BIOINFORMATICS

Vol. 17 no. 7 2001
Pages 581–586

LDB2000: sequence-based integrated maps of the human genome

Xiayi Ke, William Tapper and Andrew Collins

Human Genetics Research Division, University of Southampton, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK

Received on January 5, 2001; revised on March 20, 2001; accepted on March 23, 2001

ABSTRACT

Motivation: Integrated maps are useful for gene mapping and establishing the relationship between recombination and sequence. In this paper we describe algorithms and their implementation for constructing sequence-based integrated maps of the human chromosomes, which are presented in LDB2000, a web based resource. Gene mapping efforts are now focussing on linkage disequilibrium mapping and extension of the integrated map to represent the extent of linkage disequilibrium in different genomic regions would further increase the utility of these maps.

Results: Sequence-based integrated maps have been completed for chromosomes 21 and 22. These maps provide locations for genes and polymorphic markers in sequence and on genetic linkage, radiation hybrid and cytogenetic scales. Single nucleotide polymorphisms associated with genes in the maps are also included and their sequence locations indicated. Related locus information, such as aliases and expression information, can be searched on the WWW site.

Availability: http://cedar.genetics.soton.ac.uk/public_html/LDB2000.html

Contact: arc@soton.ac.uk

INTRODUCTION

The draft human genome sequence permits sequence-based map integration. For chromosomes with a covering sequence and well characterized gaps, we have, for the first time, confidence in the order of map objects (loci). The Location DataBase (LDB), (http://cedar.genetics.soton.ac.uk/public_html/; Morton *et al.*, 1992; Collins *et al.*, 1996a) has maintained integrated maps of the human genome for several years but has depended upon construction of approximate physical locations for loci based on evidence from alternative maps (genetic linkage, radiation hybrid, cytogenetic, YAC-based). As the sequencing nears completion sequence-based physical maps are becoming available. Establishing the relationship between the genetic linkage map and the sequence-based physical map is important for gene mapping studies and for the study of recombination generally. The first analyses of the sequence and genetic maps are starting to appear (for example, Yu *et al.*, 2001) although there has been little effort so far to construct high-resolution genetic maps constrained to the order determined in sequence. The development of a new sequence-based resource (LDB2000) has initially focused on integrated sequence-based maps of chromosomes 21 and 22. These maps are available at http://cedar.genetics.soton.ac.uk/public_html/LDB2000.html and are based on the recently released near-complete sequences of chromosomes 21 (Hattori *et al.*, 2000) and 22 (Dunham *et al.*, 1999).

Draft sequences of the entire human genome have now been published (Lander *et al.*, 2001; Venter *et al.*, 2001). Although not 'finished' sequences like those of chromosomes 21 and 22 the draft sequence undoubtedly offers the most reliable physical locations to date. For this reason we are using draft sequence to construct integrated maps of polymorphic markers to support our gene mapping and linkage disequilibrium studies. These maps will be made available as they are completed. More inclusive integrated maps will be developed as more finished sequence becomes available.

Integrated maps are of considerable value in several areas of active research. For the mapping of genes for common diseases precise maps are required, especially for multilocus approaches to linkage and Linkage Disequilibrium (LD) mapping with Single Nucleotide Polymorphisms (SNPs) (Collins *et al.*, 1999). The presence of recombination hot and cold spots influences the density of SNPs required for adequate coverage of a candidate region. A second area of research is the study of the relationship between the sequence and genetic recombination, specifically the determination of sequence motifs that influence the level of recombination in different chromosome regions and on different chromosomes. Also of interest are cytogenetic bands and their properties and the relationship of the sequence to radiation breakage

(Holmquist, 1992). We describe here the approach and algorithms used and applied in LDB2000 to construct sequence-based integrated maps.

SYSTEM AND METHODS

The software for automated data capture from online databases is written in Java 2 SDK and the software for map integration is written in C. Common Gateway Interface (CGI) programs for database searching from the WWW interface are also written in C, for efficient processing of queries. The map database is based on flat files with a WWW interface. The emphasis of LDB2000 is to represent location, particularly for genes and polymorphisms. It is not our intent to develop high resolution sequence annotation as presented at other sites (for example, Ensembl, http://www.ensembl.org/). Loci not already localized in the maps can be located using a CGI program which determines a map position using BLAST (Altschul *et al.*, 1990). Larger chromosomes may be presented in the form of well defined regions based, for example, on low resolution cytogenetic bands. LDB2000 has been developed under UNIX on a SUN Enterprise server.

ALGORITHM AND IMPLEMENTATION

In the integrated (summary) map all loci are represented by their approximate midpoints in the sequence relative to an origin at the p-telomere. At higher resolution intragenic polymorphisms can be accessed through hypertext links and a search interface. The loci represented in the map comprise genes, EST clusters representing genes, genes predicted from sequence, polymorphisms, and sequence tagged sites. An effort to cross-reference loci in the online databases was undertaken, facilitated by the development of a database of 'clusters' which define synonyms and associate many (intragenic) polymorphisms with expressed sequences. This database was constructed from the genome database, GDB (http://gdbwww.gdb.org/), UniGene (http://www.ncbi.nlm.nih.gov/UniGene/) and Locuslink (http://www.ncbi.nlm.nih.gov/LocusLink/) using a Java-based web page reader. Names for genes appearing in the maps representing a cluster (primary names) are selected following the priority: HUGO Gene Nomenclature Committee (http://www.gene.ucl.ac.uk/nomenclature/) > UniGene > GDB. Existing D-numbers are used for polymorphisms not located within genes and GenBank accession numbers for other DNA sequences including Sequence-Tagged Sites (STSs).

Figure 1 illustrates the procedure for developing the sequence-based integrated summary maps together with gene and SNP specific maps. The starting point for sequence-based map construction is the compilation of a comprehensive set of symbols starting from the existing LDB summary map. Other loci are then added from the Genemap99 radiation hybrid map (http://www.ncbi.nlm.nih.gov/genemap/) and GDB together with those determined through sequence analysis of individual chromosomes, for example by the Sanger Centre (http://www.sanger.ac.uk/). Once a list of non-redundant symbols has been obtained the list is submitted to GDB, UniGene, and LocusLink to extract related information associated with each locus, including synonyms and aliases, genetic and cytogenetic locations, primer sequences, GenBank and RefSeq sequences, and expression information. All loci are identified, where possible, in sequence through local BLAST and/or e-PCR (Schuler, 1997) searches to generate an ordered table of midpoint locations representing distance from the p-telomere in kilobases. Uncertainty about the size of unsequenced heterochromatic regions of the genome, for example 22p, means that we currently retain previous estimates (Morton, 1991) and have added a corresponding offset to all locations. An 'alias-map', containing all symbols and available location information, is constructed based on these Internet searches and the e-PCR and BLAST results. The symbols are then cross-referenced to construct an 'alias–cluster map' and the location variances within each data type are calculated. Clusters with high variances are investigated to determine the source of the errors (typically these are nomenclature problems) and incorrect clusters, which do not represent a single expressed sequence or set of synonyms, are prevented from being formed. Based on our experience we will in future adopt a stringent criteria for accepting clusters. This will presumably lead to some genes being represented more than once but this would be a small penalty for complete automation and more frequent revision. From the alias–cluster map, an intermediate summary map is built by selecting the primary name and its location information from each cluster. An information file is also constructed that contains the alias and expression information, and links to nucleic acid sequences, and this supporting information is made available through the web interface.

For our purposes the reliability of the genetic map and its relationship to sequence is of greatest importance. We have, therefore, rebuilt the genetic linkage maps, constrained to the sequence-based order, incorporating pairwise lods from CEPH version 9 (http://www.cephb.fr/cephdb) and lods from Genatlas (http://bisance.citi2.fr/GENATLAS). Genetic maps were constructed using the multiple pairwise approach implemented in **map+** (Collins *et al.*, 1996b) which allows map construction under an estimable level of chiasma interference and with typing error filtration (Shields *et al.*, 1991).

We did not rebuild the radiation hybrid maps, feeling that these were of more marginal interest that would not

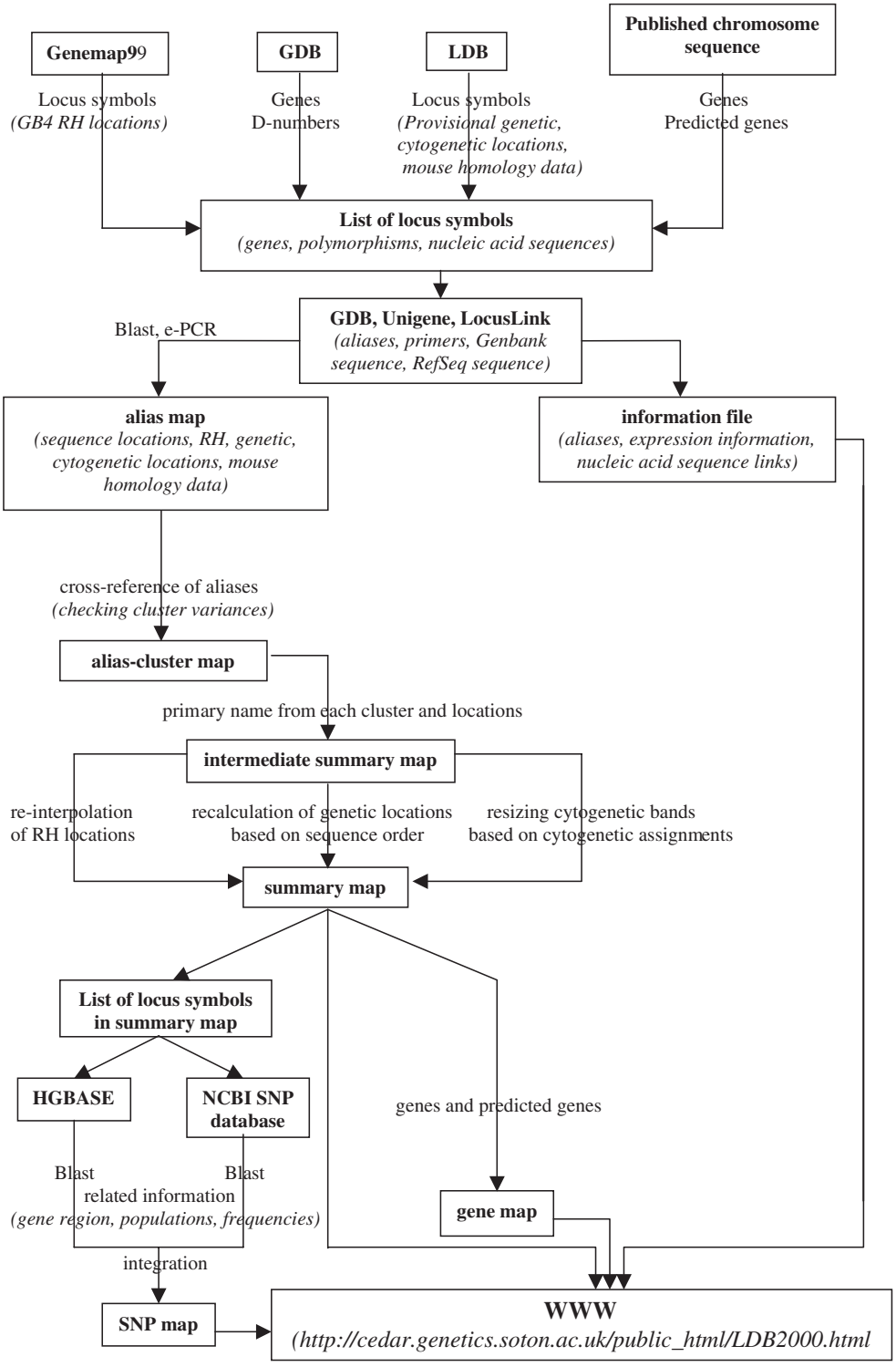

Fig. 1. Sequence-based map integration.

extend much beyond the end of the sequencing effort although they are clearly useful in evaluating draft sequence (Olivier *et al.*, 2001). However, we applied a rule-based ranking algorithm to resolve order discrepancies between the Genemap99 radiation hybrid map and the sequence order so that each locus is assigned an inferred RH location.

The procedure re-calculates locations for the loci with the most discrepant order positions by interpolation between flanking loci that are correctly ordered according to the sequence. The method for interpolation follows Morton *et al.* (1992). By ranking loci according to the sequence order the loci with the most discrepant positions have their locations re-computed first, followed by re-ranking and the resolution of the next most discrepant loci, and so on.

We have re-estimated cytogenetic band border locations so that bands correspond more closely with observed cytogenetic assignments and we have assigned a cytogenetic location to every map object. The algorithm iteratively repositions a cytogenetic band border to minimize the number of cytogenetic assignment errors. This procedure is conservative being weighted to discount loci with errors that are far from the border, since these are largely errors in their cytogenetic assignment rather than errors in border position (Tapper *et al.*, 2001). The summary map gives locations, given the sequence-based order, for all loci on the genetic scale (MCM, FCM—male and female centimorgans), radiation hybrid scale in centirays (RHCR, cR_{3000}) and cytogenetic band (BD). Corresponding original map locations are also given (mcm, fcm, rhcr and lbd and rbd for left and right cytogenetic bands).

A gene map is derived from the summary map comprising the subset of loci that are genes or predicted genes. A SNP map is constructed by submitting the mapped loci to both the NCBI SNP database and HGBASE database (http://www.hgbase.de/). The NCBI SNP database contains SNPs submitted or deposited by various contributors, such as the SNP Consortium Ltd (http://snp.cshl.org/) and laboratories associated with the National Human Genome Research Institute (NHGRI) (http://www.nhgri.nih.gov/) grants program. From the NCBI and HGBASE SNP databases, information on all SNPs associated with each locus is extracted, including primer sequences and upstream/downstream sequences. These sequences are used in a BLAST search against the chromosome sequence to obtain the individual SNP sequence locations. Intermediate SNP maps, one from each of the two database searches, are constructed together with information on populations, and allele frequency. These two maps are combined into a single SNP map and the SNP entries belonging to each locus are grouped in order according to their sequence locations. Those SNP entries associated with a particular locus, as indicated by the two databases, but having weak or multiple BLAST results or located outside the locus boundary according the BLAST results are retained pending further investigation and revisions.

The integrated summary map, gene map and SNP map are represented at the LDB2000 WWW site. Selecting any locus in the maps launches a CGI program which reports other objects in the cluster, details of intra-locus polymorphisms, links to GenBank and NCBI together with the location of any single nucleotide polymorphism and their primers derived from NCBI SNP database and HGBASE. The integrated maps can also be sampled through a search facility offering various regional views of the map, such as a region between two loci, a region between two sequence locations, a region around a locus and electronic PCR for pairs of primers.

This text-based (rather than graphical) representation of the integrated maps is designed to deliver maximum information content in the form of maps for multilocus disease gene mapping. The integrated maps are useful in the study of recombination and the role of aberrant recombination in disease, examination of cytogenetic band properties and the relationship between sequence, recombination, radiation breakage and other processes.

RESULTS AND DISCUSSION

Integrated maps have been completed for chromosomes 21 and 22. Table 1 lists the loci represented in the integrated maps for the two chromosomes. Arguments about the total number of genes are likely to continue for some time, since, even with near complete sequence, gene finding and genome annotation remains an inexact process. Chromosome 21 is known to be relatively gene poor compared with chromosome 22 and it is possible that our approach has minimized the difference between the two, with some of the predicted genes for 21 representing multiple clusters derived from the same gene and perhaps some of the chromosome 22 clusters being composites. Clearly ongoing annotation and revision of the maps is required and this can be achieved by automation of data retrieval and synthesis. A future target is for automated updating of the maps as genome annotation progresses. Starting from sequence accession numbers and/or primers it would not be difficult for a remote user to determine locations for newly characterized genes and markers and recover an updated map of a candidate disease gene region through a WWW query interface launching a BLAST search of the underlying sequence. In this way more complete annotation of a chromosome region can be achieved.

The general relationship between the genetic and physical scales for the two chromosomes is now established (Figures 2 and 3). The pattern of recombination is rather different with chromosome 22 showing more recombination hot-spots. The tendency for increased male recombination in the sub-telomeric region is evident for both chromosomes. Recombination hot and cold spots may be related to distinct sequence motifs and certain tandem repeat sequences, particularly GT/CA tracts, seem to be important (Majewski and Ott, 2000), especially in male meiosis (Tapper *et al.*, 2001). Analysis of sequence-based maps of more of the genome will be of considerable interest to confirm or refute these findings.

Most of the effort in disease gene mapping to date

Table 1. Classification of loci in the integrated maps of Chromosomes 21 and 22

Type	Class	Description	Chromosome 21	Chromosome 22	Symbol
1	G	Gene: transcribed region of the genome	214	252	HUGO > UniGene > GDB
1	g	Predicted gene based on UniGene clusters or sequence analysis	146	264	GenBank accession no.
2	P	Polymorphic: multiple polymorphisms*	6	15	D-number
2	V	Variable Number of Tandem Repeats (VNTR)	1	2	
2	Q	Tetra-nucleotide repeat	32	20	
2	T	Tri-nucleotide repeat	7	7	
2	D	Di-nucleotide repeat	137	83	
2	R	Restriction Fragment Length Polymorphism (RFLP)	0	1	
2	U	Unknown polymorphism	3	7	
3	N	Nucleic acid**	523	317	GenBank accession no.
Total			1069	968	

Type: 1: genes, 2: polymorphisms, 3: DNA sequence.
*Multiple polymorphisms with the same D-number.
**Nucleic acid sequences which may include ESTs not yet associated with predicted genes and sequence tagged sites.

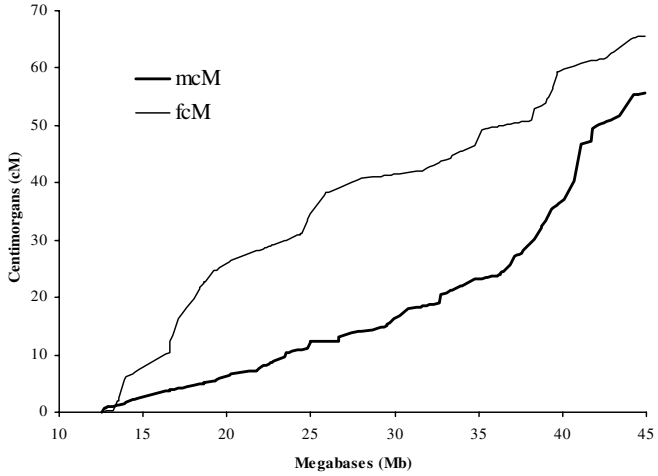

Fig. 2. Relationship between the genetic and physical scales of chromosome 21.

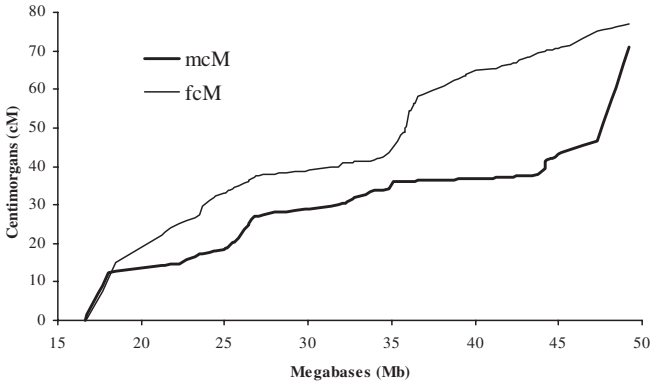

Fig. 3. Relationship between the genetic and physical scales of chromosome 22.

has been in linkage mapping but most successes have mostly come from the mapping of major genes, avoiding the complexities posed by genes for common diseases. With the attention now focussed on the latter much higher density maps of polymorphisms, particularly of SNPs, are being developed in the hope that linkage disequilibrium mapping will be more successful. Linkage disequilibrium appears, on average, to extend to several hundred kilobases in real data (Collins *et al.*, 1999), despite simulations which suggest it is far less extensive (Kruglyak, 1999a). However, it is evident that the extent of LD is highly variable and depends on the chromosome region, the population being considered and the type and age of polymorphisms. To exploit LD mapping it will be useful to establish an LD map of the genome representing regions with high and low LD, relative to the sequence map, in different populations (Kruglyak, 1999b). This has an important bearing on, for example, the density of polymorphisms required to adequately cover a region. A fairly close relationship to the linkage map is possible but not guaranteed given the influence on LD of stochastic factors over many generations. A natural extension of sequence-based map integration is to include the linkage disequilibrium map and this should improve the prospects for localization of these genes which have proven so difficult to identify to date.

REFERENCES

Altschul,S.F., Gish,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Collins,A., Frezal,J. *et al.* (1996a) A metric map of humans: 23 500

loci in 850 bands. *Proc. Natl Acad. Sci. USA*, **93**, 14 771–14 775.
Collins,A., Teague,J. *et al.* (1996b) Linkage map integration. *Genomics*, **36**, 157–162.
Collins,A., Lonjou,C. *et al.* (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA*, **96**, 15 173–15 177.
Dunham,I., Shimizu,N. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature*, **402**, 489–495.
Hattori,M., Fujiyama,A. *et al.* (2000) The DNA sequence of human chromosome 21. *Nature*, **405**, 311–319.
Holmquist,G.P. (1992) Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.*, **51**, 17–37.
Kruglyak,L. (1999a) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
Kruglyak,L. (1999b) Genetic isolates: seperate but equal? *Proc. Natl Acad. Sci. USA*, **96**, 1170–1172.
Lander,E.S., Linton,L.M. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
Majewski,J. and Ott,J. (2000) GT repeats are associated with recombination on human chromosome 22. *Genome Res.*, **10**, 1108–1114.
Morton,N.E. (1991) Parameters of the human genome. *Proc. Natl Acad. Sci. USA*, **88**, 7474–7476.
Morton,N.E., Collins,A. *et al.* (1992) Algorithms for a location database. *Ann. Hum. Genet.*, **56**, 223–232.
Olivier,M., Aggarwal,A. *et al.* (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science*, **291**, 1298.
Schuler,G.D. (1997) Sequence mapping by electronic PCR. *Genome Res.*, **7**, 541–550.
Shields,D.C., Collins,A. *et al.* (1991) Error filtration, interference and the human linkage map. *Proc. Natl Acad. Sci. USA*, **88**, 6501–6505.
Tapper,W.J., Morton,N.E. *et al.* (2001) A sequence-based integrated map of chromosome 22. *Genome Research*, in press.
Venter,J.C., Adams,M.D. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
Yu,A., Zhao,C. *et al.* (2001) Comparison of human genetic and sequence-based maps. *Nature*, **409**, 951–953.