Transcriptome and Genome Conservation of Alternative Splicing Events in Humans and Mice

C.W. Sugnet, W.J. Kent, M. Ares Jr., and D. Haussler

Pacific Symposium on Biocomputing 9:66-77(2004)

TRANSCRIPTOME AND GENOME CONSERVATION OF ALTERNATIVE SPLICING EVENTS IN HUMANS AND MICE

C.W. SUGNET, W.J. KENT
Center for Biomolecular Science and Engineering,
University of California, Santa Cruz
CA 95064, USA

M. ARES JR.
Department of Molecular, Cell, and Developmental Biology,
University of California, Santa Cruz
CA 95064, USA

D. HAUSSLER
Howard Hughes Medical Institute;
Department of Biomolecular Engineering,
University of California Santa Cruz
CA 95064, USA

Abstract

Combining mRNA and EST data in splicing graphs with whole genome alignments, we discover alternative splicing events that are conserved in both human and mouse transcriptomes. 1,964 of 19,156 (10%) loci examined contain one or more such alternative splicing events, with 2,698 total events. These events represent a lower bound on the amount of alternative splicing in the human genome. Also, as these alternative splicing events are conserved between the human and mouse transcriptomes they should be enriched for functionally significant alternative splicing events, free from much of the noise found in the EST libraries. Further classification of these alternative splicing events reveals that 1,037 (38.4%) are due to exon skipping, 497 (18.4%) are due to alternative $3'$ splice sites, 214 (7.9%) are due to alternative $5'$ splice sites, 75 (2.8%) are due to intron retention and the other 875 (32.4%) are due to other, more complicated, alternative splicing events. In addition, genomic sequences nearby these alternative splicing events display increased sequence conservation. Both the alternatively spliced exons and the proximal intron show increased levels of genomic conservation relative to constitutively spliced exons. For exon skipping events both intron regions flanking the exon are conserved while for alternative $5'$ and $3'$ splicing events the conservation is greater near the alternative splice site.

1 Introduction

Researchers have been using mRNAs to study alternative splicing for decades [1]. Large mRNA and EST sequencing projects [2] and the recent sequencing of both the human [3] and mouse [4] genomes have facilitated a number of computational surveys of alternative splicing [6,7,8,9,10,11]. Many genes have been predicted to be alternatively spliced using computational methods. However, the poor quality of ESTs makes it difficult to distinguish functionally significant alternative splicing, aberrant transcripts from cell lines, cancers, incomplete splicing, chimeric clones, etc. that have made their way into the databases. Additionally, it is not clear that every observed transcript necessarily encodes a functional product. In fact, transcription and splicing are sufficiently error-prone processes that pathways such as nonsense mediated decay have evolved to find and degrade errors that do occur [12,13].

Both individual exons and larger gene structures are very similar in both human and mouse [14]. Recently researchers have begun to use comparative genomics [15,16] to look for alternative splicing events that are conserved in both human and mouse, and thus more likely to be biologically significant to the organisms. If an alternative splicing event is conserved between these two transcriptomes then it is likely it provides some advantage to the organism. The work presented here extends the work using comparative genomics by Thanaraj et al.[15] and Sorek and Ast [16] by:

1. Using novel whole genome alignments [17] of the human and mouse genomes to find large numbers of high confidence orthologous loci for comparison.
2. Classifying alternative splicing events and analyzing them separately according to the four distinct classes of alternative splicing displayed in Figure 1. This is accomplished by examining the graph topology of the splicing graphs like the one illustrated in Figure 2.
3. Examining the patterns of conservation for different classes of splicing events separately.

Examining the 19,156 human loci for which we could identify an orthologous mouse locus we find that 1,964 (10%) contain at least one alternative splicing event expressed in both organisms. The different classes of alternative splicing vary in relative abundance, with the most common being exon skipping, and the rarest being intron retention. In addition to being conserved in both the human and mouse transcriptomes, the alternatively spliced exons and flanking intronic regions are more conserved in the genome than constitutively expressed exons. The high level of genomic conservation might indicate the presence of *cis*-elements that help regulate alternative splicing.

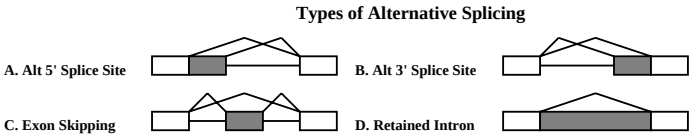

Figure 1: **Types of Alternative splicing.** Four basic classes of alternative splicing are presented. All classes are presented from $5' \rightarrow 3'$. Constitutive exons are white boxes, alternatively spliced regions are gray boxes and splice junctions are represented by arcs joining boxes. **A.** Alternative $5'$ exon with two possible $5'$ splice sites connecting to next exon. **B.** Alternative $3'$ exon with two possible $3'$ splice sites connecting to upstream exon. **C.** Exon skipping event where entire exon is included or excluded from final transcript. **D.** Retained intron event where intron is not always spliced out of final transcript.

2 Methods

Aligning mRNAs and ESTs to the genome provides both the exon-intron boundaries for a particular gene and the order and orientation of the exons. Once the order and orientation of exons is known, the $5' \rightarrow 3'$ directionality of transcription is naturally modeled using a directed acyclic graph (DAG). When looking for alternative splicing, it is useful to create a DAG where the vertices are the $5'$ and $3'$ splice sites (ss), start and end of spliced intron respectively, and the edges of this graph represent the exons and introns. Whether an edge is an intron or exon depends on whether the edge is a $5'$ss$\rightarrow 3'$ss (intron) edge or a $3'$ss$\rightarrow 5'$ss (exon) edge. Alternative splicing events are then easily found by looking for splice site vertices that connect to more than one other splice site.

In order to construct splicing graphs, we have written a program called `altSplice` which uses mRNA and EST evidence to construct splicing graphs. Only spliced mRNAs and ESTs with consensus splice sites are used, as they tend to be of higher quality and can be oriented by examining the splice sites in genomic sequence. Human and mouse splicing graphs are constructed independently for each organism, using only ESTs and mRNAs native to that particular organism. For the experiments described in this paper the human genome version NCBI Build33 and the mouse genome version NBCI Build30 were used.

2.1 Constructing Splicing Graphs.

The algorithm implemented by `altSplice` is as follows:

- Align mRNAs and ESTs to the genomic sequence using BLAT [18]. A near best in genome filter is applied where only alignments with

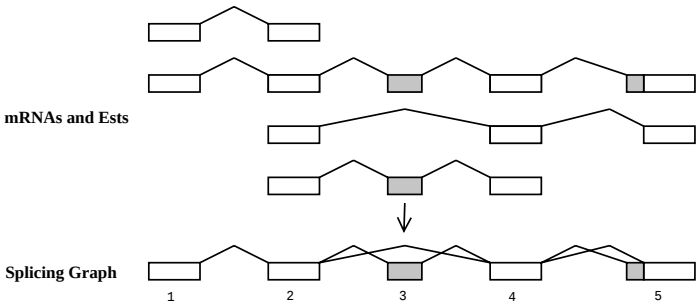

Figure 2: **Assembling mRNA and ESTs using `altSplice`.** Splice sites are determined from exons with consensus splice sites as mRNAs and ESTs are built up to form larger splicing graphs that may contain alternative splicing. Regions that are alternatively spliced are shaded gray. Note that while the exons and introns are represented explicitly and the splice sites implicitly, the splice sites are actually the vertices in the graph.

97% identity over 90% of the transcript and with a score no more than .5% lower than the best score are kept.

- Retrieve genomic sequence and use it to orient ESTs using consensus splice sites, GT→AG, and the less common GC→AG.

- Cluster alignments together by sequence overlap in exons. As new splice sites are discovered, they are entered into the graphs as vertices, and the exons and introns connecting them are recorded as edges. This graph is built into a DAG data type called `geneGraph`. Each graph is considered to be a single locus, although they may be fragments of an actual gene structure. The supporting mRNA and EST accessions for each edge are also stored.

- Extend truncated transcripts by overlap with other transcripts to the next consensus splice site. This avoids keeping vertices in the graph that are not true splice sites.

- Convert `geneGraph` records to `altGraphX` structure, which is more compact, for storage.

A visualization of the construction process of `altSplice` can be seen in Figure 2. Also, the splicing graphs are browseable interactively in the UCSC Human Genome Browser [5]. An example of the display on the browser for the ARVCF gene can be seen in Figure 3.

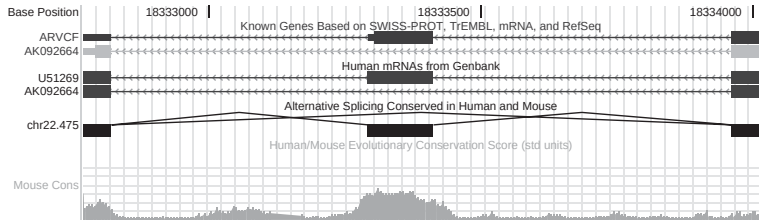

Figure 3: **Example of exon skipping from ARVCF gene as seen in the UCSC Human Genome Browser.** Transcripts can be seen which both contain and skip the exon centered in the "Human mRNA" track and visualized in the "Alternative Splicing" track. High levels of conservation are found both upstream and downstream introns in addition to the coding region of the exon as seen in the "Mouse Cons" track. The Mouse Cons[4] track shows the similarity between this region and the orthologous region in mouse.

2.2 Comparing Orthologous Splicing Graphs.

Once the splicing graphs have been generated separately for both the human and mouse genomes, orthologous graphs are found using large scale genomic alignments provided by Kent et al.[17]. Briefly, the human and mouse genomes are divided into segments and aligned all against all using `BLASTZ`[19]. The resulting alignments are then chained together into larger structures. The chaining algorithm requires that the order of aligned blocks within the chain must be consistent with the genomic sequence order in both species alignment of the two genomes. This is equivalent to preferring the alignments that increase the synteny between the two genomes. In order to model inversions and duplications, there can be large gaps in the chains, and other chains are allowed to nest inside the gaps. For the purposes of determining orthologous regions, we use only the maximally scoring chain for a region to map between human and mouse genomes. This is analogous to using a base pair resolution synteny map of the two genomes which allows us to map with confidence to orthologous regions from human to mouse. These maximally scoring chains are referred to as "nets" and can be found on the UCSC Human Genome Browser as the "Mouse Net" track[17].

We have written a program to analyze the splicing graphs from orthologous loci called `orthoSplice` which implements the following algorithm:

- Inputs are altGraphX records for two genomes and chains to map between those genomes.

- For each altGraphX record on the human genome look up the orthologous altGraphX records in the mouse genome via the maximal chain for that region.
- Using the chains, create a mapping between the splice sites (vertices) of the two altGraphX records. Then use this mapping to compare the actual splice graphs for the two records.
- Examine properties of the common splicing subgraph including conserved exons, introns, and alternative splicing. Only alternative splicing involving internal exons is considered, not alternative promoters or polyadenlylation sites.
- Output the subset of the human altGraphX record that was also observed in the mouse altGraphX record. Report results locus by locus and also edge by edge in the graph.

The resulting subset of the human splicing graphs that was also conserved in the mouse transcriptome were examined to discover and classify alternative splicing events. Representing splicing as a graph facilitates this process, as it is relatively straightforward to examine the graph topology for patterns that correspond to functional classes of alternative splicing illustrated in Figure 1. Additionally, exons that are constitutively expressed are recorded and used for controls to examine the upstream and downstream intronic regions for conservation.

2.3 Calculating Conservation Per Base

If alternative splicing events are biologically significant to an organism, it is reasonable to hypothesize that these events would be regulated at the sequence level, and that those regulatory sequences would be conserved between human and mouse genomes. To investigate the conservation in the genomic sequences we examined individual bases adjacent to splice sites and calculated the percentage of times they were conserved when alignable to mouse. Percent identity was calculated using only bases that were aligned; inserts and deletions were excluded from the calculation. This is a more conservative measurement than counting unaligned bases as non-conserved because bases may not be aligned due to other factors such as the draft nature of the mouse genome. Alignments used were the same chains of `BLASTZ` alignments that were used to find the orthologous gene structures. This analysis resulted in a per base conservation profile for each class of alternative splicing event.

3 Results

Even requiring that alternative splicing be observed independently in both human and mouse transcripts we find that 10% (1,964) of the

Table 1: **Relative abundance and size of human alternatively spliced regions conserved in mouse transcriptome.** The alternatively spliced regions are those that are shaded in Figure 1. For exon skipping events the alternatively spliced region corresponds to the entire exon. For alternative $5'$ and $3'$ events the alternative spliced region is the area between the two alternative splice sites, excluding regions of length 3bp for alternative $3'$ events. Size of retained introns is the length of the intron that can be spliced out.

Class of Alt. Splice	Number	Percentage	Mean±Sd (Med)
Alt. $5'$	214	7.9%	44.1±63 (21)
Alt. $3'$	497	18.4%	51.4±174 (18)
Exon Skipping	1,037	38.4%	104±140 (84)
Intron Retention	75	2.8%	220.3±272.0 (110)
Other	875	32.4%	(NA)
Constitutive Exons	113,549	NA	140±125 (122)

19,156 loci for which we could find a mouse ortholog had alternative splicing. We discovered 2,698 different alternative splicing events that were conserved between human and mouse.

Other studies of alternative splicing using mRNAs and ESTs have reported alternative splicing in the human transcriptome to range from 35–55% [6,8,9,11]. The 2,698 splicing events conserved between the human and mouse transcriptomes reported here represent a lower bound on the number of alternative events that are present in the human transcriptome and should be enriched for events that are biologically significant. The requirement that human alternative splicing events be conserved in the mouse transcriptome is stringent, and dependent on the depth of both human and mouse transcript libraries. There are many human alternative splicing events that are probably functionally significant and present in the mouse transcriptome, but for which no transcript has yet been sequenced.

As seen in Figure 4, exons and splice junctions that are conserved tend to have more representative transcripts. This is to be expected as highly expressed genes are more likely to be included in the EST libraries and thus found in both the human and mouse transcriptomes. After four transcripts are observed to contain a splice junction or exon, it is twice as likely to be from the conserved distribution as from the non-conserved distribution. By adding back human splice junctions and exons that are not conserved in mouse, but are observed in 4 or more human transcripts, we find 4,528 (22.3%) alternatively spliced loci out of 19,945 total loci in the human transcriptome.

By relaxing our requirement for inclusion to single transcript coverage, even if not conserved in mouse, we find that 11,929 (37.6%) loci are

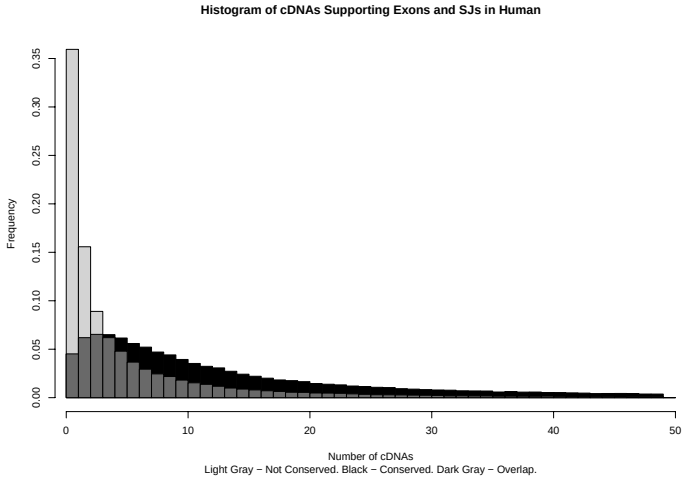

Figure 4: **Histogram of transcripts supporting conserved and not conserved splice junctions and exons.** Normalized frequency of number of transcripts that contain a given exon or splice junction for conserved (black), not conserved (light gray) and the overlap (dark gray) of exons and splice junctions examined. In general splice junctions and exons that are conserved between human and mouse transcriptome have more transcript coverage than those that are not conserved.

alternatively spliced out of 31,752 loci total. Some of these events could be conserved or could represent human-specific splicing events. It may be interesting in future studies to look at alternative splicing events that are seen in multiple transcripts but are not conserved in other organisms and may be enriched for species specific alternative splicing. Resolving whether an exon does not exist in mouse or is simply not yet represented in the transcript databases will require further work. Evolutionary implications of highly expressed alternative splicing events that are not conserved is explored further in Modrek et al [20].

Most of the conserved alternative splicing events can be described as falling into the four classes of alternative splicing described in Figure 1 and summarized in Table 1. The size distributions of the alternatively spliced regions (gray areas in Figure 1) differs for each of the separate classes (Table 1). Alternative $3'$ and $5'$ events have, on average, a much smaller number of base pairs alternatively spliced than the skipped exons. Also, the skipped exons are shorter than the constitutively spliced exons, consistent with a report that considered human transcripts only [11].

Further examination of the alternative $3'$ splicing events revealed al-

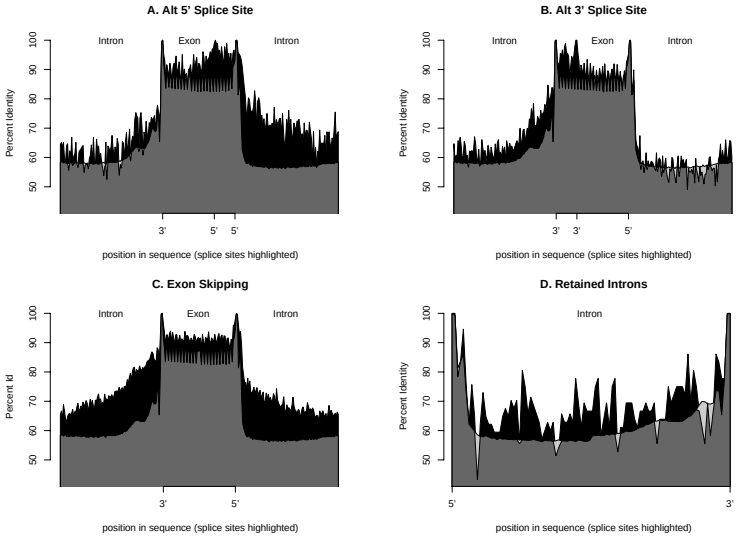

Figure 5: **High genomic conservation in alternatively spliced regions.** Alternatively spliced regions and the intronic sequences proximal to them exhibit high levels of genomic conservation between mouse and human. Average base identity for aligned bases is presented for generic representatives of alternative splicing classes described in Figure 1. Conservation for alternatively spliced regions is shown in black. Conservation for constitutively spliced regions is filled in with light gray, overlaps between the two are illustrated in dark gray. Going from left to right in panels A-C it is possible to observe the conservation of the polypyrimidine track, the $3'$ splice site, the coding exon, the $5'$ splice site for both constitutive (gray) and alternative (black) exons. The splice sites are marked for each alternatively spliced exon. **Individual Panels: A.** Alternative $5'$ event: Regions illustrated are 100bp into upstream intron, 25 bp into exon from $3'$ splice site, 25bp upstream from first $5'$ splice site, 20bp upstream from second $5'$ splice site, and 100bp downstream from second $5'$ splice site. Only data from $5'$ events with more than 20bp presented. **B.** Alternative $3'$ event: Regions illustrated are 100bp upstream from first $3'$ splice site, 20bp downstream of the first $3'$ splice site, 25bp downstream from the second $3'$ splice site, 25bp upstream from the $5'$ splice site, and 100bp into the downstream intron. Only data from $3'$ events with more than 20bp presented. **C.** Exon skipping event: Regions presented are 100bp upstream from $3'$ splice site, first 35bp of exon, last 35bp of exon, and 100bp downstream. **D.** Retained intron event: First and last 100bp of retained intron compared to introns proximal to constitutive exons.

most half (45.4%) had only 3 nucleotides separating one alternatively spliced 3′ site from the other. The fact that many of these very small alternative splices have multiple transcripts supporting both isoforms, and don't disrupt the coding frame, indicates that many of these alternative splices are real rather than an artifact. However, while they appear to be biologically real, it is not clear that they have any functional effect on the resulting protein. For the analysis shown in Figure 5B these 3bp splicing events were not included.

3.1 Conservation of Genomic Sequences Near Alternative Splicing Events.

As previously reported [16], using a smaller set of skipped exons, the upstream and downstream flanking intronic regions of exon skipping events are conserved relative to constitutively expressed exons. Using the whole genome alignments it is possible to calculate a percent identity at positions relative to the 3′ and 5′ splice sites for our set of skipped exons (Figure 5C). The 50bp upstream and downstream flanking intronic regions of exon skipping events have an average percent identity of 80% and 75% respectively while the average percent identity of the constitutive exons is 65% and 61% respectively. The increased conservation of flanking exon skipping events is consistent with that reported by Sorek et al [16] although the conservation calculation is slightly different.

Alternative 5′ and 3′ splicing events exhibit higher conservation in the proximal flanking intronic sequence, but not as much in the distal flanking exon. For alternative 5′ splicing events there is more conservation in the 50bp of flanking downstream intron (77%) than in the 50bp of flanking upstream intron (69%). The polypyrimidine tract is well conserved (65% vs 61%) even in constitutive exons (Figure 5A). The proximal intron is also better conserved near alternative 3′ splice site (Figure 5B) with the last 50bp of upstream intron (72%) greater than the downstream first 50bp of intron (62%).

It is also interesting to note that the regions of exons that are alternatively spliced are also better conserved than constitutive exons. The first and last 20bp of the skipped exons have an average percent identity of 92% while the consituative exons have a percent identity of 87%. The same is also true of the first 20bp of the regions spliced out by alternative 5′ (96%) and alternative 3′ (95%) splicing events. Such high levels of conservation suggest the presence of regulatory motifs within the exon.

While the retained introns examined did have more conservation than introns flanking constitutively expressed exons, they do not appear to have conservation levels near those of the constitutive exon sequences themselves. It is interesting to note that the median size for the retained introns of 110bp is much smaller than normal introns. Further

examination will have to be done to determine if these retained introns have a function.

4 Conclusions

By Examining alternative splicing events that are conserved in both the human and mouse transcriptomes, we have generated a set likely to be enriched for those that confer a selective advantage via some biologically significant function. Thus, this set contains minimal amounts of alternative splicing that may due to aberrant transcription or splicing.

We have shown that genomic regions nearby these alternatively spliced sequences are highly conserved. The high levels of conservation in the introns proximal to these alternative splicing events, as well as within the alternatively spliced regions implies that there are *cis*-elements present that have been conserved. It is possible that these *cis*-elements are necessary for the regulation of alternative splicing events and have been selected for in evolution. Future work will involve both computational efforts to identify regulatory elements and experiments at the bench to profile alternative splicing events.

5 Availability of Data and Programs

The subset of human splicing graphs conserved in the mouse transcriptome is browseable interactively at the UCSC Genome Browser [5], as well as downloadable in bulk. A special entry point to the browser with the list of alternatively spliced regions can be found at: `http://www.soe.ucsc.edu/~sugnet/psb2004/altGraphXCon.html` The source code for the `altSplice` and `orthoSplice` programs can be found under Jim Kent's CVS source tree under kent/src/hg/altSplice. The source tree is available at: `http://www.soe.ucsc.edu/~kent/src/.`

Acknowledgments

We would like to thank the International Human Genome Sequencing Consortium and the Mouse Genome Sequencing Consortium for providing the genomic sequence data. We would also like to thank the researchers who have contributed their cDNA sequences to Genbank. C. Sugnet is a Howard Hughes Medical Institute Predoctoral Fellow. W.J. Kent, M. Ares and D. Haussler were supported by NGHRI grant 1P41H. D. Haussler was also supported by the Howard Hughes Medical Institute.

References

1. J.W. Tamkun, J.E. Schwarzbauer, and R.O. Hynes, *PNAS.* **81**, 16 (1984).
2. M.S. Boguski, T.M. Lowe, and C.M. Tolstoshev, *Nat. Genetics* **4**, 4 (1993).
3. E.S Lander et. al., *Nature.* **409**, 6822 (2001).
4. R.H. Waterson et al., *Nature.* **420**, 6915 (2002).
5. W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler, *Genome Res.* **12**, 6 (2002).
6. A.A. Mironov, J.W. Fickett, and M.S. Gelfand, *Genome Res.* **9**, 12 (1999).
7. D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, and P. Bork, *FEBS Lett.* **474**, 1 (2000).
8. Z. Kan, E.C. Rouchka, W.R. Gish, and D.J. States, *Genome Res.* **11**, 5 (2001).
9. B. Modrek, A. Resch, C. Grasso and C. Lee, *Nucl. Acids Res.* **29**, 13 (2001).
10. Q. Xu, B.Modrek and C. Lee, *Nucl. Acids Res.* **30**, 17 (2002).
11. F. Clark and T. A. Thanaraj *Human Mol. Gen.* **11**, 4 (2002).
12. C. Gonzalez, A. Bhattacharyaa, W. Wanga and S.W. Peltz, *Gene.* **274**, 1-2 (2001).
13. B. Lewis, R. Green, and S. Brenner, *PNAS.* **100**, 1 (2003).
14. S. Batzoglou, L. Pachter, J.P. Mesirov, B. Berger, and E.S. Lander, *Genome Res.* **10**, 7 (2000)
15. T. Thanaraj, F. Clark and J. Muilu, *Nucl. Acids Res.* **31**, 10 (2003).
16. R. Sorek and G. Ast, *Genome Res.* **13**, 7 (2003).
17. W.J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, *PNAS **in press**.*, (2003).
18. W.J. Kent, *Genome Res.* **12**, 4 (2002).
19. S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler and W. Miller *Genome Res.* **13**, 1 (2003).
20. B. Modrek and C. Lee, *Nat. Genetics* **34**, 2 (2003).