Institution: The University of Manchester

Unit of Assessment: 5 (Biological Sciences)

Title of case study: The miRBase microRNA database – driving the development of commercial microRNA research tools, diagnostics and therapeutics

Period when the underpinning research was undertaken: 2007 - 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Sam Griffiths-Jones	Professor of Computational Biology Senior Lecturer Research Fellow	Aug 2016 – present Sept 2010 – July 2016 Jan 2007 – Aug 2010
Ana Kozomara	Research Associate	Dec 2009 – Jan 2019
Maria Birgaoanu	Research Assistant	Nov 2018 – present

Period when the claimed impact occurred: August 2013 – July 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

miRBase is a free online database of microRNA genes, developed and maintained by University of Manchester researchers. miRBase is the sole authority of what is and is not a microRNA, and therefore impacts **all** microRNA research worldwide, including in commercial and clinical sectors. The website attracts 20,000-30,000 monthly users, approximately 10% of whom are from commercial companies. miRBase data underpins commercially-successful research tools, diagnostics and therapeutics. More than 15 companies worldwide sell microRNA resources built from miRBase sequences. Novel therapeutics based on microRNAs defined by miRBase are in development and clinical trial. International genomic policy groups adopt miRBase standards on microRNAs.

2. Underpinning research

MicroRNAs are non-protein-coding RNA genes that regulate the expression of most protein-coding genes in all animal and plant genomes. Their prevalence was unknown before 2001. The growth of the microRNA field has been extraordinary: approximately 90% of known microRNAs have been discovered since 2007. There is therefore a strong ongoing need for an authoritative group to name newly discovered microRNA genes, and provide data and analysis tools to the community. The miRBase database (http://mirbase.org/) was established by Sam Griffiths-Jones in 2003 (then called the miRNA Registry) at the Wellcome Trust Sanger Institute; first major publication in 2004. All research and development of the miRBase resource since 2007 has been conducted by Griffiths-Jones' group at the University of Manchester (UoM). The impact described here is a result of that work.

The primary roles of miRBase throughout this time are to act as the trusted naming authority for novel microRNA genes as they are discovered, and to freely distribute microRNA sequences and associated information. miRBase research and development is ongoing and involves two main activities: 1. Expansion to include newly discovered microRNA genes. 2. Development of new features and analysis for the 20,000-30,000 monthly website users.

The UoM team has expanded miRBase by >34,000 new microRNA genes since 2007. This increase represents around 90% of the current dataset. Data added in the last five years

include microRNA annotations in 65 new animal, plant and virus species, including commercially important crops such as sugarcane, strawberry, orange and cucumber.

The development of new features and analyses in miRBase starts from UoM microRNA research. The UoM team has published >25 papers on microRNA discovery, evolution, and function, the experience of which drives priorities for the development of miRBase [1-3]. This expertise and body of work (see e.g. [4-6]) allows UoM researchers to identify unmet needs that miRBase can fulfil for the community. For example, the reliability of newly-discovered microRNAs in the scientific literature varies. UoM researchers therefore developed a pipeline to assess the quality of microRNA annotations using publicly-available deep sequencing datasets [1,2]. They collected and analysed small RNA deep sequencing data, and have developed and described a set of criteria that allow them to classify a subset of microRNA annotations as being of 'high confidence'. The high confidence set currently represents 20% to 65% of the annotations in different well-studied animal genomes [3]. Researchers can then choose to prioritise work on higher confidence microRNAs, for example in the study of their function or clinical relevance.

Another recent development addresses a major bottleneck in microRNA research, to access and analyse information about the biological function of a microRNA. To this end, the UoM team has developed a pipeline to extract microRNA functional information from the text of scientific articles [3]. Of 1,200,000 full-text articles they have analysed, >18,000 contain microRNA gene names. From these articles, they have collected 555,000 sentences that contain information on the function of 12,500 microRNAs. They have also incorporated data from a number of computational tools to predict targets of microRNAs. These rich textual data and the computational predictions of targets provide starting points for understanding microRNA function, including in commercial and clinical settings.

3. References to the research

The six indicative papers referenced here have amassed more than 6,000 citations between them (Web of Science, 22 October 2020). The core miRBase papers [1-3] are each in the top 1% of all scientific literature by citations (Web of Science). The development of miRBase has been continuously funded by the BBSRC from 2009 to 2021.

1. **Kozomara A**, **Griffiths-Jones S** (2011). miRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.* 39:D152-D157. DOI: 10.1093/nar/gkq1027
2. **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
3. **Kozomara A**, **Birgaoanu M, Griffiths-Jones S** (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47:D155-D162. DOI: 10.1093/nar/gky1141
4. Ninova M, Ronshaugen M, **Griffiths-Jones S** (2016). MicroRNA evolution, expression, and function during short germband development in *Tribolium castaneum*. *Genome Res.* 26:85-96. DOI: 10.1101/gr.193367.115
5. Bleazard T, Lamb JA, **Griffiths-Jones S** (2015). Bias in microRNA functional enrichment analysis. *Bioinformatics* 31:1592-1598. DOI: 10.1093/bioinformatics/btv023
6. **Griffiths-Jones S**, Hui JHL, Marco A, Ronshaugen M (2011). MicroRNA evolution by arm switching. *EMBO Rep.* 12:172-177. DOI: 10.1038/embor.2010.191

4. Details of the impact

Reach and significance of the impact

The miRBase group at UoM is the single authority on what is and what is not a microRNA, and the miRBase database defines the names and sequences of microRNA genes. **All** microRNA research therefore starts from sequence and annotation data from miRBase. The

research and development of the database has worldwide reach, impacting all microRNA work, including in industrial and pharmaceutical settings.

All miRBase data are distributed free and without restriction, as a service to the worldwide microRNA communities in academic, commercial and clinical sectors. The primary use is through a publicly-available web interface, hosted at UoM (http://mirbase.org/), which is updated regularly with new microRNA sequence data and new features and views for those data [1,2,3]. Since 2013, the website has consistently attracted 40,000-50,000 visits per month, from 20,000-30,000 unique users. Approximately 10% of website use is from the commercial sector.

miRBase data are redistributed by >20 different resources. miRBase is a founding member of the RNAcentral consortium, which unifies global RNA sequence resources. MicroRNA annotations from miRBase are incorporated into the international genome browsers for human and other vertebrates, NCBI, UCSC and Ensembl, but also model organism databases including FlyBase and WormBase. Each of these resources has its own user communities and impact, including in the commercial sector.

Commercial and clinical impact

The size of the market for microRNA experimental tools and resources was estimated at USD160,000,000 in 2017 [A]. The leading companies in this field include Agilent, Invitrogen, LC Sciences, Merck, Qiagen, Sigma-Aldrich, and Thermo Fisher Scientific, all of whom produce and sell experimental kits and resources that rely on miRBase data [B-F]. The availability of miRBase, and its continued development and expansion, directly benefits the sales of products from these companies, which in turn underpin experimental microRNA research, including in commercial and clinical settings.

Qiagen's entire portfolio of more than 265 microRNA products, including PCR arrays, and microRNA inhibitors and mimics, is built from miRBase sequences. Associate Directors for R&D of MicroRNA Products and Next Generation Sequencing at Qiagen state that "*All product lines I am responsible for rely heavily on the miRBase database*" [B], and sequencing product lines "*heavily rely on miRBase for product design … as well as analysis*" [B]. Thermo Fisher Scientific sell over 15,000 assays, each based on a microRNA sequence from miRBase. Their "MicroRNA Analysis Using TaqMan Assays" landing page states that "*The content for our standard TaqMan MicroRNA Assays and the new TaqMan Advanced miRNA Assays is aligned with release 22 of the miRBase database*" [C]. Over 10,000 researchers worldwide use TaqMan microRNA assays. A set of reactions for a single microRNA costs around USD250, and an active microRNA group might use 10s to 100s of sets annually. Bioinformatics tools sold by companies such as Qiagen also rely on miRBase, and "*would not be possible to provide without the miRBase resource*" (Global Product Manager OmicSoft, Qiagen [D]). Qiagen's commercial bioinformatics software, CLC, typically costs around USD5,000 per academic licence, and is used by >1,000 groups around the world for microRNA research (>5,000 citations in the microRNA literature [D]).

After a new miRBase release, companies race to release new iterations of their tools and resources. For example, in April 2018, LC Sciences announced that they had updated their microRNA array platform with miRBase v22 datasets [E]. The qPCR Product Manager from Thermo Fisher Scientific states "*We constantly update assays based on miRBase and rely on the information disseminated from it*" [F].

Amongst the authors of >4,500 papers that cite miRBase papers [1,2,3], there are >250 commercial organisations and public health bodies, including GSK, AstraZeneca, Eli Lilly, Boehringer Ingelheim, Monsanto, and Merck. These publications highlight the use of miRBase data in a wide range of commercial applications. Boehringer Ingelheim identified 14 microRNAs that provide a reliable urinary signature of diabetic nephropathy, for development of novel biomarkers [G]. An Affymetrix study used their Axiom miRNA Target Site Genotyping Array (based on miRBase data) to identify genetic variation in microRNA target sites associated with colorectal cancer risk [H].

The availability of miRBase facilitates the ongoing development of novel therapeutics for human disease, including cancer. Regulus Therapeutics is developing chemically modified microRNAs for metabolic and cardiovascular diseases. The starting point microRNAs are defined by miRBase. The Director of Infectious Disease and Immunology confirmed that miRBase "*is the primary source of microRNA sequences information that is used by our R&D team*", and that the therapeutic development pipeline relies on the aforementioned miRBase-derived products from Qiagen and Thermo Fisher [I]. In 2016, MIRagen Therapeutics started phase I clinical trials for a mimic of miR-29b for scleroderma, and an antagonist of miR-155 for a type of T-cell lymphoma; the definition and sequence of the microRNA from which the potential therapeutic is derived comes from miRBase.

Policy impact

miRBase policy decisions have global reach across all research sectors. miRBase is recognised as the official authority on microRNA gene annotation by a number of publicly-funded bodies, including the human (HUGO/HGNC) and mouse (Mouse Genome Informatics) gene nomenclature committees, and the RefSeq resource at NCBI. These bodies "*would be unable to provide up-to-date gene symbols for human microRNA genes without the invaluable miRBase resource*" [J]. miRBase microRNAs are also distributed by the primary human genomics resources at NCBI, UCSC and Ensembl. These browsers underpin all human genomic research, including in the commercial sector.

5. Sources to corroborate the impact

A. Grand View Research. MicroRNA Market Size, Share & Trends Analysis Report. https://www.grandviewresearch.com/industry-analysis/microrna-market (1 May 2018), estimating the size of the market for microRNA experimental tools and resources.
B. Letters from Associate Director of MicroRNA Products (8 January 2020) and Associate Director of Next Generation Sequencing (20 December 2019), Research and Development, Qiagen, confirming the dependence of their product lines on miRBase.
C. Thermo Fisher Scientific. MicroRNA Analysis Using TaqMan Assays. https://www.thermofisher.com/uk/en/home/life-science/pcr/real-time-pcr/real-time-pcr-assays/mirna-ncrna-taqman-assays.html (26 January 2020), stating that the content for their assays is aligned with the latest miRBase release.
D. Letter from Global Product Manager OmicSoft, Qiagen (6 January 2020), confirming the dependence of their product lines on miRBase.
E. LC Sciences. miRBase Version 22 Released – Content Available now on our miRNA Microarrays. https://www.lcsciences.com/news/mirbase-version-22-released-content-available-now-on-our-mirna-microarrays/ (12 April 2018).
F. Letter from Product Manager, Thermo Fisher Scientific (13 January 2020), confirming the dependence of their product lines on miRBase.
G. An example of the use of miRBase data in a commercial application by Boehringer Ingelheim: Delić, D, et al. (2016). Urinary exosomal miRNA signature in type II diabetic nephropathy patients (2016) PLoS ONE, 11(3):e0150154.
H. An example of the use of miRBase data in a commercial application by Affymetrix Inc.: Schmit SL, et al. (2015). MicroRNA polymorphisms and risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev. 24(1):65-72.
I. Letter from Director of Infectious Disease and Immunity, Regulus Therapeutics (17 January 2020), confirming the reliance of their therapeutic development pipeline on miRBase-derived products.
J. Letter from HUGO Gene Nomenclature Committee Advisor (26 July 2020), confirming that official human microRNA gene names are assigned from miRBase identifiers.