SnoPatrol: How many snoRNA genes are there? – Supplementary materials.

Paul P. Gardner*[1], Alex G. Bateman[1] and Anthony M. Poole[2,3]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK
[2]Department of Molecular Biology & Functional Genomics, Stockholm University, SE-106 91 Stockholm, Sweden
[3]School of Biological Sciences, University of Canterbury, Christchurch 8140, New Zealand

Email: Paul P. Gardner*- pg5@sanger.ac.uk; Alex G. Bateman*- agb@sanger.ac.uk; Anthony M. Poole - anthony.poole@canterbury.ac.nz;

*Corresponding author

Supplementary Methods

In order to investigate the taxonomic distribution of the known snoRNAs and highlight where potential new discoveries can be made we have gathered data from the Pfam, Rfam, Genomes Online Database (GOLD) and EMBL databases. To locate the essential protein components of snoRNPs we used domains annotated by the Pfam database. We chose 6 snoRNP associated domains Nop, Nop10p, Gar1, SHQ1, Fibrillarin and TruB_N, the counts of which are shown in the blue bars in Figure 2. We also ran the control experiment of ensuring that the genomes for each taxa contain rRNAs by using the annotation for the partial SSU model from the Rfam database, the counts of which are shown in the green bars in Figure 2 for each taxa. These two datasets indicate that the snoRNP machinery is present in all the known Archaea and Eukaryotic taxa, the notable exception being the absence of snoRNP protein domains in the Oömycetes, stramenopiles group. However, it appears that this lack is due to the protein sequences not being included in the public sequence databases yet rather than a bona fide vestigialization of this machinery. Yet for many major taxonomic clades there are few or no known snoRNAs. To investigate these we used species counts of the snoRNAs annotated by the Rfam database the results of these are displayed in the red bars in Figure 2. Of these there are many experimentally validated snoRNA sequences published in the literature and therefore submitted to one of the nucleotide archives Genbank, EMBL or the DDBJ, for further verification we mined the annotation of those sequences in the Rfam database for terms associated with snoRNAs, the counts of these are shown in the pink bars in Figure 2. Finally, in order to be certain that

the counts we observed aren't due to an under-sampling of sequences from certain taxonomic groupings we used data from the GOLD genomes database, genomes annotated as either "Complete", "Draft" or "In progress"; assuming these genome sequences have all been submitted to the public nucleotide archives (the annotation of which is included in the UniProt database searched by Pfam) then this should give a good representation of where sequences are available and therefore annotation of snoRNAs possible.

Supplementary Table 1

- Col. 1: taxon string (top three levels of the NCBI taxonomy)
- Col. 2: number of sequence regions annotated by the Pfam Nop, Nop10p and Gar1 families
- Col. 3: number of sequence regions annotated by the Rfam SSU rRNA family
- Col. 4: number of genomes annotated as either completed, draft or in progress from the GOLD database
- Col. 5: number of sequence regions annotated by all the Rfam snoRNA families
- Col. 6: number of sequence regions annotated by all the Rfam snoRNA families where EMBL sequences have been published as snoRNAs.

TaxString	Pfam	SSU	GOLD	predSnos	knownSnos
Archaea;Crenarchaeota;Thermoprotei;	157	475	27	128	11
Archaea;Euryarchaeota;Aciduliprofundum.	10	3	1	2	0
Archaea;Euryarchaeota;Archaeoglobi;	6	43	1	8	3
Archaea;Euryarchaeota;Halobacteria;	59	1526	15	0	0
Archaea;Euryarchaeota;Methanobacteria;	23	1486	5	0	0
Archaea;Euryarchaeota;Methanococci;	45	107	6	26	0
Archaea;Euryarchaeota;Methanomicrobia;	64	1383	11	10	0
Archaea;Euryarchaeota;Methanopyri;	6	5	1	1	0
Archaea;Euryarchaeota;Thermococci;	48	300	7	324	19
Archaea;Euryarchaeota;Thermoplasmata;	14	220	3	0	0
Archaea;Korarchaeota;Candidatus	3	1	1	0	0
Archaea;Nanoarchaeota;Nanoarchaeum.	3	2	1	0	0
Archaea;Thaumarchaeota;Cenarchaeales;	4	0	0	0	0
Archaea;Thaumarchaeota;marine	18	0	0	0	0
Eukaryota;Alveolata;Apicomplexa;	132	2675	25	58	1
Eukaryota;Alveolata;Ciliophora;	28	1438	1	10	0
Eukaryota;Alveolata;Perkinsea;	20	175	1	29	0
Eukaryota;Amoebozoa;Archamoebae;	16	326	3	3	0
Eukaryota;Amoebozoa;Mycetozoa;	13	239	1	2	0
Eukaryota;Amoebozoa;Tubulinea;	1	84	0	0	0
Eukaryota;Choanoflagellida;Codonosigidae;	9	18	1	6	0
Eukaryota;Cryptophyta;Pyrenomonadales;	3	119	1	0	0

Eukaryota;Diplomonadida;Hexamitidae;	9	90	2	0	0
Eukaryota;Euglenozoa;Euglenida;	4	458	0	0	0
Eukaryota;Euglenozoa;Kinetoplastida;	70	669	3	174	26
Eukaryota;Fungi;Dikarya;	546	11898	154	3591	181
Eukaryota;Fungi;Microsporidia;	18	540	2	0	0
Eukaryota;Metazoa;Annelida;	1	1078	1	0	0
Eukaryota;Metazoa;Arthropoda;	208	10738	40	2032	120
Eukaryota;Metazoa;Brachiopoda;	1	74	0	0	0
Eukaryota;Metazoa;Chordata;	188	17465	90	85287	952
Eukaryota;Metazoa;Cnidaria;	9	1167	2	144	0
Eukaryota;Metazoa;Echinodermata;	2	254	1	168	0
Eukaryota;Metazoa;Echiura;	1	6	0	0	0
Eukaryota;Metazoa;Mollusca;	1	2045	1	140	0
Eukaryota;Metazoa;Nematoda;	31	2420	12	304	49
Eukaryota;Metazoa;Nematomorpha;	1	19	0	0	0
Eukaryota;Metazoa;Placozoa;	9	15	1	19	0
Eukaryota;Metazoa;Platyhelminthes;	22	1521	2	38	0
Eukaryota;Metazoa;Porifera;	1	281	0	0	0
Eukaryota;Metazoa;Priapulida;	1	23	0	0	0
Eukaryota;Metazoa;Sipuncula;	1	104	0	0	0
Eukaryota;Parabasalidea;Trichomonada;	12	350	1	1	0
Eukaryota;Rhizaria;Cercozoa;	3	866	1	0	0
Eukaryota;Viridiplantae;Chlorophyta;	48	2288	6	73	20
Eukaryota;Viridiplantae;Streptophyta;	194	5791	44	15529	680
Eukaryota;stramenopiles;Bacillariophyta;	21	1125	2	9	0
Eukaryota;stramenopiles;Oomycetes;	0	261	4	75	0

Supplementary Methods

The Trichomonas vaginalis G3 C/D box snoRNA (RF00335, Small nucleolar RNA Z13/snr52) discussed in the text of the article.

The results of the cmsearch [1] match between the RF00335 model and the T. vaginalis snoRNA candidate:

```
>AAHC01000432.1/36420-36361

 Query = 1 - 101, Target = 76 - 135
 Score = 26.12, E = 104.1, P = 3.938e-09, GC =  33

           ::<<<-----------------..-<<<<~~~~~~>>>>-------------------->>
         1 uucucuaaaAUGAUGAAAUAUU..uCGGA*[47]*UCCGUUGCCUUCCUUCUGAUuaaga 99
            +C::U+A AUGAUG A  A U   :GGA      UCC:UUGCCUUCCUU UGAUU++::
        76 GACAAUUAGAUGAUGCAUAAAUgaGUGGA*[ 5]*UCCAUUGCCUUCCUU-UGAUUUUUU 133

           >:
       100 gu 101
           GU
       134 GU 135
```

A Stockholm format alignment showing the alignment between the known Fungal snR52/Z13 snoRNAs and the T.vaginalis snoRNA candidate:

```
# STOCKHOLM 1.0

#=GF AC   RF00335
#=GF ID   snoZ13_snr52
#=GF DE   Small nucleolar RNA Z13/snr52
#=GF AU   Moxon SJ
#=GF SE   Moxon SJ, INFERNAL
#=GF SS   Predicted; PFOLD; Moxon SJ
#=GF GA   20.16
#=GF TC   22.70
#=GF NC   undefined
#=GF TP   Gene; snRNA; snoRNA; CD-box;
#=GF BM   cmbuild  -F CM SEED; cmcalibrate --mpi -s 1 CM
#=GF BM   cmsearch  -Z 169604 -E 1000  --toponly  CM SEQDB
#=GF DR   SO:0000593 SO:C_D_box_snoRNA
#=GF DR   GO:0006396 GO:RNA processing
#=GF DR   GO:0005730 GO:nucleolus
#=GF RN   [1]
#=GF RM   12007400
#=GF RT   Small nucleolar RNAs: an abundant group of noncoding RNAs with
#=GF RT   diverse cellular functions.
#=GF RA   Kiss T;
#=GF RL   Cell 2002;109:145-148.
#=GF CC   Z13 or snr52 is a member of the C/D class of snoRNA which contain the C
#=GF CC   (UGAUGA) and  D (CUGA) box motifs. Most of the members of the box C/D
#=GF CC   family function in directing site-specific 2'-O-methylation of substrate
#=GF CC   RNAs [1].
#=GF WK   http://en.wikipedia.org/wiki/Small_nucleolar_RNA_Z13/snr52

#=GS S.pombe_Z13_snoRNA        CC  AJ223843.1/1-95,            score=89.65   E value=3.11e-16
#=GS S.pombe_snr52_snoRNA      CC  AJ251862.1/1-95,            score=89.65   E value=3.11e-16
#=GS N.crassa_CD42_snoRNA      CC  EU780965.1/2-92,            score=79.46   E value=2.01e-13
#=GS A.fumigatus_Af293_snoRNA  CC  AM921928.1/6-91,            score=79.30   E value=2.22e-13
#=GS S.pombe_snr52_snoRNA2     CC  AJ617322.1/3-83,            score=67.73   E value=3.46e-10
#=GS T.vaginalis_snoRNA        CC  AAHC01000432.1/36420-36361, score=26.12   E value=1.04e+02
#=GS S.cerevisiae_Z13_snoRNA   CC  AJ223033.1/15-106,          score=9.86    E value=3.17e+06
#=GS S.cerevisiae_snR52_snoRNA CC  AF064264.1/1-92,            score=9.86    E value=3.17e+06

#=GS S.pombe_Z13_snoRNA        CC  S.pombe      Schizosaccharomyces pombe Z13 small nucleolar RNA ge
#=GS S.pombe_snr52_snoRNA      CC  S.pombe      Schizosaccharomyces pombe snr52 gene for small nucle
#=GS N.crassa_CD42_snoRNA      CC  N.crassa     Neurospora crassa box C/D snoRNA CD42, complete sequ
#=GS A.fumigatus_Af293_snoRNA  CC  A.fumigatus  Aspergillus fumigatus Af293 putative C/D box snoRNA
#=GS S.pombe_snr52_snoRNA2     CC  S.pombe      Schizosaccharomyces pombe snR52 snoRNA
#=GS T.vaginalis_snoRNA        CC  T.vaginalis  Trichomonas vaginalis G3, whole genome shotgun seque
#=GS S.cerevisiae_Z13_snoRNA   CC  S.cerevisiae Saccharomyces cerevisiae Z13 small nucleolar RNA gen
#=GS S.cerevisiae_snR52_snoRNA CC  S.cerevisiae Saccharomyces cerevisiae snR52 small nucleolar RNA,
```

```
S.pombe_Z13_snoRNA                AUUUUGAAAAUGAUGAAAAU..AA.cGCGGAUGAAAUAUAU-------GU
S.pombe_snr52_snoRNA              AUUUUGAAAAUGAUGAAAAU..AA.cGCGGAUGAAAUAUAU-------GU
N.crassa_CD42_snoRNA              AC------CCUGAUGAAAUA..UU.cUCGGAUGUAAAAUUUUACUUUUGU
A.fumigatus_Af293_snoRNA          ----------UGAUGAAAUA..UU..UCGGAUGUAAACUCACAGACCUGU
S.pombe_snr52_snoRNA2             ------AAAAUGAUGAAAAU..AA.cGCGGAUGAAAUAUAU-------GU
T.vaginalis_snoRNA                GACAAUUAGAUGAUGCAUAA..AUgaGUGGA--------------------
S.cerevisiae_Z13_snoRNA           UA-----CUAUGAUGAAUGAcaUU.aGCGUGAACAAUCUCUGAUACAAAA
S.cerevisiae_snR52_snoRNA         UA-----CUAUGAUGAAUGAcaUU.aGCGUGAACAAUCUCUGAUACAAAA
#=GC SS_cons                      ::<<<---------------..--..-<<<<_____________________
#=GC MT_cons                      ..........CCCCCC..........|<25S_rRNA>|............

S.pombe_Z13_snoRNA                CCGAGA.G.............CGCAAAAAAUGUGUGGAUAUUUCCGUUGC
S.pombe_snr52_snoRNA              CCGAGA.G.............CGCAAAAAAUGUGUGGAUAUUUCCGUUGC
N.crassa_CD42_snoRNA              CUGAAAaG.............CGCAAAAACAAUGUUGAGAUUUCCGUUGC
A.fumigatus_Af293_snoRNA          UCUGAAuG.............CGCAAAACC-GUGUAGAGAUUUCCGUUGC
S.pombe_snr52_snoRNA2             CCGAGA.G.............CGCAAAAAAUGUGUGGAUAUUUCCGUUGC
T.vaginalis_snoRNA                ------.g.............................uauuUCCAUUGC
S.cerevisiae_Z13_snoRNA           UCGAAA.Gauuuuaggauuag...aaaa..........acuUAUGUUGC
S.cerevisiae_snR52_snoRNA         UCGAAA.Gauuuuaggauuag...aaaa..........acuUAUGUUGC
#=GC SS_cons                      ______..........._...______________________>>>>----
#=GC MT_cons                      ..........................................|<18S_

S.pombe_Z13_snoRNA                CUUCCUUCUGAUUCAAAAU
S.pombe_snr52_snoRNA              CUUCCUUCUGAUUCAAAAU
N.crassa_CD42_snoRNA              CUUCCUUCUGAUC------
A.fumigatus_Af293_snoRNA          CUUCCUUCUGAUGA-----
S.pombe_snr52_snoRNA2             CUUCCUUCUGA--------
T.vaginalis_snoRNA                CUUCCUU-UGAUUUUUUGU
S.cerevisiae_Z13_snoRNA           CUUCCUUCUGAA--A----
S.cerevisiae_snR52_snoRNA         CUUCCUUCUGAA--A----
#=GC SS_cons                      ---------------->>>:
#=GC MT_cons                      rRNA.>|DDDD........
//
```

References

1. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**(10):1335–7.