



## Multi-species comparative mapping in silico using the COMPASS strategy

Lei Liu<sup>1,2,\*</sup>, George Gong<sup>1</sup>, Yong Liu<sup>1</sup>, Shreedhar Natarajan<sup>3</sup>, Denis M. Larkin<sup>2</sup>, Annelie Everts-van der Wind<sup>2</sup>, Mark Rebeiz<sup>2</sup> and Jonathan E. Beever<sup>2</sup>

<sup>1</sup>The W. M. Keck Center for Comparative and Functional Genomics, <sup>2</sup>Department of Animal Sciences and <sup>3</sup>Biophysics and Computational Biology Program, University of Illinois at Urbana-Champaign, 1201 W. Gregory Drive, Urbana, IL 61801, USA

Received on November 9, 2003; accepted on November 11, 2003

### ABSTRACT

**Motivation:** The completion of human and mouse genome sequences provides a valuable resource for decoding other mammalian genomes. The comparative mapping by annotation and sequence similarity (COMPASS) strategy takes advantage of the resource and has been used in several genome-mapping projects. It uses existing comparative genome maps based on conserved regions to predict map locations of a sequence. An automated multiple-species COMPASS tool can facilitate in the genome sequencing effort and comparative genomics study of other mammalian species.

**Results:** The prerequisite of COMPASS is a comparative map table between the reference genome and the predicting genome. We have built and collected comparative maps among five species including human, cattle, pig, mouse and rat. Cattle–human and pig–human comparative maps were built based on the positions of orthologous markers and the conserved synteny groups between human and cattle and human and pig genomes, respectively. Mouse–human and rat–human comparative maps were based on the conserved sequence segments between the two genomes. With a match to human genome sequences, the approximate location of a query sequence can be predicted in cattle, pig, mouse and rat genomes based on the position of the match relatively to the orthologous markers or the conserved segments.

**Availability:** The COMPASS-tool and databases are available at <http://titan.biotech.uiuc.edu/COMPASS/>

**Contact:** leiliu@uiuc.edu

### INTRODUCTION

Comparative studies have been a very powerful approach in biological sciences (Harvey and Pagel, 1991). Comparative genomics provides insight into the genomic structure in the evolutionary context and has become an invaluable extension of the Human Genome initiative (O'Brien *et al.*, 1997, 1999).

Completion of sequencing of the human and mouse genomes has produced a wealth of resources and a strong basis for comparing ordered gene maps of other mammalian species (Lander *et al.*, 2001; Waterston *et al.*, 2002). Information from gene-rich species maps of human and mouse can be transferred to moderate-resolution gene maps of other species, such as cattle and pig. Those comparative genome maps can provide a basis for understanding the rates and patterns of genome evolution (O'Brien *et al.*, 1993, 1999; Andersson *et al.*, 1996; Band *et al.*, 2000; Larkin *et al.*, 2003). To this end, moderately dense comparative gene maps with gene markers (Type I markers) on each chromosome in different mammalian genomes are required to demarcate boundaries of conserved synteny among representative species (O'Brien, 1991; Copeland *et al.*, 1993; Womack and Kata, 1995; Schibler *et al.*, 1998; Yang and Womack, 1998; Murphy *et al.*, 1999a,b; Watanabe *et al.*, 1999). Based on the comparative genome maps, traits mapped in one species may be mapped in other species *in silico* (O'Brien *et al.*, 1999).

It has been shown that existing knowledge of comparative chromosome organization can be used to predict the map location *in silico*. This comparative mapping by annotation and similarity (COMPASS) strategy has been widely used in generating comparative genome maps of several mammalian species (Band *et al.*, 1998, 2000; Ma *et al.*, 1998; Rebeiz and Lewin, 2000; Ozawa *et al.*, 2000; Murphy *et al.*, 2000). Using COMPASS strategy in combination with radiation hybrid (RH) mapping (Larkin *et al.*, 2003) has recently created the bacterial artificial chromosome (BAC) clone-based comparative map of cattle and human and demonstrated the power of generating high-resolution comparative chromosome maps.

To implement the COMPASS strategy, Perl or other script pipelines can be written to automate the steps of process, including sequence similarity searches using the basic local alignment search tool for nucleotides (BLASTN) (Altschul *et al.*, 1990), parsing BLASTN output, querying comparative map tables, calculating the predicted position and finally composing the prediction reports (Rebeiz and Lewin, 2000;

\*To whom correspondence should be addressed.

Larkin *et al.*, 2003). In this paper, a database solution to COMPASS is presented, which can be easily extended to multiple genome map comparisons among mammalian species. The implemented COMPASS comparative tables can be updated following the re-build of reference genome assembly (e.g. human and mouse genome) and the new results from genome mapping and sequencing efforts of other species, such as cattle and pig. A web portal is developed to streamline the COMPASS process, which makes the COMPASS prediction process transparent to users and allow users to make multiple species prediction at one time. The COMPASS strategy and COMPASS-tool presented here can also be applied to other vertebrates such as fish if well-defined comparative maps are available.

## DATA SOURCES

The list of 768 cattle Type I markers mapped on the RH panel with their chromosome locations were obtained from the published results of Band *et al.* (2000). The sequences of these markers were extracted from GenBank using their accession numbers. The list of chromosome locations for 680 RH-mapped pig Type I markers were obtained (J.E. Beever, personal communication). Human (Build 33) and mouse (Build 30) genome sequences were retrieved from the National Center for Biotechnology Information (NCBI). The lists of human and mouse markers and their chromosome locations with corresponding base pair coordinates were obtained from Ensembl database (<http://www.ensembl.org>). Human(hg16)/mouse(mm3) and human(hg16)/rat(rn3) genome sequence comparisons were downloaded from University of California Santa Cruz Genome Bioinformatics site (<http://genome.ucsc.edu/index.html>).

## METHODS

### Building a synteny table

To build an effective COMPASS-tool a synteny table anchoring low-resolution maps with reference genome needs to be established. A precise identification of orthologous markers between the genomes is the first step towards building this table.

Using cattle and pig gene maps as examples, we illustrated two ways to obtain this anchored synteny table. The Human Genome Organization (HUGO) gene names have been adopted by other mammalian genome projects to name orthologous genes based on sequence similarity, function and homologous position in the genomes. In this study, 680 pig Type I markers and 467 cattle Type I markers share the same name with the HUGO gene names. The synteny relationships can be directly obtained through these orthologous genes. This method is quite specific to comparison with the human genome. For comparison with other genomes such as mouse or with non-mammalian genomes, the BLASTN search method described below is more appropriate.

For markers, which did not share the name with human genes, such as some of the cattle expressed sequence tags (ESTs), the comparison with human and mouse genome was performed, after repeat masking, using BLASTN with cut-off value of  $E^{-10}$ . For the top significant hit in the reference genome sequence the coordinates were extracted from the BLASTN output and corresponding gene name was obtained from the sequence annotation. Once the orthologous gene pairs were identified, the synteny table can be built by combining the reference genome sequence information (e.g. chromosomes, base pair coordinates), the chromosome of species of interest, and the conserved synteny markers.

### COMPASS prediction—*in silico* mapping

The chromosome location of a query sequence [e.g. an EST or a BAC end sequence (BES)] can be predicted on the map of specie of interest, by BLASTN search against the reference genome sequence and the data from the synteny table. For instance, for any cattle sequence with significant BLASTN match with build 33 of the human genome sequence the position of the closest marker mapped in both human and cattle genomes can be found (Fig. 1). If the position of the sequence of interest falls into a conserved segment on the cattle and human comparative map, then the position of this sequence can be predicted in the cattle genome. If the position of the sequence falls outside of a conserved segment, then no prediction can be made at this time. The closest marker can be defined as the marker of which the BLASTN hit overlaps with or using the following calculation in the case that the BLASTN hit lies between two markers:

if  $(HS - ME) < (M'S - HE)$ ,  $M$  is the closest marker;

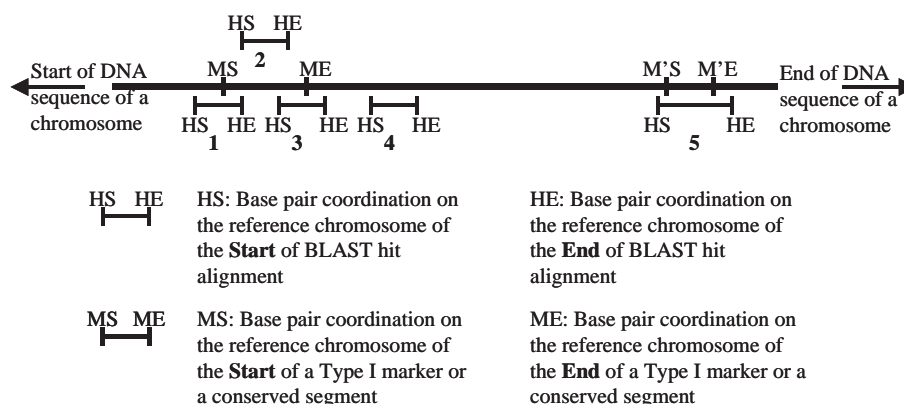
if  $(HS - ME) > (M'S - HE)$ ,  $M'$  is the closest marker.

HS, HE, ME and M'S are denoted as shown in Figure 1. In the case of prediction on mouse or rat genome, the exact conserved sequence segments were used. A prediction can be made only when the alignment of a hit overlaps with the conserved sequence segment. In this case, the prediction will not only show the chromosome but also the base pair coordination on the chromosome.

The majority of prediction process is to identify the spatial relation between the matching sequence segments on the reference genome and the conserved synteny group or conserved sequence segments between the reference genome and the predicting genome. Since the chromosome map is one dimension, the calculation is straightforward. Once the prediction for each species is complete, the results are concatenated for the same query sequence. In this study, we used two reference genomes, human and mouse, to predict map location on cattle, pig, rat, mouse or human genomes.

## RESULTS

Human and mouse were considered as reference genomes and their Type I markers were obtained from Ensembl



**Fig. 1.** Illustration of different scenarios of BLASTN hits on reference chromosome and the definition of overlap between a marker and its closest marker. M and M' represent anchored markers and H represents a BLASTN hit alignment of a query sequence. Scenario 1, 2, 3 and 5; overlap with an anchored marker; Scenario 4: close to an anchored marker.

**Table 1.** Example entries in the Synteny table

| gene_marker | ref_species_id (human) | ref_gene_marker | reference_chr | ref_chr_start | ref_chr_end | species_id (pig) | chromosome |
|-------------|------------------------|-----------------|---------------|---------------|-------------|------------------|------------|
| PDE6B       | 1                      | PDE6B           | 4             | 613592        | 658761      | 3                | 8          |
| FGFR3       | 1                      | FGFR3           | 4             | 1775687       | 1790662     | 3                | 8          |
| GNA12       | 1                      | GNA12           | 7             | 2412112       | 2528169     | 3                | 13         |
| TNNT3       | 1                      | TNNT3           | 11            | 2023069       | 2040428     | 3                | 2          |
| IGF2        | 1                      | IGF2            | 11            | 2241901       | 2247980     | 3                | 2          |
| RAD52       | 1                      | RAD52           | 12            | 899593        | 989453      | 3                | 5          |
| FGF6        | 1                      | FGF6            | 12            | 4435706       | 4447178     | 3                | 5          |
| ADCYAP1     | 1                      | ADCYAP1         | 18            | 892989        | 898656      | 3                | 6          |
| TGIF        | 1                      | TGIF            | 18            | 3439546       | 3446280     | 3                | 6          |
| AMH         | 1                      | AMH             | 19            | 2318416       | 2321365     | 3                | 2          |
| NFIC        | 1                      | NFIC            | 19            | 3435892       | 3532243     | 3                | 2          |
| EEF2        | 1                      | EEF2            | 19            | 4045332       | 4054729     | 3                | 15         |
| CSNK2A1     | 1                      | CSNK2A1         | 20            | 411734        | 472355      | 3                | 17         |
| PDYN        | 1                      | PDYN            | 20            | 1907402       | 1922703     | 3                | 17         |
| ADRA1D      | 1                      | ADRA1D          | 20            | 4149816       | 4177662     | 3                | 16         |

database (Ensembl Human v.16.33.1 and Ensembl Mouse v.16.30.1) with 23 299 and 24 948 genes, respectively. Comparing cattle Type I markers obtained from the published results (Band *et al.*, 2000) with human and mouse genome sequences, 652 cattle–human and 554 Cattle–mouse orthologous markers were obtained using BLASTN search. Using HUGO gene names, 680 pig–human orthologous markers were retrieved. These results were stored in the synteny table with example records shown in Table 1, which shows the genes in pig–human comparative map.

The data in 'gene\_marker' column are the mapped orthologous genes to the reference genome on the predicting genome. One may notice that the gene names are identical to the gene names in the 'ref\_gene\_marker' column. This is because for pig–human comparative map, we used the HUGO gene names for both species. For other comparative

map such as cattle–human, the gene names could be different.

We used 15 pig EST sequences retrieved from NCBI to test the COMPASS prediction. Among the test query sequences, 11 query sequences return single significant BLASTN hit against human genome sequences. Using these 11 sequences, our COMPASS-tool produced mapping predictions on cattle, pig, mouse and rat genome as shown in Tables 2 and 3. Table 2 shows the detailed information about the prediction on each individual genome. Each block in Table 2 contains the prediction results from one query sequence. Table 3 shows the final compiled predictions across all five species. Each row represents predictions from one query sequence. For predictions in cattle and pig, the resolution of the predictions was very low. Only chromosome and closest markers were predicted. The resolution of mouse and rat

**Table 2.** The detailed COMPASS prediction

| query_id | hit_start | hit_end  | ref_species        | ref_chr | ref_start | ref_end  | Marker   | predicted_species        | predicted_chr | predicted_start | predicted_end |
|----------|-----------|----------|--------------------|---------|-----------|----------|----------|--------------------------|---------------|-----------------|---------------|
| 3065956  | 12387061  | 12387502 | <i>Homo sapien</i> | 3       | 12387058  | 12387502 | PPARG    | <i>Bos Taurus</i>        | 22            |                 |               |
| 3065956  | 12387061  | 12387502 | <i>Homo sapien</i> | 3       | 12322438  | 12467696 | PPARG    | <i>Sus scrofa</i>        | 13            |                 |               |
| 3065956  | 12387061  | 12387502 | <i>Homo sapien</i> | 3       | 12387190  | 12387531 |          | <i>Mus musculus</i>      | 6             | 116110420       | 116110752     |
| 3065956  | 12387061  | 12387502 | <i>Homo sapien</i> | 3       | 12387190  | 12387531 |          | <i>Rattus norvegicus</i> | 4             | 151806706       | 151807037     |
| 15182871 | 28212006  | 28212423 | <i>Homo sapien</i> | 4       | 28211985  | 28212177 | UBE2D3   | <i>Bos taurus</i>        | 6             |                 |               |
| 15182871 | 28212006  | 28212423 | <i>Homo sapien</i> | 4       | 26171366  | 26180390 | CCKAR    | <i>Sus scrofa</i>        | 8             |                 |               |
| 15182871 | 28212006  | 28212423 | <i>Homo sapien</i> | 4       | 28211676  | 28212214 |          | <i>Mus musculus</i>      | 5             | 54389777        | 54390317      |
| 15182871 | 28212006  | 28212423 | <i>Homo sapien</i> | 4       | 28211712  | 28212182 |          | <i>Rattus norvegicus</i> | 14            | 52139429        | 52139901      |
| 7841991  | 6245751   | 6245920  | <i>Homo sapien</i> | 6       | 6245745   | 6245890  | F13A1    | <i>Bos taurus</i>        | 23            |                 |               |
| 7841991  | 6245751   | 6245920  | <i>Homo sapien</i> | 6       | 6134316   | 6308896  | F13A1    | <i>Sus scrofa</i>        | 7             |                 |               |
| 7841991  | 6245751   | 6245920  | <i>Homo sapien</i> | 6       | 6250575   | 6250763  |          | <i>Mus musculus</i>      | 13            | 36444472        | 36444660      |
| 7841991  | 6245751   | 6245920  | <i>Homo sapien</i> | 6       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 29277612 | 6245582   | 6245751  | <i>Homo sapien</i> | 6       | 6245745   | 6245890  | F13A1    | <i>Bos taurus</i>        | 23            |                 |               |
| 29277612 | 6245582   | 6245751  | <i>Homo sapien</i> | 6       | 6134316   | 6308896  | F13A1    | <i>Sus scrofa</i>        | 7             |                 |               |
| 29277612 | 6245582   | 6245751  | <i>Homo sapien</i> | 6       | 6250575   | 6250763  |          | <i>Mus musculus</i>      | 13            | 36444472        | 36444660      |
| 29277612 | 6245582   | 6245751  | <i>Homo sapien</i> | 6       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 437919   | 32439445  | 32439686 | <i>Homo sapien</i> | 9       | 32260716  | 32260857 | STXBP1   | <i>Bos taurus</i>        | 11            |                 |               |
| 437919   | 32439445  | 32439686 | <i>Homo sapien</i> | 9       | 32374619  | 32440835 | ACO1     | <i>Sus scrofa</i>        | 10            |                 |               |
| 437919   | 32439445  | 32439686 | <i>Homo sapien</i> | 9       | 32439993  | 32440107 |          | <i>Mus musculus</i>      | 4             | 40076880        | 40076994      |
| 437919   | 32439445  | 32439686 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 3757498  | 29085631  | 29086261 | <i>Homo sapien</i> | 9       | 29085469  | 29085937 | NR5A1    | <i>Bos taurus</i>        | 11            |                 |               |
| 3757498  | 29085631  | 29086261 | <i>Homo sapien</i> | 9       | 32374619  | 32440835 | ACO1     | <i>Sus scrofa</i>        | 1,10          |                 |               |
| 3757498  | 29085631  | 29086261 | <i>Homo sapien</i> | 9       | 29085111  | 29085252 |          | <i>Mus musculus</i>      | 4             | 36677282        | 36677423      |
| 3757498  | 29085631  | 29086261 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 6945242  | 33591957  | 33592009 | <i>Homo sapien</i> | 9       | 33591934  | 33592007 | KIAA0169 | <i>Bos taurus</i>        | 11            |                 |               |
| 6945242  | 33591957  | 33592009 | <i>Homo sapien</i> | 9       | 35672008  | 35679927 | TPM2     | <i>Sus scrofa</i>        | 1,10          |                 |               |
| 6945242  | 33591957  | 33592009 | <i>Homo sapien</i> | 9       |           |          |          | <i>Mus musculus</i>      |               |                 |               |
| 6945242  | 33591957  | 33592009 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 6962339  | 35679272  | 35679396 | <i>Homo sapien</i> | 9       | 35795507  | 35795668 | NPR2     | <i>Bos taurus</i>        | 8             |                 |               |
| 6962339  | 35679272  | 35679396 | <i>Homo sapien</i> | 9       | 35672008  | 35679927 | TPM2     | <i>Sus scrofa</i>        | 1             |                 |               |
| 6962339  | 35679272  | 35679396 | <i>Homo sapien</i> | 9       | 35679133  | 35679271 |          | <i>Mus musculus</i>      | 4             | 42365548        | 42365686      |
| 6962339  | 35679272  | 35679396 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 11072697 | 28936823  | 28936995 | <i>Homo sapien</i> | 9       | 28936949  | 28937039 | EST0660  | <i>Bos taurus</i>        | 11            |                 |               |
| 11072697 | 28936823  | 28936995 | <i>Homo sapien</i> | 9       | 32374619  | 32440835 | ACO1     | <i>Sus scrofa</i>        | 1,10          |                 |               |
| 11072697 | 28936823  | 28936995 | <i>Homo sapien</i> | 9       | 28937805  | 28937918 |          | <i>Mus musculus</i>      | 4             | 36552374        | 36552487      |
| 11072697 | 28936823  | 28936995 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |
| 18962526 | 37856182  | 37856451 | <i>Homo sapien</i> | 9       | 37856151  | 37856465 | EST1238  | <i>Bos taurus</i>        | 8             |                 |               |
| 18962526 | 37856182  | 37856451 | <i>Homo sapien</i> | 9       | 35672008  | 35679927 | TPM2     | <i>Sus scrofa</i>        | 1,10          |                 |               |
| 18962526 | 37856182  | 37856451 | <i>Homo sapien</i> | 9       | 37857513  | 37857703 |          | <i>Mus musculus</i>      | 4             | 44225243        | 44225427      |
| 18962526 | 37856182  | 37856451 | <i>Homo sapien</i> | 9       | 37856054  | 37856204 |          | <i>Rattus norvegicus</i> | 5             | 61950738        | 61950891      |
| 29282743 | 28936992  | 28937163 | <i>Homo sapien</i> | 9       | 28936949  | 28937039 | EST0660  | <i>Bos taurus</i>        | 11            |                 |               |
| 29282743 | 28936992  | 28937163 | <i>Homo sapien</i> | 9       | 32374619  | 32440835 | ACO1     | <i>Sus scrofa</i>        | 1,10          |                 |               |
| 29282743 | 28936992  | 28937163 | <i>Homo sapien</i> | 9       | 28937805  | 28937918 |          | <i>Mus musculus</i>      | 4             | 36552374        | 36552487      |
| 29282743 | 28936992  | 28937163 | <i>Homo sapien</i> | 9       |           |          |          | <i>Rattus norvegicus</i> |               |                 |               |

query\_id is the NCBI gi number. hit\_start and hit\_end are the start and end of the aligned region between the query sequences and the reference genome. ref\_start and ref\_end are the start and end of the base pair coordinates of either the gene marker or the conserved sequence segment on the reference genome. predicted\_start and predicted\_end are the start and end of the base pair coordinates of the conserved sequence segments on the predicting genome.

predictions was at the base pair level, but there were many unpredictable regions because of either incompleteness of the genomes or lack of conserved sequences (Table 3). The blank cells in the tables are either from unpredictable results or missing data.

Two pairs of query sequences (7841991/29277612 and 11072697/29282743) gave two identical predictions,

respectively. From the detailed alignments, we noticed that these EST sequences were overlap with each other and spanned over the same orthologous markers and conserved regions. For the predictions on cattle and pig genomes, we observed that only two query sequences showed the same closest gene markers, which were PPARG and F13A1. But the base pair coordination of the same marker on human

**Table 3.** The final COMPASS prediction results

| query_<br>id | human_<br>chr | h_start  | h_end    | cattle_<br>chr | gene_<br>marker | pig_<br>chr | gene_<br>marker | mouse_<br>chr | m_start   | m_end     | rat_<br>chr | r_<br>start | r_<br>end |
|--------------|---------------|----------|----------|----------------|-----------------|-------------|-----------------|---------------|-----------|-----------|-------------|-------------|-----------|
| 3065956      | 3             | 12387061 | 12387502 | 22             | PPARG           | 13          | PPARG           | 6             | 116110420 | 116110752 | 4           | 151806706   | 151807037 |
| 15182871     | 4             | 28212006 | 28212423 | 6              | UBE2D3          | 8           | CCKAR           | 5             | 54389777  | 54390317  | 14          | 52139429    | 52139901  |
| 29277612     | 6             | 6245582  | 6245751  | 23             | F13A1           | 7           | F13A1           | 13            | 36444472  | 36444660  |             |             |           |
| 7841991      | 6             | 6245751  | 6245920  | 23             | F13A1           | 7           | F13A1           | 13            | 36444472  | 36444660  |             |             |           |
| 11072697     | 9             | 28936823 | 28936995 | 11             | EST0660         | 1,10        | ACO1            | 4             | 36552374  | 36552487  |             |             |           |
| 29282743     | 9             | 28936992 | 28937163 | 11             | EST0660         | 1,10        | ACO1            | 4             | 36552374  | 36552487  |             |             |           |
| 3757498      | 9             | 29085631 | 29086261 | 11             | NR5A1           | 1,10        | ACO1            | 4             | 36677282  | 36677423  |             |             |           |
| 437919       | 9             | 32439445 | 32439686 | 11             | STXBP1          | 10          | ACO1            | 4             | 40076880  | 40076994  |             |             |           |
| 6945242      | 9             | 33591957 | 33592009 | 11             | KIAA0169        | 1,10        | TPM2            |               |           |           |             |             |           |
| 6962339      | 9             | 35679272 | 35679396 | 8              | NPR2            | 1           | TPM2            | 4             | 42365548  | 42365686  |             |             |           |
| 18962526     | 9             | 37856182 | 37856451 | 8              | EST1238         | 1,10        | TPM2            | 4             | 44225243  | 44225427  | 5           | 61950738    | 61950891  |

h\_start and h\_end are the start and end of base pair coordinates of human genome sequences matched with the query sequences. m\_start and m\_end are the start and end of base pair coordinates of mouse genome sequences of the conserved region between human and mouse. r\_start and r\_end are the start and end of base pair coordinates of rat genome sequences of the conserved region between human and rat.

genome was different for cattle and pig. We noticed that the span of the same marker in the comparative map for pig–human was much larger than the one for cattle–human. This may be due to the relatively lower resolution of the pig–human comparative map. For the other nine query sequences, the predicted closest orthologous markers were different between cattle and pig. Detailed prediction results from Table 2 revealed that the predictions for cattle genomes were mostly the scenarios 1, 2, 3 in Figure 1, but the predictions for pig genomes were mostly the scenario 4. Obviously, the scenario 4 is not a very accurate prediction. The corresponding human genome regions of predicted regions on mouse and rat genomes were much closer because the resolution of the comparative map between human and mouse and human and rat were at the base pair level. But the conserved regions still showed slight differences. This is probably due to the difference of completeness between the two genomes. Rat genome seemed much less completed than mouse genome. There were only three query sequences predicted in rat genome, but 10 query sequences predicted in mouse genome.

Seven query sequences were aligned with sequences on human chromosome 9 and gave us the chance to observe the rearrangement of the chromosome across several species. From Table 3, we can clearly see that this region of human chromosome 9 spanning about 8.9 million bp corresponds to a portion of cattle chromosome 11, a portion of cattle chromosome 8, and a continued region of mouse chromosome 4 spanning about 7.6 million bp, even though there was an unpredictable region in the middle. In pig, the predictions were somewhat ambiguous despite the fact that the query sequences were from pig. Only two orthologous markers were mapped in this region. One was TPM2 on pig chromosome 1 and the other was ACO1 on pig chromosome 10. Thus, we could not predict confidently

which chromosome the query sequences belonged to. We simply put both chromosomes in the predicted results. Only two query sequences (437919 and 6962339) were predicted with confidence because they overlapped with the mapped markers. It was difficult to know which rat chromosome corresponding to this region because there was only one prediction.

## DISCUSSION

### Accuracy of COMPASS

Previously, Band *et al.* (2000) reported 95% accuracy when sequences were predicted to cattle chromosome locations using COMPASS strategy and human RH maps. Recently, Larkin *et al.* (2003) reported about 90% accuracy in prediction of cattle chromosome locations based on the human genome sequence. The accuracy of pig–human COMPASS predictions was not tested, but was expected to be similar to what was reported in Band *et al.* (2000) and Larkin *et al.* (2003). The accuracy of prediction depends on sequence similarity of the reference and target genomes and the quality of their comparative maps. For example, if we use mouse genome sequences as the reference, we will expect fewer predictions because less sequence similarity is observed between mouse and other mammalian genomes such as cattle. In fact, fewer anchored markers were observed between mouse and cattle (Larkin *et al.*, 2003). Less sequence similarity will first result in lower resolution of COMPASS input tables, which will cause the lower accuracy of COMPASS predictions. Second, less sequence similarity will reduce the number of query sequences having significant matches in the reference genome, which will also cause lower efficiency of the COMPASS-tool. But in multiple species COMPASS, one may find improvement in prediction using more than one reference genome because the

prediction results may complement each other with different reference genomes. For example, the four query sequences that did not have a match to human genome sequences may have a match to mouse genome sequences and therefore can be predicted using mouse genome as a reference genome.

### Mammalian genome evolution

Multiple species comparison reveals different patterns of conserved segments among mammalian genomes (Murphy *et al.*, 2003). Several conserved segments were observed across the five mammalian genomes in our test dataset. Genome comparison based on genome sequences and high-resolution maps of non-sequenced genome will help to discover the fine patterns and the boundaries of the conserved segments among different species. Using BAC end sequences, the COMPASS strategy, and RH mapping, Larkin *et al.* (2003) showed the conservation of boundaries of the conserved segments between cattle and mouse using HSA11 as a reference. Recent studies on human and mouse genomic sequences (Pevzner and Tesler, 2003a,b) revealed extensive breakpoint reuse and accurately estimate the extent of rearrangement events in mammalian evolution. With increasing number of mapped markers among different species and the complete sequencing of rat, cattle and pig, we will be able to trace back mammalian genome evolution by detailed multiple species comparison and solve the 'original synteny' problem (Nadeau and Sankoff, 1997).

### CONCLUSION

Using the methods and strategy of the multiple species COMPASS described here, we attempted to develop a framework tool to take advantage of the resources of completed genomes and facilitate genome mapping and sequencing of other species. With more genomes are sequenced, the power and accuracy of the prediction will increase dramatically. This framework tool will also help study important traits such as genetics diseases across multiple species. For example, we may be able to map a trait of a rear genetic disease in cow *in silico* to human genome and discover the vulnerable region in human.

This framework can also be extended to other vertebrate species beyond mammals with finished genomes and high-quality comparative map based on conserved synteny group. For example, for fish species such as catfish and salmon, we can identify anchored markers using the sequenced zebra fish and Fugu genomes. The design and implementation of the current synteny table can be easily expanded with minimal modifications.

### ACKNOWLEDGEMENTS

We would like to thank Harris A. Lewin for his insightful discussion and suggestion for this paper. We would like to acknowledge the organizations from which we have obtained

data and software, including UCSC Genome Bioinformatics, ENSEMBL and NCBI.

### REFERENCES

- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Andersson,L., Archibald,A., Ashburner,M., Audun,S., Barendse,W., Bitgood,J., Bottema,C., Broad,T., Brown,S. and Burt,D. (1996) Comparative genome organization of vertebrates. The first international workshop on comparative genome organization. *Mamm. Genome*, **7**, 717–734.
- Band,M., Larson,J.H., Womack,J.E. and Lewin,H.A. (1998) A radiation hybrid map of BTA23: identification of a chromosomal rearrangement leading to separation of the cattle MHC class II subregions. *Genomics*, **53**, 269–275.
- Band,M.R., Larson,J.H., Rebeiz,M., Green,C.A., Heyen,D.W., Donovan,J., Windish,R., Steining,C., Mahyuddin,P., Womack,J.E. *et al.* (2000) An ordered comparative map of the cattle and human genomes. *Genome Res.*, **10**, 1359–1368.
- Copeland,N.G., Jenkins,N.A., Gilbert,D.J., Eppig,J.T., Maltais,L.J., Miller,J.C., Dietrich,W.F., Weaver,A., Lincoln,S.E. and Steen,R.G. (1993) A genetic linkage map of the mouse: current applications and future prospects. *Science*, **262**, 57–66.
- Harvey,P.H. and Pagel,M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.
- Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Larkin,D.M., Everts-van der Wind,A., Rebeiz,M., Schweitzer,P.A., Bachman,S., Green,C., Wright,C.L., Campos,E.J., Benson,L.D., Edwards,J. *et al.* (2003) A cattle-human comparative map built with cattle BAC-Ends and human genome sequence. *Genome Res.*, **13**, 1966–1971.
- Ma,R.Z., van Eijk,M.J.T., Beaver,J.E., Guérin,G., Mummery,C.L. and Lewin,H.A. (1998) Comparative analysis of 82 expressed sequence tags from a cattle ovary cDNA library. *Mamm. Genome*, **9**, 545–549.
- Murphy,W.J., Menotti-Raymond,M., Lyons,L.A., Thompson,M.A. and O'Brien,S.J. (1999a) Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics*, **57**, 1–8.
- Murphy,W.J., Shan,S., Chen,Z.Q., Pecon-Slattery,J. and O'Brien,S.J. (1999b) Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Res.*, **9**, 1223–1230.
- Murphy,W.J., Sun,S., Chen,Z.Q., Yuhki,N., Hirschmann,D., Menotti-Raymond,M. and O'Brien,S.J. (2000) A radiation hybrid map of the cat genome: Implications for comparative mapping. *Genome Res.*, **10**, 619–702.
- Murphy,W.J., Fronicke,L., O'Brein,J.S. and Stanyon,R. (2003) The origin of human chromosome I and its homologs in placental mammals. *Genome Res.*, **13**, 1880–1888.
- Nadeau,J.H. and Sankoff,D. (1997) Landmarks in the Rosetta Stone of mammalian comparative maps. *Nat. Genet.*, **15**, 6–7.
- O'Brien,S.J. (1991) Mammalian genome mapping: lessons and prospects. *Curr. Opin. Genet. Dev.*, **1**, 105–111.

- O'Brien,S.J., Womack,J.E., Lyons,L.A., Moore,K.J., Jenkins,N.A. and Copeland,N.G. (1993) Anchored reference loci for comparative genome mapping in mammals. *Nat. Genet.*, **3**, 103–112.
- O'Brien,S.J., Wienberg,J. and Lyons,L.A. (1997) Comparative genomics: Lessons from cats. *Trends Genet.*, **13**, 393–399.
- O'Brien,S.J., Menotti-Raymond,M., Murphy,W.J., Nash,W.G., Wienberg,J., Stanyon,R., Copeland,N.G., Jenkins,N.A., Womack,J.E. and Graves,J.A.M. (1999) The promise of comparative genomics in mammals. *Science*, **286**, 458–481.
- Ozawa,A., Band,M.R., Larson,J.H., Donovan,J., Green,C.A., Womack,J.E. and Lewin,H.A. (2000) Comparative organization of cattle chromosome 5 revealed by COMPASS and radiation hybrid mapping. *Proc. Natl Acad. Sci. USA*, **97**, 4150–4155.
- Pevzner,P. and Tesler,G. (2003a) Genome rearrangements in mammalian evolution: lesions from human and mouse genomes. *Genome Res.*, **13**, 37–45.
- Pevzner,P. and Tesler,G. (2003b) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl Acad. Sci. USA*, **100**, 7672–7677.
- Rebeiz,M. and Lewin,H.A. (2000) COMPASS of 47,787 cattle ESTs. *Anim. Biotechnol.*, **11**, 75–241.
- Schibler,L., Vaiman,D., Oustry,A., Giraud-Delville,C. and Cribiu,E.P. (1998) Comparative gene mapping: a fine-scale survey of chromosome rearrangements between ruminants and humans. *Genome Res.*, **8**, 901–915.
- Watanabe,T.K., Bihoreau,M.T., McCarthy,L.C., Kiguwa,S.L., Hishigaki,H., Tsuji,A., Browne,J., Yamasaki,Y., Mizoguchi-Miyakita,A. and Oga,K. (1999) A radiation hybrid map of the rat genome containing 5,255 markers. *Nat. Genet.*, **22**, 27–36.
- Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
- Womack,J.E. and Kata,S.R. (1995) Bovine genome mapping: Evolutionary inference and the power of comparative genomics. *Curr. Opin. Genet. Dev.*, **5**, 725–733.
- Yang,Y.P. and Womack,J.E. (1998) Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. *Genome Res.*, **8**, 731–736.