The limits of human microRNA annotation have been met

BASTIAN FROMM,[1] XIANGFU ZHONG,[2] MARCEL TARBIER,[3] MARC R. FRIEDLÄNDER,[4] and MICHAEL HACKENBERG[5,6,7]

[1]The Arctic University Museum of Norway, UiT-The Arctic University of Norway, 9006 Tromsø, Norway
[2]Department of Biosciences and Nutrition, Karolinska Institute, 14183 Huddinge, Sweden
[3]Science for Life Laboratory, Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, 17165 Solna, Sweden
[4]Science for Life Laboratory, Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, 10691 Stockholm, Sweden
[5]Department of Genetics, Faculty of Sciences, MNAT Excellence Unit, University of Granada, 18071 Granada, Spain
[6]Biotechnology Institute, CIBM, 18100 Armilla (Granada), Spain
[7]Biohealth Research Institute (ibs. GRANADA), University Hospitals of Granada, University of Granada, 18014 Granada, Spain

ABSTRACT

Over the last few years, the number of microRNAs in the human genome has become a controversially debated issue. Several publications reported thousands of putative novel microRNAs not included in the curated microRNA gene database MirGeneDB and the repository miRBase. Recently, by using sequencing of ∼300 human tissues and cell lines, the human RNA atlas, an expanded inventory of human RNA annotations, was published, reporting thousands of putative microRNAs. We, the developers of established microRNA prediction tools and hosts of MirGeneDB, raise concerns about the frequently applied prediction and functional validation strategies, briefly discussing the drawbacks of false positive detections. By means of quantifying well-established biogenesis-derived features, we show that the reported novel microRNAs essentially represent false-positives and argue that the human microRNA complement, at about 550 microRNA genes, is already near complete. Output of available tools must be curated as false predictions will misguide scientists looking for biomarkers or therapeutic targets.

Keywords: MirGeneDB; annotation; miRBase; microRNAs; ncRNA prediction

Since the discovery of the first human microRNAs in 2000 (Pasquinelli et al. 2000) and 2001 (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001), respectively, microRNAs have taken a center stage in human biology, medical research and, in particular, as biomarkers (Wang et al. 2016). Because the microRNA field substantially expanded in the last years, reaching a record ∼16,000 annual publications in 2021 (Kilikevicius et al. 2022), thousands of novel microRNA candidates have continuously been deposited in the microRNA candidate repository miRBase, expanding near proportionally to the number of publications for more than a decade (Fig. 1A; Supplemental File 1). Driven mainly by the advent of next-generation sequencing (NGS) and the development of bioinformatics prediction tools with flexible stringency thresholds (Friedländer et al. 2008; Hackenberg et al. 2009), this expansion in human microRNAs created an imbalance of microRNA complements between closely related species (e.g., human and macaque) that appeared improbable to many researchers given the high conservation and shared biological functions of microRNAs in animals (Fromm et al. 2020).

While it is difficult to define annotation criteria for many RNA classes such as lincRNAs or piRNAs, structure and sequence features of microRNAs are distinct and have been used to develop a system for their annotation already in 2003 (Ambros et al. 2003). However, because these clear and mechanistically well understood features are not all easily implementable in computational prediction (e.g., evolutionary conservation), there is a risk that the number of false positive annotations increases as the community analyzes more data.

With the introduction of stricter criteria for the inclusion of new microRNAs in 2014, miRBase stopped the microRNA expansion for human at a large ∼1900 putative

Corresponding author: Bastian.Fromm@uit.no

Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.079098.122.

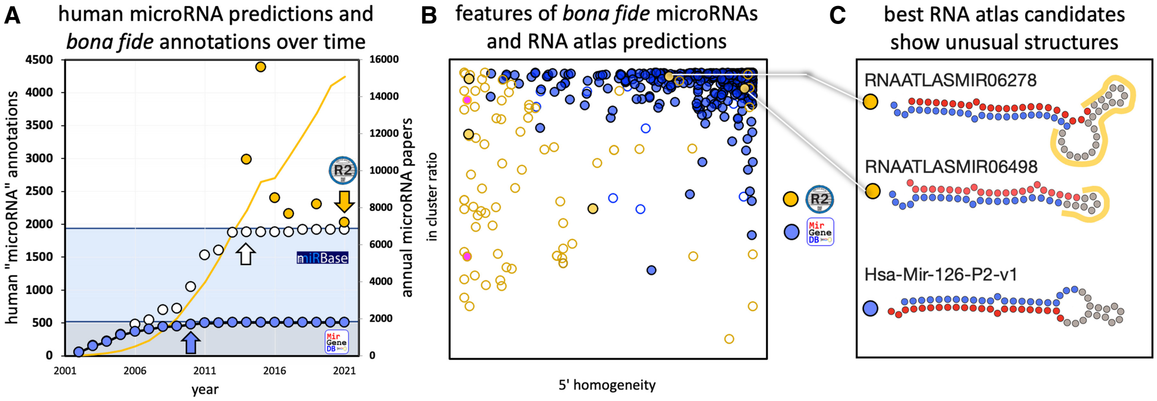

FIGURE 1. (*A*) microRNA predictions in miRBase (white), MirGeneDB (blue), and in selected papers (yellow) over time in comparison to the number of annual papers (yellow line) (see Supplemental File 1). A clear near correlation of the number of published papers and annotated microRNAs is detected until 2014 (white arrow, $R^2 = 0.876$), when miRBase introduced high-confidence criteria. The human microRNA complement annotated in miRBase was since then stable at ∼1900 microRNA candidates (light blue area). Note that the curated human microRNA complement (∼500 bona fide microRNAs from MirGeneDB [gray area]) is recovered at 95.2% (480 genes) already by 2010 in miRBase (blue arrow). Several studies predicted substantial numbers of putative novel microRNAs that differ significantly from the putative microRNAs in miRBase and the bona fide microRNAs of MirGeneDB, respectively (yellow dots, see Supplemental File 1; not shown, miRCarta [Backes et al. 2018] with ∼12,000 predictions). Yellow arrow depicts the corresponding numbers for "validated" microRNAs in the RNA atlas from Lorenzi. (*B*) Comparison of human bona fide microRNA genes as annotated in MirGeneDB (blue) and the novel "validated" microRNA candidates from the RNA atlas (yellow) by mapping of all RNA atlas reads to all precursors (see Supplemental File 2 for summary, Supplemental File 3 for detailed results). Three important annotation criteria based on microRNA biogenesis features were tested: "5′ homogeneity" (*x*-axis, bona fide microRNAs typically show values beyond 90%), "in cluster ratio" (*y*-axis, bona fide microRNAs typically show values beyond 90%), and the detection of two precursor arm products that show characteristic RNases derived 2 nt offsets (empty circles: one arm detected, filled circles: two arms detected). Note the few empty circles for MirGeneDB entries that represent either deeply conserved (Hsa-Mir-1307, eutherian), generally lowly expressed (Hsa-Mir-5579), or extremely tissue-specific entries (Hsa-Mir-217, pancreas specific and also deeply conserved) that are detected with both arms in the MirGeneDB data. Pink RNA atlas predictions represent fragments of a previously annotated Y-RNA fragment (R4). (*C*) The two best RNA atlas candidates as measured by the 3′ feature comparison show strong secondary loop structures (RNAATLASMIR06278), or unusually short loop sizes (RNAATLASMIR06498) in comparison to a representative example from MirGeneDB (Hsa-Mir-126-P2-v1) (red dots mark mature microRNA nucleotides and blue dots star microRNA nucleotides; unusual loop regions are highlighted in yellow).

microRNAs (Fig. 1A, white arrow), but kept more than a thousand *low-confidence* annotations among them (total 295 high and 1586 low confidence) (Kozomara and Griffiths-Jones 2014). Therefore, in 2015, some of us manually curated and reannotated the human microRNA complement by establishing and applying refined criteria for annotation of microRNAs with NGS data (Fromm et al. 2015). Confirming previous experimental work in mouse microRNAs (Chiang et al. 2010), we discovered that two-thirds of human entries were false-positive entries and obtained ∼500 bona fide microRNA genes now hosted in the manually curated microRNA gene database MirGeneDB (Fromm et al. 2015, 2022). When checking previous releases of miRBase, 99% of all bona fide human microRNA genes were already contained in miRBase a decade ago (Fig. 1A), and, with few exceptions, all human microRNAs conserved at least to fish were indeed already found 15 yr ago (Supplemental File 1, "fish human microRNAs in miRBase"; Bartel 2018). Most importantly, despite a range of publications claiming substantial numbers of novel microRNAs in human (Fig. 1A; Jha et al. 2015; Londin et al. 2015; Alles et al. 2019), miRBase and MirGeneDB have only added a few microRNAs to their human complements in the last ∼8 yr, respectively. This held true with the recent release of the telomere-to-telomere assembly of the human genome, where no new microRNAs were found in the previously unassembled regions (Patil et al. 2021).

The idea of creating and updating compendia for gene-products, including noncoding genes such as microRNAs, based on the most recent technological advances is a great service for the scientific community when shared as an online community resource such as the RNAatlas project (Lorenzi et al. 2021). In their "resource paper" Lorenzi et al. used RNA sequencing of small and poly(A) RNA, as well as total RNA, from ∼300 human tissues and cell lines, including cancer cells, to describe at a "more comprehensive atlas of the human transcriptome" including many coding and noncoding RNAs, such as microRNAs. For the prediction of novel microRNAs they used our software miRDeep2 (Friedländer et al. 2012). However, this was done in default settings and without applying the recommended score cut-offs. miRDeep2 output includes numerous statistics, and the score cut-off should be chosen so that the signal-to-noise ratio is above 10:1. Not applying this threshold and not curating the output yielded a staggering number of 3567 novel microRNA predictions (Lorenzi et al. 2021). These were further filtered not by applying well-established annotation criteria (see above, Fromm et al. 2015), but by using "*ncRNA target inference algorithms*" to identify negative correlations with mRNA to pre-mRNA ratios, obtaining 111 putatively novel microRNAs with inferred targets. While profiling of naïve transcription as compared to overall mRNA abundances has previously been used to distinguish transcriptional and post-transcriptional effects (Tarbier et al. 2020), and anti-correlation of microRNA and overall target expression has been shown to be detectable that is, in single cells (Nielsen and Pedersen 2021), these approaches have their own short-comings and cannot reliably confirm individual microRNA-target interactions. The presence of a few individual pairs showing negative coexpression between microRNA candidate levels and either mRNA levels or proxies for post-transcriptional regulation is to be expected by chance and should not be considered *validations* for direct microRNA action on its targets and even less as a proof for microRNA-functionality in general. Other approaches such as physiological alterations of microRNA levels, or genetic editing of microRNA target-sites are more likely to capture direct effects (see Huberdeau and Simard 2019; Kilikevicius et al. 2022); however, it is important to note that in either case a functional effect neither validates, nor does its absence disprove microRNA identity.

Because even these 111 filtered novel human microRNAs, as proposed by Lorenzi et al. seemed surprising to us, we systematically tested them for three well-established bona fide microRNA annotation criteria: (A) The "in cluster ratio" which we defined as the proportion of small RNA sequencing reads mapping to the mature microRNA relative to all other reads mapping to a precursor candidate + 10 nt flanking nucleotides (0%–100%), (B) "5′ homogeneity" which is defined as the fraction of mature reads starting at the same 5′ position (0%–100%), and (C) whether reads from both arms of the precursor are detected that show characteristic 2 nt overhang (for review, see Ambros et al. 2003; Fromm et al. 2015; Bartel 2018 on features). To have a comparable and quantitative baseline for these analyses, we took advantage of the manually curated human microRNA complement in MirGeneDB (510 genes excluding variant and noncanonical annotations) and profiled these bona fide microRNAs in all 298 RNA atlas samples, as well.

The results are summarized in Figure 1B (Supplemental Files 2, 3). Briefly, there is an obvious difference between bona fide annotations (MirGeneDB, blue dots) and the RNA atlas predictions. Only two RNA atlas candidates fulfil all three criteria for microRNA annotation (RNAATLASMIR06498 and RNAATLASMIR06278), but all other candidates show either only the expression of one arm, or below average values for 5′ homogeneity and in cluster ratio. However, the two candidates show extremely low expression values (below 5 reads/sample, not shown)—far below what is widely accepted as a biological relevant level (Witwer and Halushka 2016)—and aberrant secondary structures (Fig. 1C): with 51 nt in length, RNAATLASMIR06498 is shorter than any known canonical human microRNA (smallest known are 52 nt in length Mir-374-P1/P2) and with a hitherto not reported short loop of 7 nt. RNAATLASMIR06278 shows exceptional secondary structures in the loop, deviating from the majority of canonical hairpins and warranting further validations (such as the behavior in Drosha- or Dicer-knockdown data [Kim et al. 2016]). In terms of novelty of these rather unlikely microRNA candidates, Lorenzi identified RNAATLASMIR06278 as being annotated in miRCarta, the repository of possible small noncoding microRNA candidates (Backes et al. 2018) (ID hsa-2734-4932.1), but missed that also RNAATLASMIR06498 is found there (ID hsa-5546.1), and thus neither should be seen as *novel* microRNA candidates.

At first glance it might seem advantageous to include more annotations when looking for biomarkers or therapeutic targets, regardless whether or not they are riddled with false-positives. However, poorly curated microRNA reference and unclear criteria for microRNA annotation can cause substantial downstream issues. For instance, when using a non bona fide microRNA in a model organism as template for the search for partially homologous sequences with supposed biological functions in human (Blanco-Domínguez et al. 2021); when conducting sequence motif searches on a subset of supposed microRNAs that include other RNA fragments (Garcia-Martin et al. 2022); or when interpreting the functions of a snoRNA in the light of microRNA biology (Chinnappa et al. 2022). Further, because small RNA bulk sequencing provides only relative abundance and therefore requires data normalization of expression values, interpretations will be affected by false positives, especially when they are highly expressed (Hamaguchi et al. 2021). Among the RNA atlas candidates are fragments of HY4, a known Y-RNA (Fig. 1B, pink), which show a fivefold higher mean expression than any other candidate. Y-RNAs do not enter the microRNA pathway and therefore can be considered clear false positives (Nicolas et al. 2012). Application of expensive and time-consuming techniques downstream will be in vain when assuming these fragments behave like microRNAs and, thus, the use of a microRNA reference essentially free of false positives is highly recommended for profiling and mechanistic experiments.

In summary, we show that Lorenzi et al., like many other studies before, report large numbers of putatively novel microRNAs which do not represent bona fide human microRNAs, and hence do not advance the field. Prediction tools were developed many years ago when only few data were available and sensitivity for microRNA candidates, but not specificity to bona fide genes was their focus. Because the total number of false-positives increases with the amount of input data, three conclusions can be drawn: First, the output of prediction tools such as miRDeep2 and sRNAbench must not directly be interpreted as novel microRNAs, but instead require stringent interpretation in terms of score cut-off and also manual curation with considerations of conservation, microRNA biogenesis and structural features. Second, for the future, prediction tools need to be adapted to large multitissue input data sets to reduce the number of false positives, thus lowering the downstream manual curation effort; and third, not the repeated resequencing of similar organs, tissues and cell-types, including cancer samples with aberrant expression profiles, but the sequencing of rarely analyzed biological samples or developmental time points might be the source of few additional, hitherto undiscovered microRNAs in human.

We therefore strongly advocate for the careful use of microRNA prediction software for de novo discovery of microRNAs in any organism and reinforce our claim that the human microRNA complement, with ∼550 microRNA genes, is quasi complete.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

The authors thank Marc Halushka and Kevin J. Peterson for discussions. B.F. is supported by the Tromsø forskningsstiftelse (TFS) (20_SG_BF "MIRevolution"). M.H. is supported by the Spanish Government (AGL2017-88702-C2-2-R); University of Granada (A-BIO-481-UGR18, FEDER 18).

Received January 3, 2022; accepted February 12, 2022.

REFERENCES

Alles J, Fehlmann T, Fischer U, Backes C, Galata V, Minet M, Hart M, Abu-Halima M, Grässer FA, Lenhof H-P, et al. 2019. An estimate of the total number of true human miRNAs. *Nucleic Acids Res* **47:** 3353–3364. doi:10.1093/nar/gkz097

Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277–279. doi:10.1261/rna.2183803

Backes C, Fehlmann T, Kern F, Kehl T, Lenhof H-P, Meese E, Keller A. 2018. miRCarta: a central repository for collecting miRNA candidates. *Nucleic Acids Res* **46:** D160–D167. doi:10.1093/nar/gkx851

Bartel DP. 2018. Metazoan microRNAs. *Cell* **173:** 20–51. doi:10.1016/j.cell.2018.03.006

Blanco-Domínguez R, Sánchez-Díaz R, de la Fuente H, Jiménez-Borreguero LJ, Matesanz-Marín A, Relaño M, Jiménez-Alejandre R, Linillos-Pradillo B, Tsilingiri K, Martín-Mariscal ML, et al. 2021. A novel circulating microRNA for the detection of acute myocarditis. *N Engl J Med* **384:** 2014–2027. doi:10.1056/NEJMoa2003608

Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, et al. 2010. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24:** 992–1009. doi:10.1101/gad.1884710

Chinnappa K, Cárdenas A, Prieto-Colomina A, Villalba A, Márquez-Galera Á, Soler R, Nomura Y, Llorens E, Tomasello U, López-Atalaya JP, et al. 2022. Secondary loss of miR-3607 reduced cortical progenitor amplification during rodent evolution. *Sci Adv* **8:** eabj4010. doi:10.1126/sciadv.abj4010

Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26:** 407–415. doi:10.1038/nbt1394

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40:** 37–52. doi:10.1093/nar/gkr688

Fromm B, Billipp T, Peck LE, Johansen M, Tarver JE, King BL, Newcomb JM, Sempere LF, Flatmark K, Hovig E, et al. 2015. A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annu Rev Genet* **49:** 213–242. doi:10.1146/annurev-genet-120213-092023

Fromm B, Domanska D, Høye E, Ovchinnikov V, Kang W, Aparicio-Puerta E, Johansen M, Flatmark K, Mathelier A, Hovig E, et al. 2020. MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res* **48:** D132–D141. doi:10.1093/nar/gkz885

Fromm B, Høye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, Umu SU, Chabot PJ, Kang W, Aslanzadeh M, et al. 2022. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. *Nucleic Acids Res* **50:** D204–D210. doi:10.1093/nar/gkab1101

Garcia-Martin R, Wang G, Brandão BB, Zanotto TM, Shah S, Kumar Patel S, Schilling B, Kahn CR. 2022. MicroRNA sequence codes for small extracellular vesicle release and cellular retention. *Nature* **601:** 446–451. doi:10.1038/s41586-021-04234-3

Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM. 2009. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* **37:** W68–W76. doi:10.1093/nar/gkp347

Hamaguchi Y, Zeng C, Hamada M. 2021. Impact of human gene annotations on RNA-seq differential expression analysis. *BMC Genomics* **22:** 730. doi:10.1186/s12864-021-08038-7

Huberdeau MQ, Simard MJ. 2019. A guide to microRNA-mediated gene silencing. *FEBS J* **286:** 642–652. doi:10.1111/febs.14666

Jha A, Panzade G, Pandey R, Shankar R. 2015. A legion of potential regulatory sRNAs exists beyond the typical microRNAs microcosm. *Nucleic Acids Res* **43:** 8713–8724. doi:10.1093/nar/gkv871

Kilikevicius A, Meister G, Corey DR. 2022. Reexamining assumptions about miRNA-guided gene silencing. *Nucleic Acids Res* **50:** 617–634. doi:10.1093/nar/gkab1256

Kim Y-K, Kim B, Narry Kim V. 2016. Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. *Proc Natl Acad Sci* **113:** E1881–E1889. doi:10.1073/pnas.1602532113

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42:** D68–D73. doi:10.1093/nar/gkt1181

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858. doi:10.1126/science.1064921

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862. doi:10.1126/science.1065062

Lee RC, Ambros V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864. doi:10.1126/science.1065329

Londin E, Loher P, Telonis AG, Quann K, Clark P, Jing Y, Hatzimichael E, Kirino Y, Honda S, Lally M, et al. 2015. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc Natl Acad Sci* **112:** E1106–E1115. doi:10.1073/pnas.1420955112

Lorenzi L, Chiu H-S, Cobos FA, Gross S, Volders P-J, Cannoodt R, Nuytens J, Vanderheyden K, Anckaert J, Lefever S, et al. 2021. The RNA Atlas expands the catalog of human non-coding RNAs. *Nat Biotechnol* **39:** 1453–1465. doi:10.1038/s41587-021-00936-1

Nicolas FE, Hall AE, Csorba T, Turnbull C, Dalmay T. 2012. Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS Lett* **586:** 1226–1230. doi:10.1016/j.febslet.2012.03.026

Nielsen MM, Pedersen JS. 2021. miRNA activity inferred from single cell mRNA expression. *Sci Rep* **11:** 9170. doi:10.1038/s41598-021-88480-5

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, Maller B, Hayward DC, Ball EE, Degnan B, Müller P, et al. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408:** 86–89. doi:10.1038/35040556

Patil AH, Halushka MK, Fromm B. 2021. No evidence of paralogous loci or new bona fide microRNAs in telomere to telomere (T2T) genomic data. *bioRxiv* doi:10.1101/2021.12.09.471935

Tarbier M, Mackowiak SD, Frade J, Catuara-Solarz S, Biryukova I, Gelali E, Menéndez DB, Zapata L, Ossowski S, Bienko M, et al. 2020. Nuclear gene proximity and protein interactions shape transcript covariations in mammalian single cells. *Nat Commun* **11:** 5445. doi:10.1038/s41467-020-19011-5

Wang J, Chen J, Sen S. 2016. MicroRNA as biomarkers and diagnostics. *J Cell Physiol* **231:** 25–30. doi:10.1002/jcp.25056

Witwer KW, Halushka MK. 2016. Toward the promise of microRNAs: enhancing reproducibility and rigor in microRNA research. *RNA Biol* **13:** 1103–1116. doi:10.1080/15476286.2016.1236172