Into the Heart of Darkness: Large Scale Clustering of Human Non-Coding DNA

Gill Bejerano[a], David Haussler[a b], Mathieu Blanchette[c]

[a] *Center for Biomolecular Science and Engineering, Baskin School of Engineering University of California in Santa Cruz, 1156 High St., Santa Cruz, CA 95064* [b] *Howard Hughes Medical Institute* [c] *McGill Center for Bioinformatics, School of Computer Science, McGill University St., 3775 University, Montreal, QC, Canada, H3A 2B4*

ABSTRACT

Motivation: It is currently believed that the human genome contains about twice as much non-coding functional regions as it does protein-coding genes, yet our understanding of these regions is very limited.

Results: We examine the intersection between syntenically conserved sequences in the human, mouse and rat genomes, and sequence similarities within the human genome itself, in search of families of non protein coding elements. For this purpose we develop a graph theoretic clustering algorithm, akin to the highly successful methods used in elucidating protein sequence family relationships.

The algorithm is applied to a highly filtered set of about 700 000 human-rodent evolutionarily conserved regions, not resembling any known coding sequence, which encompasses 3.7% of the human genome. From these, we obtain roughly 12 000 non-singleton clusters, dense in significant sequence similarities. Further analysis of genomic location, evidence of transcription, and RNA secondary structure reveals many clusters to be significantly homogeneous in one or more characteristics. This subset of the highly conserved non protein-coding elements in the human genome thus contains rich family-like structures, which merit in-depth analysis.

Availability: Supplementary material to this work is available at http://www.soe.ucsc.edu/~jill/dark.html

**Contact:** jill@soe.ucsc.edu

1 INTRODUCTION

It has been estimated that at least 5% of the human genome is under purifying selection and thus likely to be functional [21, 23, 5]. By far the best studied functional features of the genome are protein-coding genes, but their coding exons account for only about 1.5% of the genome (2 % if UTRs are included) [13]. The remaining 3-3.5%, the dark matter of the human genome, is likely to contain mainly gene regulatory regions (both transcriptional and splicing), RNA genes and micro-RNAs, matrix attachment sites, origins of replication (all of which are reviewed by Mattick [19]), and perhaps some altogether novel functional elements (remember that micro-RNAs were unknown just a few years ago!). Efforts are underway to provide a functional annotation for non-coding elements but databases of experimentally verified loci like Transfac [20] for regulatory regions or RFAM [10] for RNA genes, contain information about only a tiny fraction of the regions under discussion.

Comparative genomics has proven to be a powerful approach for locating functional loci by identifying regions of the genome that show a significant degree of conservation in other species. Many published analyses focus on human-mouse comparisons (e.g., [6, 7] and references therein) but more recent works utilize newly available sequences, mostly from multiple mammals ([4, 18, 24] and others). Unfortunately, measuring conservation levels is of little help by itself for assigning a putative function to these phylogenetically conserved regions. Computational functional annotation of non-coding conserved elements is thus an acute bioinformatic challenge with extremely important applications.

The majority of the genes in the human genome has been initially annotated by sequence homology to genes, in human or other organisms, about which more was known at the time. Tools like psi-Blast [1] have been developed to detect remote homologs for a given protein sequence, and have resulted in a significant improvement of our understanding of gene functions. Based on these tools, various clustering algorithms have been developed for grouping together proteins with similar domains (e.g., [15, 17, 8]). This hierarchy of found relationships between the known proteins is curated in various database, such as InterPro [22].

Annotation by homology has only recently been applied, at a small scale, to putative non-coding functional elements. In an analysis of the CFTR region [18], it was found that most of the regions of interest appeared to be unique in the human genome (based on Blast similarity searches), and thus homology searches within the genome added new information only in a few cases. This may be because the homology search tools used are not capturing properly the type of sequence similarity most relevant for non-coding regions. It may also

Gill Bejerano[a], *David Haussler*[a, b], *Mathieu Blanchette*[c]
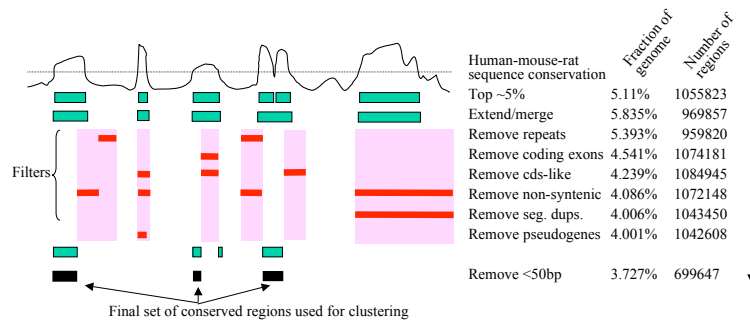
Fig. 1. Definition of the conserved non-coding regions to be clustered. Starting from the 5% most conserved sequences with respect to mouse and rat, the number of regions and their coverage of the human genome is given after each masking operation.

be because the function of some of these regions is genuinely unique in the genome. Still, this general approach has allowed the classification of some RNA genes and regulatory elements (e.g., [10], [28]).

Here, a first step is proposed to provide genome-wide classification of conserved non-coding regions of the human genome by homology. We start by comparing the human genome to the mouse and rat genomes, using stringent filters to remove many annotated regions (such as genes, pseudo-genes, repeats, etc.) to identify roughly 700 000 regions of high conservation, dissimilar to any known coding sequences, covering approximately 3.75% of the human genome. It is then shown that even using a simple sequence similarity measure (the standard affine-gap local sequence alignment method), it is possible to cluster regions with similar sequences, and thus possibly similar function. The many clusters identified have a number of interesting properties that hint at a variety of possible functions: some contain a hundred or more highly similar regions, others are located near genes of a particular family; are located predominantly in introns; or contain known or predicted structural RNA genes, etc. It is our belief that this approach is a first step in establishing a genome-wide annotation pipeline focusing on non-coding functional regions.

2 METHODS

We start by identifying a set of putative functional non-coding regions by detecting portions of the human genome that share significant similarity with their syntenic homologs in mouse and rat. To cluster these regions, we define a similarity graph $G = (V, E)$ whose vertices V are this set of human conserved regions and whose edges E are the pairs of regions that share significant sequence similarity within human. We then define a new algorithm for detecting dense clusters in this type of graph and apply it to obtain clusters of highly similar, phylogenetically conserved regions of the human genome. Finally, the clusters identified are evaluated for enrichment for an array of attributes pointing to interesting putative functions.

2.1 Defining Conserved Elements

The process of defining the non-coding conserved regions to be analyzed in this study is summarized in Figure 1. To detect regions of the human genome that are likely to be functional, we identify portions that are highly conserved with respect to their mouse and rat orthologs. A three-way multiple alignment between the genomes (NCBI human Build 34, NCBI mouse Build 32, and Baylor rat assembly version 3.1), produced by the HUMOR program (W. Miller, available at http://bio.cse.psu.edu/) was obtained from the UCSC genome browser (http://genome.ucsc.edu/), to establish orthology between the three genomes. Some 40% of the human genome is thus aligned to regions in mouse and/or rat.

The alignment was scanned with a 50bp sliding window and the conservation of each window was evaluated using a method that calculates a p-value for the degree of conservation observed, under a null model of neutral evolution, taking into consideration the phylogenetic relationships among the species considered [18]. A conservation threshold was chosen so that 5% of the whole human genome, the current estimate for functional sequences in the genome, was marked as conserved, which resulted in a set of 1055823 regions of average size 140bp. About 74% of all bases in coding exons of known genes (as defined in the knownGene annotation in [14])) are within these regions, although they account for less than 13% of the combined length of these regions (17% if UTRs are included).

In order to avoid fragmenting functional units into several conserved regions, we extended each region by 10bp on each side. We then applied a set of filters to ensure that the conserved regions retained are syntenic in mouse and rat (and thus more likely to be from alignments of orthologous DNA) and highly likely to be non-coding. The regions masked out of further consideration included various known types of repeats [27, 3], as well as coding exons from several sources, consolidated in the UCSC human genome browser [14]. From the remaining set we removed all unannotated bases with detected similarity to the known coding exons, using the sensitive
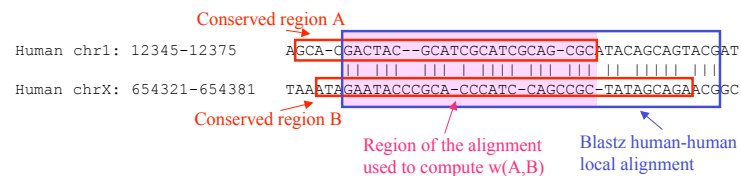
Fig. 2. Scoring the similarity between two conserved regions using a Blastz human-human local alignment. The score $w(A, B)$ of the alignment is based only on the shadowed area.

Blastz search tool[26]. Finally, to ensure that the regions used for the clustering were not the result of a primate-specific duplication, we eliminated all regions outside of a high quality synteny net to mouse [16], as well as those contained in putative pseudogenes [29, 30] and in regions suspected of being recent human specific segmental duplications [2].

As each masking phase fragments the regions of interest, filtered regions less than 50bp long were also discarded. The resulting 699647 regions, which form the vertices of our similarity graph, are not known to belong to any of the above classes, or even resemble coding sequences, and yet they exhibit high syntenic conservation between the three mammals. The average conserved region obtained is 153bp long, while the longest is 3079 bp.

2.2 Measuring intra-human similarity

To identify which pairs of human conserved regions are similar (that is, to place edges in the similarity graph), we use a precomputed Blastz set of local alignments of the repeat-masked human genome against itself (available through the UCSC genome browser). The significance threshold on sequence similarity was set very high to avoid too many false-positives.

A pair (u, v) of human regions can only be connected by an edge if a consecutive block of 15 alignment positions or more is found between u and v by means of a Blastz local alignment. Let $s(u, v)$ be the similarity score of the part of Blastz alignment located within u and v (see Figure 2), calculated using the standard affine-gap penalty method. If the alignment is too short or of too poor quality ($s(u, v) < 0$), no edge is placed between u and v. Otherwise u and v are connected by an edge of weight $w(u, v) = s(u, v)$. [1]

2.3 Identifying clusters

Similar to the result reported by Margulies *et al.* for a 1.8 Mb region around the CFTR gene [18], genome-wide we find that the large majority of conserved regions appear to be unique in human, at least based on Blastz alignments. About 96% of

the 699647 vertices of the similarity graph are not connected to any at other vertex. Nonetheless, this leaves 29349 regions similar to at least one other in the human genome. The graph contains 8333 connected components, 1446 of which are of size at least three vertices and 257 of size at least 10. The largest connected component has 823 vertices and 1673 edges.

The connected components of the similarity graph constructed constitute a first approximation to the clusters sought. They correspond to the clusters that would be produced by a single-linkage clustering algorithm. However, these connected components are often quite loose and may contain more than one dense cluster.

The problem of clustering a similarity graph to identify a dense subgraph has been studied extensively in the case where the vertices of the graph are proteins (e.g., [15, 17, 8]). It was noted that in that context, simply taking for clusters the connected component of the graph was inadequate because: (i) false-positive edges tend to collapse two dense clusters into a single large connected component, and (ii) multi-domain proteins tend to be in several different clusters, again collapsing them into one connected component. The same two problems occur with our non-coding regions: (i) false-positive edges are possible, and (ii) conserved regions made of two different but adjacent functional units play the same role as multi-domain proteins by connecting unrelated clusters. Approaches proposed to handle this situation include iteratively remove minimum-weight cuts in the graph [15]. Others rely on the identification of biconnected components and articulation points [17] or use a multi-stage approach [8]. The approach we use here is a heuristic that borrows from all three of the above approaches. To refine each connected component, we define a vertex partitioning operation and a vertex duplication operation that, when applied recursively on a connected component, yield a set of dense, edge-disjoint subgraphs.

Recall that a cut of a weighted graph $G = (V, E, w)$ is a partition of the vertices into two disjoint non-empty subsets A and B, with $A \cup B = V$. The weight of a cut (A, B) is $\sum_{(u,v) \in E, u \in A, v \in B} w(u, v)$. A low-weight cut of the graph thus separates a set of regions into two groups with little similarity between them. We are going to use minimum-weight cuts to detect false-positive edges and eliminate them.

---

[1] There may actually be two different Blastz alignments between regions u and v, one with each sequence as reference. In that case, $w(u, v)$ is defined as the maximal scores of the two alignments.

Gill Bejerano[a], *David Haussler*[a][b], *Mathieu Blanchette*[c]
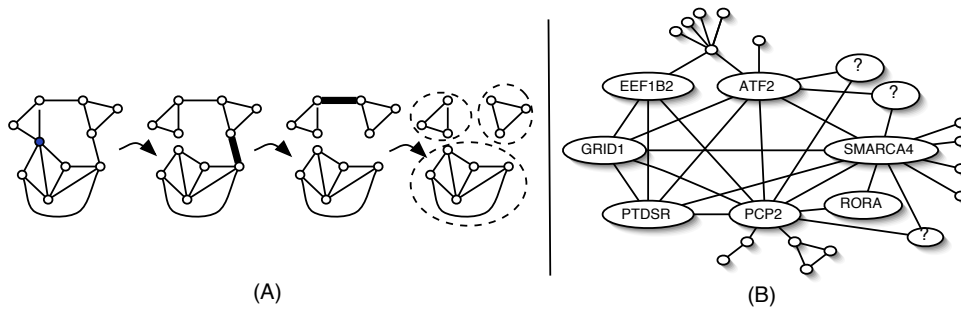
Fig. 3. A. Identification of dense subgraphs by our heuristic. Assuming all edges have weight 1, $\delta_c = 1$ and $\delta_a = 0$, the original graph has no cut of cost less than 2 but has a vertex with local-articulation score zero. This vertex is first duplicated. The resulting graph has two cuts of cost one, whose edges are removed. The resulting graph has three dense connected components. B. Example of an actual cluster (ID 652.29, see text for details). Small vertices were those removed by the algorithm.

Two approaches are used to detect and break-up multi-functional regions. First, to break-up a putative such region u, the Blastz local alignments between u and all other regions it connects to in the graph are mapped on u's sequence. If the alignments stack-up in two or more disjoint portions of u, the region u is divided into its non-overlapping portions. This is sometimes not sufficient to break all multi-functional regions and we introduce the notion of local-articulation point to handle more difficult cases. We define the local-articulation score of a vertex v as follows. Let $N(v)$ be the set of neighbors of v (excluding v itself), let $G|_X$ be the subgraph spanned by a subset of vertices X, and let $C = (A, B)$ be a minimum-weight cut of the induced subgraph $G|_{N(v)}$ spanned by the vertices of $N(v)$ (with $N(v) = A \cup B$). Then, we define local-articulation$(v) = weight(C)/|N(v)|$. In other words, vertex v will have a low local-articulation score if, when ignored, its neighbors can be partitioned into two sets with little similarity between them. Vertices with low local-articulation score are likely to correspond to conserved regions containing more than one functional unit. When such a vertex v is found, with a minimum weight cut $C = (A, B)$, it is duplicated and one copy is connected to the regions in A while the other is connected to the regions in B (see Figure 3). This approach is a generalization of the simpler articulation points method used by Kim [17]. For example, in Figure 3, graph A has no good cut and no standard articulation vertex, yet the black vertex is clearly joining two different clusters and is detected as such.

To decompose a connected component into its dense clusters, the min-cut removal and local-articulation duplication operations are executed recursively on each connected component produced until the clusters left are sufficiently dense (see an example in Figure 3). Here we use two heuristic Blastz score thresholds $\delta_c = 2000$ below which a cut is performed, and $\delta_a = 200$ below which a local-articulation vertex is duplicated. The details of the algorithm are described below.

Algorithm CUT(V, E, w)
Input: A weighted graph $G = (V, E, w)$.
Output: The minimum weight cut (A, B) of V, and its weight. Implements the Fiduccia-Mattheyses heuristic[9, 15].

Algorithm BEST-LOCAL-ARTICULATION(V, E, w)
Input: A weighted graph (V, E, w).
Output: The vertex $v \in V$ with be best local-articulation score, together with the partition (A, B) of the neighbors of v, and the weight of the cut induced.
$s_{min} \leftarrow +\infty$
for each vertex $v \in V$ **do**
    $(A, B, s) \leftarrow$CUT$(G|_{N(v)})$
    if $(s < s_{min})$ **then** $(v_{min}, A_{min}, B_{min}, s_{min}) \leftarrow (v, A, B, s)$
return $(v_{min}, A_{min}, B_{min}, s_{min})$

Algorithm GRAPH-PARTITIONING$(V, E, w, \delta_c, \delta_a)$
Input: A connected weighted graph $G = (V, E, w)$
Output: Prints a set of dense clusters of G.
$(A, B, x) \leftarrow$ CUT(V, E, w)
if $(x < \delta_c)$ **then** $E \leftarrow E - \{(u, v) \in V : u \in A, v \in B\}$
else
    $(v, A, B, y) \leftarrow$BEST-LOCAL-ARTICULATION(V, E, w)
    if $(y/|N(v)| < \delta_a)$ **then**   /*duplicate v */
        $V \leftarrow V \cup \{v'\}$ /* add vertex v' */
        $E \leftarrow E \cup \{(v', b) : b \in B\}$
        $E \leftarrow E - \{(a, b) : a \in A, b \in B\} - \{(v, b) : b \in B\}$
    else print (V', E'), **return** /* we found a dense cluster */
for each connected component (V', E') of (V, E) **do**
GRAPH-PARTITIONING(V', E', w)

Since the Fiduccia-Mattheyses heuristic [9] runs in time $O(E)$, finding the best local-articulation takes time $O(VE)$, so each iteration of graph-partitioning takes the same time. Since each partitioning iteration either removes one or more edges or duplicates a vertex, and since a vertex v can be duplicated at most $N(v)$ times, there can be at worst $O(E)$

iterations, and thus the algorithm runs in time $O(VE^2)$. In practice our 8333 connected components were partitioned in an hour on a desktop machine, with the largest fraction of the time spent on the few very large connected components.

Applying the clustering approach above yields a set of 12027 dense, homogeneous clusters whose size vary between 2 and 105 regions, with 296 clusters of size at least 5 and 84 of size at least 10. Among the 84 clusters of size at least 10, the average degree of a vertex is 6.1. Some clusters are nearly perfect cliques (e.g., ID 1758.3 has 25 vertices and 206 edges, over a possible 300), but in most case the degree of a vertex is between 4-10, irrespective of cluster size. In total, 18734 human regions are within some cluster (less than in the original graph because some vertices became singletons in the clustering process and were eliminated). The average length of the regions belonging to a dense cluster is 225 bp.

2.4   Testing significance of features shared by regions in a cluster

Assuming that members of a cluster share a common functionality, inferring a potential function for the cluster as a whole may be easier than doing so for each region individually because (i) functional annotation for one member can be mapped to other members, and (ii) statistically over-represented features shared by members may hint at function. Since very few of the members of our clusters have reliable annotation, we focus mainly on the statistical over-representation approach.

We consider the following set of boolean features of conserved regions that may help assign a putative function to the clusters:

Genomic location: For each of seven types of genomic features relative to known genes (1kb and 10kb upstream and downstream, intergenic regions, UTRs, and introns), a boolean attribute is defined on each conserved region with value 1 if the conserved region overlaps a feature of the given type and 0 otherwise.

Association to known genes. Each classification term in the Go [11] and InterPro [22] databases defines a boolean attribute. A conserved region R has value 1 for such an attribute if the closest known gene to R has that particular Go or InterPro classification, or one of its descendants in the ontology hierarchy.

Coding potential: This attribute is 1 if and only if the region overlaps a gene prediction from one of four chosen gene predictors.

Evidence of transcription: Attributes are defined for overlap with ESTs and mRNAs.

Non-coding RNAs: Attributes are defined for overlap with known RNA genes [10].

Predicted RNA secondary structure: A region has this attribute if its minimal free-energy secondary structure (computed with RNAfold [12]) is lower than that of 99% of 1000 randomly shuffled sequences with the same nucleotide composition.

Conservation in distant species: Fugu and chicken.

For each boolean attribute A from the list above, we use the set of all human conserved regions to estimate the background probability p that a given region has attribute A, except for the RNA secondary structure attribute where this would be too costly and where we set $p = 0.01$. We then obtain a p-value for the observed number of members with attribute A in a cluster of a given size, under a null model where the attribute A has value 1 independently with probability p, using the cumulative of a binomial distribution.

Since more than ten thousand clusters are to be tested, a Bonferroni type correction is necessary. Here we only report regions with uncorrected p-value below 10^{-5}.

3   RESULTS

Initial analysis of the set of clusters obtained makes it clear that we are facing a heterogenous set of clusters of a variety of classes. Table 1 lists a few of the more intriguing clusters significantly enriched for each type of features described in Section 2.4.

Considering first the overlap with known functional non-coding regions, we find 47 clusters containing exclusively members with an RNA gene annotation [10] (some of which are shown in Table 1). We find 30 unannotated regions that belong to a cluster with at least one member annotated as micro-RNA or RNA gene, suggesting a functional classification for the other members of the cluster. A subset of these novel RNA genes is currently being tested experimentally [25].

Several clusters are significantly enriched for gene predictions, and may correspond to novel protein-coding gene families. Although we have removed from consideration all known coding regions, and even went to the extent of masking all sequences that resemble even short stretches of coding exons (recall Figure 1), it is to be expected that uncharacterized gene families, if they exist, should come up in our analysis. Several other attributes reinforce the coding hypothesis. First, for many of the clusters, mRNA and EST evidence exists, attesting active transcription. Second, many of these putatively coding clusters are also conserved in chicken, further suggesting functional importance. Finally, in many cases the boundaries of the gene predictions, obtained through a conceptually different approach, match closely those of our conserved regions. Although Blastz detects no DNA sequence similarity between these regions and any known coding exons, there are a few clusters for which a more sensitive tBlastn search reveals some weak protein similarity to know genes (e.g., cluster 3089.3 in Table 1).

Several clusters are highly enriched for regions conserved in chicken, and sometimes all the way back to fugu. Besides those with good coding potential described above, many

Gill Bejerano[a], *David Haussler*[a, b], *Mathieu Blanchette*[c]

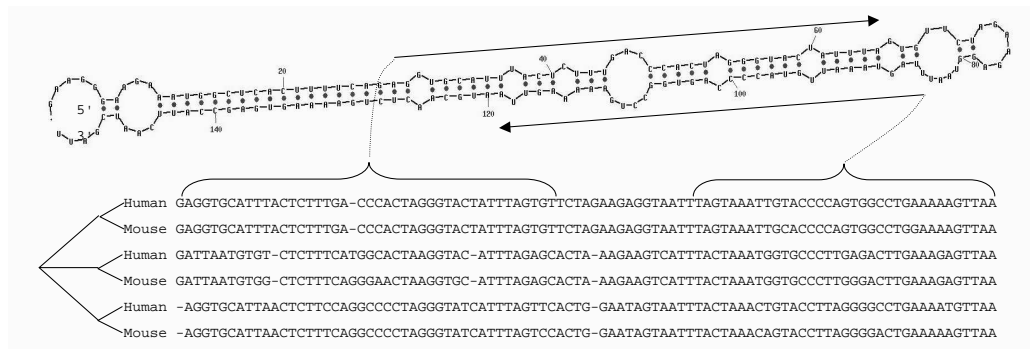| Attribute	cluster ID	#v	#Att	P-value	Comment
RNA genes	5390.1	6	6	9.7e-22	Hu-U71b snoRNAs
	2483.22	9	4	1.2e-12	miRNA mir-154. Also detected by RNA sec. struct. p-value screening.
	41 others			<1.6e-08	various RNAs and miRNAs
Chicken conservation	14.381	59	38	3.7e-13	No conservation in Fugu
	156.175	16	15	6.3e-10	Many matches to chicken EST
	1730.12	13	11	1.4e-6	5 regions have coding potential (pvalue 4.9e-4)
	2003.3	19	15	8.1e-8	10 regions have coding potential (pvalue 1.8e-8). 8 regions have RNA sec. struct (pvalue 7.2e-13)
Fugu conservation	4415.3	5	5	7.9e-11	just 5' of exons of SCNxA gene family (pvalue 8.4e-6), all are conserved in chicken (pvalue 3.7e-4)
	4290.2	4	4	8.3e-9	3' end of 5'UTR of histone H1 family
	4787.3	4	4	8.3e-9	Downstream of alt. splices exons of the NEB gene
	5602.2	4	4	8.3e-9	All are predicted genes with EST evidence
	855.1	4	4	8.3e-9	All have strong RNA sec. str (pvalue 1e-8)
	24 others			<8.6e-07	
ESTs	652.29	10,21	6	9.7e-7	6 sites are less than 1kb downstream of exons of various genes. See Figure 3 (B).
upstream	6137.8	11	10	2.6e-17	5' of genes of the ALEX family. Many other clusters are associated with the same family.
	6895.5	5	4	4.4e-7	Just 5' of genes of the PCDHB family
	1848.5	4	4	4.4e-7	Just 5' of genes of the KRTHA family
	4982.2	5	5	2.8e-7	5'UTR of genes of the SCNxA family. Many other clusters are associated with the same family
	5105.1	5	4	4.4e-7	5'UTR of genes of the GRYD family
	4 other clusters			< 5.2e-6	Various gene families
1kb intron flanks	6898.2	12	11	7.5e-11	Downstream of alternative first exons of PCDHG family. Many other clusters are associated with the same family.
	4969.6	12	9	1.2e-7	Upstream of repetitive exons of TTN
Gene predictions	7708.1	15	15	1.8e-19	Consecutive regions contained in a 12kb ORF upstream of c2orf16
	5011.6	5	5	5.6e-7	Consecutive regions contained in a 5kb ORF upstream of AK126051.
	3089.3	5	5	3.1e-8	Similar to collagen alpha 3 VI chain precursor
RNA sec. struct.	652.45	25	13	4.6e-20	8 regions overlap gene predictions
	221.127	12	9	2.1e-16	See Figure 4
	50 others			<1e-6	
Go/InterPro annotation	631	18	15	1e-18/1e-28	mostly intronic, to various homeobox transcription factors

Table 1. A sample of clusters found to be enriched for particular attributes. #v is the number of vertices in the cluster. #Att describes the number of cluster members that have a given attribute. (also see supplementary material).

are found in the vicinity of coding exons and in UTRs of gene families, like the Voltage-gated sodium channel alpha and histone H1 families. These are good candidates for transcriptional and splicing regulatory elements.

Perhaps one of the most interesting clusters (cluster 652.29 in Table 1) consists of 10 regions, 6 of which occur in introns, less than 1kb downstream of an exon, and one just upstream of a first exon. This cluster is shown in Figure 3 (B). Notice that the seven genes within or near which these regions lie form the dense core of the cluster. Although some of these genes

have related function (DNA binding), they do not appear to be paralogs, which suggests that this element family has evolved independently of a gene family, and perhaps confers a required function to the genes in which it resides. A detailed study of this cluster is underway and will be reported elsewhere.

We identify a number of clusters whose members are systematically located upstream, downstream, or in UTRs of known gene families. Although each gene in these gene families originated from a duplication that predated the primate-rodent split, the regions identified maintain a high degree of

```
             Human GAGGTGCATTTACTCTTTGA-CCCACTAGGGTACTATTTAGTGTTCTAGAAGAGGTAATTTAGTAAATTGTACCCCAGTGGCCTGAAAAAGTTAA
             Mouse GAGGTGCATTTACTCTTTGA-CCCACTAGGGTACTATTTAGTGTTCTAGAAGAGGTAATTTAGTAAATTGCACCCCAGTGGCCTGGAAAAGTTAA
             Human GATTAATGTGT-CTCTTTCATGGCACTAAGGTAC-ATTTAGAGCACTA-AAGAAGTCATTTACTAAATGGTGCCCTTGAGACTTGAAAGAGTTAA
             Mouse GATTAATGTGG-CTCTTTCAGGGAACTAAGGTGC-ATTTAGAGCACTA-AAGAAGTCATTTACTAAATGGTGCCCTGGGACTTGAAAGAGTTAA
             Human -AGGTGCATTAACTCTTCCAGGCCCCTAGGGTATCATTTAGTTCACTG-GAATAGTAATTTACTAAACTGTACCTTAGGGGCCTGAAAATGTTAA
             Mouse -AGGTGCATTAACTCTTTCAGGCCCCTAGGGTATCATTTAGTCCACTG-GAATAGTAATTTACTAAACAGTACCTTAGGGGACTGAAAAAGTTAA
```

Fig. 4. (top) Predicted RNA secondary structure for one human region belonging to cluster 221.127, at genomic position chr15:65621880-65622205 (Structure predicted by mfold). (bottom) Alignment of a portion of three human regions belonging to that cluster, each with its mouse ortholog. The first sequence is the one folded in (A).

inter-species and intra-human conservation. We hypothesize that these are probably involved in the transcriptional or splicing regulation of the respective gene family. The use of the Go functional annotation and InterPro protein domain classification allows us to examine the genes within, or next to lie the elements of a cluster. This facilitates both an independent analysis of the clusters, as well as an added perspective on any of the sets highlighted by the other attributes.

Finally, more than 50 clusters are highly enriched for regions with significant RNA secondary structures. Although some of them overlap known RNA genes and micro-RNAs, a large fraction is left unannotated, and most of them have no significant correlation with any other attribute we tested. Among the most interesting examples are ID 652.45 (see Table 1), containing 25 members, 20 of which are predicted to fold into a significant RNA secondary structure usually made of three long hairpins. Another interesting case is cluster 221.127, whose members consistently fold into a single, long hairpin. An alignment of three of these regions, together with their respective mouse orthologs is shown in Figure 4. This type of secondary structure, together with the very high degree of sequence conservation in mouse may indicate that this cluster corresponds to a novel family of micro-RNAs.

The interpretation of the function of many of the larger clusters is more problematic, with none of the attributes tested revealing statistically significant biases, except for occasional weak enrichment for RNA secondary structure. Although some of these clusters appear quite dense in terms of average pairwise similarity, it is possible that they may still contain more than one dense core that our algorithm has failed to decompose. This would obviously hamper the annotation efforts by increasing the noise level. On the other hand, it is also possible that some of these clusters do share a function that does not correlate with any of the features we tested. Another possibility is that these clusters correspond to undocumented repetitive regions, although the strict phylogenetic conservation threshold we employ should remove from consideration most of these non-functional regions.

4 DISCUSSION AND FUTURE WORK

This paper presents a first step towards genome-wide intra-species annotation of functional non-coding human regions based on sequence homology. We show that a large number of these regions can be clustered in groups of highly similar sequences, and thus are likely to consist of elements of similar function. Admittedly, these represent a relatively small fraction of all the conserved human regions. In fact, we repeated the clustering procedure with the subset of *coding* sequences found in the top 5% of the human genome aligning to mouse and rat, after filtering out non-syntenic regions, similar to our pre-processing in Figure 1. While only about 5% of the intervals (and total number of bases) in our set had any edge to another member of the set, as much as 50% of the highly conserved, syntenic coding sequences have such matches within their respective set. Nonetheless, the 18734 non-coding regions that we clustered represent a large set highly enriched for putative functional elements.

We see this as a very encouraging sign: despite the fact that the measure of similarity used was not targeted at finding one specific type of functional region, a large number of clusters were identified and many proved to provide valuable information about the function of their members. The function of the majority of the clusters we identified remains unclear but because of the strict filtering applied to the input, it is unlikely to be too similar to known features in the genome. As our understanding of our genome improves, more and more clusters will be better characterized and understood.

A number of research directions are opened up by our approach. A first, immediate goal, which we are already pursuing is the further analysis of the elucidated clusters. Several more screens can be applied to each individual cluster, as

Gill Bejerano[a], *David Haussler*[a, b], *Mathieu Blanchette*[c]

well as ancestral reconstruction to attempt and detect remote homologies, and possibly hierarchical relationships between the different clusters, as is the case in the protein world. It is also expected that with the characterization of these clusters we may be able to better define sequence similarity measures for specific types of functional regions, such as regulatory modules and classes of RNA genes, as well as improve our clustering methodology. This interplay between methods and results is bound to enrich our set of clusters as well as improve our understanding of them, in much the way that our understanding of protein sequences has evolved.

Acknowledgements

REFERENCES

[1] S. Altschul, T. Madden, A. Schaffer, J. Zhang, W. Miller, and L. D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nuc. Acids Res.*, 25(17):3389–402, 1997.

[2] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler. Recent segmental duplications in the human genome. *Science*, 297(5583):1003–1007, 2002.

[3] G. Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nuc. Acids Res.*, 27(2):573–580, 1999.

[4] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, 2003.

[5] F. Chiaromonte, R. Weber, K. Roskin, M. Diekhans, W. Kent, and D. Haussler. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. *Cold Spring Harbor Symp. Quant. Biol.*, 68, 2003.

[6] E. T. Dermitzakis, A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S. E. Antonarakis. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, 302(5647):1033–1035, 2003.

[7] C. Dieterich, H. Wang, K. Rateitschak, H. Luz, and M. Vingron. CORG: a database for COmparative Regulatory Genomics. *Nuc. Acids Res.*, 31(1):55–57, 2003.

[8] A. J. Enright and C. A. Ouzounis. GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics*, 16(5):451–57, 2000.

[9] Fiduccia, C.M. and Mattheyses, R.M. "a linear-time heuristic for improving network partitions". *Proc. 19th Design Automation Conf.*, pages 175–181, 1982.

[10] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. Eddy. Rfam: an RNA family database. *Nuc. Acids Res.*, 31(1):439–41, 2003.

[11] M. A. Harris, J. Clark, A. Ireland, and et al. The Gene Ontology (GO) database and informatics resource. *Nuc. Acids Res.*, 32(1):258–261, 2004.

[12] I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structure. *Monatsh. Chem.*, 125:167–88, 1994.

[13] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[14] D. Karolchik, R. Baertsch, M. Diekhans, T. Furey, A. Hinrichs, Y. Lu, K. Roskin, M. Schwartz, C. Sugnet, D. Thomas, R. Weber, D. Haussler, and W. Kent. The UCSC Genome Browser Database. *Nuc. Acids Res.*, 31(1):51–54, 2003.

[15] H. Kawaji, Y. Takenaka, and H. Matsuda. Graph-based clustering for finding distant relationships in a large set of protein sequences. *Bioinformatics*, 20(2):243–52, 2004.

[16] W. J. Kent, R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA*, 100(20):11484–11489, 2003.

[17] S. Kim. BAG: A graph theoretic sequence clustering algorithm. http://bio.informatics.indiana.edu/sunkim/BAG/, 2004.

[18] E. H. Margulies, M. Blanchette, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13(12):2507–2518, 2003.

[19] J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, 2003.

[20] V. Matys, E. Fricke, R. Geffers, and et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nuc. Acids Res.*, 31(1):374–8, 2003.

[21] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[22] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, and et al. The InterPro Database, 2003 brings increased coverage and new features. *Nuc. Acids Res.*, 31(1):315–318, 2003.

[23] K. M. Roskin, M. Diekhans, and D. Haussler. Scoring two-species local alignments to try to statistically separate neutrally evolving from selected dna segments. In *Proc. RECOMB*, pages 257–266. ACM Press, 2003.

[24] S. Santini, J. L. Boore, and A. Meyer. Evolutionary conservation of regulatory elements in vertebrate Hox gene clusters. *Genome Res*, 13(6A):1111–1122, 2003.

[25] P. Schattner, A. Pohl, and T. Lowe. Ongoing work. 2004.

[26] S. Schwartz, W. J. Kent, A. Smit, Z. Zhang, R. Baertsch, R. C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with BLASTZ. *Genome Res*, 13(1):103–107, 2003.

[27] A. F. Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, 9(6):657–663, 1999.

[28] K. Sumiyama and F. H. Ruddle. Regulation of Dlx3 gene expression in visceral arches by evolutionarily conserved enhancer elements. *Proc. Natl Acad. Sci. USA*, 100(7):4030–4034, 2003.

[29] D. Torrents, M. Suyama, E. Zdobnov, and P. Bork. A genome-wide survey of human pseudogenes. *Genome Res.*, 13(12):2559–2567, 2003.

[30] Z. Zhang, N. Carriero, and M. Gerstein. Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet*, 20(2):62–67, 2004.