

Assembly, Annotation, and Alignment of Genomic Sequences

Assembly of Sequence Data Sets

For each species, a single non-redundant nucleotide sequence was generated from the individual BAC sequences (note that all BAC sequences are available in GenBank; also see www.nisc.nih.gov). These compiled data sets were assembled as follows. Overlapping sequences between BACs were merged using Sequin (see www.ncbi.nlm.nih.gov/Sequin) to create a single, non-redundant sequence record. The specific sequence in the overlapping region of two neighboring BACs was chosen from the clone that contained more or higher-quality data (e.g., sequence from a finished clone was used over that of an unfinished clone; if both clones were finished and one contained an insertion relative to the other, the overlapping sequence was taken from the BAC containing the insertion). Within the compiled sequences files, gaps between BACs are represented with 50,000 Ns, whereas gaps between contigs within a clone are represented with 100 Ns. In addition to a nucleotide fasta file, a quality file was also created that contains the corresponding Phrap consensus quality scores (from the originating BAC) for each nucleotide of the assembled sequence. An agp file was also generated that contains specific information about the assembly (i.e., the exact spans of each individual BAC sequence that were ultimately used in the compiled sequence). All of these files are available at www.nisc.nih.gov/data.

Annotation of Assembled Sequence

Each assembled, non-redundant sequence was analyzed and annotated for the following features. Known repeats were detected using Repeatmasker (version 2002/05/05, run in sensitive mode; see repeatmasker.genome.washington.edu) using appropriate repeat libraries for each species (available in RepBase). Note that our sequence data allowed the development of appropriate repeat libraries for some of these species, including cow, pig, cat, dog, and chicken (A. F. Smit, unpublished data; see www.girinst.org/Repbases_Update.html). The results were parsed and annotated as repeat_region features. Next, Genscan (see genes.mit.edu/GENSCAN.html) was used to predict the locations of exons in the repeat-masked sequence; these predicted exon locations are included as gene features. Finally, genes known to be present in the human reference sequence were identified and annotated using the following computational tools and methods. For human and mouse sequences, the known genes were identified by aligning reference cDNA sequences from the NCBI RefSeq project (see www.ncbi.nlm.nih.gov/LocusLink/refseq.html) with the assembled genomic sequence using the NCBI tool Spidey (see www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey). For the remaining species (where cDNA sequences were not available), the RefSeq cDNA sequences for human and/or mouse were used. The GenBank accession numbers of the specific reference cDNA sequence(s) used are included in the annotation. In addition, genomic sequence alignments were generated using PipMaker¹ (see bio.cse.psu.edu) using either the annotated human and/or annotated mouse sequence as the reference. Exon locations were then propagated using transform-pos (see globin.cse.psu.edu/dist/piptools/all-perl/docs/piptools_ref.html#trans).

Sequin was used to import and validate all the above annotation. Known genes are annotated as gene, mRNA, and CDS (coding region) features. Protein translations are also included. Any splice-site consensus, exon structure, or protein-translation errors encountered with the above analyses were manually inspected and corrected. In addition, any annotation relating to low-quality regions, single-stranded sequence coverage, or excised transposon sequences was propagated from the individual BAC GenBank record.

The following files are available at www.nisc.nih.gov/data for each of the 12 sequenced species:

*.agp	agp file (index of how the fasta file was compiled)
*.annot.ff	GenBank-style flat file containing annotated known genes, corresponding mRNA and coding regions, repeats, and Genscan predicted exons
*.annot.sqn	ASN.1 format of annotated sequence
*.fasta	fasta nucleotide sequence
*.fasta.qual	corresponding fasta file of sequence-assembly quality scores

In addition, the following three files are also available for the human reference sequence of the greater *CFTR* region: human_T1.annot.ff, human_T1.annot.sqn, and human_T1.fasta. This human reference fasta sequence was obtained from the UCSC Genome Browser (see genome.ucsc.edu), specifically chr7: 115977709-117855134 of the April 2002 build. Finally, each species' sequence has a corresponding *_repeat directory that contains the Repeatmasker output files (including repeat-masked fasta files).

Sequence Alignments

Pair-wise sequence alignments were generated with repeat-masked sequences using blastz² (06/26/2002 build available at bio.cse.psu.edu/dist) and the following parameters: B=0 C=2 K=2500 Y=3400 T=0. When a sequence position aligned with several positions in the other species, lower scoring alignments were trimmed to eliminate the overlap.

For some analyses, Multiple Pair-Wise Alignment (MPA) files were created by joining all subsets of pair-wise alignments for each species' sequence. In order to keep each species' alignments on the same coordinate system, gaps were removed from the reference sequence in each pair-wise alignment prior to building the MPA files. This yielded a set of 13 MPA files, one with each species' sequence used as the reference. Thus, provided the underlying reference sequence used to generate the MPA, a given coordinate in one species' sequence can be correlated to any other species' sequence.

For visualizing a series of pair-wise percent-identity plots (PIPs) generated with sequence from multiple species, the program MultiPipMaker³ (see bio.cse.psu.edu) was utilized; see Fig. 1a of the main paper for an example and www.nisc.nih.gov/data for MultiPipMaker output of the entire sequence data set. In addition, MultiPipMaker was used to produce a nucleotide-level multiple sequence alignment. The alignment is computed by first creating pair-wise alignments between the reference sequence and each

of the other sequences. Elimination of overlaps (see above) permits a crude multiple alignment to be produced (analogous to the MPA file, but permitting gaps in the reference sequence). MultiPipMaker then applies an iterative improvement strategy that attempts to optimize a rigorous multiple-alignment score.

References

1. Schwartz, S. *et al.* PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577-586 (2000).
2. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103-107 (2003).
3. Schwartz, S. *et al.* MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.*, in press (2003).