Université de Sherbrooke

snoDB: An Interconnected Online Database of Human snoRNA

Par
Philia Bouchard-Bourelle
Biochimie et Génomique Fonctionnelle

Mémoire présenté à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de M.Sc.
en Biochimie

Sherbrooke, Québec, Canada
Janvier 2020

Membres du jury d'évaluation:

Michelle S. Scott, *Biochimie et Génomique Fonctionnelle*
Sherif Abou Elela, *Microbiologie et Infectiologie*
Alan Cohen, *Médecine de Famille*
Manuel Lafond, *Informatique*

ACKNOWLEDGMENTS

Thank you to Michelle Scott, Jean-Michel Garant, Gabrielle Deschamps-Francoeur and Vincent Boivin for their continued advice and support over the years. I would also like to sincerely thank Joël Simoneau, Fanny Thuriot, Gaspard Reulet, Hoang Dong Nguyen and Étienne Fafard-Couture for their camaraderie and open mindedness. And a special thank you to everyone involved in snoDBs including Clément Desjardin-Henri for developing its sister tool snoTHAW. The project wouldn't be nearly what it is without them.

I am also grateful to my research directors Michelle Scott and Sherif Abou Elela for their guidance and to all of them including my mentor Alan Cohen for their professional career advice.

It was a pleasure to get to know and work with all of you and I wish you all the best!

RÉSUMÉ

snoDB: An Interconnected Online Database of Human snoRNA

Par
Philia Bouchard-Bourelle
Biochimie

Mémoire présenté à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de maîtrise ès sciences (M.Sc.) en Biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

L'ARN est bien plus qu'une molécule transitoire entre l'ADN et les protéines. Au-delà des ARN encodant des protéines, on trouve un vaste éventail d'ARN non-codants qui demeurent encore sous-étudiés. Ces ARN ont été découverts dans les années 1960, mais ce n'est qu'au tournant du siècle que leur incroyable prévalence en cellule a pu être confirmée avec la venue de méthodes de séquençage d'ARN à haut débit. Les expériences à haut débit ont également augmenté de façon exponentielle la quantité de données sur l'ARN créant un besoin pour des outils bio-informatiques permettant leur analyse et leur stockage. Un des premiers, et des plus abondant, type d'ARN non-codant à être découvert sont les petit ARN nucléolaires (snoRNA). Canoniquement caractérisés comme guides de modifications spécifiques dans l'ARN ribosomal, ces petits ARN hautement conservés ont maintenant une liste variée de fonctions non-canoniques, notamment au niveau de l'expression génique, ainsi qu'un nombre croissant d'associations à une panoplie de maladies et de cancer. Considérant la littérature grandissante sur les snoRNA chez l'humain, ainsi que leur connexion maintenant apparente à plusieurs domaines de recherche variés, un regroupement accessible de ce large spectre d'information est maintenant indispensable. Malheureusement, les bases de données en ligne de snoRNA humain, snoRNABase, snOPY, et snoRNA Atlas, ne sont plus à jour ou sont trop pointues au niveau de leurs données. De plus, elles figurent peu ou pas de données d'interactions non-canonique et/ou d'expression. Nous avons donc créé snoDB : une base de données interactive de snoRNA humain qui contient des données sur leurs fonctions non-canoniques, trouvées à travers la littérature, des données d'expression dans une panoplie de tissus, et bien plus. Contrairement à ces prédécesseurs, snoDB offre une visualisions sélectives de son plus large éventail de données, au sein d'une table interactive aux options de recherche abondantes. Les données d'expression peuvent également être visualisées dans la même page, sous forme de carte de chaleur, grâce à l'application sœur de snoDB : snoTHAW. snoDB se démarque aussi par sa connectivité à plus d'une douzaine de ressources incluant le consortium RNAcentral, la plus grande base de données d'ARN non-codant, dont snoDB fais maintenant parti. Les données de ces ressources ont été acquises puis jointe ensemble dans une base de données relationnel postgreSQL. De plus, elles sont toutes en lien dans la table de snoDB afin de facilement pouvoir corroborer l'information visible, ainsi qu'accéder aux fonctionnalités des autres sites. Enfin, snoDB a été construit pour être facile à mettre à jour afin d'assurer ces contributions à la recherche pour de nombreuses années.

Mots clés : Petits ARN nucléolaire, Bio-informatique, Base de données (PostgreSQL), Séquençage d'ARN, Intéractions ARN-ARN, Datatables,

SUMMARY

snoDB: An Interconnected Online Database of Human snoRNA

By
Philia Bouchard-Bourelle
Biochemistry

Thesis presented to the Faculty of medicine and health sciences for the obtention of Master degree diploma maîtrise ès sciences (M.Sc.) in Biochemistry, Faculty of medicine and health sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

RNA is more than just a transitory molecule between DNA and proteins. Beyond the scope of protein-coding RNAs lies a vast underexplored landscape of non-coding RNAs (ncRNA). These RNAs have been slowly uncovered since the 1960s but it took until the turn of the century, and the advent of high-throughput RNA-Sequencing methodologies, for us to finally see how dominated by ncRNAs the transcriptome really is. High-throughput experiments also exponentially expanded the amount of data on RNA and created a need for bioinformatics tools for their analysis and storage. One of the first, and most abundant, ncRNA types to be discovered was small nucleolar RNAs (snoRNAs). Canonically pegged as guides for the modification of pre-ribosomal RNAs, these highly conserved RNAs now boast a diverse list of crucial non-canonical roles, notably in gene expression, as well as being associated to a myriad of diseases and cancers. Considering the growing body of literature surrounding snoRNAs in humans, and their increasing connections to a broad range of fields of study, having an accessible and comprehensive assessment of these data has become essential. Unfortunately, existing online human snoRNA databases, snoRNABase, snOPY, and snoRNA Atlas, are either outdated or too narrow in scope, focusing almost exclusively on canonical snoRNA interactions and lacking expression data. As such, we have created snoDB: a modern, interactive database of human snoRNAs with curated data on non-canonical snoRNA interactions, expression data in a growing range of tissues and cell lines, and more. Unlike the old snoRNA databases, snoDB features extensive visualisation and filtering capabilities, allowing for its larger array of data to be selectively viewed in an interactive and customizable table. Expression data can be further visualised in interactive heatmaps thanks to snoDB's sister tool: snoTHAW. snoDB also innovates by being much more interconnected with other resources. Data was gathered, and joined together in a relational postgreSQL database, from over a dozen resources, including the RNAcentral database consortium, the largest database of ncRNA sequences, of which snoDB is now a part of. In addition, all resources are linked to in-table, where data they provided appears, to help corroborate the data shown for transparency, as well as to grant access to interesting features housed on remote sites. Finally, snoDB is built to be easily maintainable, updatable and extensible to keep up with ongoing developments and insure that the information it contains will contribute to snoRNA research for years to come.

Keywords: snoRNA, Bioinformatics, RNA-Seq, Database, postgreSQL (PSQL), RNA-RNA Interactions, Datatables

TABLE OF CONTENT

FIGURES

ABREVIATIONS

A	Adenine, *Adénine*
cDNA	Complmentary DNA, *ADN complémentaire*
C	Cytosine, *Cytosine*
DNA	Deoxyribonucleic acid, *Acide déoxyribonucléique*
FTP	File Transfer Protocol, *protocole de transfert de fichier*
G	Guanine, *Guanine*
HGNC	HUGO Gene Nomenclature Committee
LCC	Leukoencephalopathy with Calcification and Cysts
lncRNA	Long ncRNA, Long *ARN non-codant*
miRNA	Micro RNA, *Micro ARN*
mRNA	Messenger RNA, *ARN messager*
ncRNA	Non-Coding RNA, *ARN non-codant*
NGS	Next Generation Sequencing, *Séquençage de nouvelle génération*
piRNA	PIWI-interacting RNA, *ARN intéragissant avec les protéines PIWI*
PSQL	PostgreSQL
PWS	Prader-Willi Syndrome, *Syndrome de Prader-Willi*
RDBMS	Relational Database Management System, *Système de gestion de base de données relationelle*
RNA	Ribonucleic acid, *Acide ribonucléique*
RNA-Seq	RNA-Sequencing, *Séquençage d'ARN*
RNP	Ribonucleoprotein Complex, *Complexe ribonucléoprotéique*
rRNA	Ribosomal RNA, *ARN ribosomal*
scaRNA	Small Cajal Body-Specific RNA, *Petit ARN spécifique aux corps de Cajal*
snoRNA	Small Nucleolar RNA, *Petit ARN nucléolaire*
snRNA	Small Nuclear RNA, *Petit ARN nucléaire*
SQL	Structured Query Language

T	Thymine, *Thymine*
TERC	Telomerase RNA component, *Composante ARN de la télomérase*
TGIRT	Thermotable Group 2 Intro Reverse Transcriptase, *Transcriptase inverse thermostable de group 2*
tRNA	Transfert RNA, *ARN de transfert*
U	Uracil, *Uracile*
yadcf	Yet Another Datatables Column Filter

INTRODUCTION

RNA biology has come a long way since its inception during the last century. We now have so much data about this diverse class of molecules that innumerable tools and databases have been created to aid in their analysis. But what is RNA and how do specialized bioinformatics resources, such as an online snoRNA database, help drive RNA research forward?

History of RNA Biology

Ribonucleic acids (RNAs) are a diverse class of molecules often transcribed from deoxyribonucleic acids (DNA) (Brosius and Raabe, 2016; Zhang et al., 2018). Both are long chains of nucleotides made up of a sugar phosphate spine (ribose sugar for RNA and deoxyribose for DNA) with one of four bases attached to each sugar molecule. These bases are adenine (A), thymine (T), guanine (G) and cytosine (C) with thymine being substituted for Uracil (U) in RNA. In addition, the familiar double helix structure of DNA is not seen in RNA; rather it is a single helix that folds in on itself in a wide variety of structures (Lodish et al., 2000). Cells are filled with an incredible diversity of RNA types that enact and regulate vital biological functions (Kufel and Grzechnik, 2019). Although the chemical and structural differences between the two types of nucleic acid chains have been identified for a hundred years, their individual importance, their interconnected relationship, and even their names, took several decades to be established.

Discovery & Distinction: The two nucleic acids:

The history of nucleic acids research begins in 1869 with the work of Johann Friedrich Miescher on leukocytes. These white blood cells were easily obtained with relatively good levels of purity from pus soiled bandages of a nearby clinic. Through purification of the infectious fluid, Miescher's experiment uncovered a cellular substance with unprecedented resistance to chemical methods of protein degradation. This seemingly non-proteic substance was hypothesized to originate from the nucleus which inspired the name it was given at the time: nuclein (His et al.,

1897; Miescher, 1871; Miescher, F., 1869). As it turns out, nuclein was in fact the first recorded precipitate of DNA. (His et al., 1897; Miescher, 1874; Miescher and Schmiedeberg, 1896).

These findings happened around the same time as the nucleus itself was being suggested by prominent biologist Ernst Haeckel to contain the materials responsible for heredity (Haeckel, 1866; Olby, 1969). Of note, though cell theory was technically around at the time, stating that life is built of tiny units called cells, only its bigger constituents such as the nucleus and the nucleolus it contains, could be observed with microscopes (Beale, 1858). The notion that the nucleus and the cytoplasm that surrounds it differed functionally wasn't yet popular, with the entire contents of the cell, called protoplasm, being deemed the substance of life (Welch, 1995). Miescher did theorize that nuclein could be the material responsible for fertilization, but the carrying of hereditary information was deemed an impossible role for a single molecule to possess, given the wide range of diversity observed in nature (Dahm, 2005; Miescher, 1874; Ralf Dahm, 2008). And so, the link between nuclein (DNA) and heredity was not to be confirmed until over half a century later.

In the meantime, nuclein's acidic nature would see it re-named to nucleic acid (Altmann, R., 1889). Improved methods for its purification enabled Nobel Prize laureate to be Albrecht Kossel to uncover the components of nucleic acid: adenine, guanine, cytosine, thymine and uracil along with a sugar thought to be a pentose (Hammarsten O., 1894; Jones, 1953). As these experiments were repeated, two kinds of nucleic acids emerged based on the model organisms from which they were extracted: yeast nucleic acid, obtained from plants, and thymus nucleic acid, which was abundant in the thymus gland of animals. These two types of nucleic acids, in fact RNA and DNA respectively, were found to have differing composition of bases: uracil in yeast versus thymine in thymus nucleic acid, as well as different sugars: pentose for yeast and a hypothesized hexose in thymus nucleic acid (P. A. Levene, 1910).

Polynucleic Chains & the Tetranucleotide Hypothesis

At the turn of the 20th century, nucleic acids are still seen as a soup of their constituents. The concept of polynucleic chains came around 1909 with the work of Pheobus Levene (Jacobs W.

A., and Levene P. A., 1909). The Russian physician showed how the bases contained in both types of nucleic acids bind to sugars to form nucleosides and how these subunits link up with a phosphoric acid molecule to form nucleotides (Levene and Jacobs, 1909). Despite World War I halting progress, Levene eventually established that the sugars found in yeast nucleic acid and thymus nucleic acids are ribose and deoxyribose respectively (Levene et al., 1930; Levene and Jacobs, 1909). Based on the different chemical and physical properties of these carbohydrates, they proposed that the nucleic acids be called ribonucleic acids (RNA) and deoxyribonucleic acids (DNA) (Frixione and Ruiz-Zamarripa, 2019; Levene and Tipson, 1935). Around the same time, the long held hypothesis that the two nucleic acids are exclusive to either plants or animals was finally overturned (Allen, 1941).

Despite plentiful contributions, Pheobus Levene is a name synonymous with the tetranucleotide hypothesis, the erroneous theory that nucleotides form into simple closed rings of four nucleotides known as tetradic repeats (Hunter, 1999). This infamous theory is often attributed as holding back research on nucleic acids since a tetradic repeats conformation made them predictably dull. However, historical context suggests otherwise; Levene wasn't a biologist, focusing instead on the chemical characterization of nucleic acids and leaving it up to biologist to assess their function. At the time, biologist were simply more interested in the study of proteins which they believed were the more likely carriers of genetic information in cells rather than the 'idiotic' nucleic acids (Hargittai, 2009; Olby, 1994). This mindset was likely bolstered by the successful crystallization of insulin and other key enzymes around the 1930s coupled with the unavailability of homogenous samples of nucleic acids for study (Lederberg Joshua, 1994; McPherson and Gavira, 2013).

Such pure samples would finally be available in 1944 through DNA viruses with the work of Avery, McCarty and MacLeod, on pneumococci, which went on to finally and rightfully attribute heredity as a characteristic of DNA (Lederberg Joshua, 1994).

The Path to Protein-Coding RNA & the Central Dogma of Biology

Co-inoculation in mice of harmless type 2 pneumococci with the virulent but heat-killed type 3

variant was proved to be lethal by Fredirick Griffith in 1928. The assumption was that the transformation of the inoffensive type 2 virus into its deadly type 3 counterpart was a protein mediated phenomenon (Griffith, 1928). However, Avery, McCarty and MacLeod, through extensive purification of the deoxyribonucleic acid from the heat-killed type 3 pneumococci, showed that the transformation was fundamentally attributable to DNA. This breakthrough paved the way for the DNA theory of inheritance, and renewed interest in nucleic acid research (Avery et al., 1944).

Many protein-enthusiasts were not so easily convinced, continuing to uphold proteins as the progenitors of genes up until the early 1960s (Eck, 1961). If DNA truly held all genetic information within cells, how was it turned into proteins? One theory at the time assumed the existence of a yet unknown RNA intermediate (Brachet, 1942; Capspersson, 1947). However, back then the only type of RNA to have formerly been identified was ribosomal RNA (rRNA). Known then as microsomal particles, they had been shown in an experiment to turn radiolabeled amino acids that passed through them into proteins (Hoagland et al., 1957; Roberts, 1958). Convinced of the existence of a hypothetical "adaptor" between DNA and the ribosomes, Francis Crick published the now famous central dogma of biology (Crick, 1958).

The (Actual) Central Dogma of Biology

Contrary to what many are taught in school, the central dogma of biology was never meant to be a mere roadmap of protein synthesis. Rather it outlined the possible ways for information to transit between DNA, RNA and proteins based on scientific knowledge at the time, as well as Crick's own intuition. Conversions from DNA to RNA and from RNA to protein were indeed postulated, but this was alongside established phenomena like DNA's ability to copy itself during cell division (DNA to DNA), and the ability of RNA viruses to self-replicate (RNA to RNA) (Figure 1). The flow of RNA to DNA was tentatively hypothesized as well, but it would only see confirmation in 1970 with the discovery of the "reverse transcriptase" enzyme (Coffin and Fan, 2016). Most interesting were the paths of information flow that were deemed impossible (protein to protein, protein to RNA, and protein to DNA). They lacked any known mechanisms but at the time so did the transcription of DNA to RNA, as well as the subsequent

translation of RNA into proteins. Therefore, by asserting the unfeasibility of information transits stemming from proteins, the central dogma discredited theories pinning proteins as gene progenitors. Overall, the intended takeaway was that the transfer of genetic information between DNA, RNA and proteins follows defined paths in cells and that once information reaches proteins, reversing it back into DNA or RNA is impossible (Cobb, 2017; Morange, 2008).

The central dogma of biology, and the RNA adaptor hypothesis between the ribosomes and proteins, would be confirmed shortly after the dogma's publication with the formal discovery and characterization of protein coding RNA, called messenger RNA (mRNA). Several groups published papers on mRNA describing it as a transcribed template of DNA that travels to the cytoplasm to bind with ribosomes, as well as with a specific iteration of amino acid carrying transfer RNA (tRNA), to be translated into protein (Cobb, 2015).

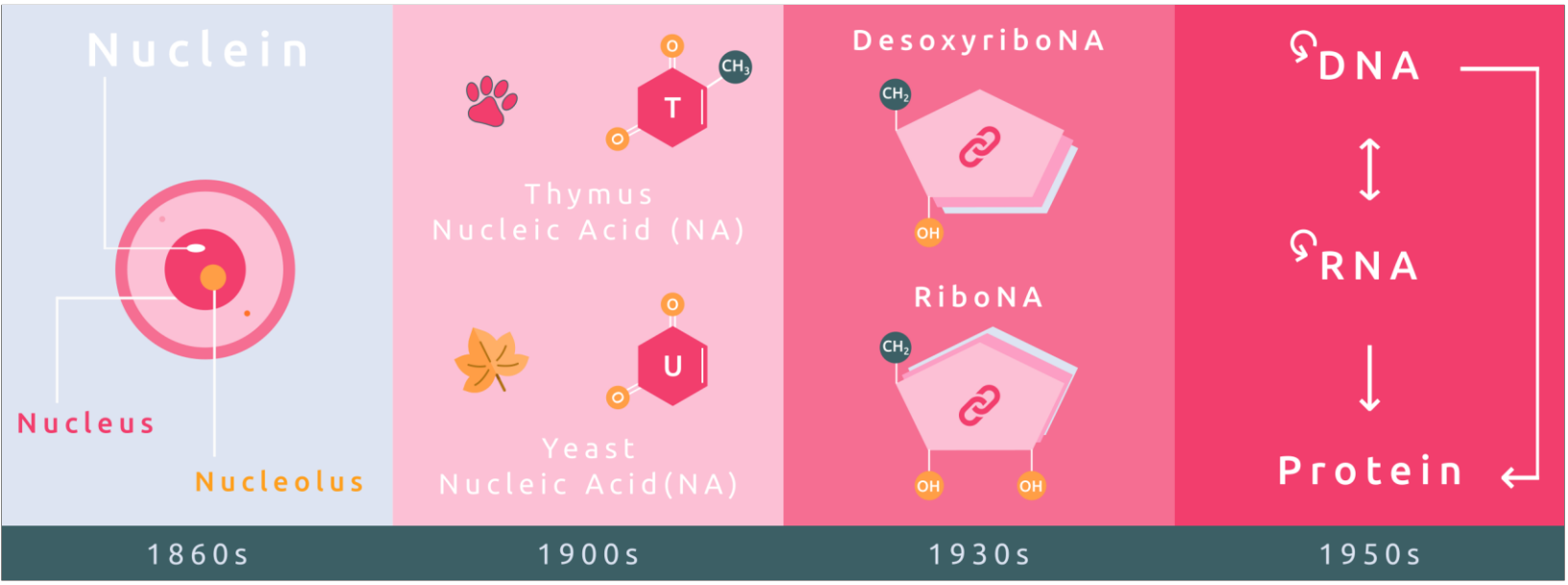

Figure 1 : Rough Timeline of Nucleic Acid Research

Nucleic acids were first discovered in the 1860s under the name nuclein. Later studies around the turn of the century categorized two types of nucleic acids based on the presence of the nucleo-bases thymine or uracil and based on their source of extraction (animal glands such as the thymus or plants such as certain yeasts). During the 1930s, the two types of sugars composing the nucleotides, found in both types of nucleic acids, had been established along with their ability to link up together to form polymeric chains. All of this paved the way for the central dogma of

biology, posited towards the end of the 1950s, which correctly mapped the flow of genetic information between what we now knew as DNA and RNA as well as proteins.

Non-coding RNA, small Nucleolar RNA & the RNA World Theory

The monumental discovery of mRNA, as vectors of genetic information for protein synthesis, was instrumental to the development of RNA biology. It also colored our view of RNA as mostly being passive elements while proteins remained the predominant functional components of the cell. Opposing this view came a string of discoveries implicating non-protein-coding RNA, simply called non-coding RNA, in various novel and essential regulatory pathways.

Analysis of HeLa cell nuclei, in the late 1960s, revealed the existence of new, uridine(U)-rich, non-coding RNA(ncRNA) species beyond rRNA and tRNA, which were named U1, U2, U3, etc. (Pene et al., 1968; Weinberg and Penman, 1968). Further studies on these new RNAs in yeast allowed for their classification into two distinct categories, based on their localization and functionalities (Riedel et al., 1986; Wise et al., 1983) . U1, U2, U4, U5 and U6 were dubbed small nuclear RNA (snRNA) for localizing to the nucleus, while U3 and U8 were dubbed small nucleolar RNA (snoRNA) for specifically localizing to a membrane-less compartment of the nucleus called the nucleolus (Busch et al., 1982). These two types of RNA both form into hairpin-like structures that average over a hundred nucleotides in length. Both RNA species were also found to individually bind with specific proteins to form ribonucleoprotein complexes (RNPs) (example: U1 snRNP and U3 snoRNP) (Busch et al., 1982). However, snRNPs are formed outside the nucleus before returning and assembling together into a big complex called the spliceosome (Patel and Bellini, 2008). The spliceosome then binds pre-mRNA and regulates a vital processing step of theirs called splicing, i.e., the excision of the non-coding regions named introns that separate the multiple coding segments of a nascent gene called exons (Lerner et al., 1980; Mount and Wolin, 2015). Meanwhile, snoRNPs form within the nucleus immediately after snoRNAs are transcribed, before making their way to the nucleolus where they individually act in rRNA processing (Schimmang et al., 1989).

Small nucleolar RNA

We now know that Small nucleolar RNAs (snoRNAs) are a group of highly structured and expressed non-coding RNAs that are conserved across all eukaryotes (Henras et al., 2004). Most snoRNAs discovered thus far have been found in humans and in yeast, but snoRNAs have also been identified in rodents, amphibians, plants and even Achaea (Bertrand and Fournier, 2013). In humans, they are mostly encoded within the introns of mRNAs and long non-coding RNAs (lncRNAs) which can be referred to as snoRNA host genes (Boivin et al., 2018b). Two thousand unique snoRNA sequences have been reported in human RNA databases to date (Bouchard-Bourelle et al., 2019).

We distinguish two main snoRNA families based on their structure and specifically conserved sequence motifs dubbed boxes: the single stem-loop (stem-bulge-stem), RUGAUGA & CUGA motif bearing box C/D family and the dual stem-loop (hairpin-hinge-hairpin-tail), ANANNA & ACA motif bearing Box H/ACA family (Figure 2). In addition, some box C/Ds sometimes possess less well conserved copies of C and D boxes dubbed C' and D'. All these conserved box motifs, along with a cascade of chaperones and other conserved proteins, enable snoRNAs to bind specific sets of proteins conserved across eukaryotes when they are transcribed in the nucleoplasm. This association prevents the degradation of the bound regions by exonucleases, and results in fully processed small nucleolar ribonucleoprotein complexes (snoRNPs) (Kishore et al., 2013).

The boxes are also responsible for the trafficking of snoRNPs from the nucleus to the nucleolus, which is where they perform their most widely recognized function as guides for specific post-transcriptional RNA modifications (Massenet et al., 2016).

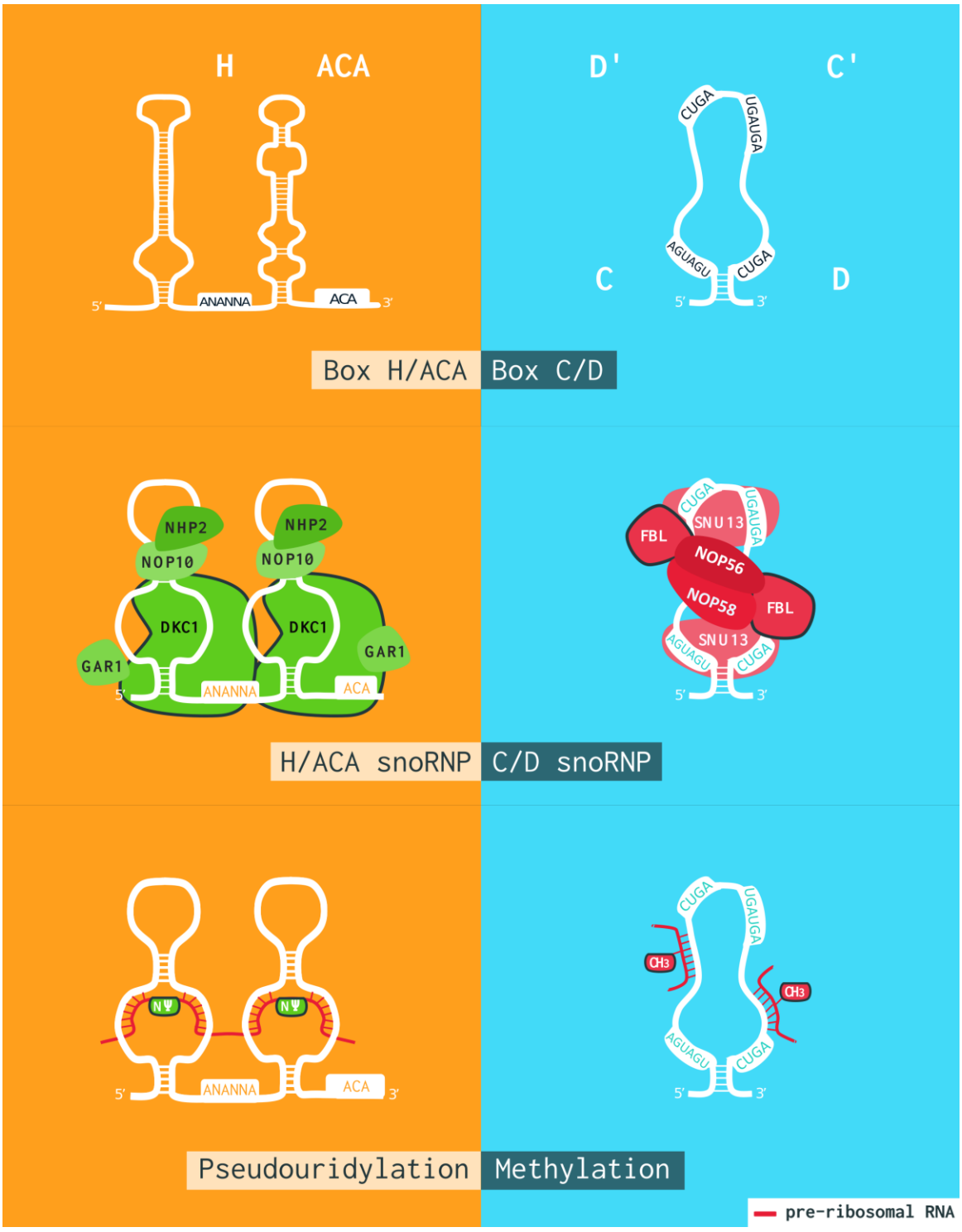

Figure 2: Canonical snoRNAs

snoRNAs are canonically divided into 2 families that bind distinct sets of proteins which catalyze specific modifications in pre-rRNA which snoRNAs bind through complementarity.

Canonical snoRNA Functions

The most well characterized function of snoRNAs is guiding multiple site-specific modifications in pre-ribosomal RNA (pre-rRNA) during ribosome biogenesis in the nucleolus. They do this through regions on the snoRNA called antisense elements which are complementary to specific regions of pre-rRNA (Henras et al., 2004; Kiss, 2001). This mechanism, and snoRNAs' role as guides for these modification, were first discovered in yeast (Philippe Ganot et al., 1997; Kiss-László et al., 1996) and confirmed in humans shortly thereafter (Balakin et al., 1996; P. Ganot et al., 1997). Orthologous snoRNAs have also been identified in organism as evolutionarily distant as Achaea, pointing to the primeval nature of canonical snoRNA functionalities (Gaspin et al., 2000; Omer et al., 2000).

Box C/D snoRNA antisense elements are composed of 10 to 21 nucleotides and are located upstream of the D' and/or D box with the modification site on complementary rRNA usually being located 5 nucleotides upstream of the D/D' box. Box H/ACA snoRNA possess antisense elements within their 3' and/or 5'-terminal hairpin domains also called pseudouridylation pockets (Philippe Ganot et al., 1997; Kishore et al., 2013).

Once the snoRNAs' antisense element binds to a complementary strand of rRNA, core proteins forming the snoRNP catalyse the modification of a specific nucleotide. Box C/D snoRNPs are canonically composed of four conserved proteins including the methyltransferase Fibrillarin (FBL) which catalyzes the 2'-O-methylation of pre-rRNA. Box H/ACA snoRNPs are canonically composed of 2 identical protein quartets, on each of their stem-loops, with Dyskerin (DKC1) being responsible for the pseudouridylation of pre-rRNA (Figure 2). In both snoRNA families, the proteins without any catalytic functions, NOP56, NOP58 and SNU13 for box C/Ds, and Nhp2, NOP10 and Gar1 for box H/ACAs, serve as scaffolds to recruit each other and stabilize the resulting complex (Filipowicz and Pogačić, 2002; Jády and Kiss, 2001).

Although the modification of specific, and often conserved, bases in rRNA by snoRNAs haven't been found to be essential, hints to their involvements in proper folding of pre-rRNA and

the interactions between rRNA and ribosomal proteins have been found pointing to a need for further research on the matter (Bachellerie et al., 2002).

However, there are snoRNAs that have for a long time been known to function as something other than guides for post-transcriptional modifications. The U3 (SNORD3A) and U8 (SNORD118) to name only two, are highly conserved box C/D snoRNAs which are not known to guide any chemical modifications in pre-rRNA while still being key players of pre-rRNA processing and accumulation (Langhendries et al., 2016).

U3 is associated with the proper folding of the 18S small ribosomal subunit. It can execute this function due to the presence within its sequence of an additional motif called a B box. This motif allows U3 to bind a protein called Rrp9 in humans which is essential for the formation of its snoRNP (Zhang et al., 2013). More specifically, it has been shown that nucleotide substitutions in the 5'ETS (external transcribed spacer), located upstream of the 18S ribosomal RNA gene, leads to cell death in yeast. This region is involved in the initial binding of the 18S pre-mRNA with U3, and when plasmids containing a mutated U3 sequence complementary to the altered 5'ETS sequence is expressed, 18S accumulation and cell growth is regained (Dutca et al., 2011). When U3 is properly bound to its core proteins, as well as to the 5'ETS region upstream of the 18S rRNA gene, it mediates early cleavage steps in pre-rRNA processing that are essential for ribosome biogenesis (Cléry et al., 2007).

U8 on the other hand is vitally involved in the accumulation of 5.8S and 28S rRNA which are components of the large ribosomal subunit. Similarly to U3, U8 binds to the 28S pre-mRNA with mutational studies showing this to be essential for maturation of the large ribosomal subunit. In this case however, although the interaction is necessary it must be undone before pre-rRNA processing is concluded to allow for the association of the 28S and 5.8S subunits (Peculis, 1997).

Small Cajal Body-Specific RNA

snoRNAs have also canonically been shown to modify specific conserved positions in snRNAs. These RNA-dependent modifications were at first thought to exclusively occur in nuclear sub-organelles known as Cajal (coiled) bodies, to which a subset of snoRNA localize, and were hence named small Cajal body-specific RNA (scaRNA). Their characterization as snoRNAs might seem odd, as they are non-nucleolar RNAs, but Cajal bodies are actually closely related to nucleoli both physically and functionally, even being originally called nucleolar accessory bodies (Trinkle-Mulcahy and Sleeman, 2016). In addition scaRNAs possess familiar C/D and/or H/ACA box motifs through which they guide Fibrillarin dependant 2'-O-methylation and Dyskerin dependant pseudouridylation, although in snRNA instead of in rRNA (Darzacq et al., 2002). However, emerging evidence points to snoRNA nomenclature potentially needing revision as both scaRNAs and snoRNAs alike can exist and be functional outside of the localization that informed their naming scheme. For example, scaRNA-dependant snRNAs modifications are still carried out in cell-types/species which lack Cajal bodies, while snoRNAs have been found to accumulate in Cajal bodies, with some also being found in the cytoplasm (Deryusheva and Gall, 2009; Michel et al., 2011). Furthermore, scaRNAs have been found to act as guides for the modification of rRNA, just like canonical snoRNAs, further blurring the divide between them (Deryusheva and Gall, 2019).

RNA World Theory

Around that same time as snRNAs and snoRNAs were begin discovered, a fundamental characteristic of non-coding RNAs' activity was being uncovered: the specific structures they fold into. It was shown to allow the t-shaped tRNAs to successfully bind amino acids, to grant ribosomes the ability to mediate protein synthesis and would be shown to be functionally important in most non-coding RNA. This is reminiscent of the need for proteins to fold into specific structures themselves to become functional, with mutated or denatured protein and RNA alike losing some or all of their activity (Bhartiya and Scaria, 2016; Rich and RajBhandary, 1976). The parallels did not stop there as the discovery of ribozymes came soon after, and showed for the first time that RNAs, just like proteins, could catalyze chemical reactions in cells

(Cech et al., 1981; Guerrier-Takada et al., 1983). Even before having discovered RNAs' catalytic functions, many researchers were pondering the hypothesis of self-replicating RNA molecules being primordial precursors to both DNA and proteins. This theory, which posits a time where RNA alone permitted life, was formally named the "RNA world hypothesis" in 1986 (Cech, 2012; Neveu et al., 2013).

1992 brought the discovery of the long non-coding RNA (lncRNA) XIST, which is involved in X-chromosome inactivation (Brockdorff et al., 1992; Brown et al., 1992). In 1993, the first micro RNA (miRNA) was discovered in yeast (Lee et al., 1993). It took until the early 2000s for these small non-coding RNAs to be formally characterized as a group and notably linked to the regulation of gene expression through gene silencing (Ambros, 2004).

In summary, ncRNAs were clearly shown to be as functionally diverse and important as proteins, if not more so, potentially enabling life independently from proteins and DNA for billions of years (Cech, 2012). Nonetheless, research into ncRNAs didn't pick-up steam until the turn of the century. In 2001, a working draft of the human genome had finally been sequenced and published, using first-generation sequencing (one nucleo-base at a time). It confirmed the genome as being overwhelmingly composed of non-coding regions (Venter et al., 2001). Around the same time, the more efficient next-generation sequencing (NGS) technologies were inaugurated, and shortly thereafter adapted, to allow for high-throughput quantification of all RNA transcripts in a population of cells (Weber, 2015).

Transcriptome

Our ability to quantify RNA transcripts using high-throughput RNA-Sequencing represented another revolution in the field of RNA biology. It was the first technique to actually give us a look at the transcriptome, that is which parts of the genome are actively transcribed into RNA, at the moment an experiment was conducted, as well as giving us insights into their cellular functions (Wang et al., 2009). But the term transcriptome had been coined years prior while only low-throughput sequencing technologies were available. The main focus of these first gene profiling experiments was mRNA, with the term transcriptome being originally defined as

representing the sum of mRNA expression in a population of cells (McGettigan, 2013; Piétu et al., 1999). This led to much of transcriptomic research being mRNA-centric, even in the era of next-generation sequencing. This despite the fact that evolutionary studies showed that non-coding RNAs can represent up to 98% of transcriptional output across higher eukaryotes (Mattick and Gagen, 2001). Such insight coupled with the growing body of literature surrounding non-coding RNAs, and their wide range of essential regulatory functions, should have been the final nail in the coffin for the notion that RNAs are merely passive intermediates between DNA and proteins. But half a century of characterizing non-coding DNA as junk (Ehret and De Haller, 1963) seems hard to shake off and an old guard persists as exemplified by Google's definition of ''*transcriptome*'' still reading:'*' the sum total of all the messenger RNA molecules expressed from the genes of an organism*.'' Relinquishing such a notion and overall adopting a more fluid view of scientific concepts new and old is more than ever imperative in this new age of fast discoveries.

For example, we have known for a long time that, in terms of the molecules present, the transcriptome is mostly composed of rRNA and tRNA. Like mRNA, these two types of non-coding RNA have also canonically been pegged as static actors in protein synthesis but have since been hinted to have other unsuspected non-canonical functions. Transfer RNA have been shown to be cleaved into tRNA-derived small RNAs (tsRNAs) which are implicated in stress-response signalization and in regulating gene expressions (Shen et al., 2018). Moreover, growing evidence points to ribosomal RNAs being highly regulated and heterogeneous in nature, leading to the preferential translation of certain mRNA, with possible implication in differential development and response to stress (Guo, 2018; Xue and Barna, 2012). mRNAs themselves have recently been attributed new roles beyond mere messengers including participating in the assembly of nuclear bodies like Cajal bodies, which are membrane-less regions of nuclei where splicing complexes are assembled (Shevtsov and Dundr, 2011). When even the most extensively studied RNA-type has secrets left for us to uncover, one can easily posit that we have in fact only begun to scratch the surface of RNA biology. This notion sees itself exemplified in small nucleolar RNAs, which were found in the early 60s and quickly pegged as actors in rRNA processing but have since gradually been ascribed unsuspected and essential multifarious functions.

Emerging Non-Canonical snoRNA Functions

snoRNAs' vital and non-vital roles in ribosome and spliceosome biogenesis have been the focus of many studies for multiple decades (Jády and Kiss, 2001; Maxwell and Fournier, 1995). This makes sense considering that the most evolutionarily stable snoRNAs across mammals are highly expressed, and located in introns of host genes related to ribosome biogenesis and translation (Hoeppner et al., 2009). Meanwhile, the less conserved snoRNAs, seen only in humans and primates, show strong evidence of being derived from transposable element, and having thus been recently retrotransposed into new genomic location/host genes (Scott and Ono, 2011). In vertebrates, most snoRNAs are co-transcribed with their host gene, while some are independently transcribed by RNA polymerase II or III (Dieci et al., 2009; Tycowski et al., 2004). Interestingly, the less conserved snoRNAs, seen only in humans and primates, do not show enrichments for canonical snoRNA functions, suggesting a broader spectrum of snoRNA functionalities (Hoeppner et al., 2009). Taken together, these studies point to a flexibility in snoRNA expression, which isn't always correlated with their host gene, because some snoRNAs likely regulate other genes involved in non-canonical functions. Indeed, in recent years a growing number of papers have begun ascribing a wide range of varied non-canonical functions to snoRNAs.

SNORD115: Alternative Splicing & Editing

These discoveries slowly began around the turn of the century with one of two clusters of cerebrally enriched snoRNAs, known then as HBII-52 (now SNORD115). This cluster is located within the locus of chromosome 15 associated with Prader-Willi syndrome (PWS), which is the leading genetic cause of obesity worldwide. The disease was shown to be caused by the deletion of the paternal locus which features genetic imprinting. Imprinting is an epigenetic methylation of a parent's gene that makes only its expression possible regardless of the presence of a copy of this gene on the complementary allele. The mechanism behind PWS was linked to the SNORD115 cluster present in the affected locus. SNORD115 lacks any known rRNA complement, but it has been found to associate with the alternatively spliced exon 5 of the serotonin receptor 5-HT2CR (Cavaillé et al., 2000). This complementarity mediated binding was

later confirmed to affect the alternative splicing of said exon, with PWS-afflicted individuals exhibiting distinct serotonin receptor mRNA isoforms when compared to healthy individuals (Kishore and Stamm, 2006). However, it remains unclear if a change in alternative splicing alone is responsible for PWS. Other studies have pointed to an additional post-transcriptional modification, mediated by an interaction between SNORD115 and the 5-HT2CR pre-mRNA, known as Adenosine-to-Inosine (A-to-I) editing, which was shown to have behavioral impacts in mice (Doe et al., 2009). Although this editing mechanism is confirmed, its link to PWS remains a matter of debate (Glatt-Deeley et al., 2010).

From Housekeeping Genes to Non-Uniform Expression & Functions

These discoveries meant that snoRNAs could no longer be thought of as constitutively expressed genes solely implicated in site-specific RNA modifications. They can have tissue-dependent expression and enact non-housekeeping functions. But with only about a quarter of snoRNAs being orphans, i.e. having no known function (Falaleeva et al., 2016), could they really have that many undiscovered roles left to uncover?

Bioinformatics-based predictive approaches have found undiscovered complementary rRNA sites for a few orphan snoRNAs in regions that were not known to be modified. This sparked the tentative connection of snoRNAs with ribosome heterogeneity, in a perhaps tissue- or condition-specific manner further (Dieci et al., 2009; Piekna-Przybylska et al., 2007; Xue and Barna, 2012). If true, this would redefine the importance of canonical snoRNA functions in stress-response, and perhaps even development, through the preferential translation of certain mRNAs by so-called specialized ribosomes. Although, it would also reduce the pool of orphan snoRNA with potential non-canonical functions even. This assumption of fewer snoRNA potentially having novel roles is however proven false on two fronts. First, new snoRNA are steadily being discovered (Jorjani et al., 2016; Kishore et al., 2013). Second, snoRNAs have been found which have both canonical and non-canonical functions, broadening the member spectrum with new functionalities to perhaps every known and yet unknown snoRNA. For example, SNORD27 has been shown to both guide rRNA methylation and mediate the alternative splicing of several protein-coding genes, including ABCA8, E2F7, FER and MAP4K3. A potential mechanistic hint

for how this functional diversity is possible came with soluble nuclear extract studies which detected two distinct SNORD27 ribonucleoprotein complexes in biochemically divisible fractions; a canonical fibrillarin-dependant snoRNP and a non-canonical fibrillarin-independent snoRNP (Falaleeva et al., 2016).

Beyond Alternative Splicing; snoRNAs` Roles in Gene Expression

Alternative splicing remains one of the best characterized non-canonical snoRNA functions, with other snoRNAs not previously listed also being implicated in it. However, snoRNAs have now been found to regulate gene expression through a much broader array of mechanisms. For example, dozens of snoRNAs show enrichment in both *Drosophila* and human chromatin, a DNA-protein complex responsible for compacting and de-compacting DNA for protection and transcription respectively. A specific protein, Df31, was shown to bind both histones, which are proteins that are part of the chromatin, and the enriched snoRNA. As a result, chromatin structure remains un-compacted allowing for genes on that segment of the genome to be transcribed (Schubert et al., 2012).

Some snoRNAs have also been found to directly regulate mRNA abundance after being degraded into smaller RNAs, such as miRNA and piRNAs, or by interfering with pre-mRNA processing. Micro RNAs (miRNAs) are small non-coding RNAs, averaging 22 nucleotides, which form a distinct microRNA ribonucleoprotein complex (miRNP), more commonly called an RNA-induced silencing complex (RISC) (Cai et al., 2009). Silencing describes any negative regulation of gene expression. Similarly to canonical snoRNAs within snoRNPs, miRNA in miRNPs/RISC possess sequence complementarity with mRNA, and serve as guides for their silencing. This resemblance with canonical snoRNA isn't a coincidence as evolutionary ties between the two types of non-coding RNA have been found, as well as the fact that some snoRNA are degraded into sno-derived RNAs (sdRNA) such as miRNA (Ender et al., 2008; Saraiya and Wang, 2008). Likewise, snoRNAs have also been shown to be processed into PIWI interacting RNAs (piRNAs). These 26 to 31 nucleotide-long RNAs are so-named for their interactions with the PIWI protein family, which is closely related to the Argonaut protein, a key component of miRNA RISC gene silencing complex. Unsurprisingly then, piRNAs derived from

snoRNAs have been found to mediate decay in targeted pre-mRNAs, which they do through recruitment of the nuclear exome (Zhong et al., 2015). Researchers have also recently stumbled upon a set of box C/D snoRNAs that associate with proteins constituting the 3' mRNA processing complex. The processing of 3' end in mRNAs is essential to their proper expression. It involves the cleavage of said 3' end followed by the addition of over 200 adenine residues, called a polyA tail, in a process named polyadenylation. Depletion of one particular snoRNA showed consistent effects on 3' polyadenylation profiles and the abundance of specific genes (Huang et al., 2017).

Recent, studies in yeasts have found two previously orphan box C/D snoRNAs to be involved in guiding the acetylation of 2 cytosine residues in rRNA. They do this by associating with the cytidine acetyltransferase Kre33, which is known to acetylate tRNAs by interacting with the adaptor protein Tan1 (Sharma et al., 2017). Considering the highly conserved nature of most snoRNAs, it would not be surprising for acetylation to also be mediated through snoRNAs in higher eukaryotes.

snoRNAs & Diseases

With so many regulatory roles to their name, snoRNA-associated diseases seem quite likely when deregulation of their functions occur. Indeed, as previously mentioned, deletion of the SNORD115 family of snoRNAs is linked to a genetic disorder called Prader-Willi syndrome. However, not only are a growing number of diseases being linked to snoRNAs, they are increasingly being investigated as diagnostic/prognostic tools in the form of biomarkers.

Metabolic Stress & Homeostasis

Genetic screenings of Chinese hamster ovary (CHO) cells under lipotoxic conditions, metabolic stress conditions which are a common complication of diabetes where lipids accumulate in non-adipose tissues, revealed that SNORD32A, SNORD33, SNORD34 and SNORD35A can localize to the cytoplasm outside the nucleus and mediate oxidative stress (Michel et al., 2011). A later study found these four snoRNAs to also regulate systemic glucose metabolism solidifying their

potential implication in the pathogenesis of diabetes (Lee et al., 2016). Further genetic loss of function screening in CHO cells found orphan snoRNAs ACA60 (SNORA60) and the U17 snoRNAs (SNORA73A & 73B) to be essential in the regulation of cholesterol trafficking and homeostasis (Brandis et al., 2013; Jinn et al., 2015). The study on the U17 snoRNAs was even able to identify their non-canonical interaction with and negative regulation of mRNA HUMMR (hypoxia upregulated mitochondrial movement regulator) as a novel mechanism in cholesterol trafficking. Meanwhile ACA11 (SCARNA22) was found to be upregulated in multiple myeloma, an incurable cancer of plasma cells, and confer resistance to chemotherapy via a suppression of oxidative stress (Chu et al., 2012).

Cancers

Links between snoRNAs and various cancers have consistently increased over the last two decades. The most recent review on the subject implicated nearly 50 snoRNAs across 11 different types of cancers as both oncogenes and/or tumor suppressors (Liang et al., 2019).

Some snoRNAs have been linked to the p53 pathway in cancer which regulates the expression of genes involved in cell cycle arrest, DNA repair and apoptosis. Overexpression of snoRNAs have been found in human breast and prostate cancer to inhibit the p53 pathway and promote tumorigenesis (Su et al., 2014). Similarly, breast cancer also sees the upregulation of the U3 and U8 snoRNAs acting as oncogenes with their depletion enabling a potent p53 anti-tumor response (Langhendries et al., 2016). SNORA42 has been found to be oncogenic in non-small cell lung cancer where it regulates the expression of p53 (Mei et al., 2012). Other experiments have confirmed p53 to directly control transcription of the snoRNA host gene GAS5. The expression of the multiple snoRNAs nested within GAS5 correlate with p53 expression in colorectal cancer, though their potential role in this pathology has yet to be determined (Krell et al., 2014).

Other cancer related pathways, such as the PI3K–AKT cascade, have also been associated with snoRNA dysfunctions in certain cancers. PI3K-AKT signaling involved in cell death, differentiation and proliferation is increased due to ACA11 over expression in hepatocellular carcinoma (HCC) (Wu et al., 2017). SNORD126 also activates the PI3K–AKT pathway by

upregulating FGFR2, which promotes cell growth in HCC as well as in colorectal cancer (Fang et al., 2017).

Sno-derived RNAs (sdRNAs) have also been found to be differentially expressed in certain cancers. SNORD17 derived miRNA miR-768-5p reportedly binds to breast cancer associated protein YB-1 (Blenkiron et al., 2013). In breast cancer still, SNORD75 derived piRNA pi-sno-75 upregulates tumor suppressor and apoptosis inducer TRAIL (He et al., 2015). And in prostate cancer, sdRNAs derived from several snoRNAs were found to be over expressed, with the expression of some sdRNAs being specifically higher in metastatic patients, hinting at their potential role as prognostic biomarkers (Martens-Uzunova et al., 2015).

snoRNAs as Biomarkers

The involvement of snoRNAs in various facets of cancer biology and the detection of many of them in serum, plasma and urine, in addition to their high stability, has led to them being studied as potential biomarkers. Indeed, as previously described, several studies have found expression variation of certain snoRNAs, or their derivatives, to be indicative of certain cancers or certain stages of the disease.

As such, eight differentially expressed snoRNAs were recently established as prognostic factors in gastric cancer (Wang et al., 2019). Several snoRNA signatures, predicting overall survivability in patients with non-small cell lung cancer, have also been found (Gao et al., 2015; Liao et al., 2010; Mannoor et al., 2014). Expression of SNORA70F & SNORD116-118 are seemingly able to distinguish two types of chronic lymphocytic leukemia with different prognostics (Ronchetti et al., 2013).

Meanwhile, six serum snoRNAs were found to be effective biomarkers of ageing joints and osteoarthritis (Steinbusch et al., 2017) and circulating levels of SNORD114-1 are markedly higher following endurance training, and may be used as a biomarker to differentiate between exercise regimens providing insight into muscle repair and recovery (Håkansson et al., 2018). These nascent findings broaden the potential of snoRNAs as biomarkers outside of cancers.

Mutations in snoRNAs & their Host Genes

The spectrum of snoRNAs-related genetic disorders has also been progressively broadened over the years beyond Prader-Willi Syndrome.

A systematic curation of the human genome published in 2012 found 151 snoRNAs which contain at least one single nucleotide polymorphism (SNP), with 298 SNPs being reported in total. Cross-referencing of these data with snoRNAs more recently found to be related to disease could prove insightful, though unfortunately the database housing the information is no longer in service (Bhartiya et al., 2012).

More recently, biallelic mutations in SNORD118 (U8) were linked to a progressive degenerative disorder called cerebral microangiopathy leukoencephalopathy with calcification and cysts (LCC). The authors speculate that these mutations give rise to a ribosomopathy, given U8's involvement in rRNA processing, which would be causal to the disease. They were however unable to establish this link, leaving the door open to potential non-canonical functions of SNORD118 being responsible for LCC (Jenkinson et al., 2016).

Mutations in the H/ACA box motif of the human telomerase RNA component (TERC) disrupt the telomere lengthening activity of telomerase, causing a rare genetic disease known as dyskeratosis congenita (DKC). Telomeres consist of nucleotide repeats at the ends of chromosomes that serve as protection against DNA damage. This protection is eroded and replenished with each mitosis in stem cell, such as the hematopoietic stem cells found in bone marrow which continually replenish our blood cells, including those comprising our immune system. Afflicted individuals are therefore more cancer-prone, often show signs of premature ageing with most eventually succumbing to bone-marrow failure (Alter et al., 2012; Trahan and Dragon, 2009). Other variants of the disease involve mutation in the core H/ACA snoRNP protein and snoRNA host gene Dyskerin (DKC1). Indeed, almost all X-linked instances of DKC (X-DKC) feature mutations in DKC1, which happens to hosts SNORA36 and SNORA56, though it remains unknown whether these mutations affect the snoRNAs thereby implicating them in X-DKC (Parry et al., 2011).

Mutations in another snoRNA host gene, the ribosomal protein RPL5, have similarly been found to cause an inherited bone marrow failure syndrome called Diamond-Blackfan anemia (DBA). RPL5 hosts SNORA66 which canonically guides the pseudouridylation of position 119 in 18S rRNA, a modification which happens to be decreased in X-DKC. A causal link between this and X-DKC or DBA has yet to be established, though these are two examples out of many where snoRNA host genes, and their intronic snoRNAs, may be relevant in diseases.

The list of potential snoRNA related diseases is much longer than those described above as seen in (Deogharia and Majumder, 2018). This is without even delving deeper into the numerous reported implications in diseases of snoRNA host genes and their often independent transcription from the snoRNAs they host (Liao et al., 2010; Ronchetti et al., 2013, 2012; Williams and Farzaneh, 2012). However, whether it be studies on host genes, snoRNAs or sdRNAs, many still lack causal links instead only finding significant up or down-regulation in disease. Further studies are therefore in order especially since the proportional and holistic detection of snoRNAs, their host and their fragments via RNA-Seq has only recently been achieved. More snoRNAs or their related entities may therefore be related to associated diseases, and new disease states may show significant differential expression of specific RNAs.

Simultaneous RNA-Sequencing of snoRNAs & their Host Genes

High-throughput RNA-Sequencing stands as a modern tool of choice to tackle many pressing biological inquiries. Its usefulness in assessing transcript abundance and splicing events has prompted researchers to continuously adapt the technique for various purposes, leaving us with a still growing number of different protocols suited to a plethora of analyses. Due to the sheer amount of data the technique generates, it also has crucial bioinformatic analytical steps with an equally if not more diverse landscape of methodologies. Despite all these variations, until recently, none of them allowed for the accurate detection of all snoRNAs (Boivin et al., 2018; Deschamps-Francoeur et al.,2019) (Figure 3). Thus, new protocols and tools suited for representative whole-transcriptome sequencing experiments have been elaborated to resolve unaddressed biases within RNA-Seq.

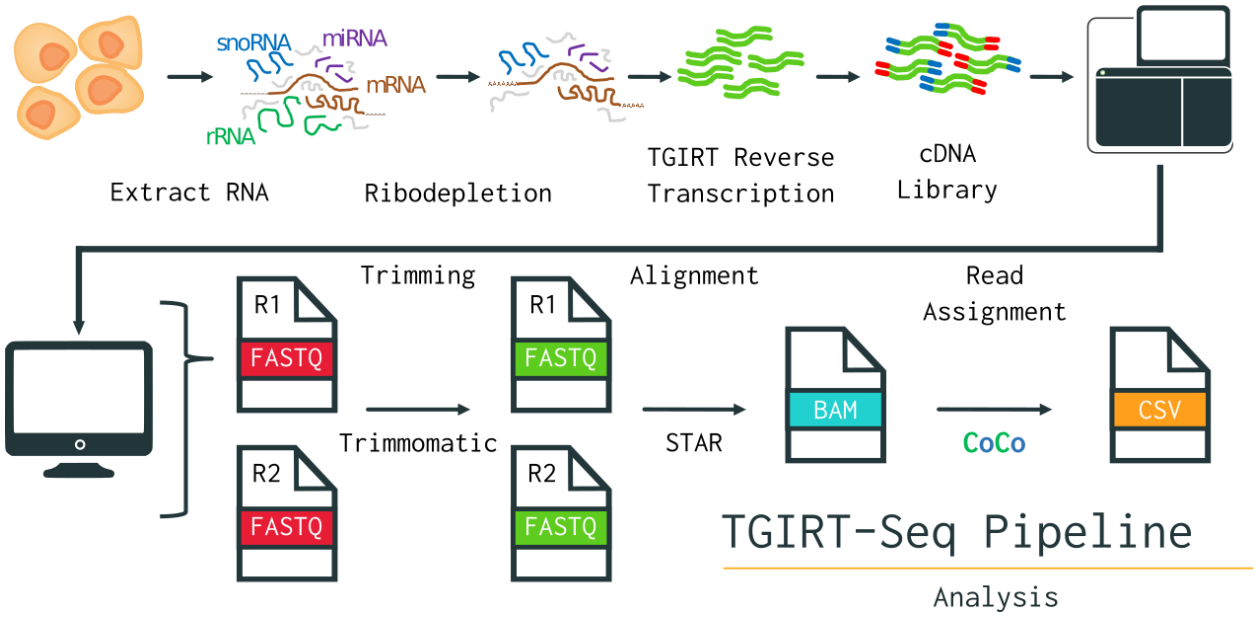

Figure 3: TGIRT-Seq + CoCo Pipeline for Holistic RNA Sequencing

Overview of the library preparation and bioinformatics analysis steps of an RNA-Seq experiment using TGIRT and CoCo.

RNA-Seq Part 1: Library Preparation and TGIRT

Sequencing experiments first begin with the isolation of total RNA from a population of cells. This RNA extract can then be treated in various ways to deplete rRNA, as their great abundance overshadows that of other RNA types. Alternatively, one may choose to ‘’pull’’ on the polyA tails, that feature mostly on mRNA as a sequence of adenine repeats at their 3’ end. This method does however create a detection bias in favor of those 3’ ends, which also makes it unsuitable for the detection of degraded or fragmented transcripts, and make it favor the detection of protein-coding RNA and polyA tail-bearing lncRNAs (Zhao et al., 2018).

Once the RNA extract has been suitably prepared, it can optionally be fragmented, using alkaline solutions or enzymes, to a specific length that fits the restriction imposed by some sequencing machines. The next step is often the reverse transcription of the RNA sample into a complementary DNA library (cDNA) (Hrdlickova et al., 2017).

TGIRT-Seq: Holistic RNA-Seq Addressing Structural Bias

RNA-Sequencing is often in fact DNA sequencing since the technology already exists, and because certain enzymes can affordably reverse transcribe RNA into DNA for sequencing to take place without new machinery being required (Wang et al., 2009). This reverse transcription step has traditionally always been carried out using retroviral reverse transcriptase (RT). This RT works well at room temperature but lacks the ability to run through highly structured RNA strands, such as long paired double stems-loops seen in H/ACA snoRNA (Figure 2) (Nottingham et al., 2016). Conducting this step at a higher temperature would denature and linearize highly structured RNAs that could then more easily be slid across by the enzyme. However, the retroviral RT (RRT) would also be denatured under these conditions rendering it unusable (Mohr et al., 2013). The laboratory of Professor Lambowitz from the University of Texas was the first to put forward an alternative enzyme to address this problem, known as TGIRT for Thermostable Group 2 Intron Reverse Transcriptase. As its name indicates, this bacterial enzyme is thermostable, a property that allows it to function at maximum efficiency at temperatures as high as 70 degrees Celsius. Unlike RRT, TGIRT also benefits from high processivity, that is the ability to polymerize long nucleic acid chains without releasing its substrate, and fidelity, which is the amount of mutations generated (Mohr et al., 2013; Nottingham et al., 2016). In a recent study, TGIRT-Seq was validated as detecting proportional abundances of highly structured RNAs, such as tRNAs and snoRNAs, without compromising the detection of other RNA types, such as mRNA or other ncRNAs, which often act as snoRNA host genes (Boivin et al., 2018).

After reverse transcription of RNA into cDNA, adaptors and primers are ligated to the ends of the cDNA fragments to differentiate their strand of origin (forward or reverse) and allow for their amplification via PCR. Typically, the amplification of cDNA is tracked by a computer in a process called base-calling, which tracks fluorescent signals, generated by the binding of labeled-nucleotides to the cDNA (Ledergerber and Dessimoz, 2011). As the amplification of PCR-generated cDNA-copies are performed in parallel, mixed signals are used to generate statistical quality scores for each base called. A popular algorithm for this is the Phred score (Ewing and

Green, 1998). The generated signals, and their quality scores, are translated to computerized data files which can then be analyzed using bioinformatic tools.

RNA-Seq Part 2: Computational Analysis & CoCo

The first processing step in the bioinformatic analysis of sequencing data is the evaluation of sequencing quality for each base, and the "trimming" of bad quality reads. This step also serves to remove the adaptor sequences used for amplification. There are several tools available for this task, one of them being Trimmomatic (Bolger et al., 2014). Typically, these tools remove reads with a Phred score below a threshold of 20, which equates to their being a 1 in 100 chance that the base was called incorrectly (Mbandi et al., 2014). Once trimming is complete, sequencing reads are often aligned to a reference genome annotation to identify which regions of the genome they correspond to. However, the presence of multiple references that don't always agree with each other is problematic. De novo transcriptome assembly, which doesn`t use an annotation, is possible though it constitutes a more daunting task without guaranteeing better results (Hölzer and Marz, 2019; Salzberg, 2019; Ungaro et al., 2017). Once again, many tools are available for alignment as well with different properties, with a popular aligner being the "ultrafast" STAR (Dobin et al., 2013). Beyond its alignment speed, STAR is also noteworthy for being one of the first competent splice-aware aligner, a property that enables it to align reads belonging to known alternative transcripts based on a database of splice-junctions (Gatto et al., 2014; Williams et al., 2014). Splice-awareness also enables mapping to be done to the genome, which contains intronic sequences that aligners such as STAR can account for, instead of aligning to a more biased transcript-based annotation (Liu et al., 2018). Regardless of the method used, once alignment is complete, the reads can finally be quantified.

CoCo: Count Corrector for Nested & Multimapped Genes

No matter which RNA-Seq protocol is used, our ability to accurately quantify RNA transcripts relies on the proper assignment of sequencing reads. This step often relies on the genomic annotation used for alignment which has the start and end positions of genes mapped to specific nucleotide coordinates. Mapping sequenced nucleotides to those found in the annotation allows

us to calculate how many times a region or gene of the genome was transcribed. However, the use of annotations presents us with certain biases, notably when trying to quantify small non-coding RNA (<200 nucleotides) like most snoRNAs. These small non-coding RNAs (sncRNA) are often found within the introns of other genes, called host genes, and/or found in many copies across the genome (Dupuis-Sandoval et al., 2015). This leads to most tools assigning sncRNA reads to their host and/or assigning reads to only one of the multiple mapped regions respectively. In both cases, sncRNA reads are being mishandled, leading to under-detection of sncRNA. A recently published tool called CoCo, for Count Corrector, addresses this bias and correctly reassigns up to 15% of sequencing reads, allowing for a more accurate quantification of RNA transcripts (Deschamps-Francoeur et al., 2019).

CoCo, coupled with TGIRT-Seq, currently stands at the cutting edge for those interested in whole-transcriptome RNA-Sequencing, including highly structured, embedded and/or multimapped RNAs. One such RNA type that often fits all three criteria is small nucleolar RNAs. Although proportional quantification of snoRNAs, along with other RNA types, yields incredibly useful information, it is by no means the only type of data that can be leveraged from high-throughput sequencing experiments. Adaptations of the methodology beyond straight RNA quantification have been elaborated, as seen with high-throughput RNA-RNA interaction experiments.

High-throughput RNA-RNA Interaction Studies

Having covered the various canonical and non-canonical interactions involving snoRNAs, we turn now to a methodology with the potential to exponentially expand said interactome. Large scale RNA-RNA interaction protocols sequence a range of RNA interactions present in a total RNA sample. These experiments incorporate crosslinking, which covalently bonds RNA-interacting strands in the sample with a loop, followed by a degradation of non-interacting single strands. Finally, a linearization and sequencing of the RNA duplexes is performed, after the ligation of primers for amplification as in standard RNA-Seq. The interacting RNAs are hence contiguous in a single strand, and using bioinformatics pipelines, their identity, interaction and abundance is recorded (Lu et al., 2018; Sharma et al., 2016). A database regrouping results from

multiple RNA-RNA interaction studies with various protocols was published in 2017 under the name RISE: RNA Interactome from Sequencing Experiments (Gong et al., 2018). RISE houses 1671 snoRNA interactions. snoRNA interactors range from the classic rRNAs and snRNAs to an abundance of mRNAs, snoRNAs and more. Despite the high potential for discovery of these techniques, they can only provide a snapshot of interactions in the cells, which no doubt explains the heterogeneous results between different studies. As more of these experiments are conducted perhaps with further improvements, an even wider array of potential interactions waiting to be confirmed will be available for research.

So, as we can see, the current snoRNA interactome is much more extensive and diverse than what was once believed, and this is just the beginning. Modern snoRNA studies are still often focusing only on canonical C/D box snoRNA functions, such as the latest paper yielding a snoRNA database, which sought to discover new snoRNAs and ascribe them methylation targets in rRNA (Jorjani et al., 2016). Turning to PubMed for statistics points to a similar bias (C/D snoRNA yields 92 PubMed articles in the last 5 years vs 42 for H/ACA snoRNA. Adding methylation and pseudouridylation to those searches brings the numbers down to 46 and 14 respectively). As more and more research orients itself towards non-canonical snoRNA functions, more of their so far unknown yet critical roles will likely emerge and further our understanding of RNA biology and health. To help pave the way towards such a future, an updated and holistic snoRNA database regrouping the wealth of available information is needed.

Databases

Databases are powerful tools for storing, organizing and searching through large quantities of data. We are often most familiar with online visual representations of databases. However, these often give limited searching powers compared to the actual database engines used in the back-end of the website. That is because many of these databases use a relational database management system (RDBMS) with SQL (Structured Query Language) to manage some or all of their data.

SQL & Relational Databases

SQL with its simple almost didactic syntax is a powerful tool for easily writing and executing queries to glean insight into and further organize large datasets (Rice et al., 2004). In contrast, online search engines must be individually programmed and tailored to fit specific needs which often leads to fewer options being considered to save time while data is made readily downloadable for import into one's own database management system for more extensive study. RDBMSs also allow for easier cross-referencing and hence large-scale studies of large datasets from multiple sources so long as they have a common unique identifier (ID) to link data entries across platforms together. These IDs are called primary keys when they are unique, and when two datasets imported into a relational database have compatible primary keys, the data they each contain can easily and selectively be joined together according to what is called a one-to-one relationship (Figure 4A). Database nomenclature also has us calling datasets inside a database ''tables''. When two tables share common columns, but one table's primary key is duplicated with distinct row information, we have what is called a one-to-many relationship, which can usually be treated as simply as a one-to-one join (Figure 4B). However, if identical identifiers are non-unique in both tables, we must deal with something called a many-to-many relationship (Figure 4C) which is slightly messier (Ferreira and Takai, 2007). The most common approach to ''solve'' these relations involves breaking up the tables into sub-tables with unique identifiers and then joining relevant information when needed, but several other context dependent methods can be employed (Figure 4D). For example, such alternative methods include joining on multiple columns that add-up to a composite primary key, aggregating linked rows together into delimiter-separated strings or arrays, pivoting the table on a column`s reoccurring data categories before the join, if one desires to have new columns based on those categories, and then performing a one-to-many join, etc. In short, dealing with large heterogeneous datasets inevitably gives rise to unwieldy relations between tables that SQL, or other database languages, help us programmatically sort out.

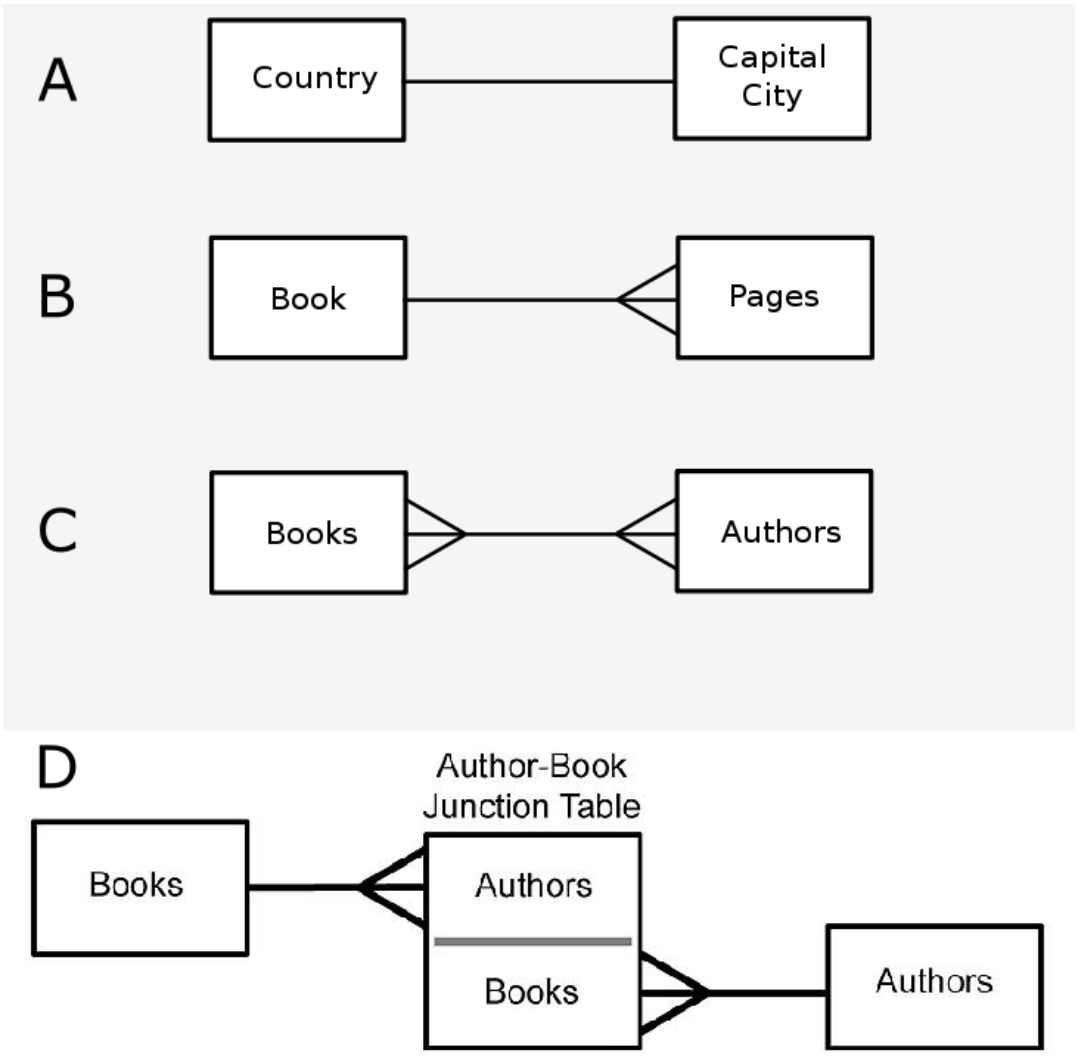

Figure 4: Data Models Found in Relational Databases

A) One-to-one relationship: A country has only one capital city and a capital city belong to only one country. B) One-to-Many relationship: A book has several pages, and those pages belong to the same unique book. C) Many-to-Many relationship: A book can be written by many authors, and an author can write many books. D) Common approach for dealing with Many-to-Many relationship: insert a junction table between them upon which a simpler One-to-Many join can be performed.

The State of Human snoRNA Data Online

Information on snoRNAs can be found in various online resources, as well as scattered across the literature. This, as with any other field of research, makes it difficult to obtain an accurate overview of our current knowledge on snoRNAs without spending tens or hundreds of hours reading research articles. Even for those who have accomplished this, staying on top of an ever-growing body of literature, as well as quickly recalling specific information on certain snoRNAs, becomes a seemingly impossible feat. This no doubt explains the plethora of online databases that exist for various subfields of research today. However, when it comes to snoRNAs, existing databases are out-of-date, incomplete or too narrow in scope to provide an accurate picture of the current snoRNA landscape. We believe desirable information in a modern snoRNA database includes: snoRNA names, genomic locations, sequence, host gene information, conservation, interactors and abundance across various tissues/cell lines (Figure 5).

snoRNABase

Published in 2006 (Lestrade and Weber, 2006) and last significantly updated in 2007, snoRNABase stands as the earliest database of human snoRNA. Cataloguing nearly 500 snoRNA entries and logging data ranging from box-type, structure, sequence, targets, host genes and referencing early articles the snoRNA is mentioned in, it served as a great resource for its time. However, snoRNABase isn't built upon a relational database model, which confers it a rather static nature, and making it much harder to periodically update or add/extract large amounts of data. Unsurprisingly then, over a decade after its publication, it exists more as a snapshot of the past which must be cross-referenced with new findings to ensure completeness and veracity of information.

Contents	snoRNA data
Identity	Symbol, ID, Synonyms , Sequence
Location	Chromosome, Start, End, Strand, Assembly
Host Genes	Symbol, ID, Synonyms, Biotype, Functions
Expression	sno & Host Abundance in Multiple Tissues
Targets	Canonical & Non-Canonical Interactions
Conservation	Orthologues, in which Species
Links	IDs and Links to Other Ressources

Contents	snoRNABase	snOPY	snoRNA Atlas
Identity	~	~	~
Location	hg18	✗	hg19
Host Genes	~	~	~
Expression	✗	✗	ENCODE
Targets	~	~	~
Conservation	✗	✓	~
Links	3	1	✗

Figure 5: Pertinent Information in a snoRNA Database

Tables showcasing a non-exhaustive list of pertinent data entries to be found in a modern snoRNA database, and the coverage, or lack thereof, of this information in existing snoRNA databases. "✓" equals good coverage, "~" means information is present but incomplete or out of date, and "✗" signifies data is either totally absent, inadequate, or severely outdated.

snOPY

Snoopy is an iconic comic-strip character created by Charles M. Schulz in 1950. The black and white beagle of the Charlie Brown universe probably inspired the name of the 2009 orthological snoRNA database snOPY with the site's url actually being spelled snoopy. Much like the character that seemingly inspired it, this database has a simple, focused design. snOPY showcases over
16 000 snoRNA orthologues across 35 species, which were found by performing sequence alignements of canonical snoRNA targets, from the 35 species, using ClustalW. Known snoRNA modification sites were then added to the alignments and matching sites qualified the snoRNAs that guide them as orthologues (Yoshihama et al., 2013). The database features data on snoRNA box-types, host genes and canonical targets in searchable tables divided by species. It holds 761 human snoRNA entries though strangely, some bear names not found in any other sources. Thankfully snOPY is still being updated and has been linked to RNAcentral, a database covering all non-coding RNAs sequences, which facilitated linking obscure snOPY entries to entries found in other resources based on sequence. This also made fetching data from snOPY more convenient since there is no good way to download the data they provide, unlike with RNAcentral (Yoshihama et al., 2013).

snoRNA Atlas

The most recent human snoRNA database, snoRNA Atlas, was published in 2016 and was seemingly never updated thereafter. It reanalysed small RNA-Seq data from the ENCODE project (Davis et al., 2018) to establish constitutive and cell type specific snoRNA expression. They later cross-referenced expression data with genomic prediction of snoRNA loci obtained with specialized tools and found new snoRNAs. Finally, the researchers also sought to find novel snoRNA methylation sites in rRNA using their new RIMSeq approach. Unfortunately, the ENCODE data they used wasn't obtained using TGIRT-Seq to correct for RNA-Seq's inherent structural bias, and they also did not use CoCo or any comparable means of accounting for multi-mapped and embedded snoRNAs. What's more, they used an outdated genomic assembly, hg19, in their analysis which contains less information than its updated counterpart, hg38, and hence

creates needless frictions for the comparison of their data with up to date resources. Although they claim to have analysed different tissues and cancer cell-lines, only a single expression column (with no units) is shown in their database along with hg19-related location data, box types, conservation data and mostly canonical target data (Jorjani et al., 2016).

General Databases

Having mentioned RNAcentral as a general database on all types of non-coding RNA, one might wonder if the reason no adequate snoRNA database has emerged is because general resources are good enough to fulfil the community's needs. Indeed, RNAcentral houses a lot of useful data from a plethora of sources such as the previously mentioned snOPY (RNAcentral Consortium, 2019). Another useful resource integrated into RNAcentral is Rfam, a database of ncRNA families. Rfam uses automated software to analyze structural alignment data that they then feed to a covariance model to find homologues of known ncRNA which they call families. While this method yields a lot of data, much of it seems to be false positives with little to no expression data supporting their existence so far (Kalvari et al., 2018). Nevertheless, considering that modern sequencing protocols have only recently been established that correctly detect many instances of ncRNA, Rfam presents an interesting predictor of ncRNA families based on sequence, structure and potential functional similarities.

However, while RNAcentral is a fantastic database, it does not describe itself as being all encompassing, choosing instead to specifically brand itself as a database of non-coding RNA sequences. The information it provides hence gravitates around information such as where sequences are in the genome, and their conservation in other organisms, while linking to other sites housing information on each sequence, and more recently what structures these sequences likely fold into. This means pertinent data specific to snoRNAs, such as what box motif they possess, if their sequence overlaps with another gene called a host gene, what interactions the snoRNAs might have, or their expression levels in various tissues or cell lines, is completely absent because it is out of RNAcentral's scope. Looking at other big general resources we see they also fall short with regards to the snoRNA data they provide, and cannot be said to serve as a well-rounded and practical information hub on the subject. To mention only the biggest one

among them, Ensembl, which provides similar data to RNAcentral while also providing a means of evaluating the confidence in transcript expression through its transcript support level system (Zerbino et al., 2018). However, this system is only applicable to protein-coding genes and pseudogenes. So again, while these databases stand as great sources of information, none of them provide enough depth of information on snoRNAs to help further research into both canonical and non-canonical snoRNA functions.

In summary, there is a wide range of snoRNA functions in humans, but information on them, and on snoRNAs in general, is scattered across the literature and in databases. The specialized databases are mostly outdated and isolated from each other and fail to catalogue the extent of known snoRNA sequences, their level of expression in various tissues, as well as their emerging functions. Considering all the fields in which human snoRNAs have been found to be functionally relevant, we believe a new, updated database of human snoRNAs would prove invaluable for the scientific community at large.

Hypothesis

snoRNAs have been attributed previously unsuspected functions for over a decade now ranging from the regulation of alternative splicing, chromatin structure, mRNA abundance, as well as being implicated in various diseases. These discoveries were often happened upon by researchers that did not have specific interests in snoRNA. There still exist snoRNAs that have no known function, and so far, only a handful of them have been attributed non-canonical functions. snoRNAs' most well-characterized function might play a role in ribosome heterogeneity, a novel concept with potentially big implications in stress response and development. This leads us to believe that we have only seen the tip of the iceberg when it comes to the snoRNA interactome. Unfortunately, current snoRNA databases feature little to no data on non-canonical functions, while also lacking other relevant information or features. Taken together, these facts outline the need for a better and updated online snoRNA resource to help further research into snoRNAs.

Objectives

Objective #1

Gather pertinent snoRNA related data from specialized and generalist resources alike and organize it all into the lab's PostgreSQL database.

Objectif #2

Connect the organized data to a web framework to publish it online and integrate various plugins to create a sleek, interactive and easy to use database that enables users to selectively view and extensively query large amounts of data that can also be downloaded and easily updated over time.

ARTICLE

snoDB: an interactive database of human snoRNA sequences, abundance and interactions

Authors: Philia Bouchard-Bourelle, Clément Desjardins-Henri, Darren Mathurin-St-Pierre, Gabrielle Deschamps-Francoeur, Étienne Fafard-Couture, Jean-Michel Garant, Sherif Abou Elela, Michelle S Scott

Status: Published (Bouchard-Bourelle P, et al., *snoDB: an interactive database of human snoRNA sequences, abundance and interactions*, Nucleic Acids Research, gkz884, https://doi.org/10.1093/nar/gkz884)

Foreword: My participation in this work encompasses the gathering of most of the data and their organization in a relational database. The code to link the database to a web framework had already been written by my predecessor Darren, but I still had to format outputted data and make them available online with the integration of various plugins to create a sleek interactive experience that remains easy to update. I also wrote the article with my supervisor Michelle Scott.

French Abstract:

Les petits ARN nucléolaire (snoRNA) forment un des types de petits ARN non-codants les plus abondants en cellule avec des fonctions qui sont conservées à travers tous les eucaryotes. On distingue deux types de snoRNA en fonction de motifs conservés les composant nommés des boites : les snoRNA à boites C/D et les snoRNA à boites H/ACA. La fonction leur étant principalement attribuée est d'agir comme guide pour la modification de sites spécifiques dans l'ARN ribosomal (rRNA) avec des implications dans la biogénèse des ribosomes. Toutefois, de plus en plus d'évidences mettent en lumière un éventail de fonctionnalités des snoRNA qui dépasse largement leur rôle canonique. Le nombre grandissant de snoRNA, leur expression non-uniforme dans différents types de cellules ainsi que leurs implications croissantes dans divers mécanismes régulateurs cruciaux d'expression génique et d'une panoplie de maladies souligne l'importance d'accroitre nos efforts de recherche sur les snoRNA. Afin de faciliter la caractérisation de ces petits ARN aux fonctions émergeantes, nous avons mis au point une base de données holistique de snoRNA humains intitulée snoDB. Cet outil en ligne consolide l'information de plusieurs autres sources sur les snoRNA humains tel que leur séquence, leur intéracteurs canoniques et non-canoniques trouvés à travers la littérature et tirés d'expériences à haut débit d'interactions ARN-ARN, ainsi que des données de séquençage d'ARN à haut débit dans divers type cellulaires qui peuvent être visualisées interactivement.

snoDB: an interactive database of human snoRNA sequences, abundance and interactions

Philia Bouchard-Bourelle[1], Clément Desjardins-Henri[1], Darren Mathurin-St-Pierre[1], Gabrielle Deschamps-Francoeur[1], Étienne Fafard-Couture[1], Jean-Michel Garant[1], Sherif Abou Elela[2] and Michelle S Scott[1]*

[1] Département de biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec J1E 4K8, Canada.

[2] Département de microbiologie et infectiologie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec J1E 4K8, Canada.

* To whom correspondence should be addressed.

Email: michelle.scott@usherbrooke.ca

Nucleic Acids Research, gkz884, https://doi.org/10.1093/nar/gkz884

Published: 10 October 2019

Keywords: box C/D snoRNA, box H/ACA snoRNA, sequence, RNA-seq expression, RNA-RNA interaction, snoRNA target, ribosomal RNA

Abstract

Small nucleolar RNAs (snoRNAs) are an abundant type of non-coding RNA with conserved functions in all known eukaryotes. Classified into two main families, the box C/D and H/ACA snoRNAs, they enact their most well characterized role of guiding site specific modifications in ribosomal RNA, through the formation of specific ribonucleoprotein complexes, with fundamental implications in ribosome biogenesis. However, it is becoming increasingly clear that the landscape of snoRNA cellular functionality is much broader than it once seemed with novel members, non-uniform expression patterns, new and diverse targets as well as several emerging non-canonical functions ranging from the modulation of alternative splicing to the regulation of chromatin architecture. In order to facilitate the further characterization of human snoRNAs in a holistic manner, we introduce an online interactive database tool: snoDB. Its purpose is to consolidate information on human snoRNAs from different sources such as sequence databases, target information, both canonical and non-canonical from the literature and from high-throughput RNA-RNA interaction datasets, as well as high-throughput sequencing data that can be visualized interactively.

Availability: http://scottgroup.med.usherbrooke.ca/snoDB/

Key points:

-snoRNA are non-coding RNA involved in ribosome biogenesis and many diverse other emerging functions

-snoDB is a novel inclusive and integrative human snoRNA database

-snoDB describes snoRNA sequence, host gene, interaction, conservation and expression data

Introduction

Small nucleolar RNAs (snoRNAs) are a conserved class of non-coding RNAs found in all eukaryotes and most extensively characterized as guiding site specific post-transcriptional modifications in ribosomal RNA (rRNA) (1,2). In addition, a small number of additional snoRNAs such as SNORD3 and SNORD118 are known to play a role in the processing and maturation of rRNA. Two types of snoRNAs have been described: box C/D and box H/ACA snoRNAs, the majority of which are encoded in introns of host genes in human (1,3). Box C/D and box H/ACA snoRNAs respectively guide the 2'-O-methylation and the pseudouridylation of their targets by direct base pairing. To do so, they require the interaction of core binding proteins, which provide stability and the catalytic activity, forming complexes known as snoRNPs (snoRNA ribonucleoprotein complexes)(4). In human, 110 rRNA residues are known to be methylated by snoRNPs and 100 are pseudouridylated (5) although recent high-throughput sequencing and systematic comparative genomics efforts have identified additional likely candidates as well as positions that are fractionally modified (6-9).

While canonical features, functionality and targets of snoRNAs are well-characterized, over the past decade, an increasingly large literature has exposed novel and unexpected aspects of snoRNA biology. High-throughput sequencing approaches give indications that snoRNAs can modify and/or otherwise interact with diverse RNAs including other snoRNAs, transfer RNAs and messenger RNAs (mRNAs) (6,10-13). As reviewed in (14), recent years have seen many potential novel functions being reported for snoRNAs including the modulation of alternative splicing (15-17), an essential involvement in stress response pathways (18-20), the regulation of pre-mRNA stability (21) and the modulation of mRNA 3' end processing (22). Moreover, high-throughput sequencing approaches and computational pipelines addressing the unique challenges of snoRNAs have been elaborated, resulting in more accurate quantification of snoRNAs, and simultaneously of their host genes, indicating that the levels of expression of snoRNAs cover a wide range and do not always mirror those of their host gene (23-25). The improved quantification and increased characterization of snoRNAs has led to increasing

numbers of snoRNAs and their host genes found to be involved in diseases. Examples of pathologies in which snoRNAs and their host genes play an important role and could be prime therapeutic targets include the Prader-Willi syndrome and diverse cancers (26-30). However, in many cases, while the involvement of snoRNAs in disease is now known, the molecular mechanism is unclear. Such is the case for SNORD118, mutations of which affect the expression, processing and protein binding of the snoRNA. But while SNORD118, like most snoRNAs, is ubiquitously expressed, germline mutations cause specific neurological phenotypes (31). The wealth of knowledge and data describing snoRNA biology requires careful management and integration to facilitate easy access and assimilation by the community. Unfortunately, much of the information regarding snoRNAs is disorganized, disseminated through disparate online platforms and peppered in the literature. For example, many RNA-RNA interactions have been detected for SNORD118 (11-13) and could be important to characterize the molecular mechanism of its involvement in disease, but mining them from high-throughput datasets from the literature is not straightforward. A central snoRNA resource would considerably facilitate the characterization of snoRNA functionality and involvement in disease.

Three dedicated snoRNA resources are currently available for human: snoRNAbase (5), snOPY (32) and snoRNA Atlas (33). However, these resources have either not been kept up to date with the new snoRNA genes annotated, new interactors and functionalities, or have a different scope (for example, snOPY is a database of snoRNA orthology). With so many key regulatory features emerging as intrinsic snoRNA functions, there is a pressing need to unify the scattered data currently available on human snoRNAs in order to optimize future research endeavors. The online interactive snoRNA database we propose, snoDB, aims to do that and more. Indeed, integrating available data is of great importance but snoDB further aspires to consolidate the above information with curated peer-reviewed high-throughput data in an effort to lead and incite research in the further characterization of the human snoRNA landscape in health and disease.

Database Content

SnoDB is based on the human hg38 reference genome assembly. It aims to be inclusive and integrate gene annotations and a wide diversity of features from all relevant available databases (Table 1). SnoRNA gene annotations were obtained from RefSeq (34), Ensembl (35) and RNAcentral (36), which in turn provides annotations from snOPY (32) and Rfam (37). Careful manual curation was carried out to consolidate the annotations and to ensure no snoRNA entries share exact same genomic coordinates. When different names are employed for a given snoRNA gene, the RefSeq name was used by default, but if absent, the RNAcentral or the Ensembl names were used. In addition to the gene symbol, genomic coordinates and gene sequence, all additional names obtained from the HUGO Gene Nomenclature Committee (HGNC) (38) are available in the 'synonym' column, and all identifiers of all above databases are provided as links. snoDB houses 2064 human snoRNAs, integrating the annotations of the above databases. In contrast, the other main snoRNA-centric resources, snoRNAbase (5), snOPY (32) and snoRNA Atlas (33), contain respectively 402, 760 and 1118 human snoRNAs (Table 1).

Table 1: Features of human snoRNA databases

Database	snoRNA count	Links to external resources	Orthology (O) and conservation (C)[a]	Host gene characteristics[b]	rRNA and snRNA target data	Non-canonical target data[c]	snoRNA expression data[d]	Host gene expression data[d]	Data available for download
snoRNAbase (5)	402	UCSC Genome Browser hg18 HGNC Genbank Literature	O (to yeast)	NCA	√	L	-	-	-
snOPY (32)	760	Refseq	O	N	√	-	-	-	-
snoRNA Atlas (33)	1118	Rfam	C	N	√	-	E	-	√
snoDB	2064	UCSC Genome Browser hg38 RefSeq HGNC Ensembl RNAcentral NCBI Rfam snoRNAbase snOPY snoRNA Atlas RISE database Literature	OC[d]	NBCA	√	LR	OPTLS	OPTLS	√

[a] In snoDB, links are provided to snOPY and Ensembl orthology pages when available and conservation data were obtained from snoRNA Atlas.

[b] Host gene characteristics: N: name; B: biotype; C: genomic coordinates; A: biological process annotation.

[c] Non-canonical target data are supported by articles in the literature (L) and by links to the RISE database (R).

[d] For snoRNA Atlas: E indicates amalgamated expression values from ENCODE. For snoDB: all expression values were obtained using the low structure bias TGIRT-seq methodology. O:

normal human ovary; P: normal human prostate; T: normal human testis; L: normal human liver; S: SKOV3ip1 human ovarian carcinoma cell line.

The snoRNA features that are available for display in snoDB also include host gene characteristics with a link to the Ensembl entry, the biotype, synonyms if relevant and genomic coordinates. In addition, snoDB features conservation data from snoRNA Atlas (33), orthology data from snOPY (32), snoRNA target data with enrichment details in select tissues from the human protein atlas (39) when available and expression data (Tables 1 and 2). Target data include known targets in rRNA annotated in snoRNAbase (5) and rRNA targets confirmed by RiboMethSeq (8). Non-canonical interactors that were experimentally validated in the literature are also included and links to the articles are available. These studies include (11,15-17,21), as described in the introduction. Finally, RNA-RNA interaction data were incorporated from the RISE:RNA Interactome, a database compiling results from multiple high-throughput RNA-RNA interaction studies (40) with the name and biotype of all RISE interactors being available. Levels of abundance of both snoRNAs and their host gene measured in various human tissues and cell lines using a low bias RNA-seq approach are also available as obtained from (23) and GEO entries from GSE126797. The snoDB back-end is built in PostgreSQL (9.5.1) as a relational database which is integrated into the Django web framework (1.6.5).

Table 2: Characteristics of snoRNAs in snoDB

	box C/D	box H/ACA	Other	Total
All snoRNAs[a]	1391	651	22	2064
Distinct snoRNA symbols[b]	461	246	21	728
Intronic snoRNAs encoded in host genes	423	318	3	744
Intergenic snoRNAs	968	333	19	1320
snoRNA-target pairs	1471	616	31	2118
• **snoRNA-rRNA target pairs**	481	255	2	738
• **snoRNA-snRNA target pairs**	113	64	7	184
• **snoRNA-noncanonical target pairs**[c]	877	297	22	1196
snoRNAs with transcriptomic data	524	469	3	996

[a] All snoRNAs include snoRNAs with the same name and/or sequence but encoded in different genomic loci.

[b] Counts every snoRNA symbol only once. Some snoRNAs bear the same symbol but have different IDs based on differences in their sequence and in the loci in which they are encoded or the length of their sequence.

[c] Non-canonical targets of snoRNAs include mRNAs and genomic regions not known to encode annotated genes.

Web Interface

The main page of snoDB is divided into four sections: 1) as shown in Figure 1A, the top of the page displays snoDB's logo adjacent to a search engine for snoRNA names. Immediately below the logo, a switch allows to toggle snoDB's sister tool snoTHAW (snoDB Table Heatmap Arrangement Widget), which enables the interactive visualization of abundance values of snoRNAs and their host gene. To the right of the logo can be found links to additional information pages on the database in the 'About', 'Tutorial', 'Statistics' and 'Experiment details' sections, as well as a link to a download page. 2) The section directly below (shown in Figure 1B) features a menu bar with options related to the table. Clicking on 'Column Options' reveals a set of buttons with 3 kinds of functionalities: toggling the visibility of single columns using the column visibility button, toggling the visibility of column groups using the color-coded buttons, and downloading data in either TSV, BED or XLSX file formats based on currently visible or selected rows in the table. The 'Advanced Search' option reveals 5 search bars that are specific to certain groups of columns as noted by their placeholder text and outline colors. The 'Reset Filters' option erases all filtering currently active on the table, whether it is from the topmost main search, the advanced search bars or the column specific search boxes in the table itself. This option, along with the 'Refresh Table' option that follows it, exist because the state of all search inputs, column visibilities and row selections are saved upon refreshing the page. Hence, 'Reset Filters' facilitates the clearing of all search fields without needing to refresh the page while 'Reset Tables' reloads the page back to its default state. 3) Below the options menu, the main table dynamically displays the snoDB data (Figure 1C). 4) The bottom of the page reveals snoTHAW when the switch at the top of the snoDB page is toggled. snoTHAW enables the visualization and interaction of RNA-seq expression data contained within snoDB (Figure 1D and Figure S2). Currently expression data are displayable for four healthy tissues (breast, liver, ovary and prostate) as well as the SKOV3ip1 ovarian cancer cell lines. In addition, box type, chromosome and conservation data also found in snoDB can be displayed on the heatmap's y-axis with the ability to re-order the columns and rows based on these features or based on the

expression data to suit the user's needs. All available expression data in snoDB was generated using the TGIRT-seq approach which allows accurate quantification and comparison of all cellular RNAs including highly structured and modified RNAs such as snoRNA (23-25), as described above. As more such datasets become available, they will also be incorporated in snoDB.

The main page features three levels of querying capabilities. The first consists of a single search-box which lies to the left of the snoDB logo atop the page (Figure 1A). Clicking and/or typing into this area reveals a drop-down menu comprised of all snoRNA symbols which reside in the table's first column of the same name. Multiple symbols can be selected making this a quick and easy way to access information on a few snoRNAs of interest. The second consists of the 5 previously mentioned search boxes located above the table upon clicking on the 'Advanced search' option. From left to right, the first one searches through the snoRNA symbols and synonyms columns, the second through all the external ID columns, the third through host symbol and synonyms, the fourth through target columns and the fifth and final search box is a global search covering the entire snoDB dataset. The first four search boxes operate on an exact-match basis while the global search supports partial search terms. All five search boxes support regular expressions as well as multiple space-separated terms making copy-pasting columns from a spreadsheet into an appropriate search engine an easy way to view numerous specific snoRNA entries. The third searching strategy is found within the table itself and provides individual column searching capabilities on select columns and it also supports multiple inputs.

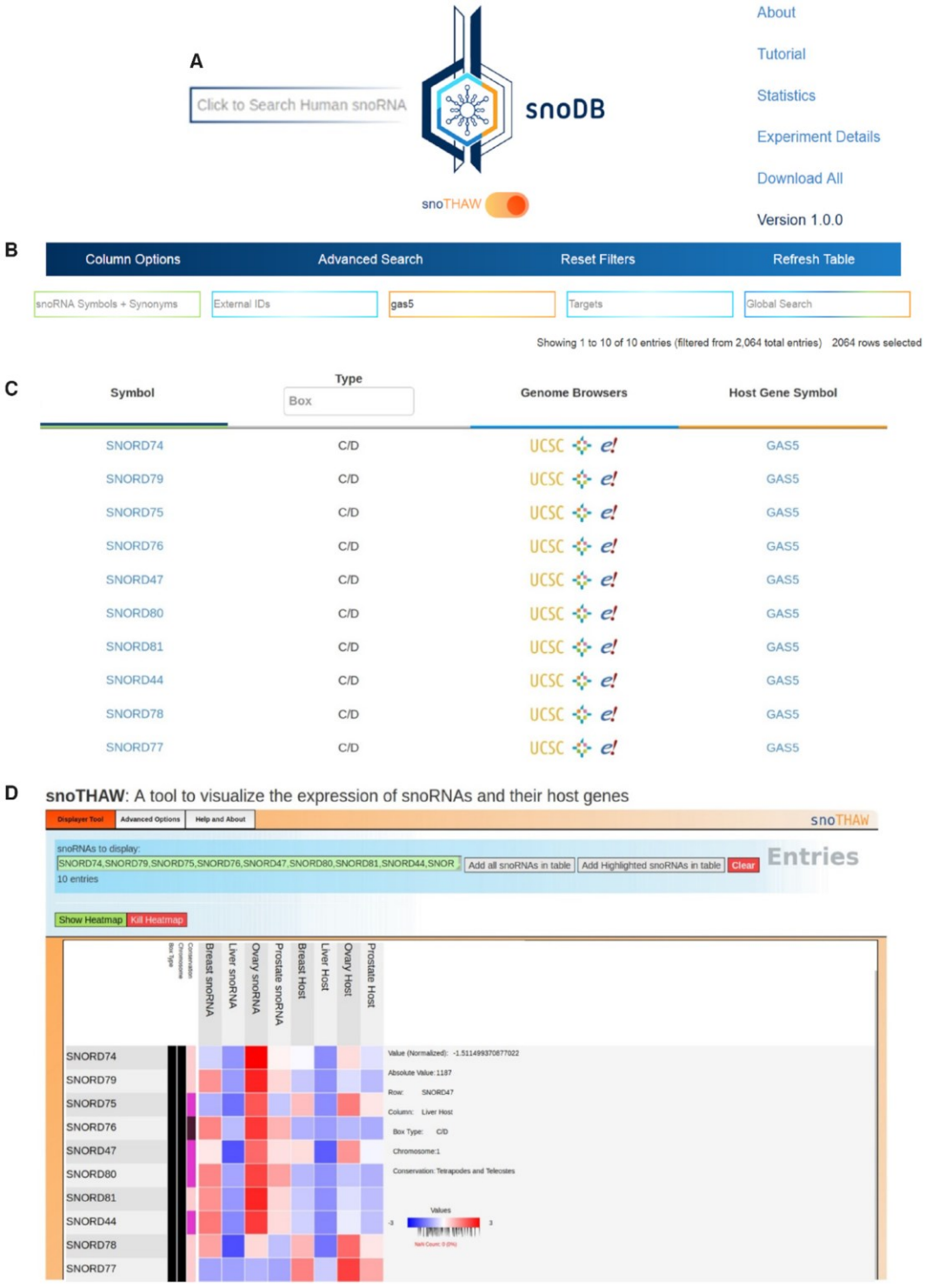

Figure 1. Screenshot of the main page of snoDB displaying the site's four sections. (A) The snoDB logo, basic search engine and links to information pages. (B) A menu bar with options to control the content and appearance of the table. (C) snoDB's main table where data are displayed and can be interacted with. By default, all 2064 snoRNA entries are shown by scrolling down. (D) The snoTHAW interface with the heatmap visualization beneath.

In addition to the interactive viewing and querying of columns, snoDB's main table contains the following features: a frozen first column for seamless horizontal scrolling through many columns, row selection upon click for visual highlights and as a means of input into snoTHAW, drag-and-drop column re-ordering, column sorting and an abundance of external links to corresponding snoRNA entries in other databases. All of these functionalities are described in the 'About' page as well as through interactive examples in the Tutorial (Figure S3).

While having all data selectively displayable in a single interactive table is a great convenience, it can also be impractical when one wishes to view all data for a single entry without needing to horizontally scroll back and forth. Therefore, clicking on any snoRNA in the 'Symbol' column opens a new tab to a page displaying all available information on that entry in a vertical format (Figure S4). These individual data hubs are divided into familiar sub-sections and feature external links to all previously mentioned sources along with additional links for interaction data, all of which can be searched through using the individual column search engines present.

Conclusion and Future Plans

The snoDB interactive web application is a holistic relational database which consolidates diverse information regarding human snoRNAs from key sources, curated articles and datasets in an attempt to facilitate further research in the field of snoRNAs. Along with minor periodic updates, additional high-throughput datasets will be incorporated in snoDB as they become available.

ACKNOWLEDGEMENTS

The authors are grateful to members of their groups, and in particular the Abou Elela group for useful discussions and to Leandro Fequino for technical support. MSS and SAE are members of the RNA group and the Centre de recherche du Centre hospitalier universitaire de Sherbrooke (CRCHUS).

Funding

This work was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant [to MSS]; a Fonds de Recherche du Québec – Nature et technologies (FRQNT) team Research grant [to MSS and SAE]; and the Fonds de Recherche du Québec – Santé (FRQS) Research Scholar Junior 2 Career Award [to MSS]. [GDF] is supported by a NSERC doctoral scholarship. [CDH] is supported by an undergraduate scholarship from the Faculty of Medicine and Health Sciences of the Université de Sherbrooke. SAE is supported by Canada Research Chair in RNA biology and Cancer Genomics.

Supplementary Data

Supplementary Data are available at NAR Online.

References

1. Dieci, G., Preti, M. and Montanini, B. (2009) Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94, 83-88.

2. Filipowicz, W. and Pogacic, V. (2002) Biogenesis of small nucleolar ribonucleoproteins. *Current opinion in cell biology*, 14, 319-327.

3. Boivin, V., Deschamps-Francoeur, G. and Scott, M.S. (2018) Protein coding genes as hosts for noncoding RNA expression. *Semin Cell Dev Biol*, 75, 3-12.

4. Massenet, S., Bertrand, E. and Verheggen, C. (2017) Assembly and trafficking of box C/D and H/ACA snoRNPs. *RNA biology*, 14, 680-692.

5. Lestrade, L. and Weber, M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic acids research*, 34, D158-162.

6. Gumienny, R., Jedlinski, D.J., Schmidt, A., Gypas, F., Martin, G., Vina-Vilaseca, A. and Zavolan, M. (2017) High-throughput identification of C/D box snoRNA targets with CLIP and RiboMeth-seq. *Nucleic Acids Res*, 45, 2341-2353.

7. Kehr, S., Bartschat, S., Tafer, H., Stadler, P.F. and Hertel, J. (2014) Matching of Soulmates: coevolution of snoRNAs and their targets. *Molecular biology and evolution*, 31, 455-467.

8. Krogh, N., Jansson, M.D., Hafner, S.J., Tehler, D., Birkedal, U., Christensen-Dalsgaard, M., Lund, A.H. and Nielsen, H. (2016) Profiling of 2'-O-Me in human rRNA reveals a subset of fractionally modified positions and provides evidence for ribosome heterogeneity. *Nucleic acids research*, 44, 7884-7895.

9. Incarnato, D., Anselmi, F., Morandi, E., Neri, F., Maldotti, M., Rapelli, S., Parlato, C., Basile, G. and Oliviero, S. (2017) High-throughput single-base resolution mapping of RNA 2-O-methylated residues. *Nucleic Acids Res*, 45, 1433-1441.

10. Kishore, S., Gruber, A.R., Jedlinski, D.J., Syed, A.P., Jorjani, H. and Zavolan, M. (2013) Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome biology*, 14, R45.

11. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. and Blencowe, B.J. (2016) Global Mapping of Human RNA-RNA Interactions. *Molecular cell*, 62, 618-626.

12. Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S. *et al.* (2016) RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, 165, 1267-1279.

13. Aw, J.G., Shen, Y., Wilm, A., Sun, M., Lim, X.N., Boon, K.L., Tapsin, S., Chan, Y.S., Tan, C.P., Sim, A.Y. *et al.* (2016) In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular cell*, 62, 603-617.

14. Dupuis-Sandoval, F., Poirier, M. and Scott, M.S. (2015) The emerging landscape of small nucleolar RNAs in cell biology. *Wiley interdisciplinary reviews. RNA*, 6, 381-397.

15. Kishore, S. and Stamm, S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*, 311, 230-232.

16. Falaleeva, M., Pages, A., Matuszek, Z., Hidmi, S., Agranat-Tamir, L., Korotkov, K., Nevo, Y., Eyras, E., Sperling, R. and Stamm, S. (2016) Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. *Proc Natl Acad Sci U S A*, 113, E1625-1634.

17. Scott, M.S., Ono, M., Yamada, K., Endo, A., Barton, G.J. and Lamond, A.I. (2012) Human box C/D snoRNA processing conservation across multiple cell types. *Nucleic acids research*, 40, 3676-3688.

18. Michel, C.I., Holley, C.L., Scruggs, B.S., Sidhu, R., Brookheart, R.T., Listenberger, L.L., Behlke, M.A., Ory, D.S. and Schaffer, J.E. (2011) Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. *Cell Metab*, 14, 33-44.

19. Lee, J., Harris, A.N., Holley, C.L., Mahadevan, J., Pyles, K.D., Lavagnino, Z., Scherrer, D.E., Fujiwara, H., Sidhu, R., Zhang, J. *et al.* (2016) Rpl13a small nucleolar RNAs regulate systemic glucose metabolism. *The Journal of clinical investigation*, 126, 4616-4625.

20. Rimer, J.M., Lee, J., Holley, C.L., Crowder, R.J., Chen, D.L., Hanson, P.I., Ory, D.S. and Schaffer, J.E. (2018) Long-range function of secreted small nucleolar RNAs that direct 2'-O-methylation. *The Journal of biological chemistry*, 293, 13284-13296.

21. Zhong, F., Zhou, N., Wu, K., Guo, Y., Tan, W., Zhang, H., Zhang, X., Geng, G., Pan, T., Luo, H. *et al.* (2015) A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. *Nucleic acids research*, 43, 10474-10491.

22. Huang, C., Shi, J., Guo, Y., Huang, W., Huang, S., Ming, S., Wu, X., Zhang, R., Ding, J., Zhao, W. *et al.* (2017) A snoRNA modulates mRNA 3' end processing and regulates the expression of a subset of mRNAs. *Nucleic acids research*, 45, 8647-8660.

23. Boivin, V., Deschamps-Francoeur, G., Couture, S., Nottingham, R.M., Bouchard-Bourelle, P., Lambowitz, A.M., Scott, M.S. and Abou-Elela, S. (2018) Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *Rna*, 24, 950-965.

24. Deschamps-Francoeur, G., Boivin, V., Elela, S.A. and Scott, M.S. (2019) CoCo: RNA-seq Read Assignment Correction for Nested Genes and Multimapped Reads. *Bioinformatics*.

25. Nottingham, R.M., Wu, D.C., Qin, Y., Yao, J., Hunicke-Smith, S. and Lambowitz, A.M. (2016) RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. *Rna*, 22, 597-613.

26. Cavaille, J. (2017) box C/D small nucleolar RNA genes and the Prader-Willi syndrome: a complex interplay. *Wiley interdisciplinary reviews. RNA*, 8.

27. Falaleeva, M., Welden, J.R., Duncan, M.J. and Stamm, S. (2017) C/D-box snoRNAs form methylating and non-methylating ribonucleoprotein complexes: Old dogs show new tricks. *Bioessays*, 39.

28. Romano, G., Veneziano, D., Acunzo, M. and Croce, C.M. (2017) Small non-coding RNA and cancer. *Carcinogenesis*, 38, 485-491.

29. Abel, Y. and Rederstorff, M. (2019) SnoRNAs and the emerging class of sdRNAs: Multifaceted players in oncogenesis. *Biochimie*.

30. Stepanov, G.A., Filippova, J.A., Komissarov, A.B., Kuligina, E.V., Richter, V.A. and Semenov, D.V. (2015) Regulatory role of small nucleolar RNAs in human diseases. *BioMed research international*, 2015, 206849.

31. Jenkinson, E.M., Rodero, M.P., Kasher, P.R., Uggenti, C., Oojageer, A., Goosey, L.C., Rose, Y., Kershaw, C.J., Urquhart, J.E., Williams, S.G. *et al.* (2016) Mutations in SNORD118 cause the cerebral microangiopathy leukoencephalopathy with calcifications and cysts. *Nature genetics*, 48, 1185-1192.

32. Yoshihama, M., Nakao, A. and Kenmochi, N. (2013) snOPY: a small nucleolar RNA orthological gene database. *BMC Res Notes*, 6, 426.

33. Jorjani, H., Kehr, S., Jedlinski, D.J., Gumienny, R., Hertel, J., Stadler, P.F., Zavolan, M. and Gruber, A.R. (2016) An updated human snoRNAome. *Nucleic acids research*, 44, 5068-5082.

34. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44, D733-745.

35. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G. *et al.* (2018) Ensembl 2018. *Nucleic acids research*, 46, D754-D761.

36. The, R.C. (2019) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic acids research*, 47, D221-D229.

37. Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D. and Petrov, A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids research*, 46, D335-D342.

38. Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S. and Bruford, E.A. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic acids research*, 45, D619-D625.

39. Uhlen, M., Fagerberg, L., Hallstrom, B.M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, 347, 1260419.

40. Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z.J., Yang, Y.T. and Zhang, Q.C. (2017) RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res*.

RESULTS

snoDB's Data: Which Data from Which Sources

Far from being a simple data aggregator, snoDB nevertheless consolidates data from an abundance of sources. Identifying which sources to gather which data from, and subsequently formatting and unifying all this information in a relational database, proved to be more of a challenge than anticipated.

Names, Synonyms & snoRNABase IDs: HUGO Gene Nomenclature Committee (HGNC)

HGNC was founded in 1979 and remains to this day the only recognized authority on human gene nomenclature (Braschi et al., 2019; Shows et al., 1979). By standardizing the nomenclature while conserving records of previous names and notable synonyms, they quickly became a crucial resource for the interoperability of other databases (Wain et al., 2002). HGNC eventually integrated snoRNABase IDs linking them to its roster of other resources like Ensembl, RNAcentral and the UCSC genome browser, which were previously detached from the snoRNA database (Casper et al., 2018; Lestrade and Weber, 2006; RNAcentral Consortium, 2019; Seal et al., 2011; Zerbino et al., 2018). This collaboration also helped to establish the familiar SNORD# and SNORA# nomenclature for snoRNAs, with SNORD being used for box C/D snoRNAs and SNORA being used for box H/ACA snoRNAs, though exceptions to this rule remain (Seal et al., 2011). HGNC focuses its effort on confirmed genes, which is why they number fewer snoRNA entries than resources integrating predicted genes such as Ensembl. This also means that many snoRNAs found in snoDB do not conform to HGNC's naming scheme. Given how widespread in use many older gene names are across the literature, exporting HGNC synonyms, for snoRNAs and their host genes, contributes to snoDBs greater usability.

Genomic Annotations: RefSeq/NCBI & Ensembl

In short, Ensembl and RefSeq are renowned public databases providing sequence information as well as their own genomic annotations. Similarly to HGNC, RefSeq/NCBI concentrate more on

experimentally validated genes while Ensembl features more predicted genes (Braschi et al., 2019; Haft et al., 2018; Zerbino et al., 2018). There are many more differences between the two but of fundamental interest to us is that these translate to Ensembl having about twice as many annotated human snoRNAs, and that all snoRNAs in RefSeq are not found in Ensembl. Some of these differences are slight, like a snoRNAs coordinates being shifted by one or two nucleotides while still having the same name. However, many entries are exclusive to both resources. Therefore, the annotation file used for our TGIRT-Seq experiments is largely composed of genes found in Ensembl with some exclusive to Refseq to broaden the amount of snoRNAs covered. Nevertheless, snoDB has shown us that there are many more snoRNAs whose expression have yet to be studied, meaning the annotation should be updated for future analysis.

Annotations & Cross-Reference Identifiers: RNAcentral

RNAcentral is the biggest database of ncRNA sequences (RNAcentral Consortium, 2019).
It integrates information from all previously named databases i.e. HGNC, RefSeq/NCBI and Ensembl in addition to snOPY and Rfam (RNAcentral Consortium, 2019). However, to build snoDB, data from those databases was still taken from their respective sources, when possible, instead of relying solely on RNAcentral to ensure cross platform accuracy. In so doing, minor errors, discrepancies and updates were found between the databases. We communicated with the resources in question who promptly corroborated our findings and made changes accordingly. Namely, we communicated with HGNC on two occasions to inform them they listed obsolete RefSeq identifiers for some snoRNA entries. The updated RefSeq ID for SNORD73B now listed it as being expressed, something our own TGIRT-Seq data also supports, prompting HGNC to remove the pseudogene tag they had previously ascribed to it. Minor issues were also found and signaled to RNAcentral such as snoRNA names for entries with only RefSeq IDs not being displaying on the website, as well as a single entry listing coordinates that should correspond to a 135 base long sequence but which listed a 137 base long sequence.

Some resources centered on names, like HGNC, group together Ensembl IDs and RefSeq IDs which have slight coordinate shifts. Meanwhile, RNAcentral's database architecture confers unique identifiers to every distinct snoRNA sequence. However, some snoRNAs with the same

sequence can be found in multiple locations across the genome, with those different locations sometimes bearing different snoRNA symbols. RNAcentral has nevertheless chosen to pool together snoRNAs with identical sequences and list their different locations under one identifier. This strategy differs from the one taken to build snoDB, which provides different identifiers to each snoRNA locus, even if its annotated sequence is identical to that of another locus. We believe this to be a better approach in our case for 3 reasons. 1) Sequencing reads sometimes don't overlap with genomic annotations meaning so-called identical sequences might in-fact be distinct, thought simply not listed as such in any current gene annotation. This is why we use CoCo for the assignment of our sequencing reads as it is the only tool to currently look at flanking reads to better match reads to the genome. In the case where the reads truly can be mapped to multiple locations, CoCo is also the only tool to equally divide reads, instead of giving them all to a single one of the two or more "copies" across the genome or discarding them altogether (Deschamps-Francoeur et al., 2019). 2) snoDB's current data structure is better-suited to displaying genomic loci data, and host gene data, in table form, something RNAcentral needs not concern themselves with. Indeed, were we to list multiple snoRNA loci under a single entry, the host gene columns and genomic location columns would be an overcrowded mess. We decided early into snoDB's development that we wanted to be able to cleanly display all information on snoRNAs in a big interactive table, to allow for the viewing and comparing of many data entries at a time. As such, we chose a data structure that allows for this type of display, which we prefer over the conventional emphasis on individual pages as seen with RNAcentral, Ensembl, Refseq, HGNC, etc. 3) The current state of snoRNA data online is more accurately represented by a loci-based data structure instead of a sequence-based one, as seen with the interaction database RISE. RISE lists different interactors for snoRNAs with identical sequences. For example, SNORD103A & SNORD103B share the same sequence yet only a single target is featured in RISE for SNORD103B vs 7 targets for SNORD103A. This is most likely a result of the problems with read assignment tools mentioned earlier, since the RNA-RNA interaction studies that produced the data contained in RISE used RNA-Sequencing but did not use CoCo. Of note, SNORD103A & SNORD103B also have different Ensembl IDs. Future updates of snoDB would benefit from interaction data being listed according to sequence rather than assigned using the identifiers listed in their study of origin.

RNA Interactions: snoRNABase, RISE, the Literature & the Human Protein Atlas

Information relating to canonical and non-canonical snoRNA interactions were downloaded from snoRNABase, several studies as well as from RISE (RNA Interactome from Sequencing Experiments). As its name implies, RISE catalogues RNA interaction data from high-throughput RNA-RNA interaction sequencing studies and its data are therefore predictions (Gong et al., 2018). However, such is the case with many snoRNA interactions regardless of their source, as few have been thoroughly validated beyond their possible base paring with rRNA (Dudnakova et al., 2018). Having a wide array of predictions from various independent sources therefore increases our confidence in those which can be traced to multiple resources. The community would certainly benefit from the painstaking curation efforts required to source all validated snoRNA interactions. However, this undertaking has not been attempted since snoRNABase, whose sources can unfortunately not be programmatically extracted (Lestrade and Weber, 2006).

As such, snoDB links back to all of the sources from which interaction data were taken, as seen on individual snoRNA pages in the "Interaction Data" section (Bouchard-Bourelle et al., 2019, Figure S4), giving users the means to assess how data was generated and ascribe them weights accordingly. To be able to list multiple sources in the in a single column online, sources were codified with a series of numbers with unique additions (1, 2, 4, 8, etc…). Summing up these numerals for each row will always correspond to a unique set of references associated to each target, which in turn allows us to program the display of sources accordingly (ex: 8= RISE, 1 + 8 = 9 = snoRNABase + RISE, etc…).

SNORD88C

Synonym	
Type	Box C/D
Location	19:50,802,328-50,802,418:- **Genome Browsers:** UCSC
Conservation	**Orthologs:** snOPY
Genomic Sequence	GGGCTCCCATGATGTCCAGCACTGGGCTCTGATCACCCCTGAGGACACAGTGCACCCCAGGACCTTTG ACACCTGGGGGTCTGAGGGGCCC

External IDs:

HGNC	Ensembl	NCBI	RefSeq	RNAcentral	Rfam	snoRNABase	snOPY	sr
HGNC:32749	ENSG00000220988			URS00006611EB	RF00604	SR0000271	Homo_sapiens300571	

Host Gene:

Symbol	Synonym	Biotype	Location	Function
C19orf48	BC006151, MGC13170	protein_coding	19:50,797,704-50,804,929:-	poorly characterized

Interaction Data:

Interactors (?)	Biotype	Tissue Enrichment	Source	Function
28S_C3680	rRNA		snoRNABase | Krogh N. et al. | RISE	
FGFR3	Protein Coding		RISE	Alternative splicing
GAS5	ncRNA		RISE	
HIPPI	Protein Coding			Alternative splicing
IL12B	Protein Coding		RISE	
RP11-299H22	lncRNA		RISE	
SNHG6	ncRNA		RISE	
SNORA70	snoRNA		RISE	

Expression Data:

Sample Type	snoRNA Abundance (TPM)	Host Abundance (TPM)
Breast	2,132	216
Liver	2,155	175
Ovarian Cancer Cell line	83	14
Ovary	4,855	248
Prostate	9,211	2,162

Figure S4: Screenshot of a detail page for SNORD88C. Basic information about the snoRNA itself populates the first table (Synonym, box type, conservation, orthologue and genomic sequence) with subsequent tables being specific to one type of data (External IDs, Host gene, Interactions and Abundance data). Individual search engines like those found in snoDB's main table are available in certain columns in the interaction and abundance tables.

Some manual curation of the literature was performed to specifically extract validated non-canonical snoRNA interactions, as they are underrepresented or even absent from previous databases. Many articles were found implicating snoRNAs in diseases, oxidative stress, the regulation of gene expression through various mechanism such an alternative splicing, chromatin organization, miRNA and piRNA derived fragments mediated mRNA silencing and decay, etc (Chu et al., 2012; Falaleeva et al., 2016; Kishore and Stamm, 2006; Michel et al., 2011; Ono et al., 2011). In particular, the literature surrounding diseases and snoRNAs as well as their host genes is extensive. Much more curation in this area would be required to integrate a disease and mutation section into snoDB. For now, over twenty non-canonical interactions feature in snoDB, with their implications and source being listed and linked in the "Function" column of the "Interaction Data" section on individual snoRNA pages (Bouchard-Bourelle et al., 2019, Figure S4).

This is in addition to the thousands of other interactions found in snoDB. They were too numerous to be effectively displayed in a single column in the main table, as many snoRNAs have multiple potential interactors. Thus, interaction data were pivoted using biotype information, which spreads interactors for every snoRNA between 10 categories. This further allowed us to highlight which snoRNAs have canonical and non-canonical interactors in the "Target types" column based on whether they have rRNA/snRNA targets or other types of targets.

Nevertheless, improvements are always possible and snoDB would benefit from developing the means to visualize interactions, as is currently the case for TGIRT-Seq expression data thanks to snoDB's sister tool snoTHAW (Bouchard-Bourelle et al., 2019, Figure S2). For interactions however, instead of heatmaps, we could envision an interactive functional network

similar to what can be achieved with protein coding genes through STRING (Szklarczyk et al., 2019).

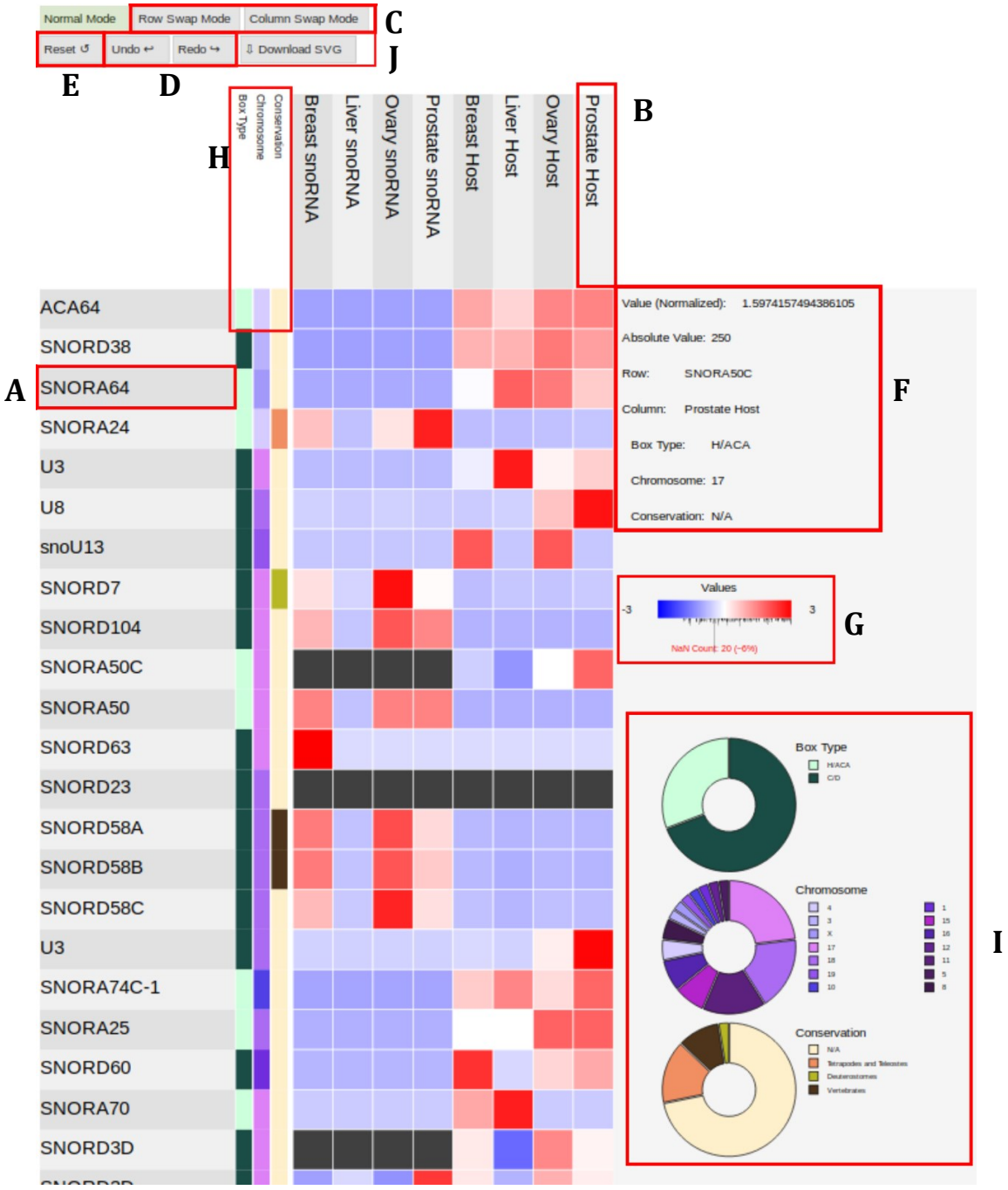

Figure S2: Screenshot of the heatmap visualization section of the snoTHAW tool. snoTHAW displays all selected snoRNAs as rows (A) and all selected datasets as columns (B) in the rendered heatmap. Clicking on the Row Swap and Column Swap buttons (C) enables dynamic

re-ordering of the heatmap. Operations on the heatmap are further facilitated using the Undo/Redo (D) and the Reset (E) buttons. If the user scrolls over the heatmap, characteristics of the current cell will be displayed to the right of the heatmap (F). The color legend of the heatmap is provided in (G). Characteristics of the snoRNAs including the box type, chromosome on which they are encoded, and conservation level can be added as additional columns (H) and doughnut charts displaying statistics of these characteristics are provided in (I). The heatmap can be downloaded (J).

In-House Data: TGIRT-Seq Datasets & Host Genes

It has by now been made abundantly clear that TGIRT-Seq currently stands as one of the best means of quantifying highly structured ncRNAs including certain snoRNAs while not compromising the detection of other RNA types (Boivin et al., 2018a). Unfortunately, the novel nature of the technique means relatively few publicly available TGIRT-Seq datasets exist yet. Thankfully, we can generate our own TGIRT-Seq data, with more datasets than what currently features in snoDB already on the way. This information is key to helping us understand and uncover preferential expression in certain tissues, as well as in a wide variety of potential disease or stress conditions, for both snoRNAs and their host genes. In addition, these data contribute to validating the existence of predicted snoRNAs at certain loci and would benefit from being extended to cover snoDBs entire catalogue in time.

The same is true of snoRNA host genes. Scripts from a past project in our group were used to gather information on the overlap of genes with snoRNAs and with snoDB now having exposed us to the wider potential array of human snoRNAs, these analyses should be done once more to update the host gene catalogue. Information on their function in tandem with their expression levels, which often don't correlate with the snoRNAs they host, will undoubtedly prove useful as the literature surrounding the involvement of snoRNA host genes in diseases continues to expand (Liao et al., 2010; Ronchetti et al., 2012; Williams and Farzaneh, 2012).

Conservation: snoRNA Atlas, snOPY & Ensembl

While snOPY is a specialized orthological snoRNA database, they offer no way to download their orthological data. Even if they did, showcasing it in a single cell for every snoRNA entry could be overwhelming for those snoRNAs with an abundance of orthologues. Meanwhile, snoRNA Atlas also houses conservation data which neatly fits into a single column that describes up to which evolutionary branch point snoRNAs are conserved. As such, conservation data proper were taken from snoRNA Atlas while snOPY's orthological data can be accessed alongside Ensembl's via links to these resources in snoDB's table (Jorjani et al., 2016; Yoshihama et al., 2013; Zerbino et al., 2018).

Joining Data Together

After gathering information from all previously mentioned sources, their data was formatted and exported to our lab PSQL databases into groups of tables by categories. Chief among those categories is the table containing all cross-reference identifiers, which enables data from all sources to be joined together. However, the consolidation of snoRNA identifiers spanning multiple databases with different data structures and annotations required careful consideration. We could not, for example, copy RNAcentral's data structure as we aimed to have each row in snoDB represent a single snoRNA for clarity's sake. Meanwhile RNAcentral often has multiple snoRNAs linked to the same ID that share a sequence but differ in the genomic location. In order to achieve our desired data structure, and in wanting to represent all sources equally, data were joined to confer distinct snoDB identifiers to every snoRNA with a different loci even if those loci partially overlap. With the table of identifiers serving as a pillar, all other data categories are joined via the common identifiers they contain to form the data table seen on snoDBs main page (Figure 6).

After some reflection, we envision to collapse partially overlapping snoRNA entries in ensuing updates to only house distinct snoRNA sequences. This could be achieved by prioritizing Ensembl's more permissive annotation when available over Refseq's, while still logging RefSeq's or any other annotation's data in a separate column in condensed form

(chromosome:start-end:strand). Further down the line, the use of our own sequencing read profile coordinates could be used as a main source to describe snoRNA loci in snoDB.

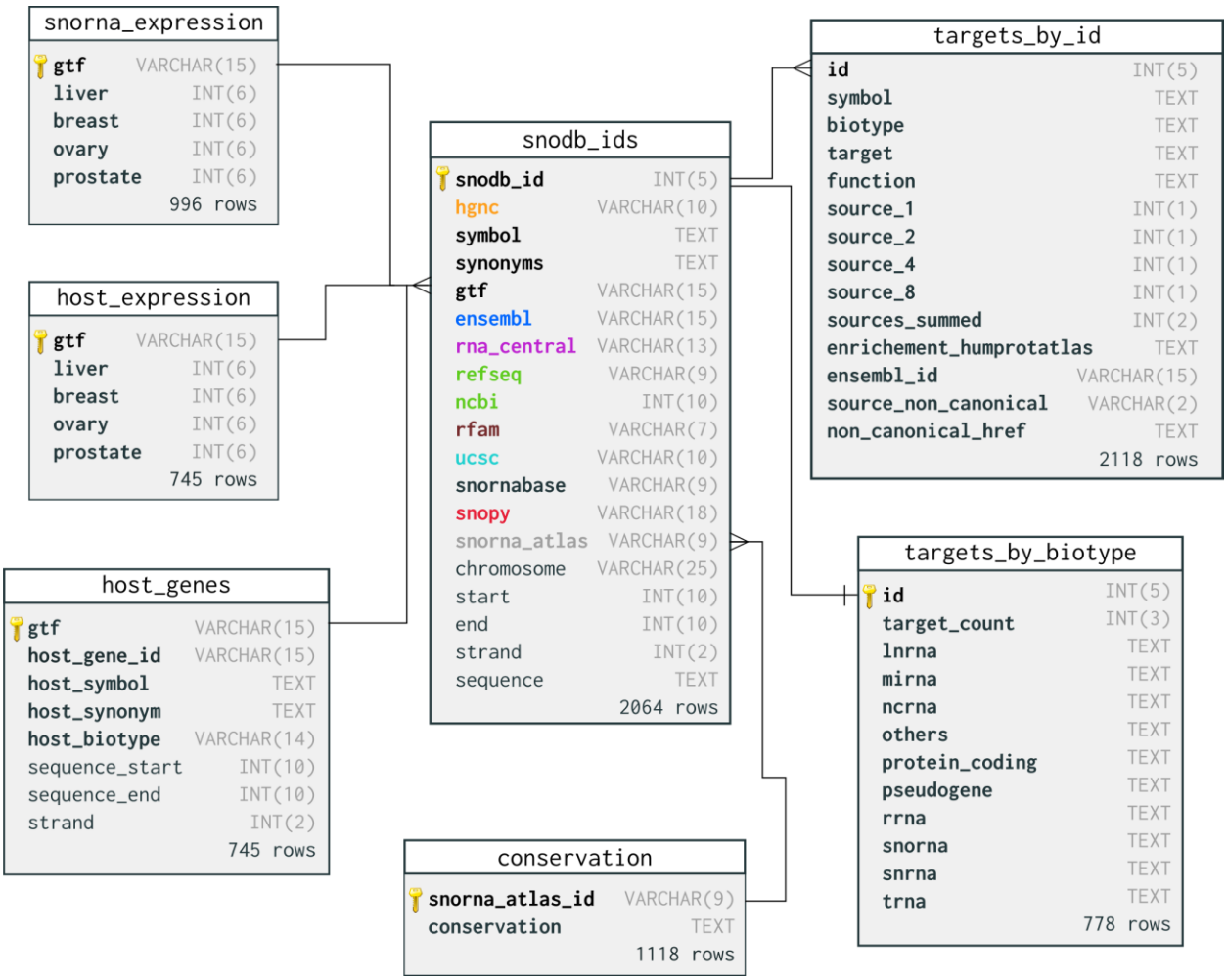

Figure 6: snoDBs Database Schema

Tables showcasing the various data categories and columns found in snoDB into which all imported data are formatted. The crow's feet indicate that a single record in one table is found multiples times in the other while the line (|) indicates that a single record in one table is also unique in another. The key symbols indicate that a column is a primary key and therefore this columns data is unique across the entire table. No capital letters are used because they do not show-up in our PSQL database. Colors are just there to emphasise the databases from which data was taken.

DISCUSSION

In this study, we showcase a novel human snoRNA database titled snoDB. Unlike previous iterations of this concept, snoDB offers a much richer experience both in term of the data it houses and in how said data are displayed in a modern interactive format. Consolidating information from a plethora of sources with links to all of them in page, featuring a wealth of querying options as well as the means to selectively visualize TGIRT-Seq expression data, in a variety of tissues, in heatmap form, snoDB constitutes a holistic resource of choice to help further research into the rapidly expanding field of snoRNAs.

What Previous snoRNA Databases Lacked Beyond their Data

In studying the older snoRNA databases for the initial data gathering step of this project, certain shortcomings in their design were noted and taken into consideration in order to improve upon them for snoDB.

Interconnectedness

The world-wide web is literally an interconnected web of pages that we access over the internet. This interconnectedness is its greatest strength as it facilitates the finding and sharing of information across time and space. Websites which fail to capitalize on this fundamental asset of the medium are therefore quickly buried and forgotten. Connecting one's scientific resource with other sources is not only of paramount importance for visibility purposes; it also increases transparency by allowing for the corroboration of information between sources at a mere click. What's more, it facilitates access to features present on other sites to not need to reinvent the wheel every time. We were therefore surprised to note that interconnectedness decreased over time with snoRNA databases.

snoRNABase, the earliest human snoRNA database, featured links to pertinent literature and general resources like HGNC and NCBI for almost all available entries. It is also linked from HGNC, meaning people perusing snoRNA entries on HGNC's website can find links to

corresponding snoRNABase pages when they exist. However, links from snoRNABase to HGNC are unfortunately no longer functional. Adding to this, though snoRNABase connects its data better externally than other resources, internally, the data featured is disjointed into multiple different pages making it difficult to gleam any kind of bird's eye view or easily draw comparisons between entries.

snOPY, the 2009 orthological snoRNA database focused on conservation, is comparatively more self-contained. Table exists for each species and all the links it contains are internal. On the individual pages for each snoRNA entry are sometimes found links to corresponding NCBI entries but no other external links are to be found. Which is surprising since snOPY's data have been integrated into RNAcentral. While RNAcentral links matching snoRNA sequences back to snOPY, linking them to other databases found within RNAcentral in the process, no individual snOPY entries are linked to corresponding RNAcentral pages on the snOPY website.

The worst offender and most recent of the snoRNA databases published in 2016, snoRNA Atlas, features a single external link across all of its pages which is for its entry of TERC (Telomerase RNA Component). The link leads to the main page of Telomerase Database which was last updated in 2013. This despite listing IDs from Rfam, a database of RNA families, which could have very easily been turned into functional hyperlinks. As for its internal structure, snoRNA Atlas can at least be commended on having all its data in a single location, as well as having individual pages for each snoRNA. The main page puts all the data in a big table which also features internal links to interaction profiles.

Meanwhile snoDB, which can also showcase all its data on its main page, features links to all three previously named snoRNA databases as well as to 7 other resources discussed above, when available, for each entry. These links are also found on each individual snoRNA page in addition to links to the literature for non-canonical interactions and to the human protein atlas when protein coding interactors show enrichment in certain tissues. snoDB has also been integrated into RNAcentral, the most comprehensive and well established resource for non-coding RNA sequences with information on over 16 million sequences from 35 different databases. This will undoubtedly expose snoDB to a wider audience over time.

Fully Downloadable Data

We have strived to make snoDB's data online as interactive and extensively queryable as possible in the hopes that it can easily be used by anyone. By also providing the option to download said data, we enable more informatic-savvy researchers to easily integrate the data into their own local databases which, regardless of their flavor, possess far greater querying power than any web-based application. This also facilitates cross-referencing of information contained in snoDB with the data researchers are currently working with. There are only advantages to making published information easily sharable in the scientific community. Yet aside from snoDB, only snoRNA Atlas features download files of all its data. snOPY has integrated some its contents into RNAcentral but otherwise, the data it lists on its main table must be manually copy-pasted to be obtained. Meanwhile snoRNABase has a single downloadable table, with its other tables needing to be copy pasted to collect their data. Additionally, the very informative interaction images found on snoRNABase's individual snoRNA pages can only be viewed one at a time with no batch download options.

Maintainability & Extensibility

Being a great resource of information means very little if said resource is not maintained and kept up to date. There are more snoRNA databases that have been published than those that have been discussed, but all of these are now inaccessible because their creators did not maintain them (Brown et al., 2003; Ellis et al., 2010; Samarsky and Fournier, 1999; Xie et al., 2007). However static maintenance is really the bare minimum as, to remain relevant, an information hub must be able to update its preexisting data categories but also extend itself to new areas, in keeping up with ongoing research. We obviously do not have access to the back-end code and/or database architecture of other snoRNA databases for comparison's sake, but here is how we went about facilitating snoDB's maintainability and extensibleness. This we hope will show how we can easily commit to Nucleic Acids Research's 5-year maintainability clause for all the tools published in its annual database issue, of which snoDB is now a part of.

The maintainability and extensibility in snoDB are both facilitated by a simple yet robust PostgreSQL codebase which translates itself to a clear relational database schema. Additionally, the database relies on popular, well documented, and regularly updated plugins, for the interactive/selective display of all its data on the main page, as well as a few of its querying options.

PostgreSQL Code & Tables

Instructions on where/how to access the various FTPs, downloadable files, and public databases queries from which snoDB gathers much of its data are all well documented. As is the process of formatting said data, with existing scripts, to fit into the various tables described in figure 6. From there, it is a simple matter of launching a bash script, which executes a cascade of PSQL commands, that updates existing tables with the new information. The relational nature of this data scheme makes it trivial to insert a new data category as its own table as long as it contains identifiers which can be linked to snoDBs cross-reference ID table.

Front-End Plugins

snoDB relies on third party plugins for the selective display of its data in a single big table as well as for some of its querying capabilities, like the individual column search boxes. These jQuery plugins, “Datatables” and “Yet Another Datatables Column Filter” (yadcf) respectively, are both open-source, are regularly updated and are relatively well documented with a slew of pre-existing forum threads on more specific user questions to help with a large range of potential applications. This once again adds itself to the snoDB specific documentation, which describes the more complex workarounds and what issues they were implemented to address. This documentation also contains instructions on how to easily add one or more columns to the table, or an entirely new section to the individual snoRNA pages, give them their own individual search engine with a few types to choose from, courtesy of yadcf (range selection, multiple search, enable regular expressions, etc), create a new page, etc. By relying on well-established plugins, snoDB’s codebase has been simplified when compared to what had been coded by the

original student who started the project several years ago. As a result, snoDB is easier to maintain, update and extend even by someone with rudimentary informatics knowledge.

Taken together, we believe these factors will undeniably contribute to snoDBs longevity and relevance as a resource that will not only be easily maintained but also extended over time in keeping with the literature and our labs own snoRNA research data.

User Experience

Using third party applications certainly has its benefits, but their use alone doesn't translate to an effective means of displaying data in a comprehensive way. Choosing which plugins and which ones of their many features to enable is key alongside forethought as to how data can be most effectively presented to users. Additionally, having a clear and concise, yet engaging, way of communicating features to users is a crucial, and frequently overlooked, element in scientific resources which all too often only offer walls of texts. snoDB approaches this latter point by utilizing a presentation framework called "reveal.js", which facilitates the creation of online PowerPoint-style slides with the ability to easily integrate working code into them. This was used to create small working examples of the many features found in snoDB, all of which can be easily navigated to, and from, thanks to an interactive table of contents that can be reached from any page via a corner banner (Bouchard-Bourelle et al., 2019, Figure S3).

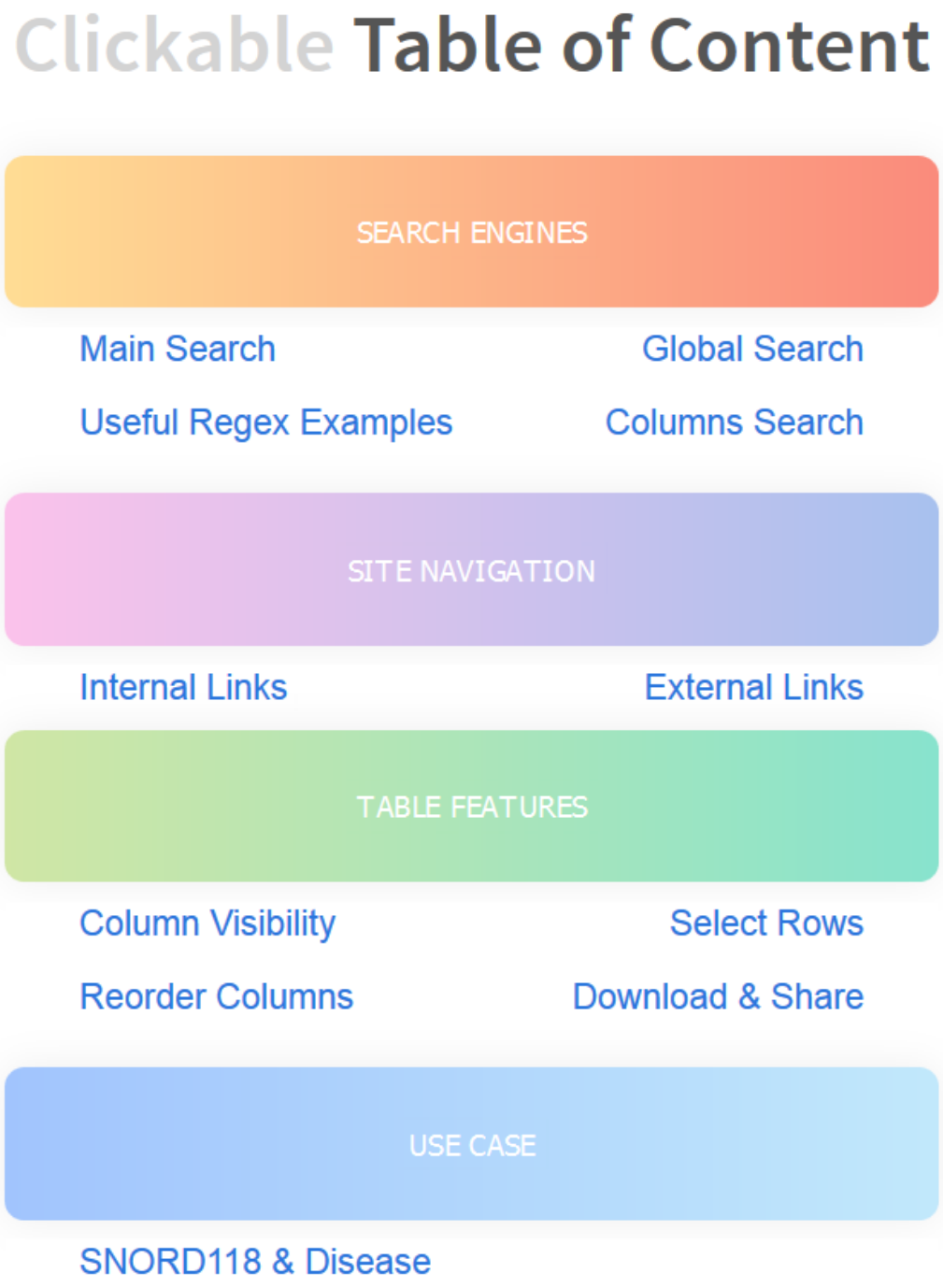

Figure S3: Screenshot of the Table of Content in snoDB's tutorial with links to interactive showcases of the site's features and content.

Previous snoRNA databases, and databases in general, shy away from offering the means of visualizing large amounts of data, prioritizing instead unitary pages dedicated to each entry, while sometimes offering small tables which give limited amounts of information. Yet we are becoming more and more interested in the analysis of biological systems in tandem with the individual parts that compose them, and our online resources should strive to reflect this emerging paradigm (Bard, 2013; Breitling, 2010). This is what motivated snoDB's modular design in terms of data visualization, with data categories and individual columns alike being highlighted and easily made to appear, or disappear, in a side-scrolling, re-orderable table, which can accommodate all available information. The potentially overwhelming nature of this type of display is offset by clear categorization, as mentioned above, but also by the specific columns. which appear by default and provide an overview of what each data category has to offer. Online scientific applications should aim to be more than simple online repositories where information is segregated and can only be easily brought together using database management systems. Even though such systems are indeed well suited to large scale data analysis, many people are not familiar with their use and we all benefit by making knowledge as freely accessible and connected as possible.

What should be Added to snoDB in Future

As previously noted, snoDB doesn't currently catalogue much information relating to snoRNA mutations and diseases. These related but distinct categories of information can be found scattered across an extensive body of literature as well as in some databases. These data are incredibly synergistic with current interaction predictions and tissue specific expression data found within snoDB as exemplified by the case of SNORD118 (U8) and LCC (Jenkinson et al., 2016). The exclusively neurological phenotype known as LCC, that was shown to arise as a result of mutation in both alleles of the U8 snoRNA, could be attributable to the predicted interactions U8 possesses with EHD3 and/or ZNF536. Both genes are shown to be specifically enriched in the cerebral cortex, according to the Human Protein Atlas, as seen in supplementary figure 1 of the article describing snoDB.

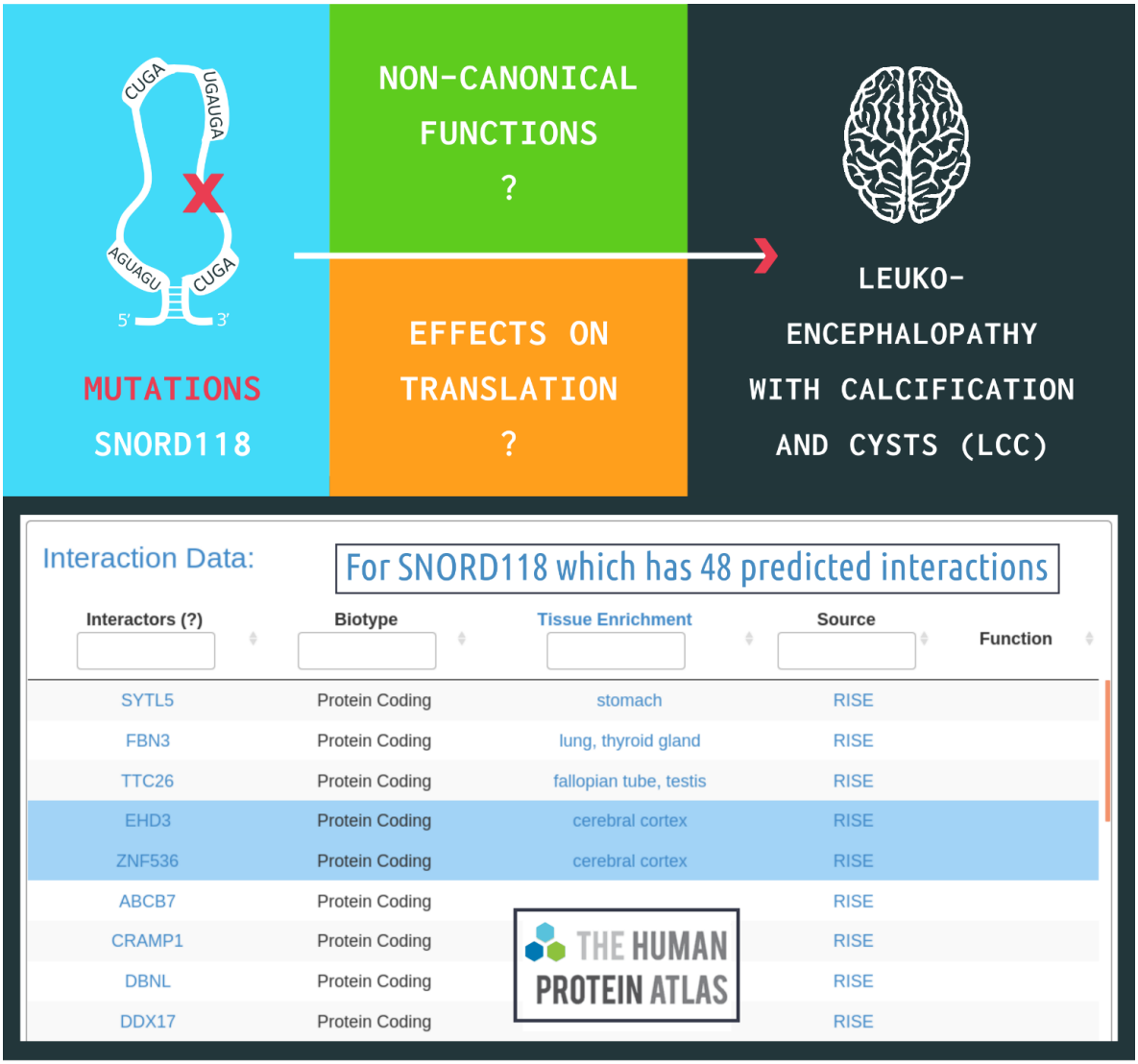

Figure S1: Case example for the use of snoDB. Jenkinson and co- authors reported that mutations in SNORD118 lead to leukoencephalopathy with calcifications and cysts (Nature Genetics (2016) 48, 1185- 92) although the exact molecular mechanism has yet to be established. Direct effects on translation were hypothesized as a cause but not yet investigated. This opens up the possibility of non- canonical interactions of SNORD118 being involved in this brain- specific illness. Interestingly, snoDB has integrated RNA- RNA interaction data from different sources and tissue enrichment data from the human protein atlas indicating that SNORD118 has two protein- coding RNA targets, EHD3 and ZNF536 with enhanced expression in the cerebral cortex. This is one example of how snoDB can lead to new avenues of exploration for researchers interested in human snoRNAs.

Conclusion

In conclusion, snoDB is a much-needed, modern, interconnected, and holistic online database of human snoRNAs. It contains an abundance of data, ranging from potential interactors, to conservation data in other species, expression in a growing number of tissues stemming from a lower bias TGIRT-Seq approach, and more. The database has been made to be extensively interactive yet simple to use, maintain and update, facilitating its future extensibility into other areas of snoRNA related research, such as mutations and diseases.

BIBLIOGRAPHY

Allen, F.W., 1941. The Biochemistry of the Nucleic Acids, Purines, and Pyrimidines. NUCLEIC ACIDS 25.

Alter, B.P., Rosenberg, P.S., Giri, N., Baerlocher, G.M., Lansdorp, P.M., Savage, S.A., 2012. Telomere length is associated with disease severity and declines with age in dyskeratosis congenita. Haematologica 97, 353–359. https://doi.org/10.3324/haematol.2011.055269

Altmann, R., 1889. Ueber nucleinsäuren. Arch. f. Anatomie u. Physiol, 1, 524-536.

Ambros, V., 2004. The functions of animal microRNAs. Nature 431, 350–355. https://doi.org/10.1038/nature02871

Avery, O.T., MacLeod, C.M., McCarty, M., 1944. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES. J Exp Med 79, 137–158.

Bachellerie, J.-P., Cavaillé, J., Hüttenhofer, A., 2002. The expanding snoRNA world. Biochimie 84, 775–790. https://doi.org/10.1016/S0300-9084(02)01402-5

Balakin, A.G., Smith, L., Fournier, M.J., 1996. The RNA World of the Nucleolus: Two Major Families of Small RNAs Defined by Different Box Elements with Related Functions. Cell 86, 823–834. https://doi.org/10.1016/S0092-8674(00)80156-7

Bard, J., 2013. Systems Biology — the Broader Perspective. Cells 2, 414–431. https://doi.org/10.3390/cells2020414

Beale, L.S. (Lionel S., 1858. The microscope in its application to practical medicine. London : Churchill.

Bertrand, E., Fournier, M.J., 2013. The snoRNPs and Related Machines: Ancient Devices That Mediate Maturation of rRNA and Other RNAs. Landes Bioscience.

Bhartiya, D., Scaria, V., 2016. Genomic variations in non-coding RNAs: Structure, function and regulation. Genomics 107, 59–68. https://doi.org/10.1016/j.ygeno.2016.01.005

Bhartiya, D., Talwar, J., Hasija, Y., Scaria, V., 2012. Systematic curation and analysis of genomic variations and their potential functional consequences in snoRNA loci. Human Mutation 33, E2367–E2374. https://doi.org/10.1002/humu.22158

Blenkiron, C., Hurley, D.G., Fitzgerald, S., Print, C.G., Lasham, A., 2013. Links between the Oncoprotein YB-1 and Small Non-Coding RNAs in Breast Cancer. PLoS One 8. https://doi.org/10.1371/journal.pone.0080171

Boivin, V., Deschamps-Francoeur, G., Couture, S., Nottingham, R.M., Bouchard-Bourelle, P., Lambowitz, A.M., Scott, M.S., Abou-Elela, S., 2018a. Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. RNA 24, 950–965. https://doi.org/10.1261/rna.064493.117

Boivin, V., Deschamps-Francoeur, G., Scott, M.S., 2018b. Protein coding genes as hosts for noncoding RNA expression. Seminars in Cell & Developmental Biology, Diversity of transcripts emanating from protein-coding genes 75, 3–12. https://doi.org/10.1016/j.semcdb.2017.08.016

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bouchard-Bourelle, P., Desjardins-Henri, C., Mathurin-St-Pierre, D., Deschamps-Francoeur, G., Fafard-Couture, É., Garant, J.-M., Elela, S.A., Scott, M.S., 2019. snoDB: an interactive database of human snoRNA sequences, abundance and interactions. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz884

Brachet, J., 1942. La localisation des acides pentosenucléiques dans les tissus animaux et les oeufs d'Amphibiens en voie de développement. Arch. Biol. 53, 207–257.

Brandis, K.A., Gale, S., Jinn, S., Langmade, S.J., Dudley-Rucker, N., Jiang, H., Sidhu, R., Ren, A., Goldberg, A., Schaffer, J.E., Ory, D.S., 2013. Box C/D Small Nucleolar RNA (snoRNA) U60 Regulates Intracellular Cholesterol Trafficking. J Biol Chem 288, 35703–35713. https://doi.org/10.1074/jbc.M113.488577

Braschi, B., Denny, P., Gray, K., Jones, T., Seal, R., Tweedie, S., Yates, B., Bruford, E., 2019. Genenames.org: the HGNC and VGNC resources in 2019. Nucleic Acids Res 47, D786–D792. https://doi.org/10.1093/nar/gky930

Breitling, R., 2010. What is Systems Biology? Front Physiol 1. https://doi.org/10.3389/fphys.2010.00009

Brockdorff, N., Ashworth, A., Kay, G.F., McCabe, V.M., Norris, D.P., Cooper, P.J., Swift, S., Rastan, S., 1992. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71, 515–526. https://doi.org/10.1016/0092-8674(92)90519-I

Brosius, J., Raabe, C.A., 2016. What is an RNA? A top layer for RNA classification. RNA Biol 13, 140–144. https://doi.org/10.1080/15476286.2015.1128064

Brown, C.J., Hendrich, B.D., Rupert, J.L., Lafrenière, R.G., Xing, Y., Lawrence, J., Willard, H.F., 1992. The human XIST gene: Analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71, 527–542. https://doi.org/10.1016/0092-8674(92)90520-M

Brown, J.W.S., Echeverria, M., Qu, L.-H., Lowe, T.M., Bachellerie, J.-P., Hüttenhofer, A., Kastenmayer, J.P., Green, P.J., Shaw, P., Marshall, D.F., 2003. Plant snoRNA database. Nucleic Acids Res 31, 432–435.

Busch, H., Reddy, R., Rothblum, L., Choi, Y.C., 1982. SnRNAs, SnRNPs, and RNA Processing. Annu. Rev. Biochem. 51, 617–654. https://doi.org/10.1146/annurev.bi.51.070182.003153

Cai, Y., Yu, X., Hu, S., Yu, J., 2009. A Brief Review on the Mechanisms of miRNA Regulation. Genomics, Proteomics & Bioinformatics 7, 147–154. https://doi.org/10.1016/S1672-0229(08)60044-3

Capspersson, T., 1947. The relations between nucleic acid and protein synthesis. Symposia of the Society for Experimental Biology 127–151.

Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., Hinrichs, A.S., Haeussler, M., Guruvadoo, L., Navarro Gonzalez, J., Gibson, D., Fiddes, I.T., Eisenhart, C., Diekhans, M., Clawson, H., Barber, G.P., Armstrong, J., Haussler, D., Kuhn, R.M., Kent, W.J., 2018. The UCSC Genome Browser database: 2018 update. Nucleic Acids Research 46, D762. https://doi.org/10.1093/nar/gkx1020

Cavaillé, J., Buiting, K., Kiefmann, M., Lalande, M., Brannan, C.I., Horsthemke, B., Bachellerie, J.-P., Brosius, J., Hüttenhofer, A., 2000. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A 97, 14311–14316.

Cech, T.R., 2012. The RNA Worlds in Context. Cold Spring Harb Perspect Biol 4. https://doi.org/10.1101/cshperspect.a006742

Cech, T.R., Zaug, A.J., Grabowski, P.J., 1981. In vitro splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. Cell 27, 487–496. https://doi.org/10.1016/0092-8674(81)90390-1

Chu, L., Su, M.Y., Maggi, L.B., Lu, L., Mullins, C., Crosby, S., Huang, G., Chng, W.J., Vij, R., Tomasson, M.H., 2012. Multiple myeloma–associated chromosomal translocation activates orphan snoRNA ACA11 to suppress oxidative stress. J Clin Invest 122, 2793–2806. https://doi.org/10.1172/JCI63051

Cléry, A., Senty-Ségault, V., Leclerc, F., Raué, H.A., Branlant, C., 2007. Analysis of Sequence and Structural Features That Identify the B/C Motif of U3 Small Nucleolar RNA as the Recognition Site for the Snu13p-Rrp9p Protein Pair. Molecular and Cellular Biology 27, 1191–1206. https://doi.org/10.1128/MCB.01287-06

Cobb, M., 2017. 60 years ago, Francis Crick changed the logic of biology. PLOS Biology 15, e2003243. https://doi.org/10.1371/journal.pbio.2003243

Cobb, M., 2015. Who discovered messenger RNA? Current Biology 25, R526–R532. https://doi.org/10.1016/j.cub.2015.05.032

Coffin, J.M., Fan, H., 2016. The Discovery of Reverse Transcriptase. Annual Review of Virology 3, 29–51. https://doi.org/10.1146/annurev-virology-110615-035556

Crick, F., 1958. On Protein Synthesis.

Dahm, R., 2005. Friedrich Miescher and the discovery of DNA. Developmental Biology 278, 274–288. https://doi.org/10.1016/j.ydbio.2004.11.028

Darzacq, X., Jády, B.E., Verheggen, C., Kiss, A.M., Bertrand, E., Kiss, T., 2002. Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. EMBO J 21, 2746–2756. https://doi.org/10.1093/emboj/21.11.2746

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., Onate, K.C., Graham, K., Miyasato, S.R., Dreszer, T.R., Strattan, J.S., Jolanki, O., Tanaka, F.Y., Cherry, J.M., 2018. The

Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res 46, D794–D801. https://doi.org/10.1093/nar/gkx1081
Deogharia, M., Majumder, M., 2018. Guide snoRNAs: Drivers or Passengers in Human Disease? Biology (Basel) 8. https://doi.org/10.3390/biology8010001
Deryusheva, S., Gall, J.G., 2019. scaRNAs and snoRNAs: Are they limited to specific classes of substrate RNAs? RNA 25, 17–22. https://doi.org/10.1261/rna.068593.118
Deryusheva, S., Gall, J.G., 2009. Small Cajal Body–specific RNAs of Drosophila Function in the Absence of Cajal Bodies. MBoC 20, 5250–5259. https://doi.org/10.1091/mbc.e09-09-0777
Deschamps-Francoeur, G., Boivin, V., Abou Elela, S., Scott, M.S., n.d. CoCo: RNA-seq read assignment correction for nested genes and multimapped reads. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz433
Dieci, G., Preti, M., Montanini, B., 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics 94, 83–88. https://doi.org/10.1016/j.ygeno.2009.05.002
Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635
Doe, C.M., Relkovic, D., Garfield, A.S., Dalley, J.W., Theobald, D.E.H., Humby, T., Wilkinson, L.S., Isles, A.R., 2009. Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. Hum. Mol. Genet. 18, 2140–2148. https://doi.org/10.1093/hmg/ddp137
Dudnakova, T., Dunn-Davies, H., Peters, R., Tollervey, D., 2018. Mapping targets for small nucleolar RNAs in yeast. Wellcome Open Res 3, 120. https://doi.org/10.12688/wellcomeopenres.14735.2
Dupuis-Sandoval, F., Poirier, M., Scott, M.S., 2015. The emerging landscape of small nucleolar RNAs in cell biology. Wiley Interdiscip Rev RNA 6, 381–397. https://doi.org/10.1002/wrna.1284
Dutca, L.M., Gallagher, J.E.G., Baserga, S.J., 2011. The initial U3 snoRNA:pre-rRNA base pairing interaction required for pre-18S rRNA folding revealed by in vivo chemical probing. Nucleic Acids Res 39, 5164–5180. https://doi.org/10.1093/nar/gkr044
Eck, R.V., 1961. Non-Randomness in Amino-Acid 'Alleles.' Nature 191, 1284–1285. https://doi.org/10.1038/1911284a0
Ehret, C.F., De Haller, G., 1963. Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium. Journal of Ultrastructure Research 9, 1–42. https://doi.org/10.1016/S0022-5320(63)80088-X
Ellis, J.C., Brown, D.D., Brown, J.W., 2010. The small nucleolar ribonucleoprotein (snoRNP) database. RNA 16, 664–666. https://doi.org/10.1261/rna.1871310
Ender, C., Krek, A., Friedländer, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N., Meister, G., 2008. A Human snoRNA with MicroRNA-Like Functions. Molecular Cell 32, 519–528. https://doi.org/10.1016/j.molcel.2008.10.017
Ewing, B., Green, P., 1998. Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. Genome Res. 8, 186–194. https://doi.org/10.1101/gr.8.3.186
Falaleeva, M., Pages, A., Matuszek, Z., Hidmi, S., Agranat-Tamir, L., Korotkov, K., Nevo, Y., Eyras, E., Sperling, R., Stamm, S., 2016. Dual function of C/D box small nucleolar RNAs in rRNA modification and alternative pre-mRNA splicing. Proc Natl Acad Sci U S A 113, E1625–E1634. https://doi.org/10.1073/pnas.1519292113

Fang, X., Yang, D., Luo, H., Wu, S., Dong, W., Xiao, J., Yuan, S., Ni, A., Zhang, K.-J., Liu, X.-Y., Chu, L., 2017. SNORD126 promotes HCC and CRC cell growth by activating the PI3K-AKT pathway through FGFR2. J Mol Cell Biol 9, 243–255. https://doi.org/10.1093/jmcb/mjw048

Ferreira, J.E., Takai, O.K., 2007. Understanding Database Design. National Center for Biotechnology Information (US).

Filipowicz, W., Pogačić, V., 2002. Biogenesis of small nucleolar ribonucleoproteins. Current Opinion in Cell Biology 14, 319–327. https://doi.org/10.1016/S0955-0674(02)00334-4

Frixione, E., Ruiz-Zamarripa, L., 2019. The "scientific catastrophe" in nucleic acids research that boosted molecular biology. J Biol Chem 294, 2249–2255. https://doi.org/10.1074/jbc.CL119.007397

Ganot, Philippe, Bortolin, M.-L., Kiss, T., 1997. Site-Specific Pseudouridine Formation in Preribosomal RNA Is Guided by Small Nucleolar RNAs. Cell 89, 799–809. https://doi.org/10.1016/S0092-8674(00)80263-9

Ganot, P., Caizergues-Ferrer, M., Kiss, T., 1997. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev. 11, 941–956. https://doi.org/10.1101/gad.11.7.941

Gao, L., Ma, J., Mannoor, K., Guarnera, M.A., Shetty, A., Zhan, M., Xing, L., Stass, S.A., Jiang, F., 2015. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. International Journal of Cancer 136, E623–E629. https://doi.org/10.1002/ijc.29169

Gaspin, C., Cavaillé, J., Erauso, G., Bachellerie, J.-P., 2000. Archaeal homologs of eukaryotic methylation guide small nucleolar RNAs: lessons from the Pyrococcus genomes11Edited by M. Yaniv. Journal of Molecular Biology 297, 895–906. https://doi.org/10.1006/jmbi.2000.3593

Gatto, A., Torroja-Fungairiño, C., Mazzarotto, F., Cook, S.A., Barton, P.J.R., Sánchez-Cabo, F., Lara-Pezzi, E., 2014. FineSplice, enhanced splice junction detection and quantification: a novel pipeline based on the assessment of diverse RNA-Seq alignment solutions. Nucleic Acids Res 42, e71. https://doi.org/10.1093/nar/gku166

Glatt-Deeley, H., Bancescu, D.L., Lalande, M., 2010. Prader–Willi syndrome, Snord115, and Htr2c editing. Neurogenetics 11, 143–144. https://doi.org/10.1007/s10048-009-0209-x

Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z.J., Yang, Y.T., Zhang, Q.C., 2018. RISE: a database of RNA interactome from sequencing experiments. Nucleic Acids Res 46, D194–D201. https://doi.org/10.1093/nar/gkx864

Griffith, F., 1928. The Significance of Pneumococcal Types. Epidemiology & Infection 27, 113–159. https://doi.org/10.1017/S0022172400031879

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., Altman, S., 1983. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell 35, 849–857. https://doi.org/10.1016/0092-8674(83)90117-4

Guo, H., 2018. Specialized ribosomes and the control of translation. Biochemical Society Transactions 46, 855–869. https://doi.org/10.1042/BST20160426

Haeckel, E., 1866. Generelle Morphologie der Organismen. Allgemeine Grundzüge der organischen Formen-Wissenschaft, mechanisch begründet durch die von C. Darwin reformirte Descendenz-Theorie, etc.

Haft, D.H., DiCuccio, M., Badretdin, A., Brover, V., Chetvernin, V., O'Neill, K., Li, W., Chitsaz, F., Derbyshire, M.K., Gonzales, N.R., Gwadz, M., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Yamashita, R.A., Zheng, C., Thibaud-Nissen, F., Geer, L.Y., Marchler-Bauer, A., Pruitt, K.D., 2018. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Research 46, D851. https://doi.org/10.1093/nar/gkx1068

Håkansson, K.E.J., Sollie, O., Simons, K.H., Quax, P.H.A., Jensen, J., Nossent, A.Y., 2018. Circulating Small Non-coding RNAs as Biomarkers for Recovery After Exhaustive or Repetitive Exercise. Front Physiol 9. https://doi.org/10.3389/fphys.2018.01136

Hammarsten O., 1894. Zur kenntnis der nucleoproteide. Z. Physiol. Chem. 19, 19–37.

Hargittai, I., 2009. The tetranucleotide hypothesis: a centennial. Struct Chem 20, 753–756. https://doi.org/10.1007/s11224-009-9497-x

He, X., Chen, X., Zhang, X., Duan, X., Pan, T., Hu, Q., Zhang, Y., Zhong, F., Liu, J., Zhang, Hong, Luo, J., Wu, K., Peng, G., Luo, H., Zhang, L., Li, X., Zhang, Hui, 2015. An Lnc RNA (GAS5)/SnoRNA-derived piRNA induces activation of TRAIL gene by site-specifically recruiting MLL/COMPASS-like complexes. Nucleic Acids Res 43, 3712–3725. https://doi.org/10.1093/nar/gkv214

Henras, A.K., Dez, C., Henry, Y., 2004. RNA structure and function in C/D and H/ACA s(no)RNPs. Current Opinion in Structural Biology 14, 335–343. https://doi.org/10.1016/j.sbi.2004.05.006

His et al., 1897. Die Histochemischen und Physiologischen Arbeiten. Leipzig : F.C.W. Vogel.

Hoagland, M.B., Zamecnik, P.C., Stephenson, M.L., 1957. Intermediate reactions in protein biosynthesis. Biochimica et Biophysica Acta 24, 215–216. https://doi.org/10.1016/0006-3002(57)90175-0

Hoeppner, M.P., White, S., Jeffares, D.C., Poole, A.M., 2009. Evolutionarily Stable Association of Intronic snoRNAs and microRNAs with Their Host Genes. Genome Biol Evol 1, 420–428. https://doi.org/10.1093/gbe/evp045

Hölzer, M., Marz, M., 2019. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8. https://doi.org/10.1093/gigascience/giz039

Hrdlickova, R., Toloue, M., Tian, B., 2017. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 8. https://doi.org/10.1002/wrna.1364

Huang, C., Shi, J., Guo, Y., Huang, W., Huang, S., Ming, S., Wu, X., Zhang, R., Ding, J., Zhao, W., Jia, J., Huang, X., Xiang, A.P., Shi, Y., Yao, C., 2017. A snoRNA modulates mRNA 3′ end processing and regulates the expression of a subset of mRNAs. Nucleic Acids Res 45, 8647–8660. https://doi.org/10.1093/nar/gkx651

Hunter, G.K., 1999. Phoebus Levene and the Tetranucleotide Structure of Nucleic Acids. Ambix 46, 73–103. https://doi.org/10.1179/amb.1999.46.2.73

Jacobs W. A., and Levene P. A., 1909. Proceedings of the American Society of Biological Chemists: On nucleic acids. J. Biol. Chem. 6, xxxvi–xxxvii [WWW Document]. URL http://www.jbc.org/content/7/1/vii.full.pdf (accessed 10.7.19).

Jády, B.E., Kiss, T., 2001. A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. EMBO J 20, 541–551. https://doi.org/10.1093/emboj/20.3.541

Jenkinson, E.M., Rodero, M.P., Kasher, P.R., Uggenti, C., Oojageer, A., Goosey, L.C., Rose, Y., Kershaw, C.J., Urquhart, J.E., Williams, S.G., Bhaskar, S.S., O'Sullivan, J., Baerlocher, G.M., Haubitz, M., Aubert, G., Barañano, K.W., Barnicoat, A.J., Battini, R., Berger, A.,

Blair, E.M., Brunstrom-Hernandez, J.E., Buckard, J.A., Cassiman, D.M., Caumes, R., Cordelli, D.M., De Waele, L.M., Fay, A.J., Ferreira, P., Fletcher, N.A., Fryer, A.E., Goel, H., Hemingway, C.A., Henneke, M., Hughes, I., Jefferson, R.J., Kumar, R., Lagae, L., Landrieu, P.G., Lourenço, C.M., Malpas, T.J., Mehta, S.G., Metz, I., Naidu, S., Õunap, K., Panzer, A., Prabhakar, P., Quaghebeur, G., Schiffmann, R., Sherr, E.H., Sinnathuray, K.R., Soh, C., Stewart, H.S., Stone, J., Van Esch, H., Van Mol, C.E.G., Vanderver, A., Wakeling, E.L., Whitney, A., Pavitt, G.D., Griffiths-Jones, S., Rice, G.I., Revy, P., van der Knaap, M.S., Livingston, J.H., O'Keefe, R.T., Crow, Y.J., 2016. Mutations in SNORD118 cause the cerebral microangiopathy leukoencephalopathy with calcifications and cysts. Nat Genet 48, 1185–1192. https://doi.org/10.1038/ng.3661

Jinn, S., Brandis, K.A., Ren, A., Chacko, A., Dudley-Rucker, N., Gale, S., Sidhu, R., Fujiwara, H., Jiang, H., Olsen, B.N., Schaffer, J.E., Ory, D.S., 2015. snoRNA U17 Regulates Cellular Cholesterol Trafficking. Cell Metab 21, 855–867. https://doi.org/10.1016/j.cmet.2015.04.010

Jones, M.E., 1953. Albrecht Kossel, A Biographical Sketch. Yale J Biol Med 26, 80–97.

Jorjani, H., Kehr, S., Jedlinski, D.J., Gumienny, R., Hertel, J., Stadler, P.F., Zavolan, M., Gruber, A.R., 2016. An updated human snoRNAome. Nucleic Acids Res 44, 5068–5082. https://doi.org/10.1093/nar/gkw386

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E.P., Rivas, E., Eddy, S.R., Bateman, A., Finn, R.D., Petrov, A.I., 2018. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46, D335–D342. https://doi.org/10.1093/nar/gkx1038

Kishore, S., Gruber, A.R., Jedlinski, D.J., Syed, A.P., Jorjani, H., Zavolan, M., 2013. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biology 14, R45. https://doi.org/10.1186/gb-2013-14-5-r45

Kishore, S., Stamm, S., 2006. The snoRNA HBII-52 Regulates Alternative Splicing of the Serotonin Receptor 2C. Science 311, 230–232. https://doi.org/10.1126/science.1118265

Kiss, T., 2001. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J 20, 3617–3622. https://doi.org/10.1093/emboj/20.14.3617

Kiss-László, Z., Henry, Y., Bachellerie, J.-P., Caizergues-Ferrer, M., Kiss, T., 1996. Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. Cell 85, 1077–1088. https://doi.org/10.1016/S0092-8674(00)81308-2

Krell, J., Frampton, A.E., Mirnezami, R., Harding, V., De Giorgio, A., Roca Alonso, L., Cohen, P., Ottaviani, S., Colombo, T., Jacob, J., Pellegrino, L., Buchanan, G., Stebbing, J., Castellano, L., 2014. Growth Arrest-Specific Transcript 5 Associated snoRNA Levels Are Related to p53 Expression and DNA Damage in Colorectal Cancer. PLoS One 9. https://doi.org/10.1371/journal.pone.0098561

Kufel, J., Grzechnik, P., 2019. Small Nucleolar RNAs Tell a Different Tale. Trends in Genetics 35, 104–117. https://doi.org/10.1016/j.tig.2018.11.005

Langhendries, J.-L., Nicolas, E., Doumont, G., Goldman, S., Lafontaine, D.L.J., 2016. The human box C/D snoRNAs U3 and U8 are required for pre-rRNA processing and tumorigenesis. Oncotarget 7, 59519–59534. https://doi.org/10.18632/oncotarget.11148

Lederberg Joshua, 1994. The Transformation of Genetics by DNA: An Anniversary Celebrationof AVERYM, ACLEODand MCCARTY(1944) 4.

Ledergerber, C., Dessimoz, C., 2011. Base-calling for next-generation sequencing platforms. Brief Bioinform 12, 489–497. https://doi.org/10.1093/bib/bbq077

Lee, J., Harris, A.N., Holley, C.L., Mahadevan, J., Pyles, K.D., Lavagnino, Z., Scherrer, D.E., Fujiwara, H., Sidhu, R., Zhang, J., Huang, S.C.-C., Piston, D.W., Remedi, M.S., Urano, F., Ory, D.S., Schaffer, J.E., 2016. Rpl13a small nucleolar RNAs regulate systemic glucose metabolism. J Clin Invest 126, 4616–4625. https://doi.org/10.1172/JCI88069

Lee, R.C., Feinbaum, R.L., Ambros, V., 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75, 843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., Steitz, J.A., 1980. Are snRNPs involved in splicing? Nature 283, 220. https://doi.org/10.1038/283220a0

Lestrade, L., Weber, M.J., 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. 34, D158-162. https://doi.org/10.1093/nar/gkj002

Levene, P.A., Jacobs, W.A., 1909. Über die Hefe-Nucleinsäure. Berichte der deutschen chemischen Gesellschaft 42, 2474–2478. https://doi.org/10.1002/cber.190904202148

Levene, P.A., Mikeska, L.A., Mori, T., 1930. On the Carbohydrate of Thymonucleic Acid. J. Biol. Chem. 85, 785–787.

Levene, P.A., Tipson, R.S., 1935. The Ring Structure of Thymidine. J. Biol. Chem. 109, 623–630.

Liang, J., Wen, J., Huang, Z., Chen, X., Zhang, B., Chu, L., 2019. Small Nucleolar RNAs: Insight Into Their Function in Cancer. Front Oncol 9. https://doi.org/10.3389/fonc.2019.00587

Liao, J., Yu, L., Mei, Y., Guarnera, M., Shen, J., Li, R., Liu, Z., Jiang, F., 2010. Small nucleolar RNA signatures as biomarkers for non-small-cell lung cancer. Mol Cancer 9, 198. https://doi.org/10.1186/1476-4598-9-198

Liu, X., MacLeod, J.N., Liu, J., 2018. iMapSplice: Alleviating reference bias through personalized RNA-seq alignment. PLoS One 13. https://doi.org/10.1371/journal.pone.0201554

Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J., 2000. Structure of Nucleic Acids. Molecular Cell Biology. 4th edition.

Lu, Z., Gong, J., Zhang, Q.C., 2018. PARIS: Psoralen Analysis of RNA Interactions and Structures with High Throughput and Resolution. Methods Mol Biol 1649, 59–84. https://doi.org/10.1007/978-1-4939-7213-5_4

Mannoor, K., Shen, J., Liao, J., Liu, Z., Jiang, F., 2014. Small nucleolar RNA signatures of lung tumor-initiating cells. Mol Cancer 13, 104. https://doi.org/10.1186/1476-4598-13-104

Martens-Uzunova, E.S., Hoogstrate, Y., Kalsbeek, A., Pigmans, B., Vredenbregt-van den Berg, M., Dits, N., Nielsen, S.J., Baker, A., Visakorpi, T., Bangma, C., Jenster, G., 2015. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. Oncotarget 6, 17430–17444.

Massenet, S., Bertrand, E., Verheggen, C., 2016. Assembly and trafficking of box C/D and H/ACA snoRNPs. RNA Biol 14, 680–692. https://doi.org/10.1080/15476286.2016.1243646

Mattick, J.S., Gagen, M.J., 2001. The Evolution of Controlled Multitasked Gene Networks: The Role of Introns and Other Noncoding RNAs in the Development of Complex Organisms. Mol Biol Evol 18, 1611–1630. https://doi.org/10.1093/oxfordjournals.molbev.a003951

Maxwell, E., Fournier, M., 1995. THE SMALL NUCLEOLAR RNAs. Annual Review of Biochemistry 64, 897–934. https://doi.org/10.1146/annurev.bi.64.070195.004341
Mbandi, S.K., Hesse, U., Rees, D.J.G., Christoffels, A., 2014. A glance at quality score: implication for de novo transcriptome reconstruction of Illumina reads. Front Genet 5. https://doi.org/10.3389/fgene.2014.00017
McGettigan, P.A., 2013. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 17, 4–11. https://doi.org/10.1016/j.cbpa.2012.12.008
McPherson, A., Gavira, J.A., 2013. Introduction to protein crystallization. Acta Crystallogr F Struct Biol Commun 70, 2–20. https://doi.org/10.1107/S2053230X13033141
Mei, Y., Liao, J., Shen, J., Yu, L., Liu, B., Liu, L., Li, R., Ji, L., Dorsey, S., Jiang, Z., Katz, R., Wang, J.-Y., Jiang, F., 2012. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. Oncogene 31, 2794–2804. https://doi.org/10.1038/onc.2011.449
Michel, C.I., Holley, C.L., Scruggs, B.S., Sidhu, R., Brookheart, R.T., Listenberger, L.L., Behlke, M.A., Ory, D.S., Schaffer, J.E., 2011. Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. Cell Metab. 14, 33–44. https://doi.org/10.1016/j.cmet.2011.04.009
Miescher, F., 1874. Die Spermatozoen einiger Wirbelthiere. Ein Beitrag zur Histochemie. Verh Nat Forsch Ges Basel 6, 138–208.
Miescher, F., 1871. Ueber die chemische Zusammensetzung der Eiterzellen. Medicinisch-chemische Untersuchungen 4, 441–460.
Miescher, F., 1869. Letter I; to Wilhelm His; Tübingen, February 26th, 1869. Die Histochemischen und Physiologischen Arbeiten von Friedrich Miescher–Aus dem Wissenschaftlichen Briefwechsel von F. Miescher, His W et al, 33-38.
Miescher, F., Schmiedeberg, O., 1896. Physiologisch-chemische Untersuchungen über die Lachsmilch. Archiv f. experiment. Pathol. u. Pharmakol 37, 100–155. https://doi.org/10.1007/BF01966284
Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., Polioudakis, D., Iyer, V.R., Hunicke-Smith, S., Swamy, S., Kuersten, S., Lambowitz, A.M., 2013. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. RNA 19, 958–970. https://doi.org/10.1261/rna.039743.113
Morange, M., 2008. What history tells us XIII. Fifty years of the Central Dogma. J Biosci 33, 171–175. https://doi.org/10.1007/s12038-008-0034-7
Mount, S.M., Wolin, S.L., 2015. Recognizing the 35th anniversary of the proposal that snRNPs are involved in splicing. Mol Biol Cell 26, 3557–3560. https://doi.org/10.1091/mbc.E14-10-1486
Neveu, M., Kim, H.-J., Benner, S.A., 2013. The “Strong” RNA World Hypothesis: Fifty Years Old. Astrobiology 13, 391–403. https://doi.org/10.1089/ast.2012.0868
Nottingham, R.M., Wu, D.C., Qin, Y., Yao, J., Hunicke-Smith, S., Lambowitz, A.M., 2016. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. RNA 22, 597–613. https://doi.org/10.1261/rna.055558.115
Olby, R., 1969. Cell chemistry in Miescher’s day. Med Hist 13, 377–382.
Olby, R.C., 1994. The Path to the Double Helix: The Discovery of DNA. Courier Corporation.
Omer, A.D., Lowe, T.M., Russell, A.G., Ebhardt, H., Eddy, S.R., Dennis, P.P., 2000. Homologs of Small Nucleolar RNAs in Archaea. Science 288, 517–522. https://doi.org/10.1126/science.288.5465.517

Ono, M., Scott, M.S., Yamada, K., Avolio, F., Barton, G.J., Lamond, A.I., 2011. Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Res 39, 3879–3891. https://doi.org/10.1093/nar/gkq1355

P. A. Levene, 1910. ON THE BIOCHEMISTRY OF NUCLEIC ACIDS.2 | Journal of the American Chemical Society [WWW Document]. URL https://pubs.acs.org/doi/abs/10.1021/ja01920a010 (accessed 9.23.19).

Parry, E.M., Alder, J.K., Lee, S.S., Phillips, J.A., Loyd, J.E., Duggal, P., Armanios, M., 2011. Decreased dyskerin levels as a mechanism of telomere shortening in X-linked dyskeratosis congenita. J Med Genet 48, 327–333. https://doi.org/10.1136/jmg.2010.085100

Patel, S.B., Bellini, M., 2008. The assembly of a spliceosomal small nuclear ribonucleoprotein particle. Nucleic Acids Res 36, 6482–6493. https://doi.org/10.1093/nar/gkn658

Peculis, B.A., 1997. The sequence of the 5' end of the U8 small nucleolar RNA is critical for 5.8S and 28S rRNA maturation. Mol Cell Biol 17, 3702–3713.

Pene, J.J., Knight, E., Darnell, J.E., 1968. Characterization of a new low molecular weight RNA in HeLa cell ribosomes. Journal of Molecular Biology 33, 609–623. https://doi.org/10.1016/0022-2836(68)90309-4

Piekna-Przybylska, D., Decatur, W.A., Fournier, M.J., 2007. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. RNA 13, 305–312. https://doi.org/10.1261/rna.373107

Piétu, G., Mariage-Samson, R., Fayein, N.-A., Matingou, C., Eveno, E., Houlgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M.-D., Wirkner, U., Ansorge, W., Cox, D., Nagase, T., Nomura, N., Auffray, C., 1999. The Genexpress IMAGE Knowledge Base of the Human Brain Transcriptome: A Prototype Integrated Resource for Functional and Computational Genomics. Genome Res 9, 195–209.

Ralf Dahm, 2008. Discovering DNA: Friedrich Miescher and the early years of nucleic acid research | SpringerLink [WWW Document]. URL https://link.springer.com/article/10.1007%2Fs00439-007-0433-0 (accessed 9.23.19).

Rice, M., Gladstone, W., Weir, M., 2004. Relational Databases: A Transparent Framework for Encouraging Biology Students To Think Informatically. Cell Biol Educ 3, 241–252. https://doi.org/10.1187/cbe.03-09-0012

Rich, A., RajBhandary, U.L., 1976. Transfer RNA: Molecular Structure, Sequence, and Properties. Annu. Rev. Biochem. 45, 805–860. https://doi.org/10.1146/annurev.bi.45.070176.004105

Riedel, N., Wise, J.A., Swerdlow, H., Mak, A., Guthrie, C., 1986. Small nuclear RNAs from Saccharomyces cerevisiae: unexpected diversity in abundance, size, and molecular complexity. Proc Natl Acad Sci U S A 83, 8097–8101.

RNAcentral Consortium, T., 2019. RNAcentral: a hub of information for non-coding RNA sequences. Nucleic Acids Research 47, D221. https://doi.org/10.1093/nar/gky1034

Roberts, R.B. (Richard B., 1958. Microsomal particles and protein synthesis; papers presented at the First Symposium of the Biophysical Society, at the Massachusetts Institute of Technology, Cambridge, February 5, 6, and 8, 1958. New York, Published on behalf of the Washington Academy of Sciences, Washington, D.C., by Pergamon Press.

Ronchetti, D., Mosca, L., Cutrona, G., Tuana, G., Gentile, M., Fabris, S., Agnelli, L., Ciceri, G., Matis, S., Massucco, C., Colombo, M., Reverberi, D., Recchia, A.G., Bossio, S., Negrini, M., Tassone, P., Morabito, F., Ferrarini, M., Neri, A., 2013. Small nucleolar RNAs as

new biomarkers in chronic lymphocytic leukemia. BMC Med Genomics 6, 27. https://doi.org/10.1186/1755-8794-6-27
Ronchetti, D., Todoerti, K., Tuana, G., Agnelli, L., Mosca, L., Lionetti, M., Fabris, S., Colapietro, P., Miozzo, M., Ferrarini, M., Tassone, P., Neri, A., 2012. The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. Blood Cancer J 2, e96. https://doi.org/10.1038/bcj.2012.41
Salzberg, S.L., 2019. Next-generation genome annotation: we still struggle to get it right. Genome Biology 20, 92. https://doi.org/10.1186/s13059-019-1715-2
Samarsky, D.A., Fournier, M.J., 1999. A comprehensive database for the small nucleolar RNAs from Saccharomyces cerevisiae. Nucleic Acids Res. 27, 161–164. https://doi.org/10.1093/nar/27.1.161
Saraiya, A.A., Wang, C.C., 2008. snoRNA, a Novel Precursor of microRNA in Giardia lamblia. PLoS Pathog 4. https://doi.org/10.1371/journal.ppat.1000224
Schimmang, T., Tollervey, D., Kern, H., Frank, R., Hurt, E.C., 1989. A yeast nucleolar protein related to mammalian fibrillarin is associated with small nucleolar RNA and is essential for viability. EMBO J 8, 4015–4024.
Schubert, T., Pusch, M.C., Diermeier, S., Benes, V., Kremmer, E., Imhof, A., Längst, G., 2012. Df31 Protein and snoRNAs Maintain Accessible Higher-Order Structures of Chromatin. Molecular Cell 48, 434–444. https://doi.org/10.1016/j.molcel.2012.08.021
Scott, M.S., Ono, M., 2011. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. Biochimie 93, 1987–1992. https://doi.org/10.1016/j.biochi.2011.05.026
Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W., Bruford, E.A., 2011. genenames.org: the HGNC resources in 2011. Nucleic Acids Res 39, D514–D519. https://doi.org/10.1093/nar/gkq892
Sharma, E., Sterne-Weiler, T., O'Hanlon, D., Blencowe, B.J., 2016. Global Mapping of Human RNA-RNA Interactions. Molecular Cell 62, 618–626. https://doi.org/10.1016/j.molcel.2016.04.030
Sharma, S., Yang, J., van Nues, R., Watzinger, P., Kötter, P., Lafontaine, D.L.J., Granneman, S., Entian, K.-D., 2017. Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation. PLoS Genet 13. https://doi.org/10.1371/journal.pgen.1006804
Shen, Y., Yu, X., Zhu, L., Li, T., Yan, Z., Guo, J., 2018. Transfer RNA-derived fragments and tRNA halves: biogenesis, biological functions and their roles in diseases. J Mol Med 96, 1167–1176. https://doi.org/10.1007/s00109-018-1693-y
Shevtsov, S.P., Dundr, M., 2011. Nucleation of nuclear bodies by RNA. Nature Cell Biology 13, 167–173. https://doi.org/10.1038/ncb2157
Shows, T.B., Alper, C.A., Bootsma, D., Dorf, M., Douglas, T., Huisman, T., Kit, S., Klinger, H.P., Kozak, C., Lalley, P.A., Lindsley, D., McAlpine, P.J., McDougall, J.K., Khan, P.M., Meisler, M., Morton, N.E., Opitz, J.M., Partridge, C.W., Payne, R., Roderick, T.H., Rubinstein, P., Ruddle, F.H., Shaw, M., Spranger, J.W., Weiss, K., 1979. International System for Human Gene Nomenclature (1979) ISGN (1979). CGR 25, 96–116. https://doi.org/10.1159/000131404
Steinbusch, M.M.F., Fang, Y., Milner, P.I., Clegg, P.D., Young, D.A., Welting, T.J.M., Peffers, M.J., 2017. Serum snoRNAs as biomarkers for joint ageing and post traumatic osteoarthritis. Scientific Reports 7, 43558. https://doi.org/10.1038/srep43558

Su, H., Xu, T., Ganapathy, S., Shadfan, M., Long, M., Huang, T.H.-M., Thompson, I., Yuan, Z.-M., 2014. Elevated snoRNA biogenesis is essential in breast cancer. Oncogene 33, 1348–1358. https://doi.org/10.1038/onc.2013.89

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., Jensen, L.J., Mering, C. von, 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 47, D607–D613. https://doi.org/10.1093/nar/gky1131

Trahan, C., Dragon, F., 2009. Dyskeratosis congenita mutations in the H/ACA domain of human telomerase RNA affect its assembly into a pre-RNP. RNA 15, 235–243. https://doi.org/10.1261/rna.1354009

Trinkle-Mulcahy, L., Sleeman, J.E., 2016. The Cajal body and the nucleolus: "In a relationship" or "It's complicated"? RNA Biol 14, 739–751. https://doi.org/10.1080/15476286.2016.1236169

Tycowski, K.T., Aab, A., Steitz, J.A., 2004. Guide RNAs with 5′ Caps and Novel Box C/D snoRNA-like Domains for Modification of snRNAs in Metazoa. Current Biology 14, 1985–1995. https://doi.org/10.1016/j.cub.2004.11.003

Ungaro, A., Pech, N., Martin, J.-F., McCairns, R.J.S., Mévy, J.-P., Chappaz, R., Gilles, A., 2017. Challenges and advances for transcriptome assembly in non-model species. PLOS ONE 12, e0185020. https://doi.org/10.1371/journal.pone.0185020

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M.L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N.N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor,

S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigó, R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., Zhu, X., 2001. The sequence of the human genome. Science 291, 1304–1351. https://doi.org/10.1126/science.1058040

Wain, H.M., Lush, M., Ducluzeau, F., Povey, S., 2002. Genew: the Human Gene Nomenclature Database. Nucleic Acids Res 30, 169–171.

Wang, X., Xu, M., Yan, Y., Kuang, Y., Li, P., Zheng, W., Liu, H., Jia, B., 2019. Identification of Eight Small Nucleolar RNAs as Survival Biomarkers and Their Clinical Significance in Gastric Cancer. Front. Oncol. 9. https://doi.org/10.3389/fonc.2019.00788

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57–63. https://doi.org/10.1038/nrg2484

Weber, A.P.M., 2015. Discovering New Biology through Sequencing of RNA1. Plant Physiol 169, 1524–1531. https://doi.org/10.1104/pp.15.01081

Weinberg, R.A., Penman, S., 1968. Small molecular weight monodisperse nuclear RNA. Journal of Molecular Biology 38, 289–304. https://doi.org/10.1016/0022-2836(68)90387-2

Welch, G.R., 1995. T.H. Huxley and the 'protoplasmic theory of life': 100 years later. Trends in Biochemical Sciences 20, 481–485. https://doi.org/10.1016/S0968-0004(00)89106-9

Williams, A.G., Thomas, S., Wyman, S.K., Holloway, A.K., 2014. RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis. Current Protocols in Human Genetics 83, 11.13.1-11.13.20. https://doi.org/10.1002/0471142905.hg1113s83

Williams, G.T., Farzaneh, F., 2012. Are snoRNAs and snoRNA host genes new players in cancer? Nature Reviews Cancer 12, 84–88. https://doi.org/10.1038/nrc3195

Wise, J.A., Tollervey, D., Maloney, D., Swerdlow, H., Dunn, E.J., Guthrie, C., 1983. Yeast contains small nuclear RNAs encoded by single copy genes. Cell 35, 743–751. https://doi.org/10.1016/0092-8674(83)90107-1

Wu, L., Zheng, J., Chen, P., Liu, Q., Yuan, Y., 2017. Small nucleolar RNA ACA11 promotes proliferation, migration and invasion in hepatocellular carcinoma by targeting the PI3K/AKT signaling pathway. Biomed. Pharmacother. 90, 705–712. https://doi.org/10.1016/j.biopha.2017.04.014

Xie, J., Zhang, M., Zhou, T., Hua, X., Tang, L., Wu, W., 2007. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. Nucleic Acids Res 35, D183–D187. https://doi.org/10.1093/nar/gkl873

Xue, S., Barna, M., 2012. Specialized ribosomes: a new frontier in gene regulation and organismal biology. Nature Reviews Molecular Cell Biology 13, 355–369. https://doi.org/10.1038/nrm3359

Yoshihama, M., Nakao, A., Kenmochi, N., 2013. snOPY: a small nucleolar RNA orthological gene database. BMC Res Notes 6, 426. https://doi.org/10.1186/1756-0500-6-426

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O.G., Janacek, S.H., Juettemann, T., To, J.K., Laird, M.R., Lavidas, I., Liu, Z., Loveland, J.E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D.N., Newman, V., Nuhn, M., Ogeh, D., Ong, C.K., Parker, A., Patricio, M., Riat, H.S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S.E., Kostadima, M., Langridge, N., Martin, F.J., Muffato, M., Perry, E., Ruffier, M., Staines, D.M., Trevanion, S.J., Aken, B.L., Cunningham, F., Yates, A., Flicek, P., 2018. Ensembl 2018. Nucleic Acids Research 46, D754. https://doi.org/10.1093/nar/gkx1098

Zhang, L., Lin, J., Ye, K., 2013. Structural and functional analysis of the U3 snoRNA binding protein Rrp9. RNA 19, 701–711. https://doi.org/10.1261/rna.037580.112

Zhang, Y.-Z., Wu, W.-C., Shi, M., Holmes, E.C., 2018. The diversity, evolution and origins of vertebrate RNA viruses. Current Opinion in Virology, Virus structure and expression • Viral evolution 31, 9–16. https://doi.org/10.1016/j.coviro.2018.07.017

Zhao, S., Zhang, Y., Gamini, R., Zhang, B., von Schack, D., 2018. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. Sci Rep 8. https://doi.org/10.1038/s41598-018-23226-4

Zhong, F., Zhou, N., Wu, K., Guo, Y., Tan, W., Zhang, Hong, Zhang, X., Geng, G., Pan, T., Luo, H., Zhang, Y., Xu, Z., Liu, J., Liu, B., Gao, W., Liu, C., Ren, L., Li, J., Zhou, J., Zhang, Hui, 2015. A SnoRNA-derived piRNA interacts with human interleukin-4 pre-mRNA and induces its decay in nuclear exosomes. Nucleic Acids Res 43, 10474–10491. https://doi.org/10.1093/nar/gkv954