarks that influence genes, mutation and human evolution.

As this book goes to press, the human genome is still unfinished and in the strictest sense it is likely to remain so for several years to come. For example, in April 2002 the human genome draft sequence reached 97.8% coverage, however only 63% of sequence was flagged as finished with 34.8% flagged as draft. The target date for final human sequence completion is 2003. However this may be a moving target, as a combination of contig errors and molecularly intractable regions are likely to continue to keep the genome in at least a partial draft state for many years to come. With this in mind, it is probably pragmatic to assume that the genome will remain unfinished in parts until at least 2005. Mouse genome sequencing is rapidly catching up with human sequencing, with the mouse also projected to finish in 2003. Other mammalian species such as the rat, dog and chimpanzee are further behind, although further genome sequencing will be assisted by existing genomes. The 'pioneer' genome sequences (human and mouse) will be used to span gaps and build contigs by comparison with existing contigs. This approach is already being used to accelerate the mouse and human genome sequencing projects, as both assemblies are being used to span gaps in each respective genome assembly (J. Mullikin, personal communication).

As we are becoming more aware of the difficulties of completing whole genome sequences, the role of physical and genetic maps is changing. Generation of new maps continues to be the first line of study for organisms with poorly characterized genomes. But where the genome sequencing of an organism is advanced, emphasis on maps is shifting to a role in the finishing and QC of existing sequencing maps. With this proviso in mind and with a specific focus on human maps, this chapter will review genetic and physical maps as they are being directly applied and integrated with the human genome and other sequenced mammalian genomes. We will not attempt to cover the full complexity of all forms of maps, or attempt to describe the use of these maps to enable the study of unsequenced organisms. Instead we will review the principles and informatics issues that apply to this area, with a focus on the data which is most likely to be useful to the human geneticist. For example we will examine the use of genetic and physical maps to check the order and orientation of marker maps and genomic contigs. For researchers who wish to construct new genetic and physical maps without sequence data we direct the reader to specialist texts in this research area.

7.1.1 What is a Genome Map?

At the most basic level, a genome map is a collective set of markers with known relative positions. A marker could be any genomic element with a uniquely identifiable sequence or property. Markers can exist in many different forms, such as non-polymorphic sequence tagged sites (STS) which act as a unique anchor or SNPs and short tandem repeats (STR), which act as both unique anchors and markers for differentiation between individuals. Genomic maps are divided into two broad categories. Polymorphic markers are used to construct genetic maps and either polymorphic or non-polymorphic markers are used to construct physical maps.

7.2 GENETIC MAPS

The genetic linkage map is a key concept which gives a fundamental insight into the genetic nature of the genome. Genetic linkage maps inform on more than just order of markers, they also give a measure of the underlying genetic recombination that occurs in a particular chromosomal region. Linkage maps show the relative locations of specific DNA markers along the chromosomes of related individuals. Any inherited physical or molecular characteristic that differs among individuals and is easily detectable is a potential genetic marker, for this reason polymorphic markers, such as SNPs and STRs are particularly suited to genetic map construction as they are plentiful, easy to characterize precisely and amenable to laboratory automation (see Chapter 3 for a review of SNPs and STR markers).

Genetic maps are constructed by evaluating the genotypes of a set of markers in groups of related individuals. This raw mapping data is analysed by software packages, such as MapMaker (Lander *et al.*, 1987; reviewed in Chapter 12) which construct genetic maps by observing how frequently the alleles at any two markers are inherited together. The closer the markers are, the less likely it is that a recombination event will separate the alleles, and the more likely it is that they will be inherited together. Thus, unlike physical maps, the distance between markers on a genetic map is not measured in any kind of physical unit; it is a measure of the recombination frequency between those two markers. This genetic map unit is measured in centimorgans (cM). The distance between two markers would be measured as 1 cM if both markers are separated by recombination on 1% of occasions. Genetic distance has an average correlation with the actual physical distance between markers, on average in humans 1 cM is equivalent to 1 Mb (this ratio varies widely between other species). The 1 cM:1 Mb ratio is often used as a rule of thumb, but it is important to recognize that this is a genome-wide average and can often diverge significantly from this ratio between different regions of the human genome. The genetic/physical ratio also differs considerably between genders, as recombination frequencies vary between males and females. To overcome these differences, genetic maps typically report distances for each sex and a 'sex-averaged' distance that integrates male and female recombination frequencies.

7.2.1 Human Genetic Maps

A range of genome-wide human genetic maps has now been published at various resolutions. Most genetic maps are based on STR markers, although a genome-wide SNP

linkage map has also been published recently (T. C. Matise *et al.*, unpublished data). Most genome-wide linkage maps are constructed with a marker framework spaced at 2.5–10-cM intervals. Denser marker maps have not been widely used for linkage analysis, as the focus of analysis is on a small number of meiotic events observable within a family. These meiotic events do not require a very dense map of markers to find evidence for possible co-segregation of a disease-influencing gene with marker locus alleles. Higher resolution genetic maps have been described, but they are generally restricted to specific chromosomal regions, such as the long arm of chromosome 21 (Lynn *et al.*, 2000), where they have been used to refine initial linkage analysis. Ideally, to be maximally informative, genetic markers need a relatively high level of heterozygosity (>0.6). This provides a high likelihood that a marker (or cluster of SNPs) will be different between any two copies of a chromosome. Markers with lower heterozygosity, for example, SNPs which range in heterozygosity from ~ 0.1 – 0.3 , need to be used in higher density to give a similar level of information.

The three main genetic maps were developed by Genethon, the Marshfield Institute and the SNP consortium (TSC) (see Table 7.1 for a comparison). The Genethon and Marshfield maps are widely indexed by mapping tools, such as MapViewer and GDB (see below). The newer TSC map is also likely to be available in these tools in the near future.

7.2.2 The Genethon Genetic Linkage Map

The Genethon human linkage map was the first whole genome genetic map to exclusively use STR markers; previous maps were based on less informative RFLPs (which are actually uncharacterized SNPs). The 5264 markers in the Genethon map have a mean heterozygosity of 0.7, which makes it more informative than previous maps. The map was constructed with data from eight CEPH families (comprising 186 meioses) so the fine order of markers is not well resolved, other than by localization within a particular chromosomal region. The map spans a sex-averaged genetic distance of 3699 cM. The average interval size is 1.6 cM, 59% of the map is covered by intervals of 2 cM at most and 1% remains in intervals above 10 cM. The map comprises 2335 positions, of which 2032 could be ordered with an odds ratio of at least 1000:1 against alternative orders. This high level of statistical confidence in marker order was subsequently used by DeWan *et al.* (2002), to highlight a number of discrepancies in the order and orientation of clones in the human genome draft assembly. Genethon map data can be accessed at the Genethon website (www.genethon.fr) and the Washington University, St Louis website (www.genlink.wustl.edu/genethon_frame/).

TABLE 7.1 Human Genetic Maps

Map	Genethon	Marshfield	TSC
Marker type	STRs	STRs	SNPs
Marker no.	5264	8325	2679
Av. heterozygosity	0.7	0.68	0.76
Resolution (kb)	1.6 cM	1.3 cM	2.5 cM
Reference	Dib <i>et al.</i> (1996)	Broman <i>et al.</i> (1998)	Matise <i>et al.</i> (unpublished data)

7.2.3 The Marshfield Genetic Linkage Map

The Marshfield genetic linkage map improved on the Genethon map, by offering a larger marker number and a slightly higher resolution. Like the Genethon map, the Marshfield map was constructed with data from eight CEPH families and therefore fine order is still poorly resolved. In particular, markers which are separated by little or no genetic distance generally have no recombination events separating them, and so they are presented in arbitrary order. Accurate ordering information for these markers can be obtained by cross referencing STS marker location with human physical maps, such as RH maps or the human genome sequence itself. The Marshfield database (<http://research.marshfieldclinic.org/genetics/>), provides a well-documented range of five genome scan marker panels (genome-wide screening sets 6–10), selected from the Marshfield map. These marker panels were initially developed from the first human linkage mapping screening set from the Cooperative Human Linkage Centre (CHLC) (Murray *et al.*, 1994). Each Marshfield marker panel provides a progressively higher density of markers, culminating in set 10 which consists of 405 di, tri and tetra-nucleotide repeat markers with an average spacing of 9 cM. Each marker set is also grouped by allele size so that each panel can be loaded into the same lane or capillary. Primers for marker set 10 are commercially available from Research Genetics, in unlabelled and fluorescent dye-conjugated forms (<http://www.resgen.com/>).

7.2.4 TSC SNP Linkage Map

Technology developments have brought the cost of SNP genotyping far below the cost of STR genotyping. This has led to calls for the development of a SNP-based linkage map. The only argument against the implementation of such a map is the lower heterozygosity of a single SNP compared to a polymorphic STR (Kruglyak, 1997; see Chapter 8 for a discussion of this issue). Use of single SNPs at similar densities to STRs would essentially be equivalent to the original and less informative RFLP maps. Two related solutions have been proposed to overcome this problem. The first solution is to use a 3–8-fold increase in SNP marker densities to produce an evenly spaced map (Kruglyak, 1997). The second is to use multiple clusters of two to three SNPs in linkage analysis at a similar density to STRs. These SNP clusters provide approximately the same amount of information as an STR in terms of heterozygosity (Goddard and Wijsman, 2002).

Matisse *et al.* (unpublished data) used the SNP cluster approach to construct a whole genome SNP linkage map. To do this they selected 666 physically and genetically mapped polymorphic STS anchor loci at 5-cM intervals across the human genome. Ten or more SNPs were then characterized across each STS locus. SNPs were assessed for genotyping success rates, assay quality, allele frequencies (ideally >20%), multi-SNP haplotype heterozygosities (ideally >0.6) and levels of linkage disequilibrium (SNPs in LD with each other were avoided). The three most informative markers per STS locus were then selected to maximize multi-SNP haplotype heterozygosities, to create an informative SNP cluster at each map position. Two thousand SNPs were selected and genotyped in 661 individuals from 48 CEPH reference pedigrees (<http://www.cephb.fr/>). Linkage maps were constructed without reference to any other mapping or sequence position information. This generated a map with an average resolution of 5 cM; to improve this, a further set of SNPs were identified at half-way points between the SNP clusters loci and were similarly evaluated. The single most informative SNPs at each of these positions were identified ($N = 679$) and genotyped in the CEPH pedigrees. These 'single' SNPs were added to the cluster linkage map to produce a final SNP map with a 2.5-cM resolution.

The construction of this map was supported by the SNP Consortium (TSC), all the data and results are available at the TSC website (<http://snp.cshl.org>).

7.2.5 SNP-based Haplotype and Linkage Disequilibrium (LD) Maps

As new SNPs arise at different loci and at different points in time, groups of neighbouring SNPs may show distinctive patterns of co-inheritance or LD, which are arranged into distinct haplotypes between individuals. The great abundance of SNPs across the genome creates an opportunity to exploit this haplotypic diversity in association studies by identifying SNPs which capture or 'tag' the majority of common human haplotypes. This enables the construction of very efficient maps, which capture maximal diversity with a minimal number of SNPs. Such haplotype tags have already been used to screen candidate genes. For example, Johnson *et al.* (2001) re-sequenced nine genes to identify common SNP haplotypes among 122 SNPs. Once these haplotypes were defined they were able to define just 34 SNPs or 'haplotype tags' which identified all the haplotypes across the genes. Extension of this principle across the genome would enable the construction of powerful haplotype-based maps which could capture most common haplotype diversity with a minimal number of SNP markers. At the time of going to press, such a map does not exist in the public domain, although at least one company has this data. A public domain genome-wide haplotype/LD map is likely to become available early in 2004 if not sooner.

Some data is already available publicly. Public domain LD or haplotype maps are available for three chromosomes, these have been generated by two distinct methods and consequently the exact nature of the data presented differs between the maps. Orchid Biosciences Inc. in collaboration with the TSC have published a SNP-based map of chromosome 19 which will be available from the TSC website before this book goes to press (Michael Phillips, personal communication); Dawson *et al.* (2002) published a SNP-based LD map of chromosome 22 and Perlegen Inc. published a SNP-based haplotype map of chromosome 21 (Patil *et al.*, 2001). We take a closer look at the Perlegen map data in Chapter 9.

7.3 PHYSICAL MAPS

While genetic maps display the linear order of genes or markers and the recombination between them, they do not give reliable information on the physical distance between markers and genes. By contrast a physical map has an absolute and invariant base-pair scale, which defines the physical distance between markers. Two markers may be very close genetically, i.e. very little recombination occurs between them, but very far apart physically. The difference between genetic and physical maps may seem academic, however if a trait or disease is localized on a physical map between two molecular markers it is important to identify the amount of recombination across the region, to select an appropriately dense panel of markers to detect a genetic association. Conversely if a genetic map places a trait or disease between two molecular markers, it is useful to know if that distance represents 1 kb, 1 Mb or further still, to define the likely number of genes or regulatory regions in the locus.

7.3.1 Cytogenetic Maps

There are many different types of physical maps; the first identified and lowest resolution physical map of the human genome is the cytogenetic map. This type of map is based on

the distinctive banding patterns of stained chromosomes. Detailed measurements of these patterns were originally used to define the gross physical size of human chromosomes, and led to the size-based sorting of the autosomal chromosomes from chromosome 1, the largest chromosome, to chromosome 22, the smallest. Unsurprisingly these early efforts at physical mapping were quite inaccurate and prone to distortion by differential contraction, which led to the incorrect ordering of chromosome 19 which is actually slightly smaller than chromosome 20 (Morton, 1991). Use of cytogenetic map locations is still remarkably prevalent, perhaps due to the ease of use of the vocabulary of cytobands, e.g. 1q32, 22q11, etc., to describe and cluster groups of genes and loci. Interestingly the cytobanding recognized by early biologists is not just decorative, but in fact the dark cytobands represent regions of higher average GC content, while light cytobands have a lower average GC content (Nimura and Gojobori, 2002). The region where a transition occurs between a dark and light cytoband is known as an isochore, these regions often show a remarkably increased rate of recombination (Eisenbarth *et al.*, 2000). This may make it important to pay special attention to genes and possible regulatory elements in these regions; we specifically address this issue in Chapter 10.

7.3.2 Fluorescence *In Situ* Hybridization (FISH) Mapping

At best a cytogenetic map could be used to locate a DNA fragment to a region of about 10 Mb—the size of a typical chromosome band. Fluorescence *in situ* hybridization (FISH) mapping, is a form of cytogenetic mapping that allows orientation and mapping of DNA sequences to a much higher resolution. Initially FISH resolved markers within 2 Mb, but further development of the FISH method, using chromosomes in interphase when they are less compact, increased map resolution further to around 100 kb. As FISH does not rely on a recombinant map but instead maps a chromosome directly, this has made FISH an important method for the QC of recombinant maps and clone contigs. The level of resolution achieved with interphase FISH, also makes this method directly applicable to the analysis of observable physical traits associated with chromosomal abnormalities, such as prenatal defects or cancer breakpoints. All of these applications are likely to keep the method in regular use well beyond the availability of a complete human genome.

7.3.3 Radiation Hybrid (RH) Mapping

Early physical mapping advanced considerably with the publication of the radiation hybrid (RH) mapping method. Goss and Harris (1975) irradiated human fibroblast chromosomes and fused the resulting fragments with recipient rodent cells. The observed patterns of co-transference of markers in a collection of hybrid cells allowed estimates to be made of linear order and distance between markers by assuming that distant markers are more likely to be separated in different hybrid cell lines than closer markers. The RH mapping technique was refined by Cox *et al.* (1990) who irradiated donor somatic cell hybrids, which contained just a single copy of one human chromosome, and fused the fragments with rodent cells. Several whole genome RH panels were developed in the 1990s which allowed the construction of genome maps containing thousands of STS markers (Gyapay *et al.*, 1996; Stewart *et al.*, 1997). The human RH map finally reached a high-resolution apex, with the development of the TNG panel (Lunetta *et al.*, 1996), which was used to generate an RH map of the human genome consisting of 40,322 STSs (Olivier *et al.*, 2001). From the 40,322 STSs mapped to the TNG radiation hybrid panel, only 3604 (9.8%) were absent from the unassembled draft sequence of the human genome.

7.3.4 Human RH-mapping Panels

Three main radiation hybrid panels have been used for mapping STSs and constructing RH maps, each offers a different level of resolution based on the dose of irradiation. The GB4 RH panel (constructed by using 3000 rad of X-rays) and the G3 RH panel (10,000 rad of X-rays) will resolve markers at 1-Mb and 260-kb intervals respectively, both providing a good long-range continuity for mapping (Deloukas *et al.*, 1998). In contrast, the Stanford TNG panel (50,000 rad of X-rays) allows STS resolution down to 60–100 kb with high confidence (Lunetta *et al.*, 1996). The price of this increased resolution is that a large number of STSs need to be scored to produce good long-range continuity. Olivier *et al.* (2001) found a solution to this by using the TNG panel in conjunction with the Stanford G3 panel to produce an RH map with high-resolution and contiguity. Publication of this map saw a shift in the role of human RH-mapping, from a direct role in mapping new genes to a primarily curatorial role to enable the QC and assembly of the human genome.

RH maps provide a marker order confidence supported by LOD (logarithm of the odds ratio of linkage versus no linkage) scores between adjacent markers, coupled with distance measures between markers. Calculation of distance is based on the frequency of breakage between two markers in the radiation hybrid clones which is measured in centiRays (cR). There is a direct linear correlation between cR units and physical distance in kb, which is fairly constant across any given RH panel. The kilobase equivalent of the centiRay unit differs between RH maps. 1 cR on the TNG map corresponds to an average of 2 kb of physical distance, whereas 1 cR on the G3 map corresponds to a physical distance of 24 kb and a distance of 260 kb on the GB4 map. Table 7.2 illustrates the main features of all three panels and Table 7.3 illustrates the main RH maps generated from these panels. RH panels are available from Research Genetics (<http://www.resgen.com/>).

TABLE 7.2 Human Radiation Hybrid Panels

Panel	GeneBridge4 (GB4)	Stanford G3	Stanford TNG
X-ray dosage	3000 rad	10,000 rad	50,000 rad
Cell lines	93	83	90
Average retention	30%	18%	16%
Av. Frag. Size	10 Mb	4 Mb	800 kb
Resolution	Low	Medium	High
Resolution (kb)	1000	267	60
Reference	Gyapay <i>et al.</i> (1996)	Stewart <i>et al.</i> (1997)	Lunetta <i>et al.</i> (1996)

TABLE 7.3 Human Radiation Hybrid Maps

Map	GeneMap 99-GB4	GeneMap 99-G3	Stanford TNG	NCBI Integrated
Marker panel	GB4	G3	TNG & G3	G3 & GB4
Marker type	STS	STS	STS	STS
Marker no.	45758	7061	40322	23723
Reference	Schuler <i>et al.</i> (1996)	Deloukas <i>et al.</i> (1998)	Olivier <i>et al.</i> (2001)	Agarwala <i>et al.</i> (2000)

Comprehensive RH maps generated from these panels can be viewed and integrated with other maps in GDB, MapViewer and other applications (see below). Novel STS markers can be placed on these existing frameworks by PCR, screening STSs against the three RH panels and submitting the results to a web server, several of which are available. For G3 and TNG RH maps Stanford run a server at <http://www-shgc.stanford.edu/RH/index.html>. The EBI also runs an RH map server which includes all three human panels and also mouse, rat, pig and zebrafish panels (<http://corba.ebi.ac.uk/RHdb/RHdb.html>). The Whitehead Institute also maintains a GB4 server (<http://www-genome.wi.mit.edu/cgi-bin/contig/rhmapper>). Data submissions to all three servers are in a binary format to indicate presence or absence of a PCR product in each hybrid bin, e.g. G3.STS1 11000010000101000000110011001000001100100010000011101000110000000110.

7.4 PHYSICAL CONTIG MAPS

Genetic maps and cytogenetic maps fulfilled many of the short-term goals of the human genome project—to develop low to medium resolution genetic and physical maps of the genome. They have also facilitated longer term goals by assisting in the construction of the more-precise high resolution maps at increasingly finer resolutions needed to organize systematic sequencing efforts (Korenberg *et al.*, 1999). FISH and RH mapping in particular have enabled the development of a complex hierarchy of physical YAC and BAC clone contigs at a range of resolutions (Figure 7.1). These physical maps also became an important framework for positional cloning efforts in the years preceding the availability of a draft human genome. Accurate ordering of YAC and BAC clones (and subsequent

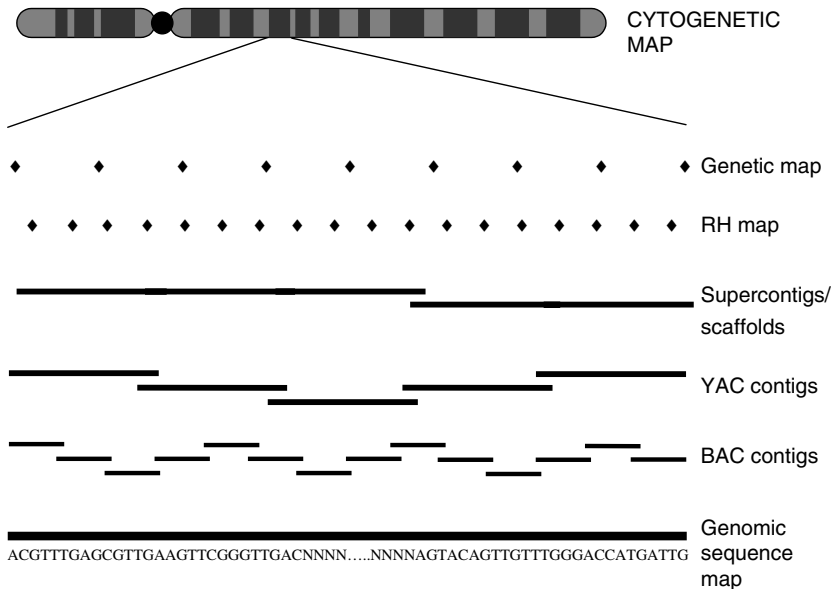


Figure 7.1 Physical and genetic maps used during the sequencing of the human genome. Many different maps were integrated to enable the construction of the framework for human genome sequencing (see Waterston *et al.* (2002) for a review).

shotgun reads) would not have been possible without existing genetic and physical maps which served as a scaffold for orientating, ordering and troubleshooting the human genome sequence assembly.

7.4.1 Yeast Artificial Chromosome (YAC) Maps

Yeast artificial chromosomes (YACs) are the lowest resolution physical clone contig maps, composed of overlapping YAC clones ranging in size from 300 kb–2 Mb. Before YACs were developed, the largest cloning vectors (cosmids) carried inserts of only 20 to 40 kb. YAC methodology drastically reduces the number of clones to be ordered; many YACs span entire human genes, making them a useful resource for further genomic study. The size of YAC inserts can often cause clone instability, which can lead to local rearrangements in the clone, this is the major drawback in the use of YACs for construction of physical contigs and underlines the need to QC YAC contigs with other available genetic and physical maps.

Several whole genome YAC maps are available, including a library of 33,000 YAC clones published by Chumakov *et al.* (1995). This library and other YAC clones can be obtained from a range of centres which are listed in the CEPH YAC library pages (http://www.cephb.fr/bio/ceph_yac.html).

7.4.2 Bacterial Artificial Chromosome (BAC) Maps

Bacterial artificial chromosomes (BACs), offer a further increase in map resolution, typically ranging in size from 100–300 kb. BAC clones are the primary vehicle of the public human genome sequencing project. Collections of human BACs estimated to represent more than a 10-fold redundancy of the human genome have been used to generate comprehensive BAC maps of the human genome. A minimally redundant set of these BACs have been assembled into physically separate contigs, representing the majority of the human genome. The sequence of these BACs is being determined by shotgun sequencing, where each BAC is digested with restriction enzymes and sub-cloned to generate a library of clones ranging from 0.5–5 kb. These clones are sequenced and assembled to form a complete BAC sequence, which are in turn assembled to form a complete chromosome.

BAC clone data can be accessed in many different ways, either directly from sequencing centres, or alternatively the NCBI have established a Human BAC resource page (<http://www.ncbi.nlm.nih.gov/genome/cyto/hbrc.shtml>). This page is a useful resource which centralizes information concerning currently available BAC maps and suppliers of BAC clones. Another useful database is GenMapDB (Morley *et al.*, 2001; <http://genomics.med.upenn.edu/genmapdb/>), which contains over 3000 mapped BAC clones spanning the genome. The database can be searched by map location or accession number. It is also possible to search for BAC clones by using BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>) to search the 'HTGS' and 'Genome' divisions of GenBank. BAC sequences can also be accessed indirectly by using tools which show contig information for the draft human genome sequence, e.g. Ensembl, Map Viewer or UCSC human genome browser (see Chapter 5).

7.5 THE ROLE OF PHYSICAL AND GENETIC MAPS IN DRAFT SEQUENCE CURATION

Sequence tagged sites (STSs) are PCR-based anchors used to define a unique genomic sequence in an RH panel, YAC or BAC contig. All that is required to generate a new STS

marker is 200–500 bp of unique sequence, this could be a sequence from the 3' UTR of a transcript or any unique genomic region. Hence STS markers have been extensively identified from characterized genes, expressed sequence tags (ESTs) and random genomic fragments (Schuler *et al.*, 1996). STS markers that include polymorphic sequences, such as microsatellites, are the central integrating force between genetic and physical maps. Common sets of such sequence-based markers can be easily screened and therefore can be used to integrate maps constructed by different mapping methods. RH panels and STS markers will play a critical role in the finishing of the human genome by providing a method to obtain markers from regions of the human genome that may be difficult to clone in conventional vector libraries. Hattori *et al.* (2000) found that up to 10% of certain gene-rich regions of human chromosome 21 were composed of such 'hard-to-clone' DNA. STSs that fail to hit available sequence can be used to screen different DNA libraries to close existing clone gaps in draft genome contigs. High resolution physical maps, such as the TNG map can also be valuable for curating draft genome contigs. Localization of RH markers to working draft sequences provides an independent measure of order and orientation for the clones underlying the draft sequence. Distances between markers can also be used to estimate the physical length of gaps between non-overlapping clones.

7.5.1 Electronic PCR (e-PCR)

Electronic PCR (e-PCR) is an *in silico* equivalent of the laboratory-based STS mapping process (Schuler, 1997; <http://www.ncbi.nlm.nih.gov/cgi-bin/STS/nph-sts>). The e-PCR tool at the NCBI maps known STSs from the dbSTS, GDB and RHdb databases to a user-submitted sequence. In a directly analogous process to PCR, e-PCR searches for sub-sequences within a query sequence that match known STS PCR primers and are in the correct order, orientation and spacing to be consistent with the PCR product size. These criteria eliminate the possibility of false positives (e.g. hits to pseudogenes or repeat sequences) that occur with other similarity searching methods such as BLAST. Electronic PCR is a valuable tool to assist in the integration of genomic sequence data with existing maps; this can be useful to assist genomic QC and to correlate genetic distances with physical distances. We offer detailed coverage of the use of this and other tools for genomic contig analysis in Chapter 9.

RH maps are playing a critical role in the QC and finishing of the human genome (see below), but once we have a finished genome, these maps may be of limited further use in humans. However, RH maps will continue to be the physical mapping method of choice for other organisms without extensive genome sequence. Human RH maps may also be of some limited use in the construction of comparative maps with other mammalian genomes (Kwitek *et al.*, 2001). But, for the purposes of this chapter, we will focus on the direct integration of genetic and physical maps, with genome sequence as the ultimate integration framework. For consideration of non-human maps we refer the reader to Chapter 6 and other specialist texts.

7.6 THE HUMAN GENOME SEQUENCE – THE ULTIMATE PHYSICAL MAP?

The complete DNA sequence of the human genome will be an accurate physical map resolved down to a single base pair resolution, but we do not have this map as yet. Geneticists will need to work with a draft assembly of the human genome for a somewhat

indeterminate number of years, until this task is truly finished. However, the draft genome assembly is still a very valuable asset for genetics, particularly if data are treated with care. With this in mind it is very important to be aware of some of the issues relating to the curation of draft sequences. Genetic and physical maps are one aid in this process.

For example, Olivier *et al.* (2001) used a 40,000-marker RH map to provide an estimate of the size and location of missing sequence in the human genome draft in relation to the existing sequence, and to provide order information for the 15,000 + clones that constitute the human genome working draft. They found that 9.8% of STS markers were absent from the October 2000 draft of the human genome. They suggest that these are likely to represent the ‘hard-to-clone’ regions of the human genome. Other studies have made similar observations (Hattori *et al.*, 2000) which suggests that a small intractable percentage of the human genome may remain in an unfinished state for longer than we may have anticipated.

Genetic maps are also playing an important role in the QC of human genomic sequence. DeWan *et al.* (2002) compared the genetic order of the Marshfield genome-scan markers (set 9 and 10) with their physical order in the April 2001 public golden path contig and the February 2001 Celera genome assembly. They found inconsistencies in 5 and 2% of the markers in the Celera assembly and the golden path assembly, respectively. The genetic order of these markers was supported with high confidence by a LOD of >3 and most discrepancies were not observed in both contigs, which suggests errors in the physical map order of both genome assemblies. Chromosome-by-chromosome breakdown of this data are available on a website: <http://linkage.rockefeller.edu/maps/>.

7.7 QC OF GENOMIC DNA – RESOLUTION OF MARKER ORDER AND GAP SIZES

The studies by DeWan *et al.* (2002) and Olivier *et al.* (2001) demonstrate the value of genetic and physical maps in the curation and QC of human genomic sequence contigs. As discussed previously the relationship between genetic distance (cM) and physical distance (Mb) is not uniform, however both genetic and physical mapping methods can resolve marker order to varying degrees of confidence, depending on the map characteristics. Marker or contig order across a locus can be validated by integrating information from different maps. The value and accuracy of different maps is not necessarily hierarchical or directly related to the density of the map. For example, one might assume that a dense RH map, or even a finished sequence map might be more accurate than a less dense genetic map, however as the studies above have shown, this does not always hold true. Using maps in an hierarchical manner may avoid the inevitable discordances between different maps, but this is not necessarily the best order for integration. In some cases for example, YAC STS content data may be more accurate than RH data, or a genetic map may be more reliable than a BAC contig. Both genetic maps and RH maps show a relative confidence in marker order by LOD scores using appropriate maximum likelihood statistical methods (Boehnke *et al.*, 1991). A LOD of 3.0 (odds 1000:1) or more is generally accepted as a strong indication of a contiguous relationship between markers. Comparison of LOD scores can help to integrate different data sources in an attempt to reach a consensus. But sometimes, all that can be done *in silico* is to flag up an unresolvable discrepancy between maps to receive special attention in the laboratory. Bioinformatic tools and databases can be a great help in the integration and evaluation of genetic and physical maps. MapViewer, GDB and UDB allow the user to compare and integrate maps; these are described below.

The UCSC human genome browser also provides graphical data on the positions of STS markers on the golden path versus their positions in other maps, including radiation hybrid, sex-averaged genetic (Marshfield), cytogenetic and YAC STS maps at the following URL: <http://genome.ucsc.edu/goldenPath/mapPlots/>.

It is also possible to view and integrate genetic and physical maps on an *ad hoc* basis using bioinformatics tools, such as Map View at the NCBI (reviewed below).

7.8 TOOLS AND DATABASES FOR MAP ANALYSIS AND INTEGRATION

There are some excellent tools which have recently become available for viewing the human genome sequence. Ensembl and the UCSC human genome browser are shining examples of the kind of biological data integration that geneticists need for their studies. But unfortunately they are lacking in functionality to enable map integration. Other more specialized tools, such as GDB and UDB exist, which allow a user to view and integrate different maps, but unfortunately these have not generally been integrated with the human genome sequence. Fortunately Entrez Map View at the NCBI, is one tool which straddles both the human genome and human genetic maps.

7.8.1 Entrez Map View

(http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search)

In Chapter 5 we reviewed Map View alongside Ensembl and the UCSC human genome browser as a tool for annotation of the human genome sequence. Map View would probably appear lower in most researchers' preferences for this purpose, although it does provide some unique gene annotation information. However, as the name suggests, Map View truly excels in its integration of a wide range of cytogenetic, genetic and physical maps with the NCBI draft and finished sequence contigs. Although this tool is sometimes a little difficult to navigate, once these idiosyncrasies are overcome, Map View becomes a complex and powerful tool.

Map View is an integrated component of the NCBI Entrez system, in the Entrez Genomes division. This division presents a unified graphical view of genetic and physical maps (including sequence maps) for over four vertebrates, including human and mouse. The tools present different genomes at four levels of detail:

- Organism home page — summarizing the resources available for that organism
- Genome View — graphical display of chromosome ideograms and search page
- Map View — presents one or more maps aligned against a master map
- Sequence View — graphically annotates the biological features in a region

7.8.1.1 Searching and Browsing Map View

Map View can be searched with almost any marker, SNP, gene or genomic element either targeted at a chromosome or genome level. Searches at the genome level return a graphic view of the location of the hit with red marks on the chromosome ideograms, this will quickly identify if a query hits multiple regions or chromosomes. A summary of the maps in which the query exists is returned in tabular format at the bottom of the page. This is the essence of the Map View tool — selection of a map from the tabular summary links to a detailed Map View of the corresponding genomic region, with the selected map as the

'master' map. The master map is presented in detail with supporting information, such as LOD scores, cM locations or gene information. To view and integrate the master map with other maps, select the 'maps & options' link at the top of the page. This will summon a pop-up window for Map View configuration. It is possible to select up to eight maps to view alongside the master map, each is presented in a compact view alongside the master map. The alignment between maps is based on common or corresponding objects. Markers or objects shared between maps are indicated by lines connecting the maps. Map View allows the user to zoom and pan into progressively more detailed views.

It is also possible to search and browse Map View by map position or cytoband. This can be achieved from the Map View of a chromosome, by entering a range of interest in the boxes in the side window. A range can be specified in base pairs, cytogenetic bands or between two gene symbols. General chromosomal browsing is possible by clicking on the region of interest in the chromosome thumbnail graphic in the sidebar, or by clicking on a region of interest on the ideograms in the genome view.

Map View is very effective for integration of genetic and physical maps on an *ad hoc* basis. Figure 7.2 shows an integrated view of the Genethon genetic map and the human genome contig for chromosome 3. In this map, the Genethon markers are mapped to sequence and a line is drawn between the marker positions on the two maps. This clearly illustrates some key map integration issues. Firstly several markers in the genetic map are seen to conflict with the order of markers on the sequence (or physical) map. This may be due to an error in either map, so further maps need to be compared to support either order and the LOD scores on the genetic map need to be examined. Figure 7.3 shows such a comparison. The red line traces the Genethon marker, AFMA121WD5, through the Marshfield, GB4, G3 and TNG maps through to the genomic contig level. In this case the marker order is confirmed by each map. Sometimes it may not be possible to conclusively determine which map is 'right', instead further laboratory work may be necessary to resolve marker order. Figure 7.2 also clearly shows the variable relationship between genetic and physical distance. In particular it highlights some of the physical properties of chromosomes, for example the genetic physical distance ratio at the telomere of the P arm of chromosome 3 is very low; the marker AFM234TF4, for example, has a genetic location of 22 cM and physical location of 8 Mb. This illustrates the higher rates of recombination that are often observed in telomere regions (Riethman, 1997). Both figures indicate the presence or absence of each marker in available maps by an array of symbolic green circles at the far right of each marker. This helps to indicate non-specific markers. For example, some markers map to multiple locations in the same chromosome, these are indicated by green circles with a strike through. Other markers map to more than one chromosome, these are indicated by yellow circles and finally some markers, map to multiple chromosomes and multiple locations, indicated by a yellow struck through circles (e.g. AFMA191ZG5 in Figure 7.2).

There are a number of somewhat idiosyncratic features in Map View which might confuse the user. Firstly if a map is viewed in low resolution, it seems to display a somewhat arbitrary selection of markers, the full marker set only becomes visible when the user zooms in. Secondly, if the locus is too large to view in one window it is broken up into pages indicated at the top of the window. This pagination feature can make it slow and difficult to assess a whole locus, but this can be overridden by altering the page size in the configuration window. Setting a page size of 100–200 will allow a very large map to be viewed in a single window, this may take some time to load but it is worth it in the end.

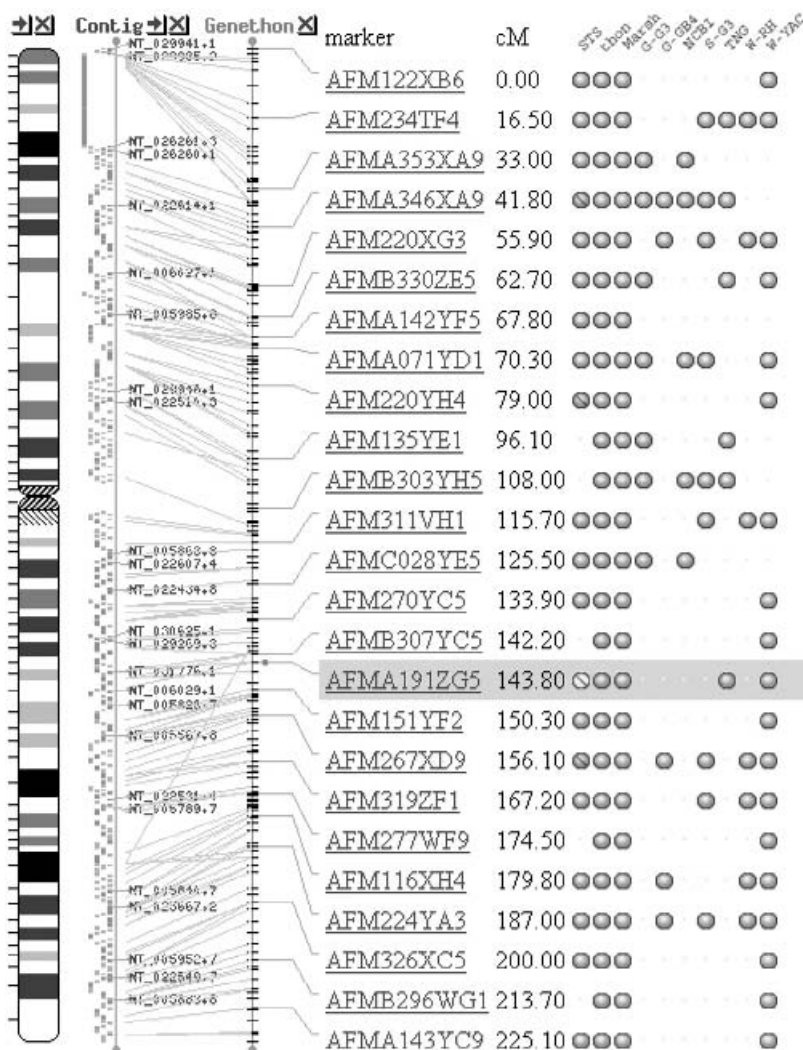


Figure 7.2 Integration of genetic maps and genome contigs. This figure shows an integrated view of the Genethon genetic map and the human genome contig for chromosome 3 generated by the NCBI Map View tool. The Genethon markers are mapped to sequence with a line drawn between the marker positions on the two maps. Lines which cross over show markers which conflict in order between the genetic map and the physical sequence map.

7.8.2 The Genome Database (GDB) (www.gdb.org)

The Genome Database (GDB) was the first web-based graphical interface to the human genome, as such it was a pioneering bioinformatic tool. Now Ensembl, UCSC and Map View present effortless graphical genome views and the GDB graphical interface is starting

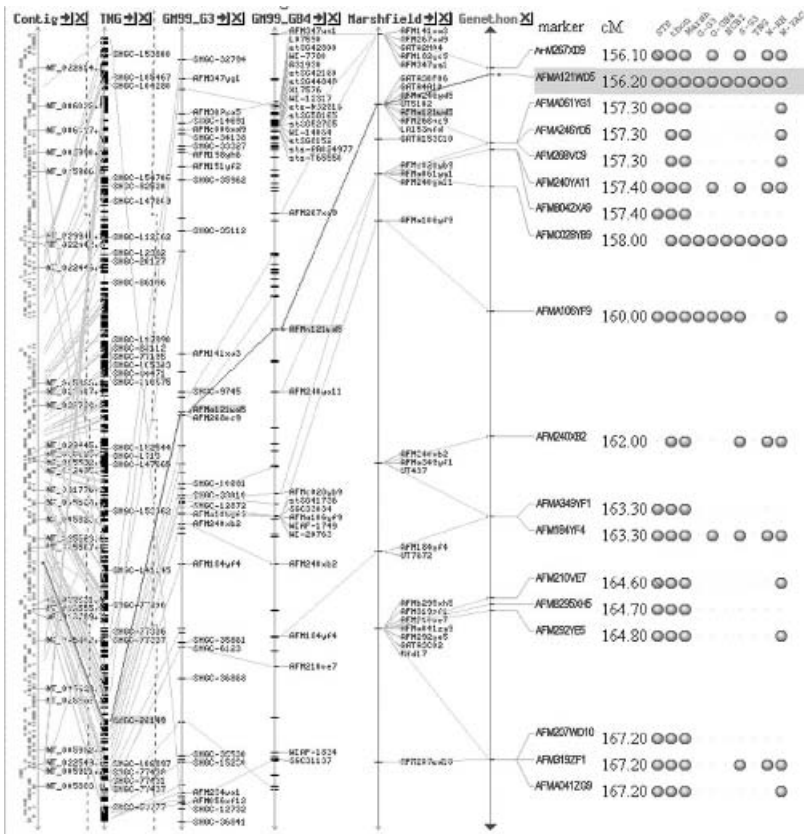


Figure 7.3 Integration of genetic and physical maps with the human genome contig on chromosome 3 using the NCBI Map View tool. The grey line traces the Genethon marker, AFMA121WD5, through though the Marshfield, GB4, G3 and TNG maps through to the genomic contig level. In this case the marker order of this marker is confirmed by each map.

to look a little tired and most of the graphical functionality is covered by Map View. But GDB does have a productive text/table-based search interface which is an improvement on Map View’s limited text-based capability. GDB is also a comprehensive source for some forms of genetic data, particularly tandem repeat polymorphisms (it contains over 18,000), and an eclectic range of information on fragile sites, deletions, disease genes and mutations, collected by a mixture of curation and direct submission. This makes GDB a valuable tool for text-based data mining to assist in the construction of marker lists and the identification of marker variables, such as primer and marker sequences.

The text-based search interface is accessible on the front page of the GDB database by following the ‘advanced search’ link. This interface allows complex queries, for example, it is possible to retrieve all known polymorphic or non-polymorphic markers between two markers or genes. Results are retrieved and ordered based on the genetic distances of the markers, along with a very roughly estimated Mb location (unfortunately actual

integration with the human genome draft is currently lacking). As the markers are ordered by genetic distance, the distances are very approximate, with no fine measure of distance or order. It may be necessary to clarify the order with another tool such as Map View.

7.8.3 The Unified Database for Human Genome Mapping (UDB) (<http://bioinformatics.weizmann.ac.il/udb/>)

The Unified Database for Human Genome Mapping (UDB) is maintained by the Weizmann Institute of Science, Israel. UDB has attempted to create an integrated map based on a diverse range of human genome mapping data retrieved from a number of public databases. The map consists of an integrated hierarchy of genetic, RH, cDNA and YAC maps down to a kilobase resolution, on a scale converted from centiRays (cR) to megabases (Mb). UDB generates its maps using data from the Whitehead/MIT STS map, GeneMap'98, the Stanford TNG map and Genethon maps. The database can be searched in several different ways. An initial search by chromosome number can be narrowed by specification of cytogenetic band, position (in Mb) or marker interval. It is also possible to search by gene or marker name. This gives the estimated location of the gene as well as links to GeneCards and the Genome Database (GDB). The database also displays the estimated boundaries (in Mb) of the cytogenetic bands of any chromosome.

The UDB database is a good starting point for constructing physical or transcript maps across a genomic region. The main benefit of the database is that it eliminates the need to look at a number of different websites and integrates markers from several different maps with genomic contigs from NCBI. Unfortunately UDB is somewhat over zealous in its map integration, sometimes this might cause problems. It assumes a hierarchical value of RH maps over genetic maps and genetic maps over YAC maps which is not necessarily the best order for integration, it may have been better to flag conflicting marker orders for laboratory-based resolution. However as the human genome map solidifies around finished sequence this approach will begin to represent the simplest and most effective use of time and resources.

7.9 CONCLUSIONS

As this chapter has described, there are many tools available to give an integrated view of genetic and physical maps across a defined chromosomal locus. Comparison of the physical and genetic distances between markers can provide a great deal of information about the underlying nature of a locus. Yu *et al.* (2001) compared the genetic and physical distances across the whole genome and found that the genetic/physical distance ratio ranged widely between 0 and 9 cM per Mb. They used this ratio to infer recombination rates and identified several chromosomal regions up to 6 Mb in length with very low or high recombination rates, which they termed recombination 'deserts' and 'jungles', respectively. Linkage disequilibrium (LD) was much more extended in the deserts than in the jungles as higher rates of recombination are likely to reduce the extent of LD.

When sequencing of the human genome is truly complete genetics will become technically much easier. Human map QC may become a distant memory, but presently we are still struggling to study complex phenotypes with draft contigs and incomplete datasets. Every piece of data and data curation may count in this struggle—in Section III of this book we review how physical and genetic map data can come together with literature data, marker data, gene data and comparative organism data to assist genetic studies in the laboratory.

REFERENCES

- Agarwala R, Applegate DL, Maglott D, Schuler GD, Schaffer AA. (2000). A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res* **10**: 350–364.
- Boehnke M, Lange K, Cox DR. (1991). Statistical methods for multipoint radiation hybrid mapping. *Am J Hum Genet* **49**: 1174–1188.
- Broman KW, Murray JC, Sheffield VC, White RL, Weber JL. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet* **63**: 861–869.
- Chumakov IM, Rigault P, Le Gall I, Bellanne-Chantelot C, Billault A, Guillou S, *et al.* (1995). A YAC contig map of the human genome. *Nature* **377**(Suppl.): 175–297.
- Cox DR, Burmeister M, Price ER, Kim S, Myers RM. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**: 245–250.
- Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, *et al.* (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* (in press).
- Deloukas P, Schuler GD, Gyapay G, Beasley EM, Soderlund C, Rodriguez-Tome P, *et al.* (1998). A physical map of 30,000 human genes. *Science* **282**: 744–746.
- DeWan AT, Parrado AR, Matisse TC, Leal SM. (2002). The map problem: a comparison of genetic and sequence-based physical maps. *Am J Hum Genet* **70**: 101–107.
- Dib C, Faure S, Fizames C, Sampson D, Drouot N, Vignal A, *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152–154.
- Eisenbarth I, Vogel G, Krone W, Vogel W, Assum G. (2000). An isochores transition in the NF1 gene region coincides with a switch in the extent of linkage disequilibrium. *Am J Hum Genet* **67**: 873–880.
- Goddard KA, Wijsman EM. (2002). Characteristics of genetic markers and maps for cost-effective genome screens using diallelic markers. *Genet Epidemiol* **22**: 205–220.
- Goss SJ, Harris H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**: 680–684.
- Gyapay G, Schmitt K, Fizames C, Jones H, Vega-Czarny N, Spillett D, *et al.* (1996). A radiation hybrid map of the human genome. *Hum Mol Genet* **5**: 339–346.
- Hattori M, Fujiyama A, Taylor TD, Watanabe H, Yada T, Park HS, *et al.* (2000). The DNA sequence of human chromosome 21. *Nature* **405**, 311–319.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, *et al.* (2001). Haplotype tagging for the identification of common disease genes. *Nature Genet* **29**: 233–237.
- Korenberg JR, Chen XN, Sun Z, Shi ZY, Ma S, Vataru E, *et al.* (1999). Human genome anatomy: BACs integrating the genetic and cytogenetic maps for bridging genome and biomedicine. *Genome Res* **9**: 994–1001.
- Kruglyak L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genet* **17**: 21–24.
- Kwitek AE, Tonellato PJ, Chen D, Gullings-Handley J, Cheng YS, Twigger S, *et al.* (2001). Automated construction of high-density comparative maps between rat, human, and mouse. *Genome Res* **11**: 1935–1943.
- Lander ES, Green P, Abrahamson P, Barlow A, Daly MJ, Lincoln SE, *et al.* (1987). MAP-MAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**: 174–181.

- Lunetta KL, Boehnke M, Lange K, Cox DR. (1996). Selected locus and multiple panel models for radiation hybrid mapping. *Am J Hum Genet* **59**: 717–725.
- Lynn A, Kashuk C, Petersen MB, Bailey JA, Cox DR, Antonarakis SE, *et al.* (2000). Patterns of meiotic recombination on the long arm of human chromosome 21. *Genome Res* **10**: 1319–1332.
- Morley M, Arcaro M, Burdick J, Yonescu R, Reid T, Kirsch I, *et al.* (2001). GenMapDB: a database of mapped human BAC clones. *Nucleic Acids Res* **29**: 144–147.
- Morton NE. (1991). Parameters of the human genome. *Proc Natl Acad Sci USA* **88**: 7474–7476.
- Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Manion F, *et al.* (1994). A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC). *Science* **265**: 2049–2054.
- Niimura Y, Gojobori T. (2002). *In silico* chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence. *Proc Natl Acad Sci USA* **99**: 797–802.
- Olivier M, Aggarwal A, Allen J, Almendras AA, Bajorek ES, Beasley EM, *et al.* (2001). A high-resolution radiation hybrid map of the human genome draft sequence, *Science* **291**: 1298–1302.
- Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, *et al.* (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719–1723.
- Riethman H. (1997). Closing in on Telomeric Closure. *Genome Res* **7**: 853–855.
- Schuler GD. (1997). Sequence mapping by electronic PCR. *Genome Res* **7**: 541–550.
- Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, *et al.* (1996). A gene map of the human genome. *Science* **274**: 540–546.
- Stewart EA, McKusick KB, Aggarwal A, Bajorek E, Brady S, Chu A, *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Res* **7**: 422–433.
- Waterston RH, Lander ES, Sulston JE. (2002). On the sequencing of the human genome. *Proc Natl Acad Sci USA* **99**: 3712–3716.
- Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, *et al.* (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409**: 951–953.