Identification of non-coding functional regions in human DNA

Donald M. Pianto

dpianto@unb.br

UFPE

References

Gill Bejerano, David Haussler and Mathieu Blanchette. “Into the heart of darkness: large-scale clustering of human non-coding DNA”. *Bioinformatics* Vol. 20 Suppl. 1 pages i40–i48, (2004), *and references therein.*

National Center for Biotechnology Information website: `http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html`

Motivation

- 5% of the human genome is under purifying selection and thus likely to be functional (Mouse Genome Sequencing Consortium, 2002; Roskin et al., 2003; Chiaromonte et al., 2003).
- protein-coding genes' exons account for only about 1.5% of the genome (2% if UTRs are included) (International HumanGenome Sequencing Consortium, 2001).
- The remaining $3 - 3.5\%$ is the dark matter of the human genome.

What is Dark Matter?

Dark matter is likely to contain mainly:

- gene regulatory regions (both transcriptional and splicing),
- RNA genes and micro-RNAs,
- matrix attachment sites,
- origins of replication,
- and perhaps some altogether novel functional elements.

Isn't this being done?

Yes, but databases of experimentally verified loci like:

- Transfac for regulatory regions (Matys et al., 2003)
- RFAM for RNA genes (Griffiths-Jones et al., 2003)

contain information about only a tiny fraction of the regions under discussion.

How to proceed?

Identify orthologs.

Homologs, Orthologs, and Paralogs

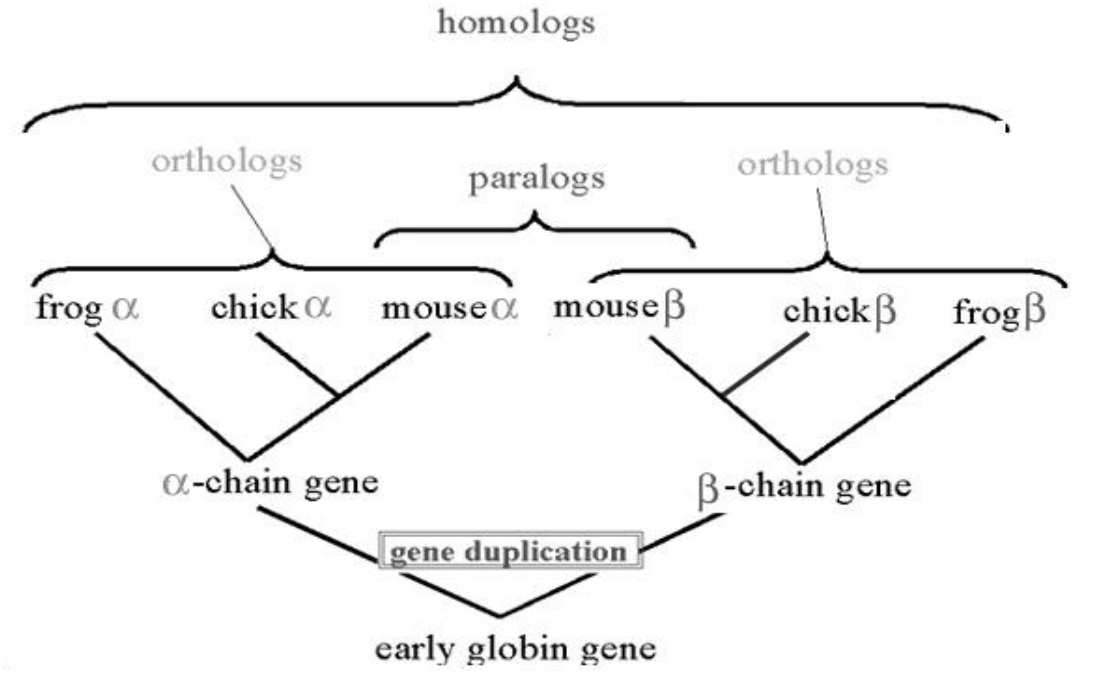


Figure 1: Homologous sequences. Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication.

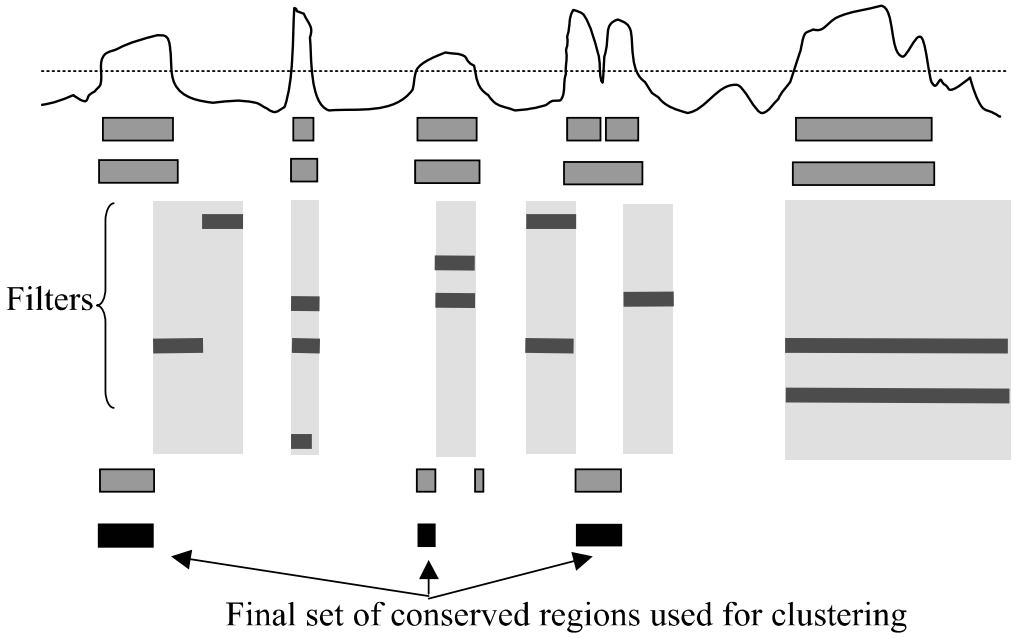
What is the use of these orthologs?

- Regions of the genome that show a significant degree of conservation in other species are probably functional.
- If we remove known protein-coding regions or orthologs of such regions from the preserved regions we should be left with functional dark matter.
- By grouping this functional dark matter into clusters with similar sequences (and likely similar functions) one can begin to more extensively annotate non-coding regions.
- This has already been performed by Margulies *et al.* (2003) for the CFTR region.

Filtering the Orthologs (1)

- Three way multiple alignment of human, rat, and mouse to establish orthology.
- Keep the 5% most conserved sequences
  - 1 055 823 regions averaging 140 bp
  - 74% of the bases in coding exons (only 13–17% of the region)
- Extend to avoid fragmenting
- Ensure synteny between mouse and rat
- Remove coding or similar
- Remove repeats
- Require synteny with mouse
- Remove putative psuedo-genes
- Remove human segmental duplications

Filtering the Orthologs (2)



Human-mouse-rat sequence conservation	Fraction of genome	Number of regions
Top ~5%	5.11%	1055823
Extend/merge	5.835%	969857
Remove repeats	5.393%	959820
Remove coding exons	4.541%	1074181
Remove cds-like	4.239%	1084945
Remove non-syntenic	4.086%	1072148
Remove seg. dups.	4.006%	1043450
Remove pseudogenes	4.001%	1042608
Remove <50bp	3.727%	699647

Figure 2: 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.

Measuring intra-human similarity (1)

- Each of the identified 699 647 regions of the genome is now considered a vertex of a similarity graph.
- Blastz local alignments of the repeat-masked genome against itself are used to determine which regions are similar.
- A pair of regions is connected by an edge of weight $s(u, v)$ if:
  - a consecutive block of 15 alignment positions is found between them and
  - $s(u, v) \geq 0$ (where $s(u, v)$ is the similarity score from the Blastz local alignment).

Measuring intra-human similarity (2)

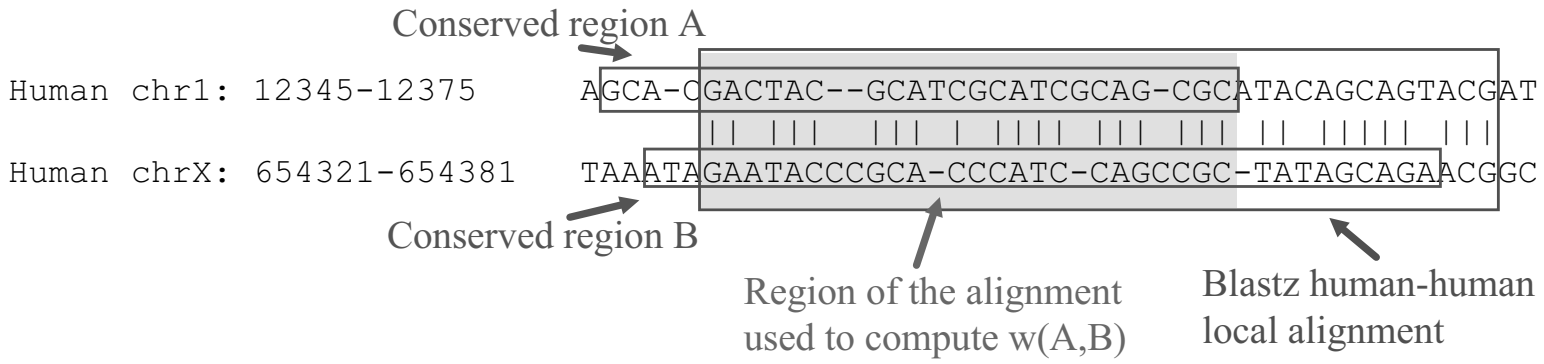


Figure 3: Scoring the similarity between two conserved regions using a Blastz human-human local alignment. The score $w(A, B)$ of the alignment is based only on the shadowed area.

Identifying Clusters (1)

- 96% of the 699 647 vertices of the similarity graph are not connected to any other vertex.
- 29 349 regions are similar to at least one other.
- 8333 connected components, 1446 with 3+ vertices, and 257 of 10+ vertices. Largest has 823 vertices and 1673 edges.
- How can we identify true clusters?

Identifying Clusters (2)

- This problem has been studied in similarity graphs with proteins as the vertices.
- Taking the connected parts of the graph as clusters has two main problems:
  1. false-positive edges tend to collapse two dense clusters into a single large connected component and
  2. conserved regions made of two different but adjacent functional units may connect unrelated clusters.
- Two algorithms are proposed to partition vertices and duplicate vertices to avoid the above mentioned problems.

Identifying Clusters (3)

- 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 separates a set of vertices into two groups with little similarity. This can be used to eliminate false-positive edges.
- The function $\mathrm{CUT}(V, E, w)$ returns the minimum weight cut (A, B) of V and its weight using the Fiduccia-Mattheyses heuristic.

Identifying Clusters (4)

- To separate a putative multi-functional region, u, first the local alignments are mapped to u's sequence. If the alignments stack into two or more disjoint portions of u then it is divided into its non-overlapping regions.
- More difficult cases require the identification of a local-articulation point.

Identifying Clusters (5)

- The local-articulation score of a vertex, v, is defined as follows.
  - Let $N(v)$ be the set of neighbors of v (excluding v).
  - Let $G|_X$ be the subgraph spanned by a subset of vertices X.
  - Let $C = (A, B)$ be a minimum-weight cut of the induced subgraph $G|_{N(v)}$.
  - local-articulation(v) =weight$(C)/|N(v)|$.
- Vertices have low local-articulation score if their neighbors can be partitioned into two sets with little similarity.
- When such vertices are found they are duplicated with one copy connected to A and the other to B.

Identifying Clusters (6)

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.
$s_{min} = +\infty$
for each vertex $v \in V$ do

- $(A, B, s) = CUT(G|_{N(v)})$
- if $(s < s_{min})$ then $(v_{min}, A_{min}, B_{min}, s_{min}) = (v, A, B, s)$

return $(v_{min}, A_{min}, B_{min}, s_{min})$

Identifying Clusters (7)

- 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.
- If the Blastz weight of a minimum weight cut is below $\delta_c = 2000$ then a cut is performed.
- If the local-articulation score is below $\delta_a = 200$, the corresponding vertex is duplicated.

Identifying Clusters (8)

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.

$(A, B, x) =$CUT(V, E, w)

if $(x < \delta_c)$ then $E = E - \{(u, v) \in V : u \in A, v \in B\}$

else

- $(v, A, B, y) =$ BEST-LOCAL-ARTICULATION(V, E, w)
- if $(y/|N(v)| < \delta_a)$ then `/* duplicate */`
  - $V = V \cup v'$ `/* add vertex v' */`
  - $E = E \cup \{(v', b) : b \in B\}$
  - $E = E - \{(a, b) : a \in A \cup v, b \in B\}$
- else print (V', E'), return

for each connected component (V', E') of (V, E) do

- GRAPH-PARTITIONING(V', E', w)

Algorithmic Example

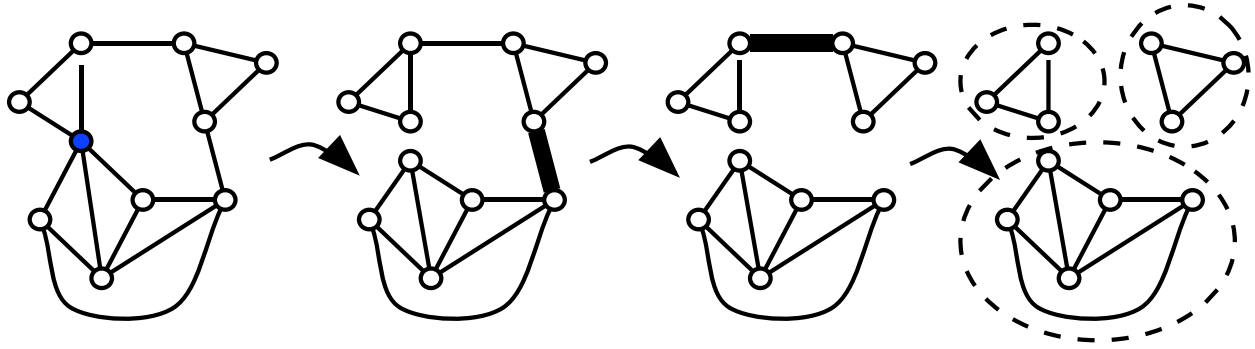


Figure 4: Identification of dense subgraphs by their heuristic. Assuming all edges have weight 1, $\delta_c = 1$ and $\delta_a = 0$, the original graph has no cut-of-cost less than 2 but has a vertex with local-articulation score zero. This vertex is first duplicated. The resulting graph has two cuts-of-cost one, whose edges are removed. The resulting graph has three dense connected components.

Empirical Clusters

- The algorithm runs in time $O(VE^2)$. 1hr on their desktop.
- Obtain 12 027 dense homogeneous clusters with from 2 to 105 vertices (regions). 296 clusters with 5+ vertices and 84 clusters with 10+ vertices.
- 18 734 human regions are within some cluster.
- Average region length is 225 bp.

Cluster Functions

- Assuming common functionality, inferring this function may be easier for the cluster because:
  - functional annotation for one member can be mapped to others;
  - statistically over-represented features may hint at function.
- Features considered include: Genomic location; Association to known genes; Coding potential; Evidence of transcription; Non-coding RNAs; RNA secondary structure; Conservation in distant species.

Results from previous annotation

- 47 clusters containing exclusively members with an RNA gene annotation
- 30 unannotated regions that belong to a cluster with at least one member annotated as micro-RNA or RNA gene, suggesting a functional classification. (Currently under analysis).

Statistical over-representation results 1

Possible novel protein-coding gene families.

- Several clusters enriched for gene predictions
- mRNA and EST evidence, implying active transcription
- also conserved in chicken, suggesting functional importance
- boundaries of gene predictions match those of the conserved regions
- some clusters (*e.g.* 3089.3) show weak protein similarity to known genes (tBlastn).

Statistical over-representation results 2

Possible transcriptional and splicing regulatory elements

- Several clusters enriched for chicken conservation
- some go back to fugu
- found in the vicinity of coding exons and in UTRs of gene families

Statistical over-representation results 3

Cluster (652.29) with 10 regions, 6 of which occur in introns less than 1kb downstream of an exon and one just upstream of a first exon.

- minimal related function (only DNA binding)
- apparently not paralogs
- suggests evolution independent of a gene family
- perhaps provides a required function to the genes where it resides
- similar clusters which evolved with gene families are likely involved in transcription or splicing regulation of the family.

Empirical Cluster

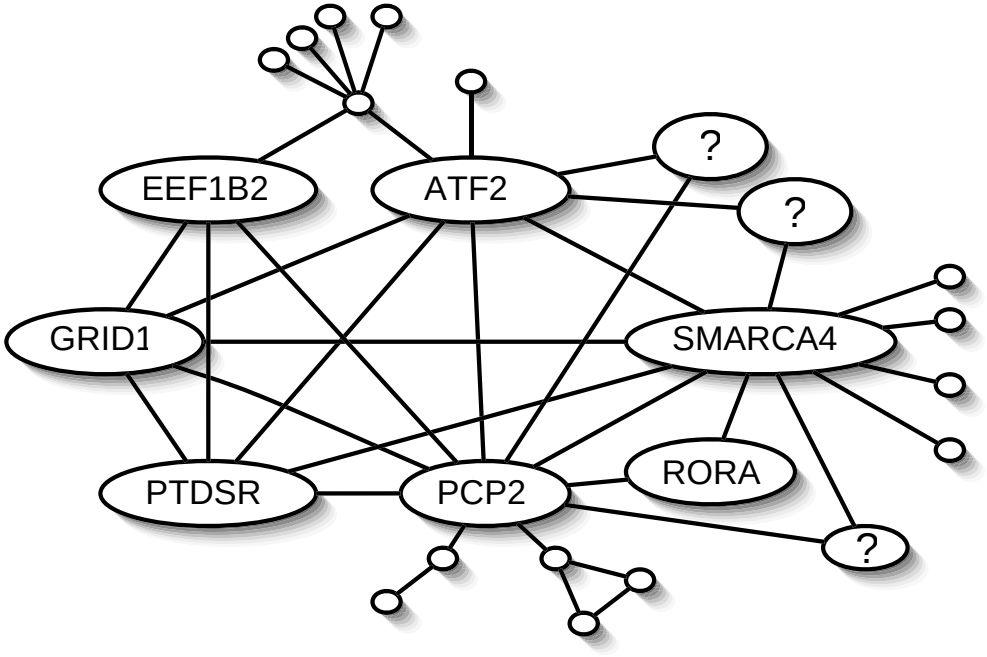


Figure 5: Example of an actual cluster (ID 652.29). Small vertices were those removed by the algorithm.

Statistical over-representation results 4

More than 50 clusters are highly enriched for regions with RNA secondary structures.

- some overlap known RNA genes and micro-RNAs
- many are un-annotated
- cluster 652.45 has 25 members, 20 of which are predicted to fold into three hairpins
- cluster 221.127 has 12 members, all of which fold into a hairpin (see figure)
- this structure and its conservation may indicate a novel family of micro-RNAs

RNA Secondary Structure

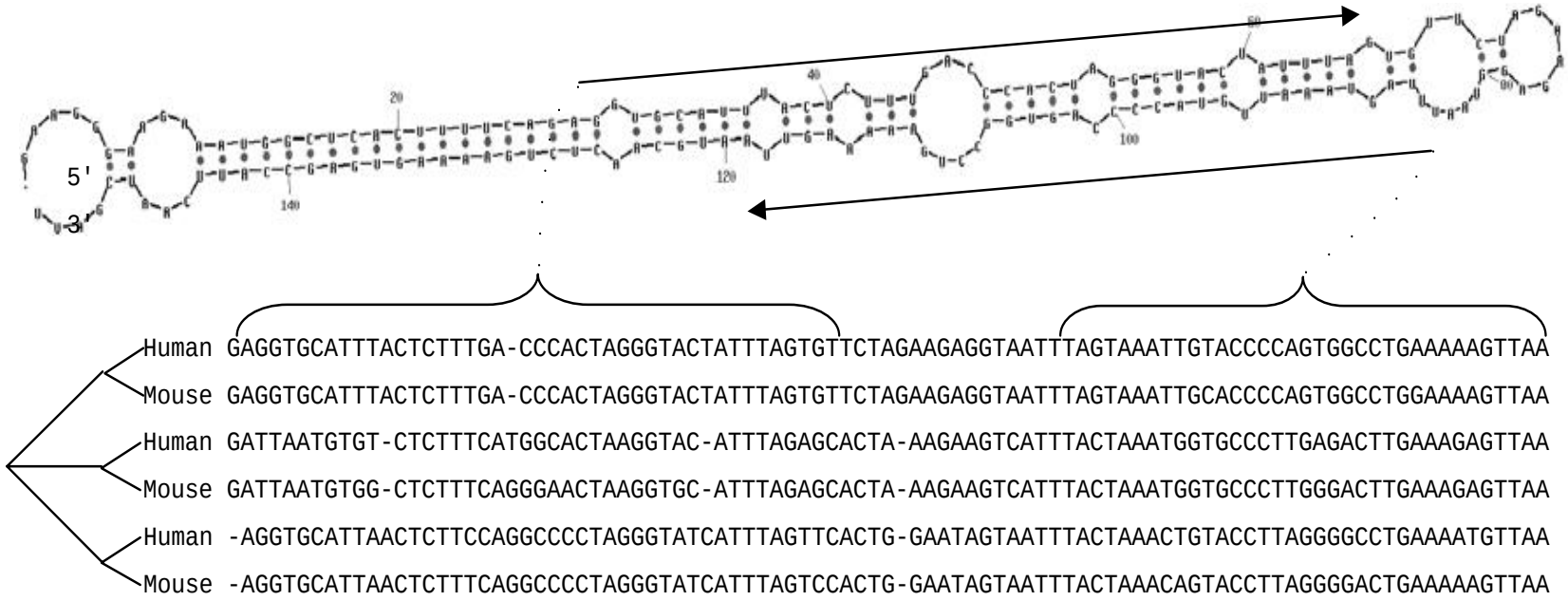


Figure 6: (Top) Predicted RNA secondary structure for one human region belonging to cluster 221.127, at genomic position chr15:65621880-65622205 (Structure predicted by mfold). (Bottom) Alignment of a portion of three human regions belonging to that cluster, each with its mouse ortholog.

Table 1: A sample of clusters found to be enriched for particular attributes.

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	Five regions have coding potential (*p*-value 4.9e−4)
	2003.3	19	15	8.1e – 8	Ten regions have coding potential (*p*-value 1.8e−8) and 8 regions have RNA secondary structure (*p*-value 7.2e−13)
Fugu conservation	4415.3	5	5	7.9e – 11	Just 5′ of exons of SCNxA gene family (*p*-value 8.4e−6), all are conserved in chicken (*p*-value 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 (*p*-value 1e−8)
	24 others			<8.6e – 07	
ESTs	652.29	10,21	6	9.7e – 7	Six sites are <1 kb downstream of exons of various genes (Fig. 3B).

Table 1: (continued)

Attribute	Cluster ID	#v	#Att	*p*-value	Comment
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
1 kb 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 12 kb ORF upstream of c2orf16
	5011.6	5	5	5.6e – 7	Consecutive regions contained in a 5 kb 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 Fig. 4
	50 others			<1e – 6	
Go/InterPro annotation	631	18	15	1e – 18/1e – 28	Mostly intronic, to various homeobox transcription factors

Large Clusters

- none of the large clusters contain statistically significant biases
- possibly are mixtures of several dense cores which the algorithm did not separate
- possible that a function which does not correlate with any of the tested features is shared among the clusters
- possibly undocumented repetitive regions (synteny should have removed these non-functional regions)