



GeneLoc: exon-based integration of human genome maps

Naomi Rosen, Vered Chalifa-Caspi, Orit Shmueli, Avital Adato, Michal Lapidot, Julie Stampnitzky, Marilyn Safran* and Doron Lancet

Weizmann Institute of Science, Rehovot, Israel

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: Despite the numerous available whole-genome mapping resources, no comprehensive, integrated map of the human genome yet exists.

Results: GeneLoc, software adjunct to GeneCards and UDB, integrates gene lists by comparing genomic coordinates at the exon level and assigns unique and meaningful identifiers to each gene.

Availability: <http://bioinfo.weizmann.ac.il/genecards> and <http://genecards.weizmann.ac.il/udb>

Supplementary information: <http://bioinfo.weizmann.ac.il/cards-bin/AboutGCids.cgi>, <http://genecards.weizmann.ac.il/GeneLocAlg.html>

Contact: marilyn.safran@weizmann.ac.il

INTRODUCTION

Whole genome databases on the World Wide Web include: NCBI, with access to RefSeq, LocusLink, and the Human Genome MapViewer (Wheeler *et al.*, 2002); the Ensembl database project, which annotates the genome, integrating data from other sources with its own predictions (Hubbard *et al.*, 2002); and the Human Genome Browser at UCSC, which provides a graphical viewer of the genome with 'tracks,' each of which shows different information about the area in question (Kent *et al.*, 2002). In parallel, we have developed the Unified Database for Human Genome Mapping (UDB) (Chalifa-Caspi *et al.*, 1997; Safran *et al.*, 2003), which sorts genomic objects by chromosomal location to create an integrated genome map on a megabase scale with genes, markers, and ESTs.

Although whole-genome mapping resources currently use NCBI's assembly and all contain large gene lists, no two lists are identical. While all of the sources use prediction programs (Burge and Karlin, 1997; Kulp *et al.*, 1997; Yeh *et al.*, 2001), different programs and parameters can produce varying results. Moreover, as in the case of Ensembl, many model genes with only weak support are omitted (Hubbard *et al.*, 2002). In contrast, in an effort

to provide a comprehensive gene list, NCBI's LocusLink contains thousands of model genes, categorized by level and type of support. Even known genes appearing in every database may have different names in each database. The biologist must move among databases to figure out which genes are the same, and which could be a novel gene sought. UCSC's Genome Browser website maps genes from several sources on the same scale, but the maps are not integrated, making it difficult to relate genes from different sources. As stated (Jongeneel, 2000), 'there is an urgent need for a human gene index that can be used to identify transcripts unambiguously.' The author contends that this index should have, among others, the following qualities: comprehensiveness, uniqueness, and stability.

We therefore developed GeneLoc, a software adjunct to GeneCards. (Rebhan *et al.*, 1998; Safran *et al.*, 2002) and UDB that unifies gene collections, eliminates redundancies, and assigns a meaningful location-based identifier to each gene in the index. GeneLoc currently works with gene sets from NCBI and Ensembl. It aims to compare genes in these collections and decide which should be unified as one entry and which are discrete. Since the gene annotations use the same assembly and coordinate scheme, GeneLoc effects this gene integration by comparing the genes' genomic locations. The resulting GeneLoc 'gene territory' reflects the range of the unified genes, taking into account all possible exons.

ALGORITHM

GeneLoc first obtains genomic information from each source, including position of each gene and exon, names and ids, and gene validation status. It then builds two separate maps. Exon Map, created by comparing positions of nearby exons, includes all possible exons of each gene from all sources, regardless of their alternative combinations in mRNA splice variants. Each group of overlapping exons, and each single non-overlapping exon, gets one 'exon group' number. Gene Map, which includes all genes from all sources with their details, is made by comparing all neighboring and overlapping genes and

*To whom correspondence should be addressed.

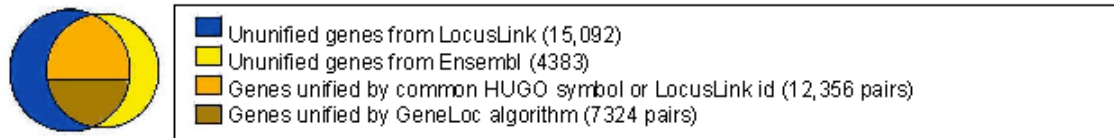


Fig. 1. The relative numbers of genes.

deciding which gene pairs can be merged. Decisions are hierarchical—each chromosome (and each unlocalized contig) is read strand by strand, performing the following steps: (I) genes are compared to find pairs whose HUGO symbols or LocusLink ids match. Genes sharing symbols indicate that each source has decided, based on evidence such as mRNA, that its gene corresponds to the same well-characterized gene. Since these decisions were based on strong evidence, they form the top of GeneLoc's decision hierarchy and a pair of genes thus correlated will be given one GeneCards identifier (GC id), immutable in later stages of the GeneLoc process. (An example of a GC id is the one for DMD, GC0XM029640, which reflects its chromosome (**X**), strand (**Minus**), and start position (**29 640 kb**)—see Supplementary Information). (II) A pair of genes not sharing an identifier, but with the same (within six base pairs) absolute base coordinates, also gets one GC id. (III) Each gene location is then compared to that of all nearby genes until no overlaps are found. A gene not overlapping with others gets its own GC id, unless already linked to another by id, as noted above. (IV) Mere overlap of two genes does not automatically merge them. If both are from one source, they get distinct GC ids, since the source deems them discrete. Otherwise, if either the starts or the ends of the two genes are within six base pairs of each other, they are currently considered one gene and get one GC id. When two genes overlap with starts and ends less closely aligned, GeneLoc checks the exons. Two genes with at least one overlapping exon get one GC id. The GeneLoc algorithm is currently being enhanced with a frame evaluation step that notes open reading frame inconsistencies between overlapping exons (see Supplementary Information). (V) When several genes overlap, or each gene in a group overlaps its neighbor (an 'overlap cluster'), GeneLoc checks which genes share exons. A gene in a cluster may have no sequence in common with other cluster members, usually when it or one of its exons lies within another gene's intron. Each gene is checked for shared exon groups with the rest of the cluster. A gene not sharing exon groups with other genes and not already linked by id gets its own GC id. Two genes (from different sources) sharing at least one exon group get one GC id. Two genes from one source sharing exon groups get separate GC ids, as does each gene in a group of three or more sharing exon groups. GeneLoc currently defines these genes as distinct, choosing to depend on

each source to maintain internal consistency. Nevertheless, the GeneLoc algorithm is currently being extended to consolidate as many genes resulting from these clusters as possible (see Supplementary Information). GeneLoc uniquely features association of genes with others, based on this overlapping-exon criterion.

Validation of this algorithm by elimination of the 'match by symbol' step showed over 98% success rate for gene matching (see Supplementary Information).

RESULTS

At this writing, there are 39 155 genes in GeneLoc. Many were unified from the LocusLink and Ensembl gene lists (33 845 and 22 980 genes, respectively) by GeneLoc's matching algorithm (Fig. 1). There are 15 092 LocusLink genes not corresponding to any Ensembl genes and 4383 from Ensembl not matching any LocusLink genes. An additional 1954 genes have overlapping exons with several other GeneLoc genes. These are part of a total of 6182 genes that GeneLoc identified in 1692 different clusters (Fig. 2). GeneCards currently has 46 179 entries, including the GeneLoc genes and 7024 genes from LocusLink with no associated coordinates.

GeneLoc's results can be seen in UDB and GeneCards. Since GeneLoc offers a combined view of genes, markers, genomic sequences, and their absolute positions, genomic areas of interest can be viewed from GeneCards via a gene-based query, or UDB via a mapping- or position-based query.

DISCUSSION AND CONCLUSION

The combined resources from UDB and GeneCards give a good picture of genes of interest while providing other useful details, and complement other resources, such as NCBI, Ensembl, and UCSC. UDB's tab-delimited display of GeneLoc's results allows viewing of genomic objects in the context of nearby objects, regardless of strand, object type, or data source. Moreover, since GeneLoc integrates genes from several other resources, its gene list is more thorough than those of other resources (Fig. 3).

UDB/GeneLoc's ability to display chromosomal regions between any two genetic markers, and to display even large genomic areas at a glance, makes it invaluable for positional cloning. Moreover, by taking advantage of our GeneNote project (Shmueli *et al.*, 2003), it will soon be possible to include gene expression information in GeneLoc. As a result, normal human tissue gene

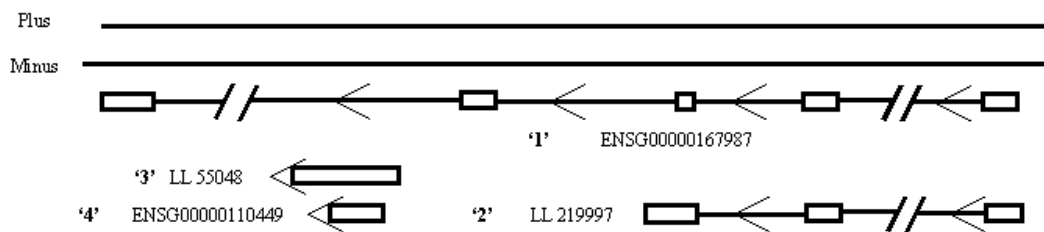


Fig. 2. An 'overlap cluster' found by the GeneLoc algorithm. Here, on the minus strand of chromosome 11, each source lists two different genes in the same range. GeneLoc merged these into one gene with several exons (merging '1' and '2') and another, with only one exon (merging '3' and '4'). The second gene lies within an intron of the first. Documented nested genes (Cawthon *et al.*, 1991; Adato *et al.*, 2002) are rare in the literature. GeneLoc has found over 100 in the human genome.

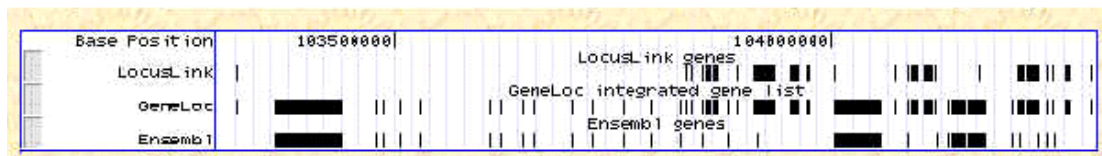


Fig. 3. Comparison of GeneLoc gene index with that of Ensembl and Locuslink in a typical genomic area (on chromosome 14), shown using UCSC Genome Browser's custom tracks option. Included are genes present in one source but excluded by the other.

expression profiles will be presented on the chromosomal map, helping in the definition of candidate disease genes in the studied linkage intervals.

The GeneLoc algorithm can be further used to integrate other gene list sets. Candidate additional sources include UCSC data and DOTS (<http://www.allgenes.org>) assemblies. Moreover, the GeneLoc software can be applied to organisms other than humans to obtain comprehensive maps of other genomes.

ACKNOWLEDGEMENTS

The authors thank Damian Conway for coding help, Ron Shamir for validation suggestions, and David Swidler for editing help. This work was funded by the Weizmann Institute Crown Human Genome Center, the Yeda Fund, and the Abraham and Judith Goldwasser Foundation.

REFERENCES

- Adato, A., Vreugde, S., Joensuu, T., Avidan, N., Hamalainen, R., Belékiy, O., Olender, T., Bonne-Tamir, B., Ben-Asher, E., Espinos, C. *et al.* (2002) USH3A transcripts encode clarin-1, a four-transmembrane-domain protein with a possible role in sensory synapses. *Eur. J. Hum. Genet.*, **10**, 339–350.
- Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
- Cawthon, R.M., Andersen, L.B., Buchberg, A.M., Xu, G.F., O'Connell, P., Viskochil, D., Weiss, R.B., Wallace, M.R., Marchuk, D.A., Culver, M. *et al.* (1991) cDNA sequence and genomic structure of EV12B, a gene lying within an intron of the neurofibromatosis type 1 gene. *Genomics*, **9**(3), 446–460.
- Chalifa-Caspi, V., Rebhan, M., Prilusky, J. and Lancet, D. (1997) The Unified DataBase (UDB): a novel genome integration concept. *Genome Digest*, **15**.
- Hubbard, T., Barker, D. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
- Jongeneel, C.V. (2000) The need for a human gene index. *Bioinformatics*, **16**, 1059–1061.
- Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
- Kulp, D., Haussler, D., Reese, M.G. and Eeckman, F.H. (1997) Integrating database homology in a probabilistic gene structure model. *Pac. Symp. Biocomput.*, 232–244.
- Rebhan, M., Chalifa-Caspi, V., Prilusky, J. and Lancet, D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.
- Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A. *et al.* (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW21, and HORDE. *Nucleic Acids Res.*, **31**.
- Safran, M., Solomon, I., Shmueli, O., Lapidot, M., Shen-Orr, S., Adato, A., Ben-Dor, U., Esterman, N., Rosen, N., Peter, I. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**.
- Shmueli, O., Horn-Saban, S., Chalifa-Caspi, V., Shmoish, M., Ophir, R., Benjamin-Rodrig, H., Safran, M., Domany, E. and Lancet, D. (2003) GeneNoTE: whole genome expression profiles in normal human tissues.
- Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L. and Rapp, B.A. (2002) Database resources of the NCBI: 2002 update. *Nucleic Acids Res.*, **30**, 13–6.
- Yeh, R.F., Lim, L.P. and Burge, C.B. (2001) Computational inference of homologous gene structures in the human genome. *Genome Res.*, **11**, 803–816.