BIOINFORMATICS ORIGINAL PAPER
Vol. 29 no. 2 2013, pages 255–261
doi:10.1093/bioinformatics/bts676

Databases and ontologies
Advance Access publication November 19, 2012

Non-redundant compendium of human ncRNA genes in GeneCards

Frida Belinky[1,*], Iris Bahir[1], Gil Stelzer[1], Shahar Zimmerman[1], Naomi Rosen[1], Noam Nativ[1], Irina Dalah[1], Tsippi Iny Stein[1], Noa Rappaport[1], Toutai Mituyama[2], Marilyn Safran[1,3] and Doron Lancet[1]

[1]Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 76100, Israel, [2]Computational Biology Research Center (CBRC), Advanced Industrial Science and Technology (AIST), 2-41-6, Aomi, koto-ku, Tokyo 135-0064, Japan and [3]Department of Biological services, The Weizmann Institute of Science, Rehovot 76100, Israel

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Non-coding RNA (ncRNA) genes are increasingly acknowledged for their importance in the human genome. However, there is no comprehensive non-redundant database for all such human genes.

Results: We leveraged the effective platform of GeneCards, the human gene compendium, together with the power of fRNAdb and additional primary sources, to judiciously unify all ncRNA gene entries obtainable from 15 different primary sources. Overlapping entries were clustered to unified locations based on an algorithm employing genomic coordinates. This allowed GeneCards' gamut of relevant entries to rise ~5-fold, resulting in ~80 000 human non-redundant ncRNAs, belonging to 14 classes. Such 'grand unification' within a regularly updated data structure will assist future ncRNA research.

Availability and implementation: All of these non-coding RNAs are included among the ~122 500 entries in GeneCards V3.09, along with pertinent annotation, automatically mined by its built-in pipeline from 100 data sources. This information is available at www.genecards.org.

Contact: Frida.Belinky@weizmann.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 19, 2012; revised on October 18, 2012; accepted on November 14, 2012

1 INTRODUCTION

In recent years, there has been a massive expansion of knowledge regarding the non-coding RNA (ncRNA) gene universe. Following the Encode project (Birney *et al.*, 2007), it is now accepted that while ~80% of the genome is transcribed, only ~2% of the genome is protein coding. Whereas an increasing body of evidence (Cawley *et al.*, 2004; Ponjavic *et al.*, 2007) supports the existence of purifying selection as well as binding of transcription factors to regulate the expression of ncRNA genes and lends strength to some ncRNA entries, it is still unclear what fraction of this immense molecular repertoire is functional (Carninci *et al.*, 2005; Cawley *et al.*, 2004; Ponjavic *et al.*, 2007; Struhl, 2007). Furthermore, well-documented ncRNA genes do not exceed 1% of all transcripts. Obviously, this state of affairs points to a potential of tremendous future augmentation of ncRNA gene counts.

Some classes of ncRNA genes (e.g. rRNA and tRNA) have been known for a very long time (Holley *et al.*, 1965; Spencer *et al.*, 1969), while the discovery of the diversity and magnitude of ncRNA genes has been accelerated in the past few years, reviewed in Mattick and Makunin (2006). Unlike protein-coding genes, which have been extensively studied in structure, orthologue identification and single-nucleotide polymorphism annotation, the study of ncRNA genes lags behind.

There exist more than a dozen ncRNA gene databases. Some are dedicated to particular ncRNA classes, such as miRBase (Kozomara and Griffiths-Jones, 2011), snoRNA-LBME-db (Lestrade and Weber, 2006) and piRNABank (Sai Lakshmi and Agrawal, 2008), while others such as Rfam (Gardner *et al.*, 2009), RNAdb (Pang *et al.*, 2007) and fRNAdb (Mituyama *et al.*, 2009) include a variety of ncRNA classes. Many ncRNA genes are also included in the conventional gene databases—Ensembl (Flicek *et al.*, 2012), NCBI's Entrez Gene (Maglott *et al.*, 2011), HGNC (Seal *et al.*, 2011) and GeneCards (Safran *et al.*, 2010; Stelzer *et al.*, 2011). Interestingly, these databases vary greatly in their human ncRNA gene counts, partly because of the use of different sources and integration mechanisms. Grand unification is thus urgently needed (Bateman *et al.*, 2011).

We have initiated some steps towards this ambitious goal based on the integration capacities within GeneCards. GeneCards is an integrated human gene compendium, which strives to consolidate information about all human genes (Safran *et al.*, 2010; Stelzer *et al.*, 2011). Its previous V3.07 included 15 118 RNA genes, mined from a limited number of sources: Ensembl, Entrez Gene and HGNC (Flicek *et al.*, 2012; Maglott *et al.*, 2011; Seal *et al.*, 2011). We have launched an expansion and integration protocol, based chiefly on the use of fRNAdb (Mituyama *et al.*, 2009), which in turn is mined from seven sources currently untapped by GeneCards. Using judicious unification protocols, we report here the augmentation of the ncRNA gene count to 79 344 in V3.09. While this 5-fold enhancement is in large part due to the addition of 21 812 non-redundant piRNA genes (Sai Lakshmi and Agrawal, 2008), our pipeline has also resulted in the addition of 36 151 genes belonging to other classes, yielding a comprehensive, upgradable compendium of ncRNA genes.

*To whom correspondence should be addressed.

2 SYSTEM, METHODS AND ALGORITHM

2.1 A genome location-based algorithm

The realm of ncRNA gene databases is a complex network of mutually linked data structures (Fig. 1). The integration effort described herein (Supplementary Fig. S1) makes use of the hierarchical position of two databases, fRNAdb and GeneCards, each with a set of mined sources and with complementary data coverage. We use fRNAdb as a major source, as it is the most comprehensive among ncRNA-exclusive database, containing data from many primary and derived sources. Further, the uniform data structures provided by fRNAdb for all its sources reduce the complexity of handling many different data sources. As RNAdb and some of its primary sources (such as piRNA, Antisense, Evofold, RNAz and ncRNAsearch) are not expected to be updated, mining of fRNAdb poses no risk. In distinction, primary sources for which mining fRNAdb might compromise periodical updates are in parallel mined directly at the most updated version (miRBase 19.0, Rfam V11.0 Hinvdb 8.0 and NONCODE V3.0 for the presently described GeneCards V3.09). Our main pipeline thus involves the content merger of these databases into a widely expanded ncRNA category within GeneCards. For this, we have opted for a genome location-based algorithm for ncRNA gene integration, as described below.

2.2. Mapping to the genome

All the mined human ncRNA sequences were downloaded from the relevant sources (Fig. 1) and analysed (Fig. 2 and Supplementary Fig. S1). Mapping to multiple locations for the same gene was allowed when an ncRNA received exactly the same score in such locations. Entries from fRNAdb annotated as IRES, SECIS and UTR were not considered as genes and thus disregarded.

2.3. Clustering overlapping entries

We strived to overcome the problem of having several entries that map to appreciably overlapping positions. To cope with such redundant ncRNA entries, and to unite presumed parallel versions of the same gene, a clustering algorithm was applied to join entries with overlaps >70% of the genomic territory of the smaller partner, when occurring on the same strand. This unification process was applied judiciously, based on the class affiliations available in the mined sources, reflecting biological evidence. Only entries reported as belonging to the same ncRNA class were united. In 102 940 cases where an entry was not annotated as belonging to a specific ncRNA class (class = 'other;' Fig. 3), such entry was joined to another class, with a small minority (220 genes) joined to two to three relevant classes. Notably, neither RNAdb nor fRNAdb performed a

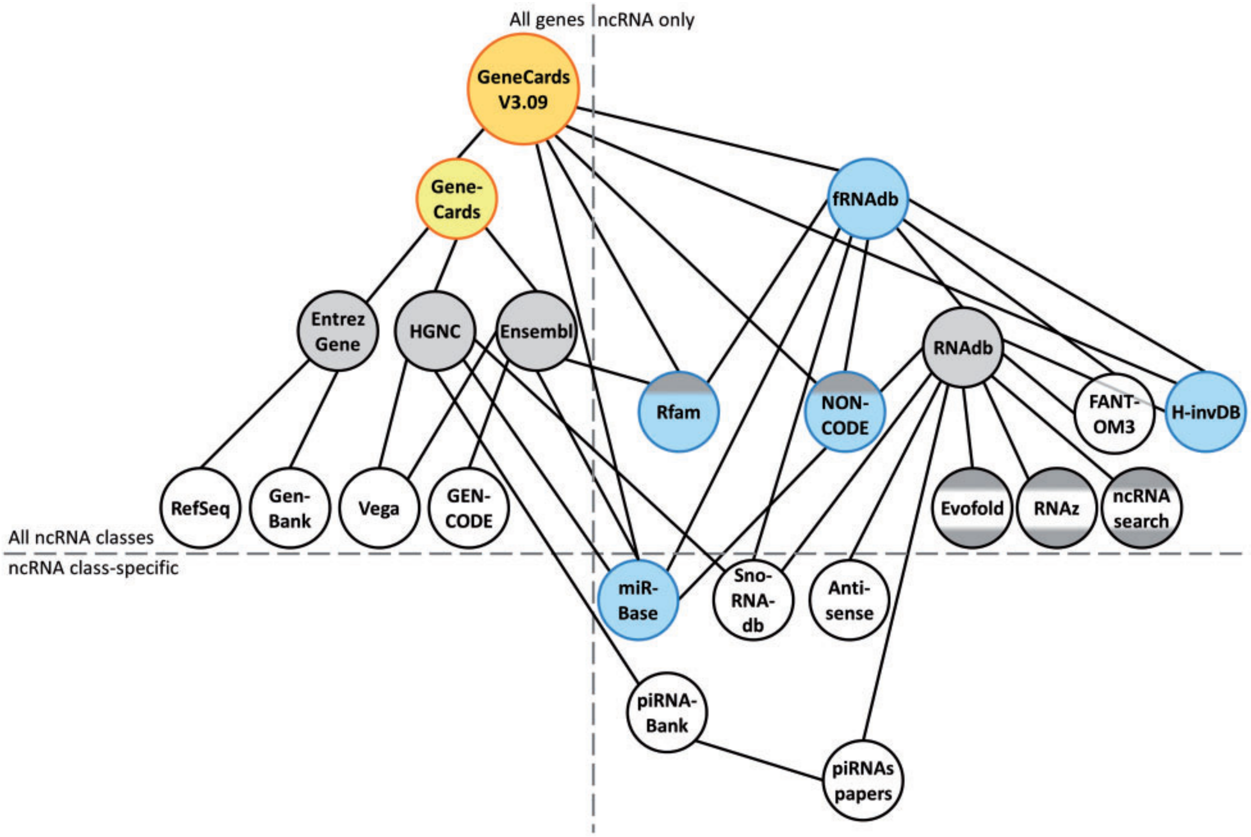

Fig. 1. The ncRNA database universe: unification of human ncRNA entries within GeneCards is based on hierarchical data mining flow as shown. Data sources are either derived (grey, also fRNAdb) or primary, with references as follows (see also Supplementary Table S1): Entrez Gene (Maglott *et al.*, 2011), HGNC (Povey *et al.*, 2001; Seal *et al.*, 2011), Ensembl (Flicek *et al.*, 2012), RNAdb (Pang *et al.*, 2007), Rfam V11.0 (Gardner *et al.*, 2009), NONCODE V3.0 (Bu *et al.*, 2012), FANTOM3 (Carninci *et al.*, 2005), H-invDB V8.0 (Yamasaki *et al.*, 2010), RefSeq (Pruitt *et al.*, 2012), GenBank (Benson *et al.*, 2011), Vega (Wilming *et al.*, 2008), GENCODE (Harrow *et al.*, 2006), Evofold (Pedersen *et al.*, 2006), RNAz (Washietl, 2007), ncRNAsearch (Torarinsson *et al.*, 2006), miRBase 19.0 (Kozomara and Griffiths-Jones, 2011), snoRNA-LBME-db 3 (snoRNAdb; Lestrade and Weber, 2006), predicted antisense ncRNAs (Antisense; Engstrom *et al.*, 2006), piRNABank (Sai Lakshmi and Agrawal, 2008) and piRNA papers (Aravin *et al.*, 2006; Girard *et al.*, 2006). The mined sources, fRNAdb V3.4, miRBase 19.0, Rfam V11.0, ncRNAs from HinvDB V8.0 and lncRNAs from NONCODE V3.0, marked blue. Purely predictive databases are indicated by two grey lines; semi predictive databases are indicated by one grey line

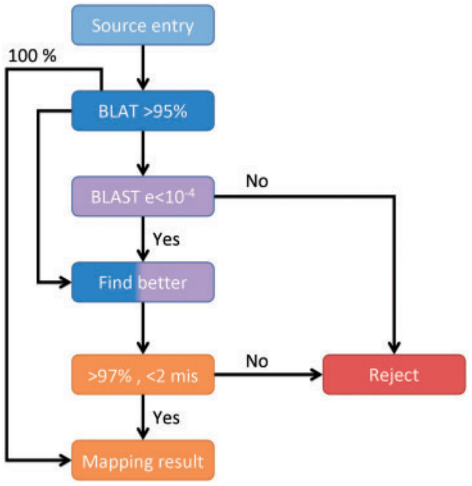

Fig. 2. For genome mapping of ncRNAs, a BLAT search was run for each source entry against the human genome build hg19 with non-default parameters: minimum score = 12, minimum match = 1, minimum identity = 95% (applied to the matched segment). The former two parameters allow more accurate detection of short ncRNAs and the latter was permissive to allow running this initial lengthy procedure only once, with liberty to select different stringencies ≥95% in subsequent steps. For searching lncRNAs from NONCODE V3.0, we used the non-default parameters: minimum score = 50, tile size = 18, allowing better efficacy for such lengthy sequences. Sequences that did not obtain a 100% BLAT match were subjected to a BLASTN search with *e*-value $\leq 10^{-4}$. Only sequences >50 nt were allowed to be interrupted by introns, based on the knowledge that short ncRNA classes (e.g. piRNAs ~30 nt and miRNAs ~20 nt) are intronless. The better of the BLAT and BLAST results was used to infer a tentative genomic location. Final filtering and determination of exact genomic coordinates were done by reanalysing the BLAT/BLAST outputs, applying a full-length identity cutoff of ≥97% or ≤2 mismatches for sequences <67 nt

similar integration process for the data obtained from the primary sources. Moreover, both allow for clear identification of the origin of the entry and provide the sequences as obtained from primary sources, without manipulation.

2.4. Unification of ncRNA clusters with GeneCards entries

Unification of ncRNA clusters with pre-existing GeneCards entries was performed with respect to all GeneCards genes, including protein coding as well as non-protein coding. For pre-existing ncRNA GeneCards entries, we used the same joining criterion as for the clustering algorithm, with the exception of not clustering piRNAs with piRNA clusters (PIRCs; see Section 4). For other GeneCards entries, joining was performed if the GeneCards entry's endpoint (start/end) is within 10 nt of the matched entry's endpoint, and the length of the shorter gene is ≥50% of the long one. Unification was actually performed on the individual ncRNA entries, and subsequently employed for the cluster as a whole. In the case that a cluster matched more than one GeneCards entry (1342 instances), multiple unification pointers were instituted. Protein-coding genes were not unified with their corresponding antisense ncRNA, since they reside on opposite strands and represent distinct functional entities. This results in an integrated GeneCards list, including all ncRNA clusters, whether joined or unjoined to existing GeneCards genes.

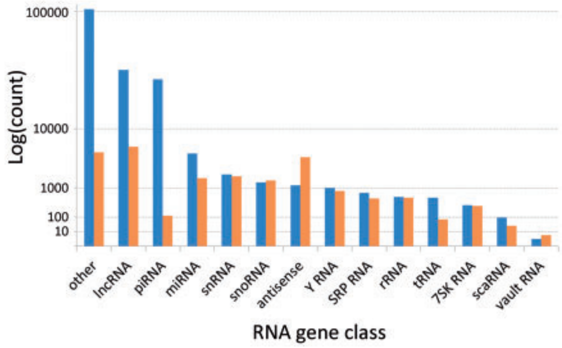

Fig. 3. ncRNA frequencies presented on a special root scale $Y = X(1/B)$, where $B = \log_2 10$ (Safran *et al.*, 2003; Shmueli *et al.*, 2003) of pre-existing GeneCards genes (orange) and human fRNAdb entries (blue) for the different ncRNA classes represented in GeneCards V3.09: piRNA—piwi-interacting RNAs, ~30 nt long RNAs derived from human testes which guide silencing of repetitive elements in germ cells (Girard *et al.*, 2006); lincRNA—in GeneCards, these are long intergenic ncRNA mined from Ensembl, identified according to methylation patterns (Flicek *et al.*, 2012; Guttman *et al.*, 2009), while in fRNAdb they are long ncRNA based on length >200 and sequence ontology (Ponting *et al.*, 2009); antisense RNA—involved in regulation of transcription (Morris and Vogt, 2010); snRNA—small nuclear RNA—including spliceosomal RNAs essential for splicing (McKeown, 1993); snoRNAs—small nucleolar RNA guide chemical modifications on other ncRNA genes (Kiss, 2001); Y RNA—required for DNA replication and involved in rRNA maturation (Christov *et al.*, 2006; Stein *et al.*, 2005); miRNAs—micro RNAs, ~20 nt long mediators of transcript silencing (Bushati and Cohen, 2007); tRNA—adaptors between codons and the coded amino acid (Kubli, 1981); rRNA—RNA components of the ribosome (Moore and Steitz, 2002) and 7SK RNA—inhibitors of RNA polymerase II (Blencowe, 2002)

2.5. Removal of certain predicted ncRNA singletons

Finally, candidate GeneCards entries stemming from a single prediction in the purely predictive ncRNA gene sources (Evofold, RNAz and ncRNAsearch), and seen in no other source, were excluded. This is to diminish the probability of false entries, stemming the estimated high false-discovery rate (50–70%; Gorodkin *et al.*, 2010; Washietl *et al.*, 2007). Singleton predictions from the semi-predictive sources Rfam and NONCODE were however maintained, as they are supported by additional information within the same source.

By the end of this four-stage process (Supplementary Fig. S1), ncRNAs not merged to existing GeneCards genes are added to the compendium as novel entries. As such, these genes do not have HGNC-approved symbols, and we opted to use the GeneCards id (Rosen *et al.*, 2003) as a newly assigned GeneCards symbol. For piRNAs, the symbol is taken from RNAdb (e.g. PIR61598). For piRNA multi-membered clusters, we use the identifier of one of the members; other members are included as aliases. In the future, every ncRNA gene assigned an HGNC official symbol will be changed accordingly.

2.6. Expression evidence and quality score

An important issue is how the accuracy and consistency of the catalogued genes can be ensured. We thus provide a quality score for each ncRNA gene, which appears along with the list of relevant identifiers in the Aliases and Descriptions section (see Section 3.5). The quality score Q is computed as the sum $Q = 10S_F + 5S_E + 0.2S_P + 0.5S_N$, where S_i denotes the count of data sources of the following kind: S_F, showing functional annotation; S_E, showing expression; S_P, reporting prediction and

S_N, none of the above. In this respect, GeneCards does not simply unify information about ncRNAs from other resources but also attempts to convey evidence parameters. GeneCards version 3.09 provides expression information for ~60% of its ncRNAs, as derived from four sources, miRBase (1706 GeneCards), NONCODE (7729 GeneCards), H-invDB (16 151 GeneCards) and fRNAdb (21 812 piRNA GeneCards). Of the remaining entries, 2506 ncRNAs are purely predicted as reported by Evofold, RNAz and ncRNAsearch. Expression evidence annotation from additional sources will be provided in upcoming versions.

2.7 Functional annotation

In terms of functional annotation, the ncRNA gene universe is still in its infancy. When it comes to protein-coding genes, GeneCards has an elaborate set of sections that provide functional information about genes. This includes Function, Pathways/Interactions, Drugs/Compounds, Expression, Genomic Variants and Disorders/Diseases (Safran *et al.*, 2010). Some of these sections provide links to laboratory products such as inhibitory RNAs, *in situ* assays, clones, SAGE tags and PCR assays. A recently established companion database, MalaCards (http://malacards.org/, cf. https://www.iscb.org/cms_addon/conferences/ismb2012/latebreakingresearch.php#LBR17), provides a disease-centric view with extensive gene-related links. This infrastructure will sub-serve the mining, dissemination and web display of pertinent information also for ncRNA genes. Currently implemented is miRTarBase (Hsu *et al.*, 2011) providing experimentally validated miRNA targets.

3 IMPLEMENTATION

3.1 Pre-integration status

To obtain an integrated compendium of ncRNA genes, we aimed to bring together such genes from two major derived sources, each representing a number of primary sources (Fig. 1) as well as several more updated primary sources (see Section 2). The first is fRNAdb with 126 406 human ncRNA entries and the second is GeneCards with 21 451 human ncRNAs, with the additional sources Rfam, miRBase, HinvDB and lncRNAs from NONCODE with respective counts of 2838, 3446, 23 407 and 33 829. Figure 3 shows the current numerical breakdown of different ncRNA gene classes in newly added ncRNA entries versus pre-existing GeneCards. We note that fRNAdb has 84 192 entries purely based on computational algorithm predictions (based on sequence conservation and secondary structure), as derived from three primary sources: Evofold, RNAz and ncRNAsearch. These were completely absent in GeneCards. Another large group of fRNAdb genes belongs to the piRNA class, constituting 32 148 entries. These were previously represented in GeneCards by a mere 114 entries with a PIRC* symbol, which are reported PIRCs (Seal *et al.*, 2011) that constitute piRNA precursor transcripts (Lin, 2007; Zamore, 2010). In a previous version of GeneCards (V3.07), there were only 15 118 entries categorized as RNA genes. This count is partially based on a restrictive definition which relies on the appearance of the string ‘RNA’ in several fields in the data sources mined. In preparation for the ncRNA unification effort, we employed a more consistent method of categorization relying more on the sources’ own category definition in the three datasets used (Ensembl, HGNC and NCBI’s Entrez Gene). This resulted in the addition of 6333 entries, previously annotated as ‘uncategorized’, to the ncRNA GeneCards roster used for unification, totaling 21 451 ncRNAs.

3.2 Mapping of fRNAdb and other source entries to the genome

Out of the total of 126 406 fRNAdb entries, we determined the genomic location for 124 676 by employing the BLAST and BLAT routines. The 1730 sequences for which the genomic location could not be determined based on our mapping criteria (Supplementary Fig. S1) were not considered further. Similarly, we were able to map 2294 (Rfam), 3444 (miRBase), 17 098 (HinvDB) and 33 712 (NONCODE) from the additional sources. Our genome mapping is very similar to the reported mapping via fRNAdb (after conversion from hg18 to hg19)—78% of the entries were mapped to the exact same set of locations per entry. However, we also determined genome mapping for 12% of the entries for which the genome mapping was either not reported by fRNAdb due to mapping suppression of multiple locations or due to inability to convert between genome builds.

3.3 Clusters of fRNAdb and other source entries

We applied the algorithm for generating clusters of overlapping RNA genes with unified location to all sequence-mapped fRNAdb entries and additional sources. This process identified 149 863 clusters, of which 135 779 were singleton ncRNA entries and 14 084 had 2–143 members (Fig. 4 and Supplementary Fig. S2). The latter encompass a total of 45 445 (30%) of the mapped ncRNA entries. Many of these overlapping entries are not unique to fRNAdb and appear also within the individual databases, as exemplified in Fig. 4 and available in GeneCards V3.09.

It is worth noting that the overlap between the predictive methods is relatively low (only ~6% purely predicted clusters were kept). The overlap between RNAz and Evofold was previously found to be only 7.2%, which is attributed to the higher sensitivity of Evofold to AU rich regions, while RNAz is more sensitive to GC rich regions, despite the fact that both were trained using the Rfam database (Washietl *et al.*, 2007).

3.4 Unifying fRNAdb and GeneCards

The 149 863 clusters of ncRNA genes were compared with all GeneCards entries. Only 17 648 of these were found to match at least one GeneCards entry, and 86% of these showed an overlap to targets included among GeneCards’ 21 451 ncRNAs. Notably, we found that 7801 pre-existing ncRNA genes in GeneCards do not match any cluster of the newly added ncRNA entries. In turn, 57 636 clusters had no match in GeneCards, including 21 812 piRNAs and 33 318 non-computationally predicted RNA genes. These genes have now been added to GeneCards version 3.09. After this unification process, GeneCards includes almost 80 000 human ncRNA genes (Fig. 5), now the largest category within this compendium, quadruple the number of recorded protein-coding genes.

3.5 Annotation

The ncRNA GeneCards entries, both originally present and integrated from fRNAdb and additional sources are annotated by the same procedures used throughout this database (Safran *et al.*, 2010). A crucial facet is the alias section, in which, when relevant, fRNAdb identifiers, as well as primary and derived database identifiers, with deep links to the relevant source data facilities.

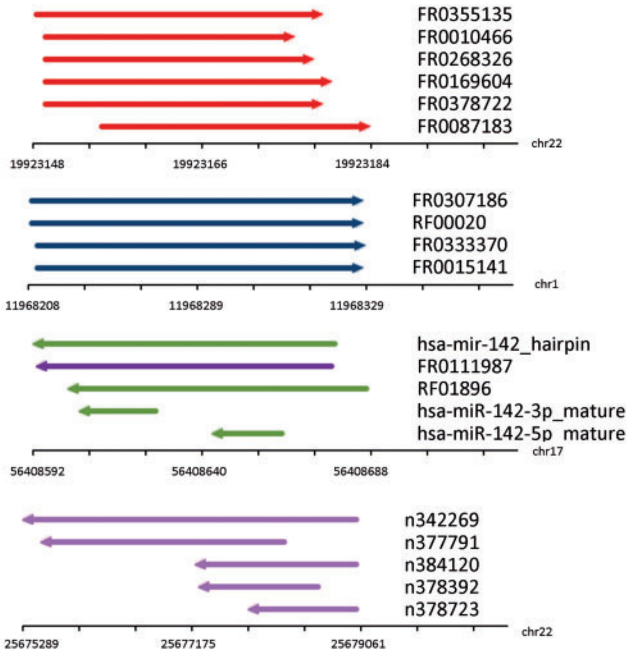

Fig. 4. Four examples of clustered overlapping RNA genes with unified location, represented by their member ncRNA entries. Red—piRNAs, Blue—snRNAs, Purple—predicted ncRNAs, Green–miRNA, Pink—lncRNA

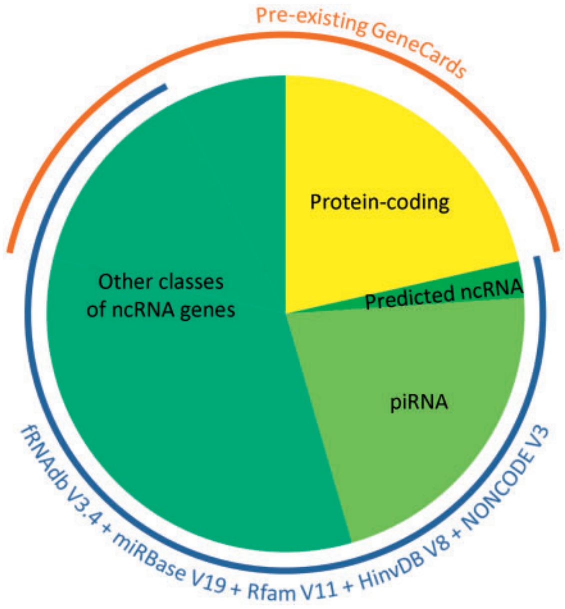

Fig. 5. The ncRNA grand unification in GeneCards 3.09. Protein coding (21 660; Yello) and ncRNAs (79 344; Green) as available in GeneCards V3.09 containing a total of 101 004 molecularly identified functional gene entries. The arcs indicate presence in the two mined sources, with an overlap of 13 907 entries. Shown types: predicted ncRNA (2506), piRNA (21 812) and others (33 318)

The ncRNA class affiliation, as well as additional functional information is shown. The genomic location of the entire cluster as well as the separate genomic locations of the cluster members are provided. Other annotations from fRNAdb, such as secondary structure, summaries, OMIM IDs and PubMed identifiers for articles, are also available in the appropriate sections.

4 DISCUSSION

The field of ncRNA genes is a dynamic one; new genes and even new classes are continuously discovered. We have strived to generate a comprehensive integrated list of human ncRNA genes within GeneCards, thereby placing them in the context of a complete human gene compendium. For that, two large and mutually complementary sources of human ncRNA genes were put to use, conveying the content of 15 more restricted databases. This resulted in the inclusion of 79 344 GeneCards for ncRNAs, when compared with 15 118 in GeneCards V3.07, and to 21 451 after the reclassification of 6333 previously ‘uncategorized’ entries as ncRNAs, respectively, ~5-fold and ~3.5-fold increases. The ~58 000 ncRNAs added in this unification process were derived from ~180 000 genome-mapped fRNAdb and other source entries. The latter were pruned to ~75 500 unique, more reliable ncRNAs (i.e. after clustering and removal of purely predicted singletons), candidates for integration into GeneCards. There was a surprisingly small overlap of ~13 900 entries between the pruned ncRNAs and the past GeneCards embodiment, probably stemming from the existence of partially orthogonal worldwide pipelines for ncRNA discovery, and the almost mutually exclusive source lists for fRNAdb and GeneCards (Fig. 1). GeneCards V3.09 thus contains 101 004 molecularly identified functional gene entries (protein coding and ncRNAs). In addition, this version has 21 409 entries for pseudogenes, gene clusters, genetic disease loci and ‘uncategorized’, providing a grand total of 122 413.

A major step in the fRNAdb pruning process was clustering of ncRNA entries with extensive genome overlaps, greatly extending the removal of redundant entries already performed within fRNAdb (Mituyama *et al.*, 2009). The arguably arbitrary clustering cutoff we have used is not much different from that employed by others, as exemplified by the case of piRNABank (Sai Lakshmi and Agrawal, 2008), whereby the 32 148 entries represented in fRNAdb have undergone local removal of redundancies to 23 439, quite similar to our procedure. Our method further eliminated sequences that were predicted only once as ncRNAs on the basis of sequence conservation and/or secondary structure alone. This is because such predictions were reported to comprise a high fraction of false positives, up to 50–70% (Gorodkin *et al.*, 2010; Washietl *et al.*, 2007). We believe that the resulting list of ~79 300 ncRNAs represents an adequately comprehensive compendium, in which purely predicted entries are only minimally presented (~3%), allowing effective future scrutiny.

GeneCards V3.07 already included 114 piRNA-related entries (with symbols PIRC*). These were mined from HGNC (and originally from piRNABank) and constitute clusters of ~40–4000 (Sai Lakshmi and Agrawal, 2008) individual piRNAs. We have decided that GeneCards will have a card for each individual piRNA, with annotation indicating cluster affiliation in GeneCards’ Aliases and Descriptions section. For continuity and comprehensive consistency with HGNC, the PIRC entries are included as well. This decision is inspired by the fact that PIRC coordinates are computationally predicted with limited hard evidence for the existence of the 114 PIRCs as transcripts (Zamore, 2010). Another reason is not omitting piRNAs that are not included in PIRC clusters. Finally, recent studies have found a link between individual piRNAs and tumour development (Cheng *et al.*, 2012, 2011), emphasizing the importance of regarding each piRNA as an individual gene. In the future, we also intend to add annotations for PIRCs according to several algorithms that provide cluster boundaries often differing from one another (Girard *et al.*, 2006; Rosenkranz and Zischler, 2012; Sai Lakshmi and Agrawal, 2008).

Some of the criteria applied in our clustering procedure are somewhat arbitrary, such as the use of 70% positional overlap. However, we verified that this criterion does not impact much the number of multi-membered clusters created (Supplementary Fig. S3).

When comparing the original GeneCards (version 3.07) to the newly added ncRNA sources (Fig. 5), we found that only ~13 900 ncRNA entries were shared, while ~7800 entries were unique to GeneCards and ~57 600 were seen only in the added sources. This is surprising, since our expectation from the much larger size of newly introduced ncRNA collection was that it would include most or all of the original GeneCards ncRNA entries (~21 500). However, many of the non-overlap cases are easily explainable; ~21 800 of these fRNAdb-unique entries are piRNAs and ~2500 are ncRNA predictions from Evofold, RNAz and ncRNAsearch, both not included in GeneCards V3.07, leaving ~33 500 (~58%) to be accounted for. Conversely, ~72% of the GeneCards-unique entries come from Ensembl (Flicek *et al.*, 2012), a data source not mined by fRNAdb, which addresses only ncRNA-specific databases. Thus, such mutual omissions may not be suspected as erroneous.

The foregoing underline the fact that our unification process results in the most comprehensive dataset of human ncRNA genes, containing 7801 entries absent in the newly added sources. These are ncRNA genes uniquely found in the ‘general’ gene databases Entrez Gene and Ensembl, but not in ncRNA-specific compendia. Further, fRNAdb and the other added sources contain a considerable number (~45 500) of redundant entries, which in our procedure are judiciously unified into ~14 000 clusters based on genomic location. Finally, the reduced compendium of ncRNA genes within GeneCards V3.09 is based on the notion that genes predicted by only one source are less likely to be valid. In the current version such genes are removed; in the future, they will be presented with a prediction score probability indicator (see below).

We make a distinction between two sets of predicted ncRNAs. The first set is predicted purely by computational algorithms, as exemplified by RNAz and Evofold entries in fRNAdb. The second set includes Rfam entries, predicted based on known ncRNA seeds, and in turn serving (along with miRBase entries) as prediction templates for Ensembl ncRNAs (Flicek *et al.*, 2012). Similarly, Ensembl relies on lincRNA discovery that includes identification of chromatin methylation outside protein-coding genes (Flicek *et al.*, 2012). Therefore, we included all Ensembl ncRNA predictions while the fRNAdb predictions were further filtered. In the future, other filtration methods will be considered, e.g. based on the reported prediction scores (Pedersen *et al.*, 2006; Torarinsson *et al.*, 2006; Washietl, 2007). In addition, we will make the predictive arsenal more comprehensive, including the use of additional algorithms such as AlifoldZ (Washietl and Hofacker, 2004), or improved versions of presently utilized algorithms such as RNAz 2.0 (Gruber *et al.*, 2010). While at present GeneCards policy is to mine existing repositories of prediction outputs, we will consider also performing prediction runs on our own when needed.

We provide here a status description for GeneCards version 3.09. Notably, GeneCards is a dynamic data structure, with three version updates each year. The mining of ncRNA genes from all sources (Fig. 1) has now become part of the standard GeneCards generation process. Thus, all inclusion and annotation aspects would be frequently revised. fRNAdb released a minor update in March 2012 and is expected to undergo a major update in December 2012, followed by subsequent releases about every half year. GeneCards in turn will retrieve new fRNAdb content at each of its own updates (three times a year) and add directly primary sources not updated by fRNAdb.

The Encode project predicts that ~80% of all genomic territories are transcribed in one fashion or another. Using integrated GeneCards information, and disregarding overlaps, we estimate that 56% of the 3.1 Gb of the genome is occupied by genes, taking into account cumulative lengths of all preprocessed mRNA, including introns. The cumulative gene territories of all ncRNA genes in our GeneCards integration effort is 25%, of which the newly added 57 636 clusters add ~15%. The total genome territory thus covered amounts to 71% of the human genome. Such difference between ENCODE and GeneCards may simply be due to dataset input and computation discrepancies. However, at least some of this difference may represent genes (mainly ncRNAs) still awaiting identification and characterization.

ACKNOWLEDGEMENTS

We thank Michal Twik for helpful discussions.

Funding: LifeMap Sciences Inc. California (USA); the Crown Human Genome Center at the Weizmann Institute of Science and the SysKid—EU FP7 project number (241544).

Conflict of Interest: None declared.

REFERENCES

Aravin,A. *et al.* (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, **442**, 203–207.

Bateman,A. *et al.* (2011) RNAcentral: a vision for an international database of RNA sequences. *RNA*, **17**, 1941–1946.

Benson,D.A. *et al.* (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.

Birney,E. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Blencowe,B.J. (2002) Transcription: surprising role for an elusive small nuclear RNA. *Curr. Biol.*, **12**, R147–R149.

Bu,D. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.

Bushati,N. and Cohen,S.M. (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol.*, **23**, 175–205.

Carninci,P. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

Cawley,S. *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. *et al*. (2011) piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells. *Clin. Chim. Acta*, **412**, 1621–1625.

Cheng,J. *et al*. (2012) piR-823, a novel non-coding small RNA, demonstrates in vitro and in vivo tumor suppressive activity in human gastric cancer cells. *Cancer Lett*., **315**, 12–17.

Christov,C.P. *et al*. (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol*., **26**, 6993–7004.

Engstrom,P.G. *et al*. (2006) Complex loci in human and mouse genomes. *PLoS Genet*., **2**, e47.

Flicek,P. *et al*. (2012) Ensembl 2012. *Nucleic Acids Res*., **40**, D84–D90.

Gardner,P.P. *et al*. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res*., **37**, D136–D140.

Girard,A. *et al*. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199–202.

Gorodkin,J. *et al*. (2010) De novo prediction of structured RNAs from genomic sequences. *Trends Biotechnol*., **28**, 9–19.

Gruber,A.R. *et al*. (2010) Rnaz 2.0: improved noncoding RNA detection. *Pacific Symp. Biocomput*., **15**, 69–79.

Guttman,M. *et al*. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.

Harrow,J. *et al*. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol*., **7** (**Suppl. 1**), 1–9.

Holley,R.W. *et al*. (1965) Structure of a ribonucleic acid. *Science*, **147**, 1462–1465.

Hsu,S.D. *et al*. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*., **39**, D163–D169.

Kiss,T. (2001) Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J*., **20**, 3617–3622.

Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*., **39**, D152–D157.

Kubli,E. (1981) The structure and function of tRNA genes of higher eukaryotes. *Experientia*, **37**, 1–9.

Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res*., **34**, D158–D162.

Lin,H. (2007) piRNAs in the germ line. *Science*, **316**, 397.

Maglott,D. *et al*. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*., **39**, D52–D57.

Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*., **15**, R17–R29.

McKeown,M. (1993) The role of small nuclear RNAs in RNA splicing. *Curr. Opin. Cell Biol*., **5**, 448–454.

Mituyama,T. *et al*. (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res*., **37**, D89–D92.

Moore,P.B. and Steitz,T.A. (2002) The involvement of RNA in ribosome function. *Nature*, **418**, 229–235.

Morris,K.V. and Vogt,P.K. (2010) Long antisense non-coding RNAs and their role in transcription and oncogenesis. *Cell Cycle*, **9**, 2544–2547.

Pang,K.C. *et al*. (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res*., **35**, D178–D182.

Pedersen,J.S. *et al*. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol*., **2**, e33.

Ponjavic,J. *et al*. (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*., **17**, 556–565.

Ponting,C.P. *et al*. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Povey,S. *et al*. (2001) The HUGO Gene Nomenclature Committee (HGNC). *Hum. Genet*., **109**, 678–680.

Pruitt,K.D. *et al*. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res*., **40**, D130–D135.

Rosen,N. *et al*. (2003) GeneLoc: exon-based integration of human genome maps. *Bioinformatics*, **19** (**Suppl. 1**), i222–i224.

Rosenkranz,D. and Zischler,H. (2012) proTRAC—a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics*, **13**, 5.

Safran,M. *et al*. (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*., **31**, 142–146.

Safran,M. *et al*. (2010) GeneCards Version 3: the human gene integrator. *Database*, **2010**, baq020.

Sai Lakshmi,S. and Agrawal,S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res*., **36**, D173–D177.

Seal,R.L. *et al*. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*., **39**, D514–D519.

Shmueli,O. *et al*. (2003) GeneNote: whole genome expression profiles in normal human tissues. *C. R. Biol*., **326**, 1067–1072.

Spencer,M. *et al*. (1969) Studies on ribosomal RNA structure. I. The isolation of crystallizable fragments. *Biochim. Biophys. Acta*, **179**, 348–359.

Stein,A.J. *et al*. (2005) Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell*, **121**, 529–539.

Stelzer,G. *et al*. (2011) In-silico human genomics with GeneCards. *Hum. Genomics*, **5**, 709–717.

Struhl,K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol*., **14**, 103–105.

Torarinsson,E. *et al*. (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res*., **16**, 885–889.

Washietl,S. (2007) Prediction of structural noncoding RNAs with RNAz. *Methods Mol. Biol*., **395**, 503–526.

Washietl,S. and Hofacker,I.L. (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol*., **342**, 19–30.

Washietl,S. *et al*. (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*., **17**, 852–864.

Wilming,L.G. *et al*. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res*., **36**, D753–D760.

Yamasaki,C. *et al*. (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res*., **38**, D626–D632.

Zamore,P.D. (2010) Somatic piRNA biogenesis. *EMBO J*., **29**, 3219–3221.