The DART Classification of Unannotated Transcription within the ENCODE Regions: Associating Transcription with Known and Novel Loci

Joel Rozowsky[1#], Daniel Newburger[1], Fred Sayward[2], Jiaqian Wu[3], Greg Jordan[1], Jan O. Korbel[1], Ugrappa Nagalakshmi[3], Jin Yang[2], Deyou Zheng[1], Roderic Guigo[4], Thomas Gingeras[5], Sherman Weissman[6], Perry Miller[2,7], Michael Snyder[3] and Mark Gerstein[1,7#].

1. Molecular Biophysics & Biochemistry Dept; Yale University, PO Box 208114, New Haven, CT 06520-8114, USA.

2. Center for Medical Informatics; Yale University, PO Box 208009, New Haven, CT 06520-8009, USA.

3. Molecular, Cellular & Developmental Biology Dept; Yale University; New Haven, CT 06520, USA.

4. Grup de Recerca en Informática Biomèdica; Institut Municipal d'Investigació Mèdica/Universitat Pompeu Fabra. Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.

5. Affymetrix, Inc.; Santa Clara, CA, 92024, USA.

6. Dept of Genetics; Yale University; New Haven, CT 06520, USA.

7. Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA.

[#]Corresponding Authors

Running title: Unannotated Transcription in the ENCODE Regions

Key words: unannotated transcription, genomic tiling microarrays, ENCODE, transcriptionally active regions

Abstract
For the ~1% of the human genome selected by the ENCODE consortium only about half of the transcriptionally active regions (TARs) identified with tiling microarrays correspond to exons of annotated genes. Grappling with, classifying and analyzing this large amount of "unannotated transcription" presents new challenges. Here we attempt an overall categorization of the 6,988 novel TARs detected. We use a number of disparate features to classify TARs -- their expression profile of array signals across 11 cell lines and conditions, their overall sequence composition, their phylogenetic profile (presence or absence of syntenic conservation across 17 mammalian species), and their location relative to gene annotation. We perform the classification stepwise. In a first pass, we filter out TARs with unusual sequence composition and those likely to result from cross-hybridization. We then associate some of those remaining with known exons based on proximity and similar expression profiles. Finally, we attempt to group unclassified TARs into putative clusters of novel transcription, perhaps representing novel loci, based on similarity in their expression and phylogenetic profiles. Storing, visualizing, manipulating, and comparing all these different groupings of TARs requires a different infrastructure than for conventional gene annotation. Therefore, we have constructed the Database of Active Regions and Tools (DART.gersteinlab.org) to develop and encapsulate our classification. DART has special facilities for rapidly handling and comparing many sets of TARs and their heterogeneous features, synchronizing across genome builds, and robustly interfacing with other resources (such as the UCSC Genome Browser). Using our DART classification, we find that ~14% of the novel TARs can be confidently associated with known genes, while ~21% can be clustered into ~200 novel loci, comprised of ~7 TARs each. We also find some novel TARs are arranged in tandem arrays of sequence-similar blocks. We survey novel TARs for protein homology and their potential to form structured RNAs, and we conclude that TARs associated with known genes are most strongly enriched for structural RNAs. Finally, we observe that many of the novel TAR clusters are associated with a nearby promoter. In a future scale-up of the ENCODE project, we anticipate that categorization of novel TARs can help systematically targeting medium-scale follow-up experiments (e.g. by RACE and RT-PCR). To benchmark this, we use the DART classification to design a set of experiments for testing the connectivity of transcripts amongst novel TARs and between them and known exons. Overall, we find that ~40% of the connections tested (18 out of 46) validate by RT-PCR and that 4 out of the 5 PCR products that were sequenced confirm connectivity unambiguously.

INTRODUCTION

In recent years there have been a number of experiments using genomic tiling microarrays that have found significantly more transcribed DNA sequences in the human genome than had been previously annotated as genes (see Kapranov et al. 2002, Rinn et al. 2003, Bertone et al. 2004 & Cheng et al. 2005). The biological function of this vast quantity of additional transcribed RNA is not yet fully understood. There have been independent experiments using complementary sequencing technologies that have also detected large amounts of previously unidentified transcription (Carninci et al. 2005). Genome tiling arrays have likewise been used for transcript mapping in a variety of organisms besides human: *A. thaliana* (Yamada et al. 2003), *D. melanogaster* (Stolc et al. 2004, Manak et al. 2006), *S. cerevisiae* (David et al. 2006) and *O. sativa* (Li et al. 2006).

One of the goals of the ENCODE (ENCyclopedia Of Dna Elements) project (ENCODE Project Consortium 2004) is to map out and determine the function of these unannotated transcripts for the one percent of the human genome selected for the pilot phase of the project. For the selected ENCODE regions, RNA transcript maps were constructed for a variety of cell lines and biological conditions ([reference to ENCODE manuscript]). Consistent with earlier studies a large fraction of the sequences identified as transcribed are not in annotated genomic regions. An important result obtained from these experiments was the discovery of tissue-specific alternative transcription start sites, found by conducting 5' RACE extensions from exons of known transcripts. Many of the transcription start sites were found to be more than 100 kb upstream of an annotated start site. Although these alternate long transcripts account for some of the novel transcribed regions detected, the majority remain unexplained. These long transcripts demonstrate that gene loci are quite complex and that there is probably a multiplicity of alternative isoforms that are transcribed from most complex loci. Even in the set of well curated genes for the ENCODE regions (the GENCODE/HAVANA annotation, Harrow et al. 2006) we see on average 5.4 alternative isoforms per locus. This number is most likely a significant underestimate of the number of distinct transcripts arising from an average locus in all cell lines, especially when all cellular conditions are considered.

Transcribed regions detected by genomic tiling arrays are known as TARs (transcriptionally active regions, see Rinn et al. 2003) or alternatively as transfrags (transcribed fragments, see Kapranov et al. 2002). Although novel transcribed regions have been observed and analyzed in previous works, in this paper we present an overall characterization and systematic classification of novel TARs. Some of this classification is briefly discussed in [reference to ENCODE manuscript] where novel TARs are categorized on the basis of their vicinity to known genes. We extend this analysis by grouping the novel TARs into a number of distinct possible categories: (i-a) novel TARs with peculiar sequence composition, (i-b) novel TARs that are probably caused by cross-hybridization on the microarray, (ii) novel TARs that are associated with known gene loci, and (iii) novel TARs that are not associated with known genes but can be grouped into clusters which may be novel transcribed loci.

Data sets for novel TARs and their associated information should not be thought of as

regular gene annotation, since unlike genes, properties such as the connectivity between novel TARs, which potentially form spliced transcripts, are not very well defined. Moreover, TARs have additional information such as the fluorescent array signal that is not usually associated with gene annotation. Thus existing databases such as the UCSC Genome Browser (Kent et al. 2002), Gene Expression Omnibus (Edgar et al. 2002) or ArrayExpress (Brazma et al. 2003) do not have the flexibility to store sets of TARs with all the associated information and make them accessible in an efficient manner. For this reason we have constructed a database (DART: Database of Active Regions and Tools) for encapsulating our classification. The database is optimized for browsing sets of TARs discovered in tiling microarray experiments. In addition the database allows the storage of sites of transcription factor binding and modifications called BARs (binding active regions), which are important to associate with TARs. We have also constructed a set of tools (Active Region Comparer) that can be used for the comparison of multiple sets of active regions with each other and with annotations from EnsEMBL (Birney et al. 2006). Both the database and tools are connected with the UCSC Genome Browser for automated visualization of custom tracks.

The DART methodology developed in this paper is a first pass analysis of the novel TAR data sets and transcript maps that are available today as part of the pilot phase of the ENCODE project. An optimal approach to understanding the biological role of the multitude of novel TARs is to couple array experiments with medium-scale follow-up experiments. As an initial iteration of this process, we used the results of our classification to design some small-scale experiments that investigated the connectivity between novel TARs and exons of known genes and the connectivity between novel TARs clustered into novel transcribed loci. This validation experiment demonstrates that ~40% of the novel TARs tested for association with either an exon of a known gene or another novel TAR can be confirmed to be connected in a transcribed RT-PCR product. When the next phase of the ENCODE project scales to the whole genome, the resulting experimental data can be used to optimize the classification procedure in future iterations

RESULTS

Novel Transcribed Regions

Transcript maps were constructed across the 44 ENCODE regions using genomic tiling microarrays for 11 different cell lines and conditions ([reference to ENCODE manuscript]). The 44 ENCODE regions span 30 Mb of genomic sequence, half of which comes from manually selected gene loci (e.g. HOX cluster & CFTR locus) and half comes from 500 kb regions chosen to stratify differing levels of both gene density and non-exonic conservation with mouse. The 11 different cell lines and conditions were a combination of both Poly(A)+ and total RNA samples. Transcript maps were constructed by hybridizing reverse transcribed double-stranded cDNA to a high-density oligonucleotide tiling array which covered one strand of the ENCODE regions.

TARs were determined by locating stretches of oligonucleotide probes with high hybridization signals compared to background. The signal thresholds used to identify

these transcribed genomic regions were determined using bacterial control sequences included on the Affymetrix tiling microarrays (Kampa et al. 2004). We note that the amount of transcription detected and the fraction that is in annotated regions are dependent on the signal threshold used (see Royce et al. 2005 & Emanuelsson et al. 2006). Using a more stringent threshold, we detect fewer overall TARs, however, the fraction that corresponds to annotated exons increases because novel TARs tend to be transcribed at lower levels than exonic TARs. A threshold was determined such that the false positive rate from bacterial negative controls was only 5 percent for each of the cell lines and conditions mapped. There has been an ongoing debate in the genomics community as to the fraction of the human genome that is transcribed. In [reference to ENCODE manuscript] it has been determined that more than ~70% of the human genome is transcribed as primary transcripts. However, the use of a stringent threshold for tiling microarray signal selects for genomic regions that are transcribed as part of processed (spliced) RNAs. Thus the large number of novel TARs detected as part of the ENCODE project's pilot phase are more likely to be components of processed transcripts rather than due to the basal level of transcribed genomic DNA. Our DART classification procedure attempts to categorize these novel transcribed regions as part of known genes and into potential novel transcribed loci. Although many of the TARs that were detected correspond to exons of known genes, this study focuses on the novel TARs that do not match exonic sequences. These novel, unannotated TARs lie either, within the introns of known genes or within the intergenic regions between known genes. Here we will use the set of GENCODE/HAVANA annotation (Harrow et al. 2006), which is a comprehensive set of all the well-curated transcripts contained within the ENCODE regions.

The initial set of all TARs generated can be classified into three basic categories: TARs corresponding to known or putative GENCODE genes, TARs overlapping annotated pseudogenes and novel TARs in unannotated regions. TARs overlapping pseudogenes are ambiguous given the homology of the pseudogene sequence to its parent gene, both of which are potentially transcribed. Other more detailed studies of pseudogene transcription have determined that a small but significant fraction are transcribed and can be distinguished from parental gene transcription (Zheng et al. 2005). However, in order to avoid these ambiguities for the purposes of this analysis the sets of TARs are filtered for those that intersect low complexity repeats or any annotated pseudogene in the ENCODE regions. Novel TARs are then classified into one of the following categories: (i) intronic TARs, (ii) intergenic TARs and (iii) TARs that match other ESTs that were not part of the GENCODE annotation (typically unspliced ESTs that do not contain a polyadenylation signal). The intergenic and intronic TAR sets are further subdivided into those that are proximal subsets that are within 5 kb of GENCODE exons and distal subsets that are further than 5 kb (see figure 1a for a diagram of this classification). The distribution of TAR locations can be seen in table 1 where we observe that nearly half of the novel TARs are in intronic regions proximal to exons of known genes. The table also includes the 195 TARs that intersect pseudogenes prior to their removal. In figure 1b we develop a strategy for a more detailed classification of sets of novel TARs, which will be described in more detail in the steps below. Each of these classified sets can also be individually partitioned as per figure 1a, on the basis of proximity to annotation.

Each novel TAR has a number of distinct features: expression profile across the biological samples mapped, genomic location relative to known GENCODE annotation, sequence composition, sequence conservation, and phylogenetic profile of conservation (see figure 2). In the following analysis we make use of some of these features when grouping the sets of novel TARs into the following distinct categories: (i) potentially artifactual TARs that are caused by peculiar sequence composition or cross-hybridization, (ii) novel TARs that can be associated with known gene loci and (iii) novel TARs that can be clustered into groups forming potential novel transcribed loci. For the remaining unclassified novel TARs with above average array signal, additional clustering is performed on the basis of vicinity and phylogenetic similarity. See figure 1b for a schematic of the stepwise classification procedure. Many of the DART classification steps use the expression profiles of individual novel TARs across the eleven different cell lines and conditions. For each novel TAR we also construct a phylogenetic profile across the species sequenced by the ENCODE consortium ([reference to ENCODE manuscript]). These profiles identify which of these species contain the novel TAR in a syntenic region. The classification uses other information as well, such as the sequence composition of novel TARs and their location relative to known genes. We also study the protein coding potential for novel TARs by searching for homologous protein sequences, and we investigate the likelihood of the various categories of TARs to form structured RNAs (using RNAz, Washietl et al. 2005). All of these features, as well as the classification sets, are stored in the DART database.

Step 1A: Filtering TARs for Peculiar Sequence Composition

Genomic tiling microarrays interrogate genomic sequences by the use of short oligonucleotide probes that tile the region of interest. There are two main effects that can cause regions to erroneously appear as transcribed. The first effect results from the basic mechanism by which array hybridization works, which is the binding of a sample's cDNA to its matching reverse complement DNA oligonucleotide probe. The amount of cDNA that hybridizes to a particular spot on the microarray, and the corresponding fluorescent signal measured, are subject to the binding affinity between the cDNA and probe, which is in turn dependent on the sequence composition of the oligonucleotide. Thus probes with higher G/C content tend to bind more tightly and show greater fluorescent signal (SantaLucia 1998). In addition short sequence motifs that bind with higher affinity would cause many probes to show abnormally high signal in genomic regions not transcribed (Rozowsky et al. unpublished). Probe sequence effects are dramatically reduced by the use of sliding window scoring, which averages the signal from multiple oligonucleotide probes in a short genomic span (see Cawley et al. 2004, Kampa et al. 2004, Royce et al. 2005). However, biases due to oligonucleotide probe sequence effects are still evident when one compares the G/C content of sequences detected to be transcribed against all annotated sequences.

In [reference to ENCODE manuscript] the di-nucleotide frequency was compared among novel TARs, exonic TARs, all exons and randomly selected sequences. This analysis showed that the di-nucleotide frequency of novel TARs was more similar to that for exons than random sequences. However, for the CC/GG and AA/TT di-nucleotides (both

the forward and reverse complement di-nucleotides are combined since TARs are not stranded) the average frequency was significantly different from the frequency for annotated exons. Figure 3 illustrates this difference where the distribution of CC/GG frequencies for novel TARs is skewed to higher frequencies than that for GENCODE exons. Thus CC/GGs occur more often in novel TARs than in known exons, while the AA/TT frequency for novel TARs is lower than for exons (see supplementary figure 1). In order to be cautious, we removed novel TARs whose CC/GG frequency was above the top one percent of CC/GG frequencies for GENCODE exons as well as those whose AA/TT frequency was below the bottom one percent for exons. There are 380 novel TARs whose CC/GG frequency is greater than 0.156 (indicated by the black arrow on figure 3), as well as 175 novel TARs whose AA/TT frequency is below 0.004. Thus 503 novel TARs were excluded from the 6988 total novel TARs leaving 6485 novel TARs which we shall consider.

Step 1B: Filtering TARs for Cross-Hybridization

The second main microarray artifact, which can lead to false positive detection of transcribed regions, is cross-hybridization. Cross-hybridization happens when oligonucleotide probes on the array hybridize to cDNA from transcripts that have partial sequence complementarily to the probe, but the transcripts originate from somewhere else in the genome. One standard approach is to take the sequences of novel transcribed regions and BLAST (Altschul et al. 1990) them against the current build of the genome to identify sites of potential cross-hybridization. However, the limitation of this approach is that once one has located a potential site of cross-hybridization, which could be either annotated as part of a known transcript or an additional putative novel TAR, the true source of transcription remains ambiguous (one or both sites could be transcribed). The approach that we propose would resolve this ambiguity.

Using the method by which novel TARs will be determined to be associated with known gene loci by use of co-expression of novel TARs with exons of known GENCODE genes, we propose the following procedure: We first identify the most likely source for cross-hybridization by using BLAST (we call the matching region a blastTAR). Only TARs that have a significant match are considered (at a BLAST e-value of less than 10^{-5} or a bit score of 54.0, which corresponds to approximately 40-50 nucleotides with greater than 90 percent sequence identity). The expression profile of the original novel TAR is then compared against exons of genes in the local genomic vicinity of the blastTAR. If the novel TAR is co-expressed with the blastTAR's surrounding exons, then the most likely explanation is that the blastTAR is the primary source of transcription and the original novel TAR was detected because of cross-hybridization. Determining the true source of transcription from two genomic locations with high degree of sequence similarity is thus made possible by using the expression profiles of the novel TARs compared with exons nearby the potential cross-hybridization site.

Of the 6485 filtered novel TARs from step 1A, 658 have matches with an e-value of 10^{-5} or better. Since the ENCODE regions only cover approximately one percent of the genome, a naïve expectation is that only about one percent of these matches would be

located within the ENCODE regions (we can only implement this procedure for blastTARs that are located within ENCODE since we need to compare them with the expression profiles of nearby exons). BlastTARs that are located in the same ENCODE region as the original TAR need to be treated separately (this is discussed in further detail later on). However, there are no novel TARs for which a blastTAR is located in a different ENCODE region. Even though we are unable to utilize this approach for the novel TARs in ENCODE, it will be applicable when tiling arrays studies that cover the entire genome become more abundant.

Step 2: Association of Novel TARs with Known Gene Loci

We want to address the question of how many of the novel TARs can be confidently assigned to known gene loci. By this we mean that the novel TARs are transcribed as parts of longer transcripts, which are as yet unannotated isoforms of transcripts from a specific gene locus or of distinct RNAs that are co-regulated with the gene of interest. In order to make these assignments we identify novel TARs that are co-expressed with exons of genes in the vicinity of the novel TARs. We do this by computing the Pearson correlation coefficient between the expression profiles of novel TARs and the expression profiles of nearby exons (see figure 4). This method is similar to how different genes are determined to be co-expressed. Here, however, we are comparing the expression profiles of individual novel TARs and exons, not those of entire transcripts. For a gene that only encodes a single transcript (i.e. has no alternative isoforms), the expression profile of the gene should be the same as that for each of its constituent exons. However, for a locus that is transcribed as multiple different isoforms, the expression profiles of the different exons may be different. Thus, a novel TAR which is co-expressed with an exon of a known gene can be assigned with some confidence to that locus as part of an alternative isoform or as part of a distinct coregulated RNA.

In order to demonstrate that this method works, we first took the set of all known GENCODE genes in the ENCODE regions and computed the expression profiles for all component exons. For each exon, we can test whether we can assign it to the correct gene by comparing its expression profile with the expression profiles of nearby exons. The assignment is made to the target gene which has an exon with the highest correlation. In figure 5 we plot sensitivity against the false positive rate using this assignment procedure. The Pearson correlation threshold for making an assignment is what parameterizes each curve. The blue curve represents the assignment to exons for genes anywhere in the ENCODE regions; the red and green curves are for assignment to exons of genes that are within 100 kb and 20kb of the exon that is being tested. As expected, we see that the accuracy of the assignment is improved by restricting attention to nearby exons. See supplementary material for a more detailed description of this simulation.

For each novel TAR (see Methods for more details) we use the above method to find the known exon within a 20kb window on either side of the TAR and from either strand that has the highest Pearson correlation between its expression profile and that of the novel TAR. We choose to use a Pearson correlation of 0.9 as a threshold, as that corresponds to a p-value of less than 0.05 (given that the correlation coefficient is computed by

comparing expression profiles which have 11 dimensions and on average each novel TAR, has approximately 19 known exons within 20 kb)[1]. Thus, we can associate 955 of the 6485 filtered novel TARs with a known GENCODE exon. From this analysis we can assign more than 13 percent of the original set of 6,988 novel TARs as part of new alternative isoforms of known transcripts.

Step 3: Clustering Novel TARs into Novel Transcribed Loci

Step 3A: Clusters based on Expression Profiles

After assigning 955 of the novel TARs to known gene loci, we have 5,530 remaining. We cluster co-expressed novel TARs into groups, which we call novel transcribed loci. However, in the assignment of novel TARs to known genes we only assigned those that were highly correlated with exons of known genes. There are likely many more novel TARs in the remaining group of 5,530 that should be assigned to known gene loci but were not because their correlations were below the chosen threshold. In order to focus attention on novel TARs that have a low likelihood of being associated with known gene loci, we first select a subset of the 5,530 novel TARs which have at most a Pearson correlation of 0.1 with any GENCODE exons within 20 kb of the novel TAR. Using this criterion we select a subset of 1846 novel TARs, which we group into novel TAR clusters as described below.

We construct a matrix of correlation coefficients between novel TARs in this set (correlations between novel TARs further than 20 kb apart are set to zero). We use k-means clustering (Hartigan et al. 1979) with a k of 102, which meets the criterion set by Hartigan (1975) (see Methods for more details). With this value of k we obtain 96 clusters that have three or more elements and are localized to one ENCODE region. The 6 remaining clusters, which are not considered, correspond to small groups of only 2 elements and one large group of novel TARs from multiple chromosomes, which is the set of remaining unclustered TARs. A summary of statistics for these novel TAR clusters is in table 3.

Step 3B: Clusters based on Phylogenetic Profiles

Following the preceding steps of the DART classification procedure (steps 1A, 1B, 2, and 3A) we have 4,748 novel TARs unassigned. We first filter out the 3,122 novel TARs with below average array signal. We then cluster the remaining 1,626 novel TARs with above average signal in a similar manner to the previous step using the phylogenetic profiles for 17 different species sequenced in the ENCODE regions instead of expression profiles (see Methods for more details). A correlation matrix is computed between phylogenetic profiles of novel TARs that are within 20 kb of each other. We then use k-means clustering on this matrix and find optimal clustering for a k of 111. This clustering yields 100 clusters of 3 or more groups of TARs containing a total of 782 novel TARs, with a

[1] This estimation of a p-value of less than 0.05 takes into account the multiple testing of the expression profile of a novel TAR with on average 19 known exons within 20 kb. The p-value for obtaining a Pearson correlation of 0.9 for two 11 dimensional vectors is less than 10^{-3}.

median cluster size of 7. Summary details of these clusters are also available in table 3. As with the k-means clustering using expression profiles the majority of the novel TARs are in one unclustered group.

DART (Database for Active Regions with Tools)

DART (DART.gersteinlab.org) has been developed to facilitate the flexible storage, visualization, and analysis of the growing number of experimentally defined sets of regions detected using genomic tiling microarrays. These are either sets of transcriptionally active regions (TARs) or sites of transcription factor binding called binding active regions (BARs) or more generally active regions (ARs). DART has been designed to address a number of challenging issues that arise when attempting to store and analyze this type of data. These challenges will clearly grow in the future, as the ENCODE project expands from the analysis of 1% of the genome to the entire genome, and as more, increasingly diverse sets of ARs are experimentally determined. The key aspects of DART include the following.

1. Dealing with heterogeneous datasets: DART needs to be able to incorporate a rapidly growing number of sets of ARs which have been derived from a wide variety of experimental conditions. The current DART design is a first step towards allowing for multiple sets of ARs to be analyzed in a flexible fashion, including analyzing the unions and intersections of multiple sets and viewing the overlap among ARs in different AR sets.

2. Flexibility for storing different AR attributes: DART allows the flexible storage of different types of attributes associated with ARs, such as sequence information and array fluorescent signal intensities, as well as the adjustable groupings of ARs into subsets or clusters (potentially forming novel transcribed loci for the case of TARs). To accommodate this diversity, we use the Entity-Attribute-Value (EAV) data storage technique (Nadkarni et al. 1998) to define the attributes of either individual active regions or sets of active regions without modifying the database structure or program. These attributes can be used to search for desired AR sets.

3. Accommodating new genome builds: DART is designed to handle problems that occur as new builds (versions) of the human genome are defined and as the annotation associated with each AR set is updated to accommodate each new genome build. DART can store multiple values for AR genome locations corresponding to different genome builds. These coordinates are updated using the UCSC LiftOver tool, which maps between genome builds (Kent et al. 2002).

4. Integrated linking to other web resources for broader visualization and analysis: DART contains a number of capabilities designed to facilitate the integrated visualization and analysis of the data. These include both the ability to pass selected AR sets to the Active Region Comparer (ARC) for comparative analysis and annotation and the ability to display overlap among the ARs of different sets. Also, as described above, DART is integrated at several levels with the UCSC Genome Browser (Kent et al. 2002).

See figure 6a for an overview of the current implementation of DART's functionality. More details of which are provided in the Methods section.

Active Region Comparer Tool

The Active Region Comparer (ARC) provides a web-based interface for comparing, filtering, and annotating multiple sets of genomic regions, such as sets of TARs. The tool facilitates the analysis of ARs by determining how the regions in each set overlap those in other sets and by generating summary statistics to describe these relationships. ARC therefore allows the user to find regions that are common to multiple sets as well as regions that are specific to one set and not another. Additionally, by interfacing with a local EnsEMBL database (Birney et al. 2006) ARC can obtain a region's genomic annotation, which includes the sequence of the region, overlapping or nearby annotated transcripts, and other details such as the lengths and coordinates of overlapping and nearby exons. ARC also has an interface for exporting and visualizing multiple data sets via the UCSC Genome Browser, which displays sets of ARs alongside sets of genomic annotation to provide a graphical overview of the selected region. A diagram of how ARC works and its connectivity with the main DART database is presented in figure 6b. ARC also has the functionality to view individual ARs together with surrounding transcription start sites, CpG islands, known transcription factor binding sites and a local G/C content map using TAR-Vis (supplementary figure 2). See Methods for further details about the inner workings of the ARC tool as well as the TAR-Vis visualizer connected to it.

Observations Concerning Novel TARs

Tandem Duplicated TARs

While attempting to remove novel TARs that were likely caused by cross-hybridization we found that none of the 658 novel TARs that had a BLAST e-value of 10^{-5} or better had a corresponding blastTAR located in a different ENCODE region. A naïve expectation would be that, given that the ENCODE regions account for one percent of the human genome approximately one percent of the BLAST matches would be within the ENCODE regions. However, we find that there are 396 blastTARs located in the same ENCODE regions as their corresponding TARs. Of these TARs, 64 are located within 1kb of the original TAR and 144 are located within 20kb. Of the 396 blastTARs, 249 of them are actually different novel TARs (this makes sense, for if they have similar sequences they would typically also be detected as transcribed by the tiling arrays). These tandem sets of matching TARs come from many of the ENCODE regions, with the following three regions being most overrepresented: ENm006 (chromosome X from 152,635,144 to 153,973,591 with respect to human genome build NCBI Build 35), ENm007 (chromosome 19 from 59,023,584 to 60,024,460) and ENr233 (chromosome 15 from 41,520,088 to 42,020,088). These three ENCODE regions have tandem arrays of paralogs likely arising from segmental duplications (e.g. the ENm007 has a family of immunoglobulin-like receptors).

These tandem sets of novel TARs might be caused by cross-hybridization. However, since they are located in regions arising from local segmental duplication, it is not clear that cross-hybridization is the cause. For this reason, we chose not to remove them from the set of novel TARs under investigation.

Comparison of sets of novel TARs with RACE products

We first compared the different sets of novel TARs against the so-called `RACEfrags' or RACE (Rapid Amplification of cDNA Ends) fragments generated by hybridization of cloned 5' RACE products off exons of known genes in the ENCODE regions (see [reference to ENCODE manuscript] for more details). The RACEfrags like transfrags or TARs, are identified as transcribed regions; however, they also indicate the connectivity of the extended 5' RACE products to the indexed exon from which the primer was selected. Thus, all 5' RACEfrags upstream of an annotated transcription start site correspond to a novel 5' end. The RACE reactions were done using RNAs from 12 tissues, different from the 11 cell lines and conditions that were used in mapping the TARs. We find that the set of all novel TARs has a 6 percent overlap with the RACEfrags while the set of novel TARs assigned to gene loci has a 12 percent overlap (a two-fold enrichment). By comparison the set of novel TARs grouped into novel TAR clusters only has a 0.4 percent overlap with the RACEfrags, as expected (see table 4). By comparison a randomly generated set of unannotated regions only has a 1.9 percent overlap with the RACEfrags (see Methods for further details).

Structural RNAs

We also investigated the differing potential for the various sets of TARs to form structural RNAs using RNAz (Washietl et al. 2005) (see Methods). The recently submitted ENCODE companion paper (Washietl et al. 2006) deals with a comprehensive analysis of structural RNAs in the ENCODE regions and discusses approaches for detection of structural RNAs using computational approaches and transcriptional evidence. Here we take a somewhat different focus, investigating what fraction of the classified sets of novel TARs have the potential to form structural RNAs. Using a relatively stringent threshold score from RNAz of 0.95, which corresponds to structural RNA of high confidence, we find that the set of novel TARs that can be associated with known gene loci has the largest fraction with significant scores. We also note that the set of novel TARs with unusual sequence composition has above average enrichment for structural RNAs. This finding most likely reflects the fact that this set of novel TARs tends to have higher G/C content, which can affect the prediction made by RNAz (again see table 4).

Protein Homology

By design the translated sequences of the initial set of 6,988 novel TARs do not have strong similarity to known protein sequences, since we filtered out those that have BLASTx matches to annotated genes in the genome (i.e. pseudogenes). However, there

may be some novel TARs that have distant homology to gene relics. Using the profile hidden Markov model software HMMER (Eddy et al. 1998), we find that only 6 of the translated novel TAR sequences have significant matches, all of which are located in intronic regions.

Comparison of novel TAR clusters with TSSs and transcription factor binding sites

To test the validity of the 96 novel transcribed loci generated using expression profiles, we compare these clusters of novel TARs with two other datasets that were generated in [reference to ENCODE manuscript], the set of CAGE tags and paired-end-tags (ditags). These datasets been combined to form a set of 1144 known and putative transcription start sites (TSSs). We find that 6 of the 96 novel TAR clusters have a TSS within 1 kb of either end (since the strandedness of a novel TAR cluster is undetermined). An example of one of these is shown in figure 7, where we see a novel TAR cluster comprising 4 novel TARs with the rightmost TAR overlapping a putative transcription start site. This example is in a region of chromosome 2 (from 118175232 to 118198192, build NCBI Build 35) where there are no other annotated transcripts. Comparing the set of novel TAR clusters to the composite list of promoters identified in [reference to ENCODE manuscript], we find that 23 of the 96 novel TAR clusters have an end that is within 1 kb of a composite promoter[2]. When we compare the 100 novel TAR clusters grouped on the basis of similar phylogenetic profiles we find that 34 have an end within 1 kb of a TSS while 32 have an end within 1 kb of a composite promoter. We performed a simulation for random clusters of similar genomic extent to our novel TAR clusters, and found that only 9.2 out of 100 would have an end within 1 kb of a TSS while 17.5 out of 100 would have an end within 1 kb of a composite promoter (see Methods for details of the simulation).

Testing connectivity of transcripts using RT-PCR and Sequencing

As a small-scale follow-up experiment we selected 23 novel TARs that were assigned to known gene loci. These were selected such that the novel TAR and its associated exon are both expressed in Placental Poly(A)+ RNA. Using primer pairs generated from the novel TARs and their associated known exons, 23 RT-PCR reactions were performed. We found that 9 out of the 23 primer pairs (39%) yielded a PCR product on the gel (with no band in the absence of RT), which is evidence for a transcribed sequence spanning both the TAR and the known exon. In addition another 23 pairs of novel TARs that were grouped as being part of a novel TAR cluster were tested for connectivity by selecting a primer from each novel TAR sequence. Of these again 9 out of the 23 (39%) yielded a PCR product that provides experimental support for the connectivity of these novel TARs in a spliced RNA transcript. An additional two pairs of primers were selected as negative controls, neither of which showed any PCR product. The gel for some of these PCR products is presented in figure 8a. Supplementary table 1 lists all of the pairs of regions

[2] There are 828 putative composite promoters on the list from [reference to ENCODE manuscript] which is a set of both known and predicted promoters. Promoters were predicted using multiple ChIP-chip datasets for promoter specific transcription factors and modifications. This set of promoters is available at DART.gersteinlab.org.

tested for connectivity as well as the presence or absence of a RT-PCR product. When we see a PCR product generated from primers for a pair of novel TARs or for a novel TAR and an exon, it implies that both of the sequences are transcribed and that the product is likely a portion of a spliced transcript that utilizes and connects both of the sequences.

In order to verify the PCR reactions, 5 PCR products were then directly sequenced using their respective forward and reverse primers. The five PCR products yielded sequences which align to the transcribed sequences tested for connectivity, only one of which could not be counted as confirmationed. This PCR product was probably caused by cross-hybridization due to the sequence mapping better to another genomic location. Of the four PCR products which were confirmed by sequencing, one of them yielded a spliced sequence (see figures 8b and 8c) and three produced a sequence that was not spliced and included the intervening sequence between the two regions tested. These sequenced PCR products are shown in supplementary table 2. Even though not all the sequenced products were spliced, the results do confirm the RT-PCR products. Thus 4 of the 5 PCR products that were sequenced unambiguously confirm the connectivity of the associated pairs of sequences tested.

DISCUSSION

We have developed the DART system for the classification and categorization of the large quantities of novel transcribed regions that have been identified in the human genome. We can assign each novel TAR with reasonable confidence to one of the following sets: novel TARs that are likely caused by unusual sequence composition or cross-hybridization, novel TARs that can be assigned to known genes and novel TARs that can be clustered into novel transcribed loci. This last category of novel TARs possibly corresponds to entirely new transcripts.

To encapsulate our classification we have constructed DART, a database and tool set designed for the storage and visualization of large quantities of TAR sets and all of their additional features. DART is also designed to have a flexible framework that can incorporate any information associated with sets of TARs. DART and its companion tool ARC facilitate the comparison and display of multiple sets of TARs (or a set of Active Regions such as transcription factor binding sites) either though its own custom interface or via the UCSC Genome Browser.

We find that the set of novel TARs identified by the ENCODE Consortium has a number of interesting characteristics. There is enrichment in the potential for novel TARs to form structural RNAs compared to random sequences. This trend is especially prominent for the novel TARs that are associated with known gene loci. Some of these might correspond to structural RNAs that are coregulated with genes. We also find a significant overlap between the ends of clusters of novel TARs (novel transcribed loci) derived from either expression or phylogenetic profiles with both transcription start sites and promoters. There is also a significant enrichment among the novel TARs assigned to known gene loci for overlap with the 5' RACE extensions (or RACEfrags) of known genes identified in [reference to ENCODE manuscript].

We followed up our classification procedure by experimentally testing the connectivity of novel TARs that were assigned to known genes. Using RT-PCR, we found that 39% of the 23 novel TARs tested could be identified as part of a transcript that utilized the sequence of the novel TAR and at least one exon of the known gene. In principle, not all novel TARs that are assigned to known genes must be part of alternative isoforms of known transcripts. Some might correspond to other RNAs that are co-regulated with transcripts from the locus. In addition we tested the connectivity of identified clusters of novel TARs using RT-PCR. Again, we found that 39% of the 23 pairs of novel TARs yielded a PCR product, which is evidence of both the transcription and connectivity of the novel TARs within a single transcript. When a RT-PCR product is obtained from pairs of primers sourced from separated genomic regions (either two novel TARs or a novel TAR and an exon) this confirms that both regions are transcribed and utilized as part of a single spliced transcript (of which the PCR product is a piece). Of the 5 PCR products sequenced, 4 of the sequences match uniquely to the correct genomic location and further verify the results obtain by RT-PCR.

The datasets that were employed in the analysis presented in this paper were from the transcript maps derived from 11 different cell lines and conditions for the ~1% of the human genome included in the ENCODE regions. The statistical power of this procedure will increase non-linearly as the number and size of the data sets increases: as the number of data sets increases, so will the accuracy with which novel TARs can be associated with known genes. In addition, when transcript maps cover the entire genome, we will be able to more confidently remove novel TARs that are caused by cross-hybridization. In the next phase of the ENCODE project, there will be many more data sets generated that will span the entire genome. The methods developed here can be employed to initially classify the large amount of novel transcription that will be identified. This classification followed by medium-scale experiments will lead to a better understanding of the function of the multitude of RNAs that are transcribed in human cells. This iterative approach, consisting of analysis followed by more detailed experiments that feed back to improve the analytical methods, will lead to a more complete understanding of the diversity of transcripts of the human genome.

MATERIALS & METHODS

Experimental Testing of Connectivity of Genomic Regions by RT-PCT and Sequencing

Primer pairs were selected for 23 novel TARs that are expressed in placental RNA and are assigned to known gene loci. The primer sequences were selected from each novel TAR as well as from the exon of the gene with which the novel TAR had the strongest correlation. An additional 23 primer pairs were selected from pairs of different novel TARs that are present in placental RNA and could be clustered together using their expression profiles. An additional 2 pairs of primers were selected as negative controls from novel TARs that are located on different chromosomes. The regions selected and the corresponding primer sequences are available in supplementary table 1. 1μg of

Human placenta poly(A)+ RNA was used in a final volume of 20μl Reverse Transcription (RT) reaction (50ng/μl). RT reactions were primed by Oligo dT using Superscript™ II reverse transcriptase 200U in 20μl reactions (Invitrogen, CA, USA). In parallel, reactions without reverse transcriptase (RTase minus) were also performed as the negative controls for genomic contamination. RT was followed by PCR amplification using the Advantage™ 2 PCR Enzyme System (Clontech, CA, USA). The 2μl RT reaction and the 2μl RTase minus negative control from the above were used side by side in 50μl PCR reactions. The PCR program was started at 95°C for 30 seconds, followed by 35 cycles of 95°C for 15 seconds, 68°C for 1 minute, and concluded by an extension cycle of 72°C for 3 minutes. The PCR products were visualized on a 1% agarose gel. Five of the PCR products were then sequenced using both the forward and reverse primers.

Expression Profiles for Sets of Novel TARs and Known Exons

For each of the 11 different cell lines and conditions, a transcript map corresponds to fluorescent intensities for 755,457 25mer oligonucleotide probes tiling the non-repetitive sequence of one strand of the ENCODE regions. The array hybridizations in [reference to ENCODE manuscript] were done using double stranded cDNA, thus the signal maps correspond to the signals from both strands. The 11 cells lines and conditions are: GM06990 Poly(A)+ RNA, HeLa Poly(A)+ RNA, HL60 Poly(A)+ RNA (0 hour after treatment with retinoic acid, 2 hour after treatment with RA, 8 hour after treatment with RA, 32 hour after treatment with RA), Placental Poly(A)+ RNA, Neutrophil Total RNA, NB4 Total RNA (untreated, treated with RA, treated with TPA). The transcript maps are first scaled to each other using quantile normalization (Bolstad et al. 2003). An expression profile is then calculated for each novel TAR as well as for each known GENCODE exon by computing the median fluorescent signal from all the oligonucleotide probes contained within the boundaries of the TAR or exon. Exons that are not in the tile path of the Affymetrix ENCODE array are excluded.

Phylogenetic Profiles for Sets of Novel TARs

Phylogenetic profiles were generated using data derived from multi-species sequence alignment constructed by the ENCODE-MSA group ([reference to ENCODE manuscript]). In this analysis, we surveyed the presence/absence of novel TARs in the orthologous genomic regions of other species. Sixteen mammals (chimp, baboon, macaque, marmoset, galago, rat, mouse, rabbit, cow, dog, rfbat, shrew, armadillo, elephant, tenrec, monodelphis) were selected for this study, since they had received better sequence coverage than the other species used by the MSA group. A TAR was considered as "present" (given a value of 1 and otherwise 0) in a species if >1/3 of its content was detected in the MSA alignment from that species. We used the alignments constructed by the program TBA (Threaded Blockset Aligner) (Blanchette et al. 2004).

K-Means Clustering of Novel TARs

We use k-means clustering to form groups of nearby novel TARs. The k-means

clustering is done using the R statistical package with the default Hartigan et al. (1979) algorithm. We choose an appropriate value of k for optimal clustering using the rule of thumb of Hartigan (1975), where we find a k such that the weighted ratio of the sum of squares is significantly greater than 10 for (k-1) compared with k.

$$\left(\frac{SS \text{ within } (k-1) \text{ groups}}{SS \text{ within } k \text{ groups}} - 1 \right) * (n - k - 2) \geq 10$$

Where SS is the sum of squares and n is the number of novel TARs being clustered. We find the ratio is 143.4 for k=102 when clustering with expression profiles and the ratio is 78.3 for k=111 when clustering with phylogenetic profiles.

Implementation of DART (Database of Active Regions and Tools)

DART includes a relational database implemented in MySQL on a Linux server. There are tables for recording basic active region information such as chromosome, location, strand, sequence, and genome build number. Other tables and relations define higher-level objects such as sets of active regions, classes of sets, and attributes describing sets. In figure 6a we provide an overview of the DART's current functionality.

At the most general level, the user is presented with a listing of sets of ARs. These AR sets may be searched and selected in various ways and then passed to the ARC (AR Comparer) tool for further analysis. Alternatively, data about a single AR set may be viewed at successive levels of detail, e.g., 1) a summary of the AR locations by chromosome, 2) a summary of the AR locations by chromosomal segment, 3) a list of ARs found within a selected chromosomal segment, and finally 4) detailed information about a single AR, including a graphical indication of its overlap with ARs in other AR sets. From various DART screens, data can be passed to custom tracks in the UCSC Genome Browser (Kent et al. 2002) so that the DART data can be viewed in a broader context.

Software and Web pages access the DART database through library routines written in Perl. These library routines have a convenient object oriented structure. They support functions such as defining a genome build number, reannotating active regions for a new genome build, inserting active regions, defining sets and their attributes, and defining classes of sets. As objects are entered into DART, the library routines assign a unique accession number to each object created or inserted. Public domain Perl libraries are used to construct and display graphs on certain DART web pages. URLs are constructed to allow DART data to be sent to public browsers such as the UCSC Genome Browser (Kent et al. 2002).

The current implementation of DART represents a first step in confronting the challenges involved in manipulating and displaying heterogeneous AR datasets. As the amount of data, as well as the heterogeneity of that data, grows rapidly in the future, we will clearly need to extend and augment DART's capabilities to keep pace with the new challenges that arise. The code base for DART is downloadable from the DART website. All the

TAR data sets from [reference to ENCODE manuscript] as well as the results of this paper are available from DART (http://DART.gersteinlab.org).

Active Region Comparer Tool

The ARC site features four pages (see figure 6b), the first of which is the ARC Home page. ARC Home accepts formatted files[3] and DART datasets for upload, and it offers options for regulating the AR analysis. These options include filtering ARs on length, adding flanking sequences to each AR, grouping ARs by strand identifier, and mapping datasets from one build to another using a local copy of the UCSC liftOver tool.

ARC initiates AR analysis by flattening each file's genomic intervals onto a single coordinate axis such that any overlapping regions are combined to form a single region. ARC then performs combinatorial operations on these datasets using an algorithm that achieves high efficiency through a hierarchical series of unions and pairwise intersections. These operations may be used to perform one of two types of analysis. The first procedure determines which nucleotides are common to at least k out of n files, where k is a number between 1 and n, while the second procedure determines which nucleotides are common to exactly k out of n files. For each permutation, ARC generates a new dataset containing the corresponding genomic intervals. ARC also performs standard subtraction operations on two files, for which it generates new datasets as well.

The combinatorial algorithm described above minimizes run time by reducing the number of intersection operations that ARC must perform. It first takes the union of the genomic intervals in all n datasets to create a file that contains each region present in at least one of the original datasets (a.k.a. all regions). It then calculates all unique pairwise intersections among the original n datasets to create n choose 2 new datasets. The union of these datasets yields a file that contains each region present in at least two of the n original datasets. The next iteration of the procedure produces n choose 3 new datasets whose union produces a file containing regions in at least 3 of the n original datasets. When carried to completion, the algorithm creates n files (one for each iteration of n choose k). To ensure good performance time, every new group of datasets is clustered as shown in supplementary figure 3 so that the fewest possible intersection operations are performed. In the case where a user wishes to see one specific permutation only, instead of all n, ARC uses the algorithm described at http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dv_vstechart/html/mth_lexicograp.asp. This method requires fewer intersection operations when applied to a single permutation.

To present the results of the above computations, ARC displays summary statistics for each dataset in the ARC Results page (see figure 6b). ARC also creates new sets for the ARs in each chromosome of each full dataset, and it provides summary statistics for these

[3] ARC accepts files in BED format and files containing inclusive intervals. The Browser Extensible Data (BED) format uses a 0-based, half-open coordinate system. It was developed for the UCSC Genome Browser and is described fully at http://genome.ucsc.edu/FAQ/FAQformat#format1. The inclusive intervals option accepts 1-based, closed coordinates as used by Ensembl.

subgroups. All sets may be downloaded directly, or they may be further analyzed by the ARC Annotations page and the UCSC Display page.

The ARC Annotations page (see figure 6b) annotates and filters AR datasets using a local Ensembl database (Birney et al. 2006). Its options include grabbing features of the interval itself (sequence, G/C content, etc.), identifying overlapping transcripts and exons, and finding neighboring transcripts. The page also filters on AR length, G/C content, and classification (exon, intron, or intergenic). Processed datasets can be downloaded or exported to the UCSC Display page (see figure 6b).

The UCSC Display page facilitates the visualization of datasets by exporting them to the UCSC Genome Browser. Each dataset received by the Display page is loaded as a custom UCSC track in an in-frame version of the Genome Browser. These tracks can be viewed in the Genome Browser using either UCSC navigation tools or ARC hyperlinks. The tracks can also be analyzed with the UCSC tools. The UCSC Display page retains a history of exported datasets, and selecting multiple datasets from the history loads each one as a custom track in the UCSC browser, allowing for their direct comparison. These features provide a graphical interface for an otherwise abstract set of data points.

TAR-Vis

TAR-Vis is a collection of Perl scripts and modules that uses the open-source Bioperl modules (Stajich et al. 2002) and Ensembl's Perl API to automatically retrieve, analyze, and display sequences of genomic DNA containing a specific TAR or set of TARs. Given a chromosomal region and a genome build, TAR-Vis fetches the sequence region (including at least 1,000bp upstream and downstream in order to avoid boundary conditions on the subsequent calculations) from Ensembl's main databases and copies it to the local machine. From there, various calculations are run on the selected region, including Eponine transcription start site detection (Down et al. 2002), Cluster-Buster (Frith et. al 2003) transcription factor binding site detection (using the JASPAR TFBS database), CpG island detection, and G/C content graphing. Finally, all surrounding gene annotations are collected from Ensembl's annotation server. The resulting calculations and gene annotations are stored in a GFF3 file and visually presented using the Bio::Graphics module of Bioperl.

Generation of Randomized Sets of Novel TARs and Novel TAR Clusters

In order to assess the significance of the overlap of the different sets of novel TARs with the set of RACEfrags, a set of 7,000 random TARs were generated (comparable in size to the set of all novel TARs). This set of random TARs was selected so as to avoid intersecting any annotated GENCODE exon, to include only non-repetitive DNA sequence and to have the same length distribution as the set of all novel TARs detected on the ENCODE tiling arrays.

In order to compare the overlap of the ends of the novel TAR clusters within 1Kb of putative TSSs or composite promoters from [reference to ENCODE manuscript], we

created a random set of 1,000 novel TAR clusters whose length distribution was the same as that for the novel TAR clusters generated using either expression or phylogenetic profiles.

Accessing Structural RNA Potential of Novel TARs using RNAz

We used the following approach to predict structural non-coding RNAs (ncRNAs) with conserved and thus potentially functional secondary structures using the RNAz tool (Washietl et al. 2005): TARs were first collected and extended by 50 nucleotides on either side (this ensures detection of tightly structured ncRNAs, which may hybridize more poorly to microarrays than unstructured RNAs). All sequences were mapped to their corresponding TBA multiple sequence alignment blocks (23-way) constructed for the ENCODE regions. In each case, the human sequence together with the five most distant sequences, each sharing an overall sequence identity of at least 70% with the human sequence, were kept and analyzed using RNAz. Alignment blocks of 120 bp were subjected to analysis by RNAz, using an offset of 40 and considering both DNA strands independently (smaller alignment blocks of a minimum size of 50bp were analyzed without offset). When comparing different TAR sets, maximum RNAz scores were calculated for each TAR (the RNAz score, from 0 to 1, denotes the probability for a DNA sequence to encode a structural RNA, calculated based on support vector machine classification, Washietl et al. 2005).

ACKNOWLEDGEMENTS

We thank the ENCODE Project Consortium for making their data publicly available, specifically to the Genes and Transcripts, Transcription Regulation and Multiple Sequence Alignment groups for providing data. JR acknowledges Thomas Royce and Olof Emanuelsson for valuable discussions. This work was supported by grants from the National Institute of Health (NIH).

TABLE LEGENDS AND FIGURE CAPTIONS

Table 1: The sizes and percentages of coverage of the GENCODE exonic, pseudogenic (exons only) and unannotated regions are shown. The number and percentage of all TARs are shown for each of these partitionings. Unannotated regions are segmented into proximal intronic regions (closer than 5 kb to an exon), distal intronic regions, proximal intergenic regions, distal intergenic regions and regions corresponding to other ESTs that are not annotated as exons of GENCODE genes (also see figure 1a). Coverage and percentage is displayed for the number of novel TARs in each of these partitions. We observe that the number of novel TARs is significantly overrepresented for the intronic proximal and EST categories compared to the percentage coverage of these partitionings.

Table 2: Counts of the number of novel TARs in each of the classification sets: novel TARs with peculiar sequence composition, novel TARs associated with known genes, TARs caused by cross-hybridization and novel transcribed loci identified either using expression profiles or phylogenetic profiles (also see figure 1b).

Table 3: Summary statistics for the novel transcribed loci identified using either expression profiles of array signals or phylogenetic profiles. Genomic length is the genomic footprint of a cluster in the genomic sequence, while the putative transcript length corresponds to the sum of the lengths of the component novel TARs.

Table 4: Overlap of the sets of novel TARs with the mapped 5' RACE fragments (RACEfrags) and the fraction that are indicated by RNAz as potentially being a structural RNA at a score of 0.95. The set of novel TARs that are associated with known genes has the greatest enrichment for overlap with RACEfrags, while the set of novel TARs in novel transcribed loci has the least. We do not expect the novel TARs assigned to known genes to completely agree with the set of RACEfrags for a number of reasons. The RACEfrags are mostly 5' extensions of known genes (a small fraction corresponds to internal novel exons), but the novel TARs associated with known genes are not necessarily alternative isoforms of the gene transcript – some may be part of distinct but co-regulated RNAs. The set of novel TARs associated with known genes has the largest fraction that potentially corresponds to structural RNAs.

Figure 1a: Schema for the partitioning of TARs on the basis of location relative to GENCODE genes and pseudogenes (also see table 1). Proximal regions are located within 5 kb of the nearest GENCODE exon. 1b: Outline of the DART classification procedure of novel TARs. Novel TARs are first filtered on the basis of sequence composition (step 1), and then a fraction of the remaining novel TARs are associated with known genes (step 2). A portion of the remaining novel TARs are clustered in novel transcribed loci on the basis of expression profiles (EPs) and phylogenetic profiles (PPs) (step3). See table 2 for the numbers of novel TARs classified by each of these steps. The singlet and ambiguous TARs are what remains at the end of the classification procedure.

Figure 2: Summary of the features that are associated with each novel TAR and that are utilized by the classification procedure.

Figure 3: Plot of the distribution of GENCODE exons (blue line) and novel TARs (red line) against CC/GG di-nucleotide frequency. The distribution of novel TARs is skewed to high CC/GG di-nucleotide frequencies. A black arrow indicates the di-nucleotide frequency (0.155) above which only ~1% of the GENCODE exons are found. This threshold is used to filter novel TARs with peculiar sequence composition (CC/GG di-nucleotide frequency higher than 0.155).

Figure 4: Illustration showing how novel TARs can be associated with known genes by identifying novel TARs that are co-expressed with exons of known genes. Co-expression is determined by computing the Pearson correlation of expression profiles of array signals between novel TARs and nearby (closer than 20 kb) exons. The sizes of the circles correspond to the fluorescent signal intensity measured on the tiling arrays for each of the 11 different cell lines indicated by S_1 through S_{11}.

Figure 5: Plot of sensitivity against false positive rate for the assignment of exons of

known genes to the correct gene on the basis of the exon being co-expressed with other exons. The blue curve is calculated where an exon is allowed to be assigned to any gene in the genome, while the red and green curves are where the assignment is limited to genes which have exons within 100 kb and 20 kb of the target exon respectively. Restricting assignment to exons of nearby genes reduces the false positive rate of the assignment. The Pearson correlation of the best possible assignment for each exon is the threshold which parameterizes each curve.

Figure 6a: The current functionality of DART is displayed. At the top level one can access all the sets of ARs (either TARs or BARs) that are in the database. Upon selecting a collection of these sets one can either transfer sets to the ARC tool or inspect each set individually. At the individual set level, ARs can be viewed either at a complete set level, chromosomal level or a more local level. Individual ARs can be viewed with all their associated attributes. For an individual AR, DART also displays how it overlaps all other ARs in the database. Additionally at multiple levels these sets can be visualized via the UCSC Genome Browser. 6b: ARC Home accepts datasets from DART and from uploaded text files. Submission of the ARC Home form leads to the ARC Results page, which displays summary statistics for uploaded and newly generated datasets. From the ARC Results page, datasets may be downloaded, annotated in the ARC Annotations page, or visualized in the UCSC Display page. The Annotations page formats its processed datasets for presentation in HTML tables, for download as text files, and for export to the ARC Display page. Datasets sent to the UCSC Display page are loaded in the UCSC Genome Browser as custom tracks. From the UCSC Display page, one can return to the ARC Results page to repeat these analyses for other datasets.

Figure 7: Plot of a novel transcribed locus identified using the expression profile (the clustered TARs are shown in blue). Other novel TARs that are not part of this cluster are shown in red. In green we see the overlap of a putative transcription start site with the likely 5' end of this cluster. There are no annotated transcripts in the region displayed (chromosome 2 from 118,175,232 to 118,198,192, NCBI Build 35). We also observe transcript maps for the 11 different cell lines and conditions (not all novel TARs are shown in this region).

Figure 8a: Image of an agarose gel of RT-PCR products results from testing the connectivity between novel TARs and exons of known genes as well as between pairs of novel TARs clustered as a novel transcribed locus. Experiments were performed using Placental Poly(A)+ RNA, where "+" indicates the presence of reverse-transcriptase and "-" indicates its absence. "L" indicates the molecular weight ladder. A table of regions tested and their corresponding ids and primer sequences is located in supplementary table 1. 8b: Example of a pair of novel TARs (id B15) predicted to be associated with each other, potentially as part of a single transcript. This was confirmed by RT-PCR using placental RNA and was also successfully sequenced. The region displayed is on chromosome 21 from 34,270,568 to 34,270,998 (NCBI build 35). The sequence obtained from the PCR product is shown in red, the two connected novel TARs are in blue and the forward and reverse primers are in black. There are no annotated transcripts in the region displayed. 8c: Alignment of the sequenced PCR product against the genomic sequence

shows that the transcript which connects the two novel TARs is spliced.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. Oct 5;215(3):403-10.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M. 2004. Global identification of human transcribed sequences with genome tiling arrays. Science. Dec 24;306(5705):2242-6. Epub 2004 Nov 11.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, et al. 2006. Ensembl 2006. Nucleic Acids Res. Jan 1;34(Database issue):D556-61.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14: 708-715.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. Jan 22;19(2):185-93.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, et al. 2003. ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. Jan 1;31(1):68-71.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, et al. 2005. The transcriptional landscape of the mammalian genome. Science 309, 1559-63.

Cawley, S., Bekiranov, S., Ng, H.H., Kapranov, P., Sekinger, E.A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A.J., et al. 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499-509.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, et al. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. May 20;308(5725):1149-54. Epub 2005 Mar 24.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM.. 2006. A high-resolution map of transcription in the yeast genome. Proc Natl Acad Sci U S A. Apr 4;103(14):5320-5. Epub 2006 Mar 28.

Down TA, Hubbard TJ. 2002. Computational detection and location of transcription start

sites in mammalian genomic DNA. Genome Res. Mar;12(3):458-61.

Manak JR, Dike S, Sementchenko V, Kapranov P, Biemar F, Long J, Cheng J, Bell I, Ghosh S, Piccolboni A, Gingeras TR. 2006. Biological function of unannotated transcription during the early development of Drosophila melanogaster. Nat Genet. Oct;38(10):1151-8. Epub 2006 Sep 3.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. Jan 1;30(1):207-10.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics. 14(9):755-63.

Emanuelsson O, Ugrappa Nagalakshmi, Deyou Zheng, Joel S. Rozowsky, Jiang Du, Zheng Lian, Alexander E. Urban, Viktor Stolc, Sherman Weissman, Michael Snyder, Mark Gerstein. 2006. Assessing the Performance of Different High-density Tiling Microarray Strategies for Mapping Transcribed Regions of the Human Genome. Genome Research, in press.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306, 636-40.

Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. Nucleic Acids Res. Jul 1;31(13):3666-8.

Hartigan, J. A. 1975. Clustering Algorithms. Wiley.

Hartigan, J. A. and Wong, M. A. (1979). A K-means clustering algorithm. Applied Statistics 28, 100-108.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, et al. 2006. GENCODE: Producing a reference annotation for ENCODE. Genome Biol. 7 Suppl 1:S4.1-9. Epub 2006 Aug 7.

Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G, et al. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res. Mar;14(3):331-42.

Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. Science. May 3;296(5569):916-9.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. Jun;12(6):996-1006.

Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, Deng XW. 2006. Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet. Jan;38(1):124-9. Epub 2005 Dec 20.

Nadkarni PM, Brandt C, Frawley S, Sayward F, Einbinder R, Zelterman D, Schacter L, Miller PL. 1998. Managing attribute-value clinical trials data using the ACT/DB client-server database system. Journal of the American Medical Informatics Association 5(2):139-151.

Rinn J, Euskirchen G, Bertone P, Martone R, Luscombe NM, Hartman S, Harrison PM, Nelson FK, Miller P, Gerstein M, Weissman S, Snyder M. 2003. The transcriptional activity of human Chromosome 22. Genes Dev. Feb 15;17(4):529-40.

Royce, T.E., Rozowsky, J.S., Bertone, P., Samanta, M., Stolc, V., Weissman, S., Snyder, M., and Gerstein, M. 2005. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet. 21: 466-475.

SantaLucia, J. 1998. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. 95: 1460-1465.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, Lehvaslaiho H, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002 Oct;12(10):1611-8.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, Bussemaker HJ, et al. 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster. Science. Oct 22;306(5696):655-60.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U.S.A. 102, 2454-2459.

Washietl S, Pedersen JS, Korbel JO, Fried C, Gruber A, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, et al. 2006. Structured RNAs in the ENCODE Selected Regions of the Human Genome. Submitted.

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, Pham P, et al. 2003. Empirical analysis of transcriptional activity in the Arabidopsis genome. Science. Oct 31;302(5646):842-6.

Zheng D, Zhang Z, Harrison PM, Karro J, Carriero N, Gerstein M. 2005. Integrated pseudogene annotation for human chromosome 22: evidence for transcription. J Mol Biol. May 27;349(1):27-45.

Table 1

Locations of all TARs

	Exonic	Pseudogenes	Unannotated Regions
Size of ENCODE Regions (bp)	1,776,157	144,745	28,077,158
Percentage of all ENCODE	5.9%	0.5%	93.6%
Number of TARs	3,666	195	6,988
Percentage of all TARs	33.8%	1.8%	64.4%

Locations of Novel TARs

	ESTs not in Exons	Intronic Proximal	Intronic Distal	Intergenic Proximal	Intergenic Distal
Size of Unannotated Regions (bp)	2,477,910	8,522,559	5,536,879	2,434,101	9,250,454
Percentage of Unannotated Regions	8.8%	30.2%	19.6%	8.6%	32.8%
Number of Novel TARs	1,194	3,006	864	772	1,300
Percentage of all Novel TARs	16.7%	42.1%	12.1%	10.8%	18.2%

Table 2: Sets of Classified Novel TARs

	Number	Percentage
Total	6,988	100.0%
With peculiar sequence composition	503	7.2%
Assigned to known genes	955	13.7%
Caused by cross-hybridization	-	-
In novel transcribed loci using expression profiles	681	9.7%
In novel transcribed loci using phylogenetic profiles	782	11.2%

Table 3

Summary Statistics for 96 Clusters of Novel TARs using Expression Profiles

	Minimum	Median	Average	Maximum
Number of TARs	3	6	7.1	21
Genomic Length (bp)	2315	21,819	23,225	76,683
Putative Transcript Length (bp)	213	533	786	3791

Summary Statistics for 100 Clusters of Novel TARs using Phylogenic Profiles

	Minimum	Median	Average	Maximum
Number of TARs	3	7	7.8	14
Genomic Length (bp)	1,354	22,594	24,331	39,810
Putative Transcript Length (bp)	208	664	894	2,159

Table 4: Features of Novel TAR Sets

	Number	Overlap with RACEfrags	Percentage overlap with RACEfrags	Overlap with RNAz	Percentage overlap with RNAz
All Novel TARs	6988	434	6.2%	270	3.9%
TARs with peculiar sequence composition	503	30	6.0%	22	4.4%
TARs assigned to known genes	955	116	12.1%	55	5.8%
TARs in novel transcribed loci using expression profiles	681	3	0.4%	19	2.8%
Tandem repeat TARs	249	26	10.4%	5	2.0%

Figure 1a

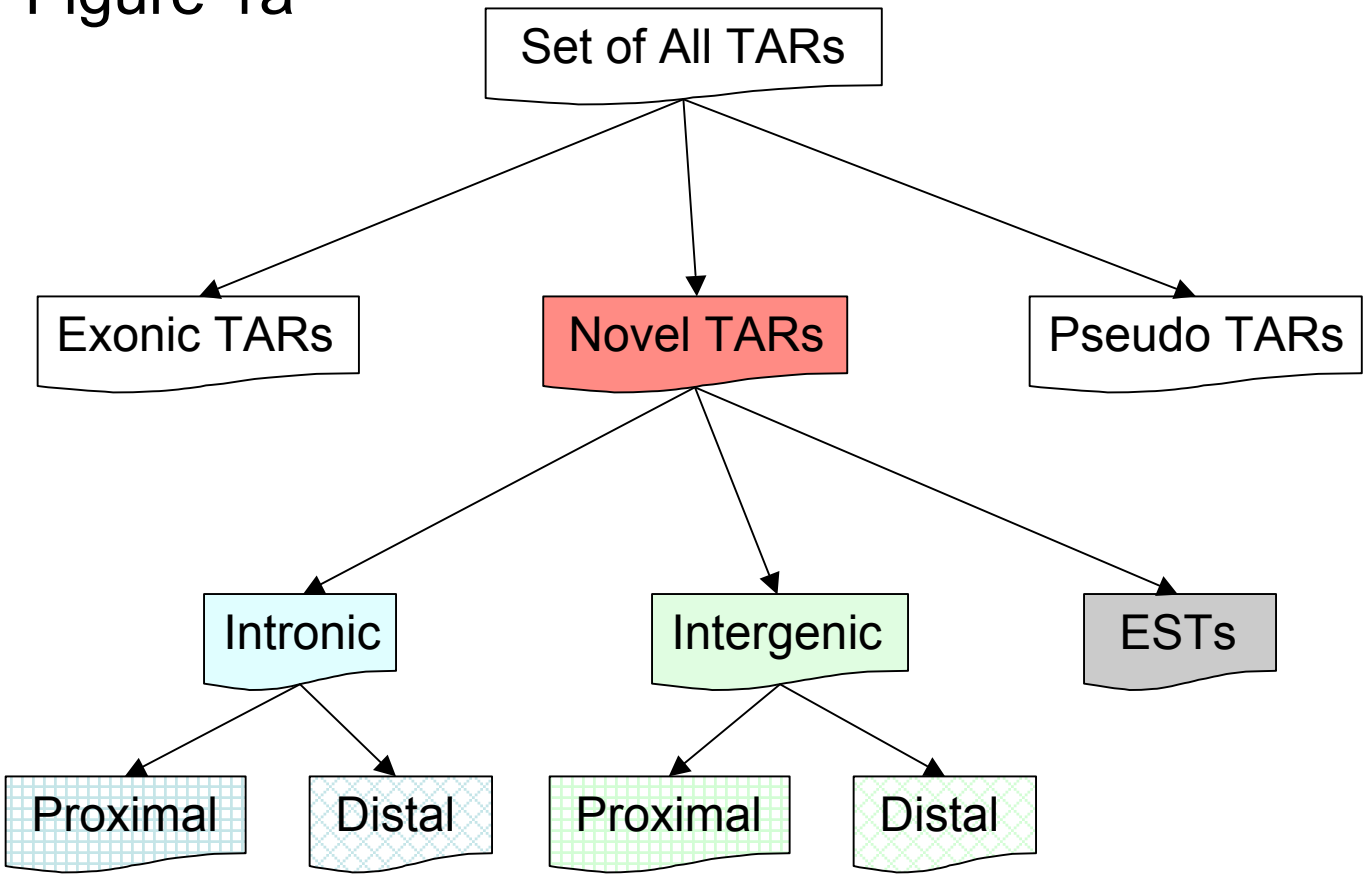

Set of All TARs
Exonic TARs
Novel TARs
Pseudo TARs
Intronic
Intergenic
ESTs
Proximal
Distal
Proximal
Distal

Figure 1b

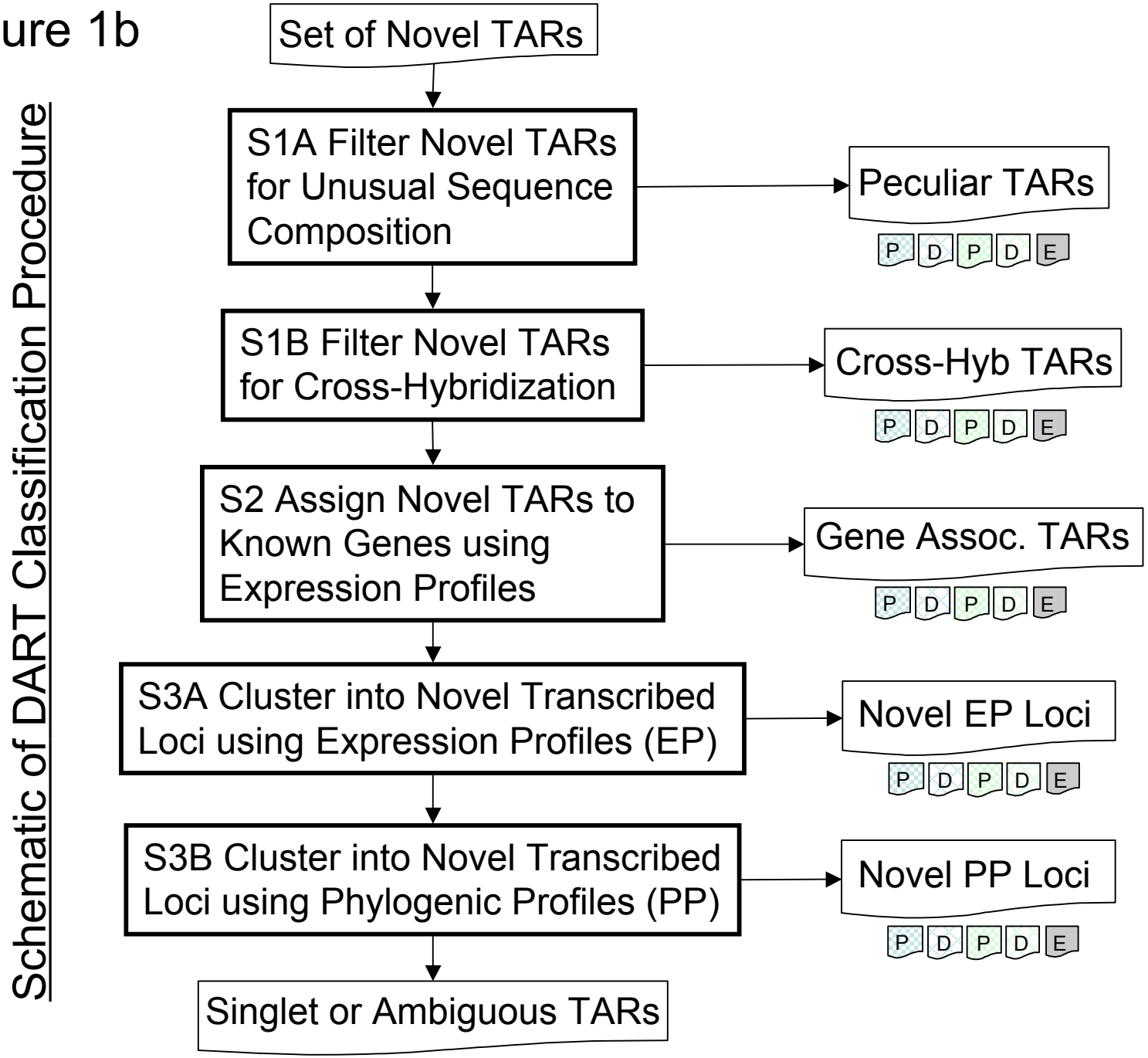

Schematic of DART Classification Procedure
Set of Novel TARs
S1A Filter Novel TARs for Unusual Sequence Composition
Peculiar TARs
P D P D E
S1B Filter Novel TARs for Cross-Hybridization
Cross-Hyb TARs
P D P D E
S2 Assign Novel TARs to Known Genes using Expression Profiles
Gene Assoc. TARs
P D P D E
S3A Cluster into Novel Transcribed Loci using Expression Profiles (EP)
Novel EP Loci
P D P D E
S3B Cluster into Novel Transcribed Loci using Phylogenic Profiles (PP)
Novel PP Loci
P D P D E
Singlet or Ambiguous TARs

Figure 2

Characteristics of TARs	
Key Features:	• Expression profile of array signals for 11 cell lines and conditions • Genomic location relative to GENCODE/HAVANA genes
Relationship to Genomic Features	• Vicinity to TSSs from CAGE tags and ditags • Overlap with TARs from other array experiments • Vicinity to promoters identified by ChIP-chip/ChIP-PET
Sequence Features	• Sequence composition of TARs • Phylogenic profile of TARs
Functional Assignment	• Sequence similarity to protein sequences (using HMMER) • Potential for being a structural RNA (using RNAz)

Figure 3

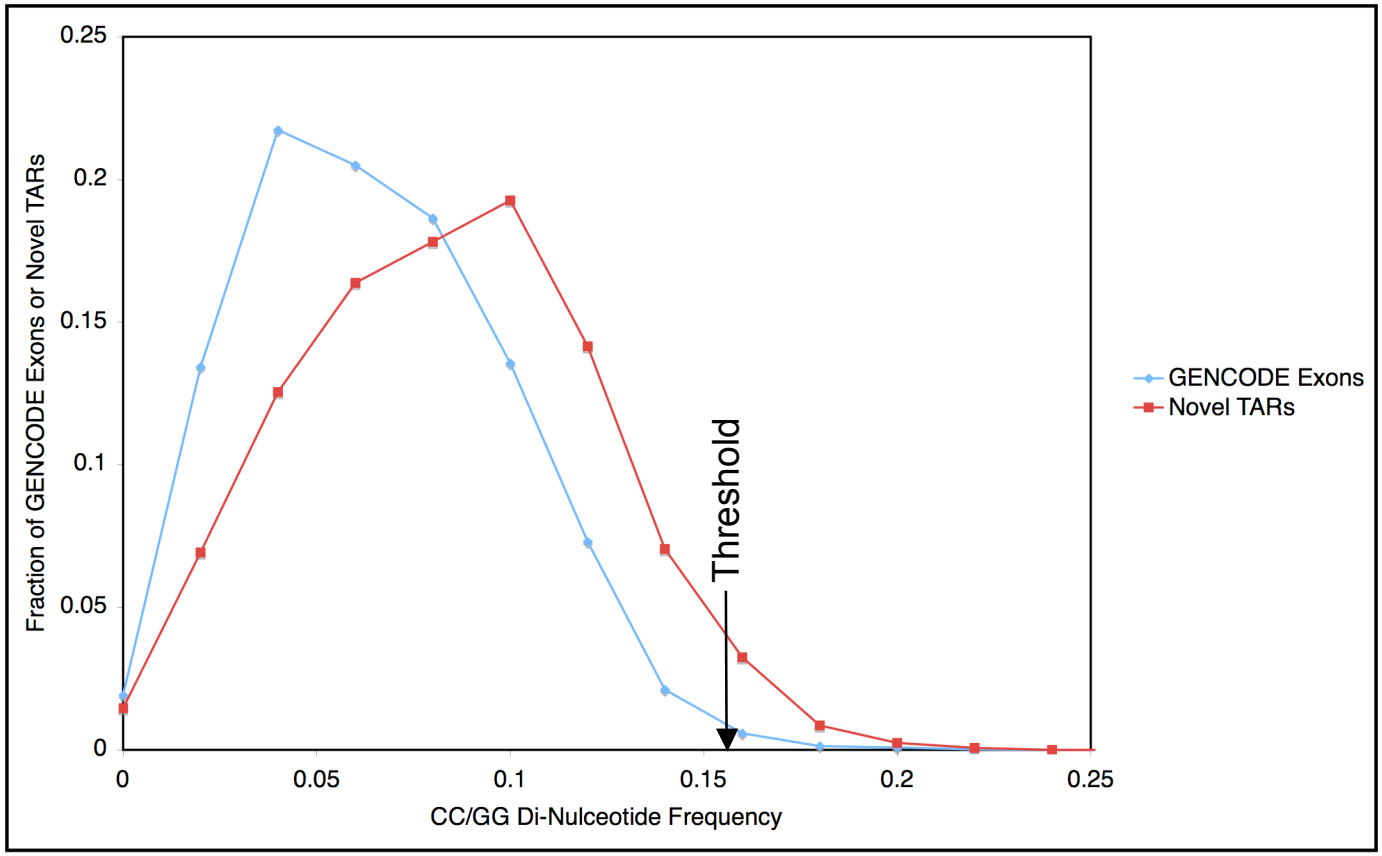

0.25
0.2
0.15
0.1
0.05
0
Fraction of GENCODE Exons or Novel TARs
0
0.05
0.1
0.15
0.2
0.25
CC/GG Di-Nulceotide Frequency
Threshold
GENCODE Exons
Novel TARs

Figure 4

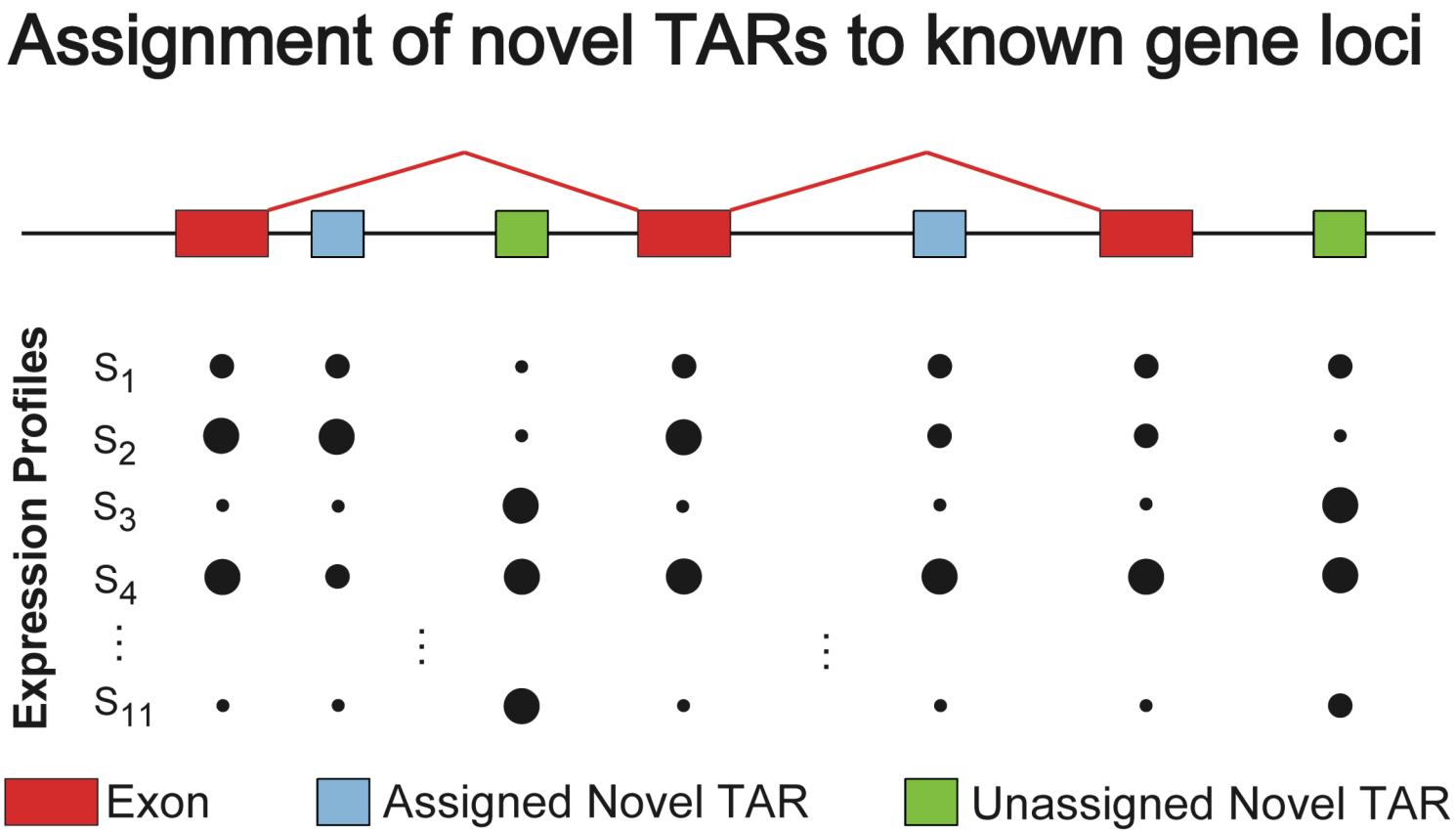

Assignment of novel TARs to known gene loci
Expression Profiles
S1
S2
S3
S4
S11
Exon
Assigned Novel TAR
Unassigned Novel TAR

Figure 5

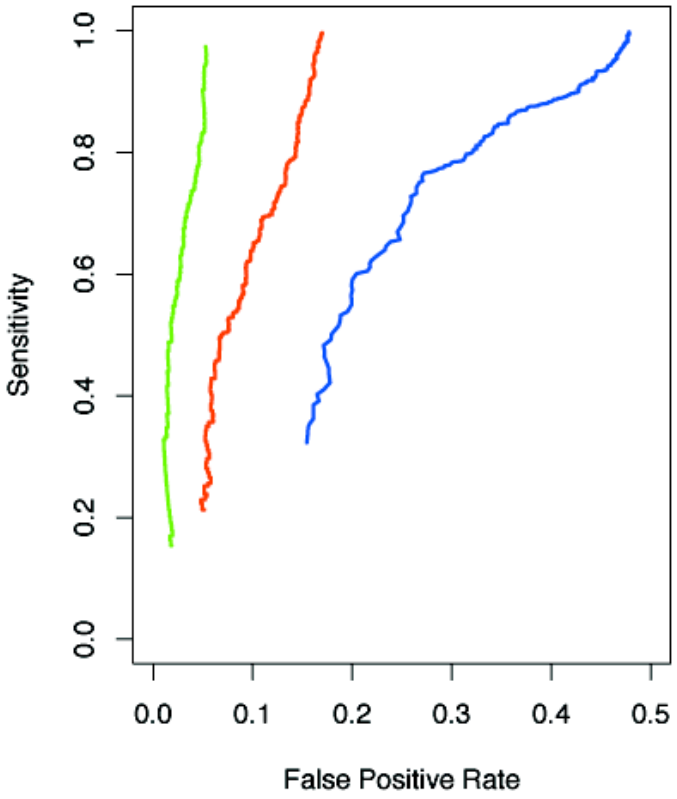

Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
False Positive Rate

Figure 6a

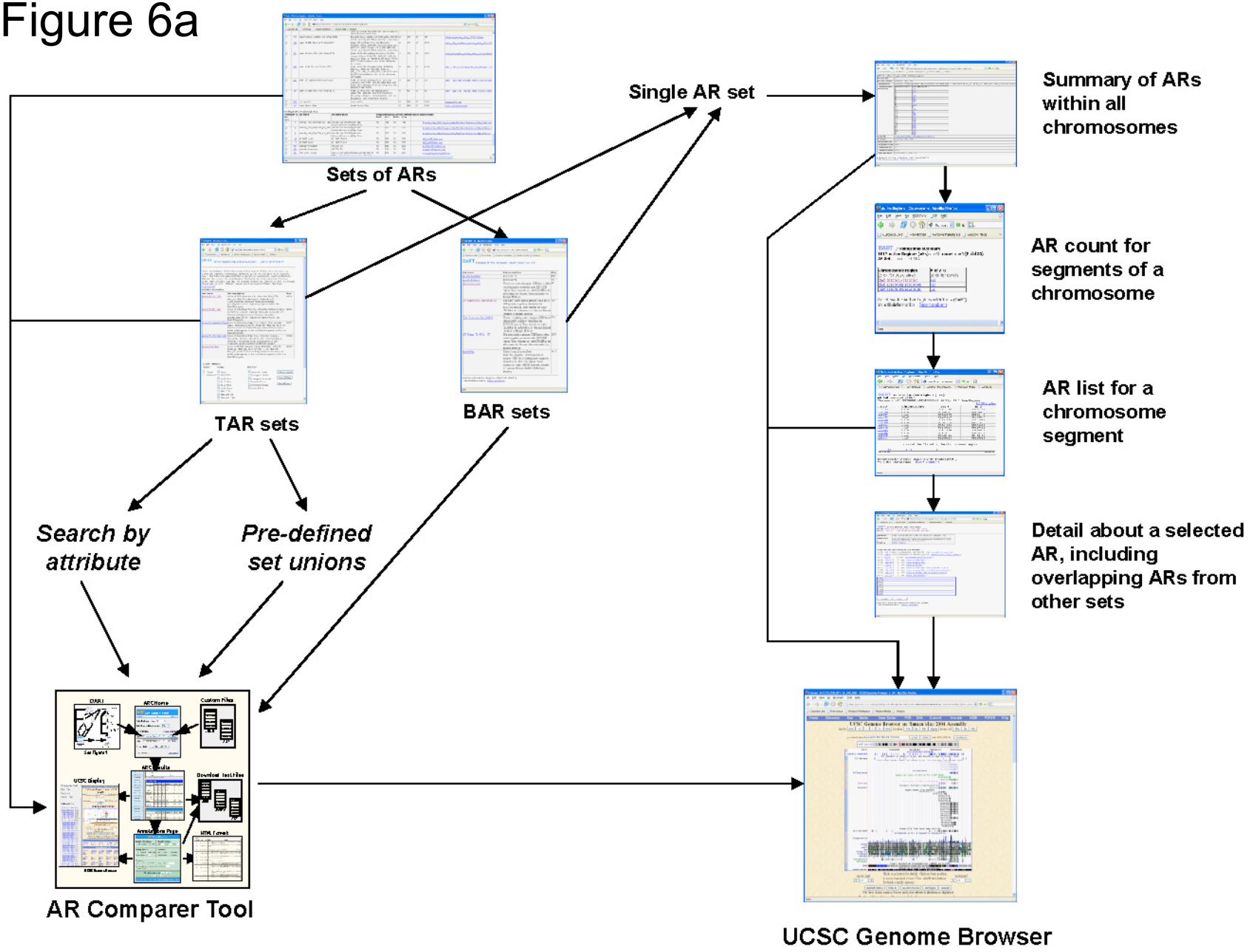

Sets of ARs
Single AR set
Summary of ARs within all chromosomes
AR count for segments of a chromosome
AR list for a chromosome segment
Detail about a selected AR, including overlapping ARs from other sets
TAR sets
BAR sets
Search by attribute
Pre-defined set unions
AR Comparer Tool
UCSC Genome Browser

Figure 6b

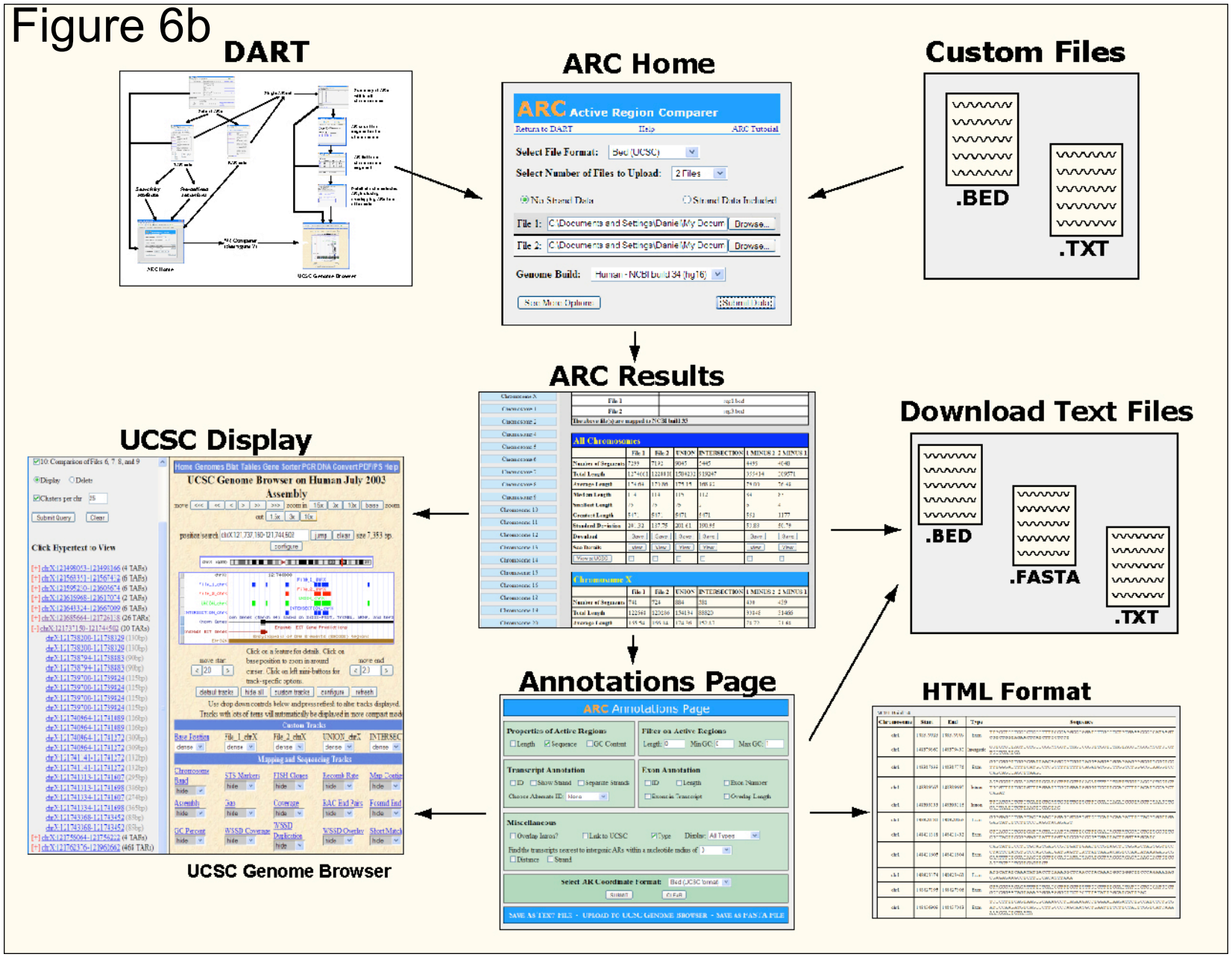

DART
ARC Home
ARC Active Region Comparer
Custom Files
.BED
.TXT
ARC Results
Download Text Files
.BED
.FASTA
.TXT
UCSC Display
UCSC Genome Browser on Human July 2003 Assembly
UCSC Genome Browser
Annotations Page
ARC Annotations Page
HTML Format

Figure 7

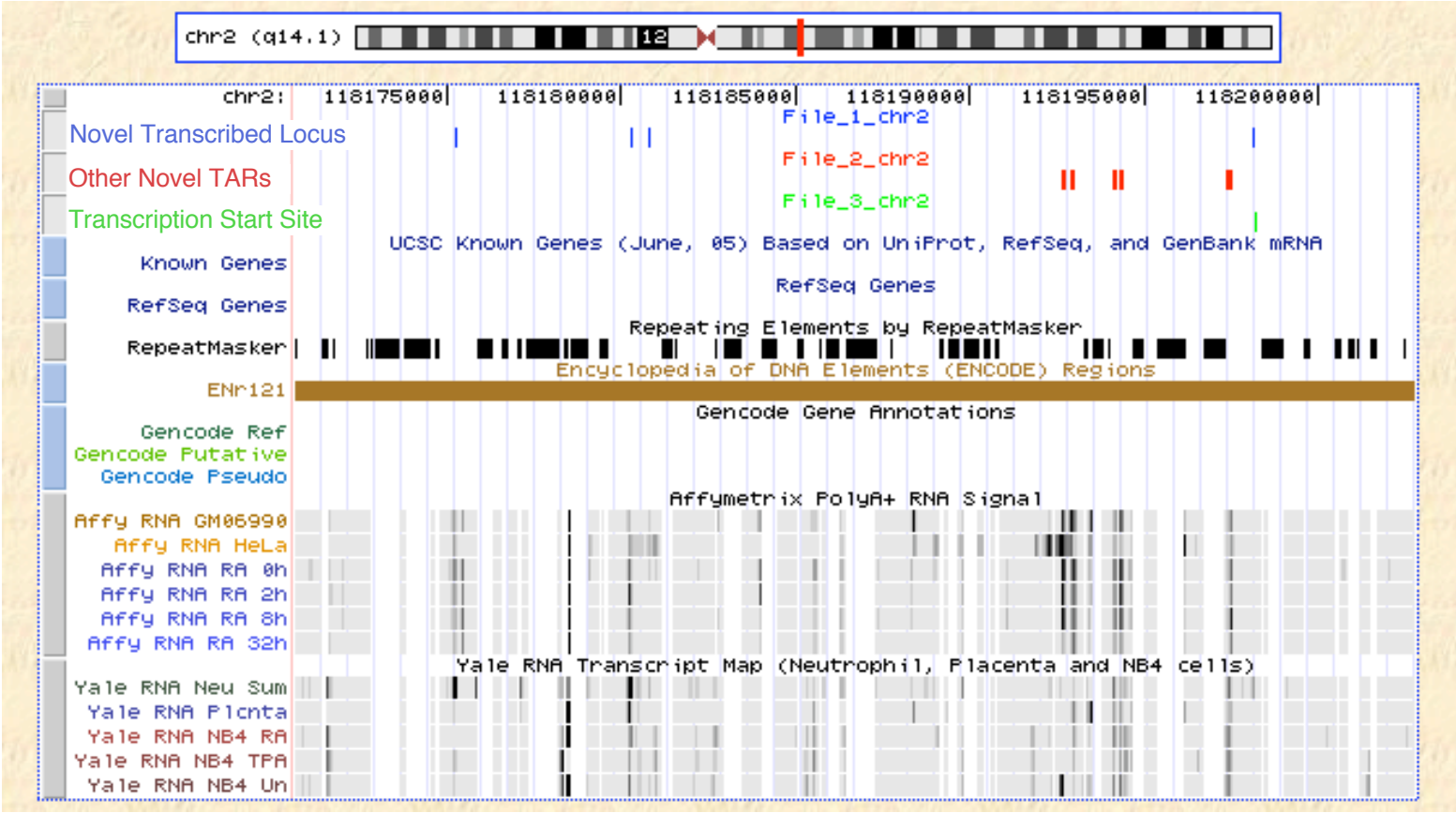

chr2 (q14.1)
chr2:
118175000
118180000
118185000
118190000
118195000
118200000
Novel Transcribed Locus
Other Novel TARs
Transcription Start Site
File_1_chr2
File_2_chr2
File_3_chr2
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
Known Genes
RefSeq Genes
RefSeq Genes
Repeating Elements by RepeatMasker
RepeatMasker
Encyclopedia of DNA Elements (ENCODE) Regions
ENr121
Gencode Gene Annotations
Gencode Ref
Gencode Putative
Gencode Pseudo
Affymetrix PolyA+ RNA Signal
Affy RNA GM06990
Affy RNA HeLa
Affy RNA RA 0h
Affy RNA RA 2h
Affy RNA RA 8h
Affy RNA RA 32h
Yale RNA Transcript Map (Neutrophil, Placenta and NB4 cells)
Yale RNA Neu Sum
Yale RNA Plcnta
Yale RNA NB4 RA
Yale RNA NB4 TPA
Yale RNA NB4 Un

Figure 8

A)

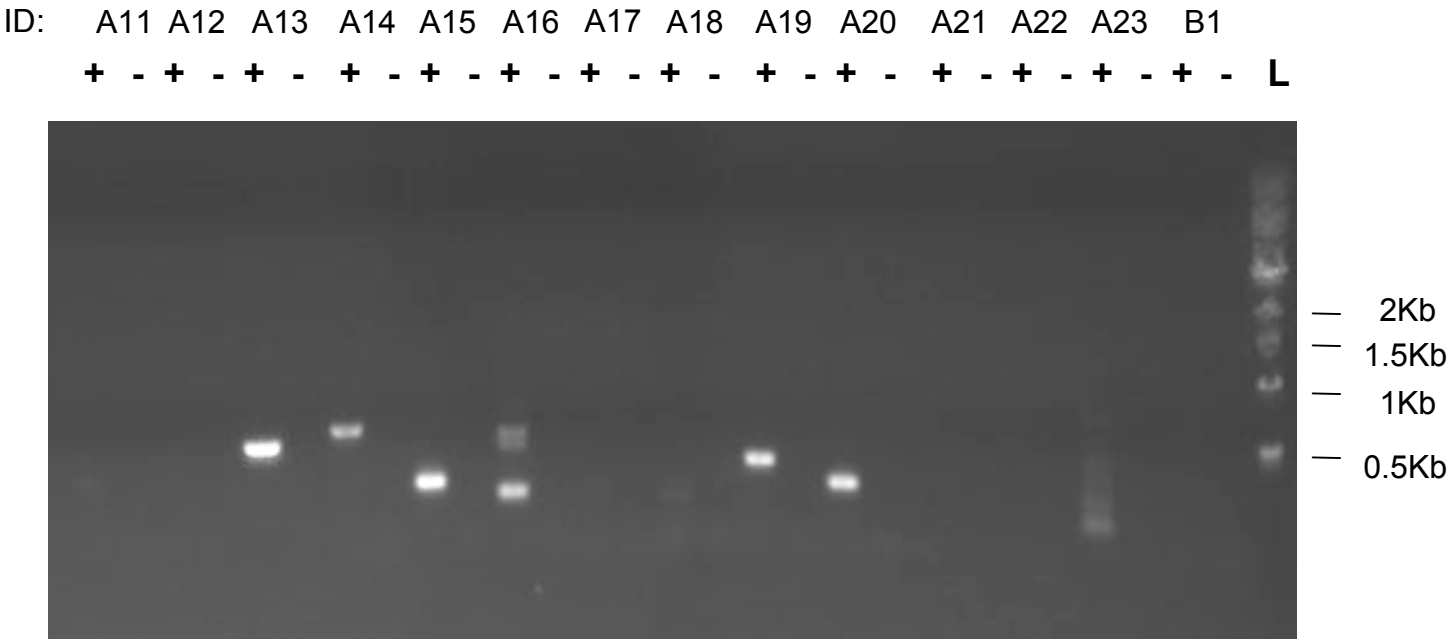
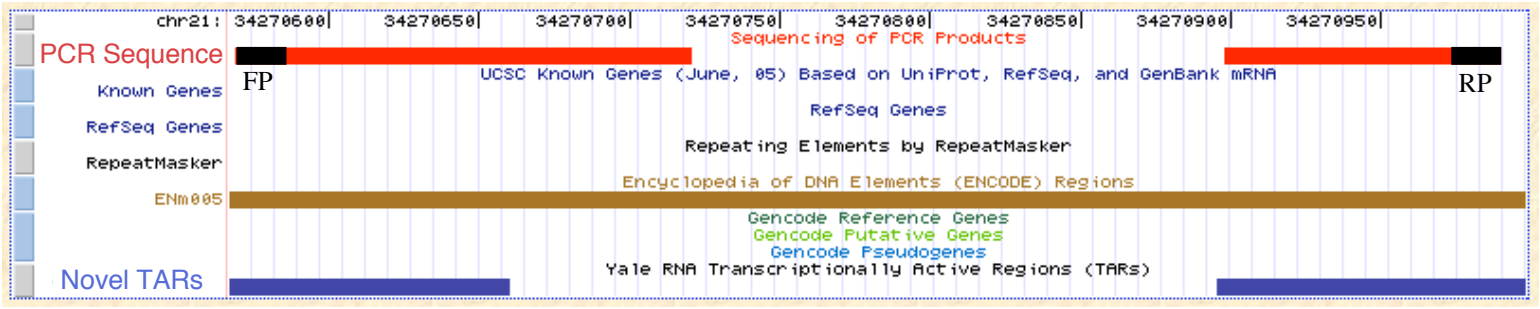

ID: A11 A12 A13 A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 B1
+ - + - + - + - + - + - + - + - + - + - + - + - + - + - L
2Kb
1.5Kb
1Kb
0.5Kb

Figure 8

B)



C)

TAR 1

Forward Primer

```
PCR Sequence      1 ttcttcggaaaagcacatgaactctttggagtctcctgttccacttggtaaatttcctat 60
                    ||||||||||||||||||||||||||| |||||||||||||||||||||||||||||||||
Chr21   34,270,569 ttcttcggaaaagcacatgaactcttcggagtctcctgttccacttggtaaatttcctat 34,270,628

PCR Sequence     61 agctccgcactgaaagtccctgctgccctccttcctctgagcttgtggggcccacagatc 120
                    |||  ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Chr21   34,270,629 agccacgcactgaaagtccctgctgccctcctcctctgagcttgtggggcccacagatc 34,270,688

PCR Sequence    121 ccctgctccacttcctgcttcatttcagctgat 153
                    |||||||||||||||||||||||||||||||
Chr21   34,270,689 ccctgctccacttcctgcttcatttcagctgat 34,270,721
```

TAR 2

```
PCR Sequence    154 ggatgacactccctcgttctaataccatctgaatgcctgagcaattacatcttacaacct 213
                    ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Chr21   34,270,898 ggatgacactccctcgttctaataccatctgaatgcctgagcaattacatcttacaacct 34,270,957

PCR Sequence    214 catgaaaaacacagcagcttgtcacgatgaatg 246
                    |||||||||||||||||||||||||||||||
Chr21   34,270,958 catgaaaaacacagcagcttgtcacgatgaatg 34,270,990
```

Reverse Primer