

# Development of novel analysis and data integration systems to understand human gene regulation

Dissertation

zur Erlangung des Doktorgrades

Dr. rer. nat.

der [Fakultät für Mathematik und Informatik](#)  
der [Georg-August-Universität Göttingen](#)

im [PhD Programme in Computer Science \(PCS\)](#)  
der [Georg-August University School of Science \(GAUSS\)](#)

vorgelegt von

[Raza-Ur Rahman](#)

aus Pakistan

Göttingen, April 2018

**Betreuungsausschuss:** Prof. Dr. Stefan Bonn,  
Zentrum für Molekulare Neurobiologie (ZMNH),  
Institut für Medizinische Systembiologie, Hamburg

Prof. Dr. Tim Beißbarth,  
Institut für Medizinische Statistik, Universitätsmedizin,  
Georg-August Universität, Göttingen

Prof. Dr. Burkhard Morgenstern,  
Institut für Mikrobiologie und Genetik Abtl. Bioinformatik,  
Georg-August Universität, Göttingen

**Prüfungskommission:**

Referent: Prof. Dr. Stefan Bonn,  
Zentrum für Molekulare Neurobiologie (ZMNH),  
Institut für Medizinische Systembiologie, Hamburg

Korreferent: Prof. Dr. Tim Beißbarth,  
Institut für Medizinische Statistik, Universitätsmedizin,  
Georg-August Universität, Göttingen

Weitere Mitglieder  
der Prüfungskommission: Prof. Dr. Burkhard Morgenstern,  
Institut für Mikrobiologie und Genetik Abtl. Bioinformatik,  
Georg-August Universität, Göttingen

Prof. Dr. Carsten Damm,  
Institut für Informatik, Georg-August Universität, Göttingen

Prof. Dr. Florentin Wörgötter,  
Physikalisches Institut Biophysik,  
Georg-August-Universität, Göttingen

Prof. Dr. Stephan Waack,  
Institut für Informatik, Georg-August Universität, Göttingen

Tag der mündlichen Prüfung: der 30. März 2018

# Contents

<b>List of Figures</b>	<b>v</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>Abstract</b>	<b>1</b>
<b>List of publications and softwares</b>	<b>3</b>
<b>Thesis structure</b>	<b>5</b>
<b>1 Biological Background Knowledge</b>	<b>6</b>
1.1 Deoxyribonucleic acid . . . . .	6
1.2 Gene expression . . . . .	6
1.2.1 Transcription start site . . . . .	7
1.2.2 RNA polymerase II . . . . .	7
1.2.3 Promoter . . . . .	8
1.2.4 Enhancers . . . . .	8
1.2.5 Transcription factors . . . . .	8
1.3 Alternative Splicing . . . . .	8
1.4 Small RNA (sRNA) . . . . .	10
1.4.1 MicroRNAs . . . . .	10
1.4.2 PIWI-interacting RNAs . . . . .	11
1.4.3 Small nucleolar RNAs . . . . .	12
1.4.4 Small interfering RNA . . . . .	12
1.4.5 Small nuclear RNAs . . . . .	13
1.5 Next generation sequencing . . . . .	15
1.5.1 RNA sequencing . . . . .	15
1.5.1.1 Method . . . . .	16
<b>2 Bioinformatics Background Knowledge</b>	<b>18</b>
2.1 Database management systems . . . . .	18
2.1.1 DBMS Architecture . . . . .	19
2.2 Types of databases . . . . .	20
2.2.1 Relational database systems . . . . .	20
2.2.1.1 Constraints . . . . .	22
2.2.1.2 Entity relationship model (ER model) . . . . .	23
2.2.2 Non-relational database systems . . . . .	23
2.2.2.1 Types of NoSQL databases . . . . .	24

---

2.3	Standard workflows for NGS data analysis . . . . .	26
2.3.1	Raw data (FASTQ) . . . . .	26
2.3.2	Quality control (QC) . . . . .	27
2.3.2.1	FastQC . . . . .	27
2.3.3	Adapter trimming . . . . .	28
2.3.4	Alignment and counting . . . . .	29
2.3.5	Differential expression (DE) analysis . . . . .	29
2.4	Biological ontologies . . . . .	30
2.5	Principles of supervised machine learning methods . . . . .	30
2.5.1	Classification . . . . .	31
2.5.1.1	Biological example . . . . .	31
2.5.1.2	Random forest . . . . .	32
2.6	Thesis related existing resources and research . . . . .	33
2.6.1	sRNA-seq analysis tools . . . . .	33
2.6.1.1	sRNA workbench . . . . .	33
2.6.1.2	CAP-miRSeq . . . . .	34
2.6.1.3	omiRas . . . . .	34
2.6.1.4	mirTools 2.0 . . . . .	34
2.6.1.5	MAGI . . . . .	34
2.6.1.6	Chimira . . . . .	34
2.6.1.7	sRNAtoolbox . . . . .	34
2.6.2	sRNA expression databases . . . . .	35
2.6.2.1	miRmine . . . . .	35
2.6.2.2	DASHR . . . . .	35
2.6.2.3	Miratlas . . . . .	35
2.6.2.4	YM500v3 . . . . .	36
2.6.3	Mutually exclusive splicing of exons . . . . .	36
2.7	Goals of the Thesis . . . . .	36
2.7.1	Online analysis of small RNA deep sequencing data (Oasis) . . . . .	36
2.7.2	sRNA expression atlas (SEA) . . . . .	37
2.7.3	Mutually exclusive splicing of exons . . . . .	38
<b>3</b>	<b>Results, Discussion and Outlook</b> . . . . .	<b>39</b>
3.1	Online analysis of small RNA-seq data (Oasis 2) . . . . .	39
3.1.1	Oasis 2's module . . . . .	39
3.1.2	OasisCompressor . . . . .	42
3.1.3	Quality Control (QC) . . . . .	44
3.1.4	Functional enrichment analysis . . . . .	45
3.2	Small RNA expression atlas (SEA) . . . . .	47
3.2.1	System design . . . . .	48
3.2.2	Annotation tool . . . . .	49
3.2.2.1	Annotation criteria . . . . .	50
3.2.3	SEA web application . . . . .	51
3.3	Mutually exclusive splicing of exons . . . . .	52
3.3.1	Data sources . . . . .	52
3.3.2	Prediction of MXE candidates . . . . .	53
3.3.3	Validation of MXE candidates . . . . .	53

---

3.3.4	Spatio-temporal expression of MXEs . . . . .	54
3.3.5	Disease pathology prediction . . . . .	55
3.4	Conclusion and outlook . . . . .	57
<b>References</b>		<b>67</b>
<b>Appendices</b>		<b>68</b>
<b>A</b>	<b>Article 1</b>	<b>69</b>
<b>B</b>	<b>Article 2</b>	<b>80</b>
<b>C</b>	<b>Article 3</b>	<b>95</b>

# List of Figures

1.1	DNA structure . . . . .	7
1.2	Gene expression . . . . .	7
1.3	Promoter, enhancers and TFs . . . . .	9
1.4	Forms of alternative splicing . . . . .	10
1.5	miRNA biogenesis . . . . .	11
1.6	piRNA biogenesis . . . . .	13
1.7	snoRNA biogenesis . . . . .	14
1.8	siRNA biogenesis . . . . .	15
1.9	RNA-seq library preparation workflow . . . . .	17
2.1	Three-level DBMS architecture . . . . .	19
2.2	DBMS architecture along with different ways of querying the DBMS . . . . .	21
2.3	ERD representation . . . . .	22
2.4	Standard workflow for NGS data analysis (RNA-seq,sRNA-seq) . . . . .	26
2.5	FastQ format . . . . .	27
2.6	FastQC per-base quality . . . . .	28
2.7	FastQC sequence quality . . . . .	28
2.8	Disease ontology . . . . .	30
2.9	Supervised machine learning . . . . .	31
2.10	Illustration of random forest algorithm . . . . .	32
3.1	Oasis 2 modules and workflow . . . . .	40
3.2	OasisCompressor . . . . .	43
3.3	Browser view of the primary output of sRNA detection module . . . . .	44
3.4	Assessment of Oasis 2' (QC) outlier detection . . . . .	46
3.5	SEA system architecture . . . . .	49
3.6	SEA data integration workflow . . . . .	50
3.7	Annotation tool . . . . .	51
3.8	SEA home page . . . . .	52
3.9	MXE illustration . . . . .	54
3.10	Spatio-temporal expression of MXEs . . . . .	55
3.11	MXE-ratio expression predicts disease pathology . . . . .	56

## *Acknowledgements*

First, I would like to thank Prof. Dr. Stefan Bonn for his guidance and helpful suggestions, who helped me to expand on my bioinformatics skills, and guided me to be able to manage teams. I would also like to thank my Thesis Committee, Prof. Dr. Tim Beißbarth and Prof. Dr. Burkhard Morgenstern, who gave me advice regarding my various projects from time to time. I would like to thank the entire Bonn lab, who were very helpful and encouraging. I would especially like to thank Abhivyakti Gautam and Abdul Sattar, who helped me in the development of these projects. Finally, I would like to dedicate my phd to my mother Shams-un Nahar for her ongoing love and support and to my father Atta Ur Rahman who could not see this thesis completed.

## Abstract

This thesis covers a very broad range of bioinformatics methods ranging from the development of the analysis pipeline to the data integration and development of an expression atlas (database and web application development). In addition, an in silico method was developed to annotate genome with novel features, and predicting diseases based on the expression profiles.

### Development of online analysis of small RNA sequencing data

Small RNA (sRNA) are biomolecules that play important roles in organismal health and disease; as such, sRNA dysregulation can cause severe diseases. The modern method of choice for sRNA expression profiling is sRNA sequencing (sRNA-seq). There are several sRNA-seq analysis platforms available that differ in their analysis portfolio, performance, and user-friendliness. However, these analysis platforms lack one or more important features such as disease biomarkers identification, detection of viral and bacterial infections in sRNA-seq samples, storage of novel predicted miRNAs, multivariate differential expression (DE) analysis and automated submission of jobs via an application programming interface (API).

To this end, we developed an online analysis tool called as **Oasis 2**, a fast and flexible web application which provide many different sRNA-seq analysis options on a single platform. Its major functionalities include quantification of different sRNA species, multivariate differential expression (DE), identification of biomarkers for disease, prediction and storage of novel miRNAs with proper universally accepted nomenclature, identification of infection or contamination, functional/enrichment analysis. Additionally Oasis 2 enables users to perform all these different analysis over the web application, as well as over API for automatic submission. Oasis 2 generates downloadable interactive web reports for easy visualization, exploration, and analysis of data on a local system. In future, small RNA editing, modification, and mutation events can be implemented in Oasis 2. Additionally the reported output for bacterial and viral infections and contaminations can be enhanced.

### Development of small RNA expression atlas (SEA)

As discussed in Section 2 that sRNAs have crucial role in organismal health and disease, yet the number and scope of the currently available sRNA-seq expression repositories are very limited. For example, most of the sRNA-seq repositories support one or two organisms and none of these databases provide search by ontological terms.



Considering these shortcomings, we developed sRNA expression atlas (**SEA**), a data repository to store sRNA expression profiles along with the experimental details such as organism, tissue, cell type, disease, age, gender and technical details like sequencer, kit and barcode etc. Additionally we built a web application that allows end users to query and visualize sRNA expression profiles in an interactive manner. SEA allows users to search for ontology-based queries, supporting single or combined searches for five pre-defined terms such as organism, tissue, disease, cell type, and cell line across different experiments. Currently it contains expression and meta-information of over 2,500 sRNA-seq samples across 10 organisms. As far as we are aware, SEA is the only sRNA-seq database that supports ontology-based queries. In the future, additional available meta-information such as age, gender, developmental stage, genotype as well as technical experimental details can be standardized (connect to ontologies) and the search could be enhanced to allow users to query sRNA expression profiles based on them. Moreover, further sRNA-seq datasets should be incorporated into SEA. Lastly, one can store DE and biomarker prediction results for all the sRNA-seq datasets having at-least two groups (such control and diseased) and make them query-able and comparable across different datasets.

### **Prediction and validation of mutually exclusive splicing of exons**

Mutually exclusive splicing of exons (MXEs) is a mechanism of functional gene and protein diversification with important roles in organismal development and diseases, such as in SNAP-25 as part of the neuroexocytosis machinery [1]. Additionally mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the CACNA1C gene) [2, 3]. Despite their important roles, the current knowledge of human MXEs is very limited, that is to say, that the human genome annotation (GenBank v. 37.3) contains only 158 MXEs in 79 protein-coding genes.

To this end, an *in silico* method was developed to predict MXEs based on sequence similarity, similar lengths, and reading frame conservation; predicted MXEs were validated using the publicly available billions of RNA-seq reads. Based on this method the current knowledge of human MXEs is increased by almost an order of magnitude from 158 to 1,399 MXEs. These MXEs show tissue and developmental stage specific expression and also have potential roles in diseases. As a heuristic approach was used for the prediction of MXEs in this thesis, in the future a machine learning approach can be used for the prediction of MXEs, which may increase the predicting power of the method and could result in further novel MXEs.

# List of publications and softwares

## Published

1. **Raza-Ur Rahman**, Abhivyakti Gautam, Jörn Bethune, Abdul Sattar, Maksims Fiosins, Daniel Sumner Magruder, Vincenzo Capece, Orr Shomroni and Stefan Bonn. (2018). Oasis 2: improved online analysis of small RNA-seq data. BMC Bioinformatics (volume19).
2. **Raza-Ur Rahman**, Abdul Sattar, Maksims Fiosins, Abhivyakti Gautam , Daniel Sumner Magruder, Jörn Bethune, Sumit Madan , Juliane Fluck , and Stefan Bonn. (2017). SEA: The small RNA Expression Atlas. bioRxiv preprint. <https://doi.org/10.1101/133199>.
3. Hatje, Klas and **Rahman, Raza-Ur** and Vidal, Ramon O and Simm, Dominic and Hammesfahr, Björn and Bansal, Vikas and Rajput, Ashish and Mickael, Michel Edwar and Sun, Ting and Bonn, Stefan and Kollmar, Martin (2017). The landscape of human mutually exclusive splicing. Molecular Systems Biology (volume 13).
4. Vincenzo Capece, Julio C. Garcia Vizcaino, Ramon Vidal, **Raza-Ur Rahman**, Tonatiuh Pena Centeno, Orr Shomroni, Irantzu Suberviola, Andre Fischer and Stefan Bonn. . (2015). Oasis: online analysis of small RNA deep sequencing data. Bioinformatics 31, 1–3
5. Rashi Halder, Magali Hennion, Ramon O. Vidal, Orr Shomroni, **Raza-Ur Rahman**, Ashish Rajput, Frauke van Bebber, Anna-Lena Schuetz, Susanne Burkhardt, Eva Benito, Julio C. Garcia Vizcaino, Vincenzo Capece, Tonatiuh Pena Centeno, Magdalena Navarro Sala, Sanaz Bahari Javan, Christian Haass, Bettina Schmid, Andre Fischer, Stefan Bonn. DNA methylation changes in plasticity genes accompany the formation and maintenance of memory. Nature Neuroscience, 19(1), 102–110.
6. Tonatiuh Pena Centeno, Orr Shomroni, Magali Hennion, Rashi Halder, Ramon Vidal, **Raza-Ur Rahman**, Andre Fischer, Stefan Bonn. Genome-wide chromatin

and gene expression profiling during memory formation and maintenance in adult mice. Scientific data.

### **In preparation**

1. Eugenio F. Fornasiero, Sunit Mandad, **Raza-Ur Rahman**, Tonatiuh Pena Centeno, Ramon O. Vidal, Hanna Wildhagen, Burkhard Rammner, Sarva Keihani, Felipe Opazo, Inga Urban, Till Ischebeck, Koray Kirli, Eva Benito, André Fischer, Sven Dennerlein, Peter Rehling, Ivo Feussner, Henning Urlaub, Stefan Bonn, Silvio O. Rizzoli. The codon sequences predict protein lifetimes and other parameters of the protein life cycle in the mouse brain. eLife
2. Eugenio F. Fornasiero, Sunit Mandad, Hanna Wildhagen, Burkhard Rammner, Inga Urban, Till Ischebeck, Eva Benito, Koray Kirli, **Raza-Ur Rahman**, Sven Dennerlein, Peter Rehling, Ivo Feussner, André Fischer, Stefan Bonn, Henning Urlaub, Silvio O. Rizzoli. The analysis of protein lifetimes in the mouse brain reveals basic turnover principles. Nature Neuroscience

### **Softwares**

1. **Oasis 2:** Improved online analysis of small RNA-seq data. <https://oasis.dzne.de/>.
2. **SEA:** Small RNA Expression Atlas. <https://sea.dzne.de/sea/sea.jsp>.
3. **Memory-epigenome-browser:** A genome browser for the interactive visualization of (in house) NGS data. <https://oasis.dzne.de/JBrowse-1.11.4/index.html>.

# Thesis structure

In this thesis, three main projects were developed.

1. **Oasis 2:** Improved online analysis of small RNA-seq data. The original publication is available at <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2047-z>, and the corresponding web application can be accessed at <https://oasis.dzne.de/>.
2. **SEA:** Small RNA Expression Atlas. It is submitted to biorxiv and is available at <https://www.biorxiv.org/content/early/2017/08/04/133199>, and the corresponding web application can be accessed at <https://sea.dzne.de/sea/sea.jsp>.
3. **Prediction and validation of mutually exclusive splicing of exons:** The original publication is available at <http://msb.embopress.org/content/13/12/959>.

There are three main chapters in the thesis followed by the three above mentioned articles.

- Chapter 1 : Provides biological background knowledge required for this thesis.
- Chapter 2 : Provides bioinformatics background knowledge required for this thesis.
- Chapter 3 : This chapter summarizes the three aforementioned articles, their development, results and the outlook of the projects.
- Appendix [ A, B, C ] : All the three aforementioned articles are provided in the appendix.

# Chapter 1

## Biological Background Knowledge

This chapter explains the biological background required for this thesis including: the process of gene regulation, exon splicing, role of small RNAs (sRNAs) in gene regulation and the basic mechanism of latest technologies such as next generation sequencing (NGS) to obtain gene expression as well as sRNA expression data.

### 1.1 Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) carries the genetic code that is used in the development and growth of living organisms and also some viruses. DNA is a double-stranded molecule that is composed of four bases: adenine (A), thymine (T), cytosine (C) and guanine (G). In order to hold the double-stranded structure of DNA, these molecules bind to each other in a particular order such as cytosine (C) binds to guanine (G) and adenine (A) binds to thymine (T) as shown in Figure 1.1. In the double stranded structure of DNA, the strands are anti-parallel (the direction of nucleotides is opposite). The ends of these strands are named, three prime (3') end having a terminal hydroxyl group and five prime (5) end having a terminal phosphate group. These DNA molecules are used to make various ribonucleic acid (RNA) and protein molecules required by living organisms to carry out different biological functions.

### 1.2 Gene expression

DNA is made up of nucleotides. Some strings of nucleotides form genes which convey units of functionality. Genes are transcribed to a particular RNA molecule called as messenger RNA (mRNA), which can further be translated into a protein as show in

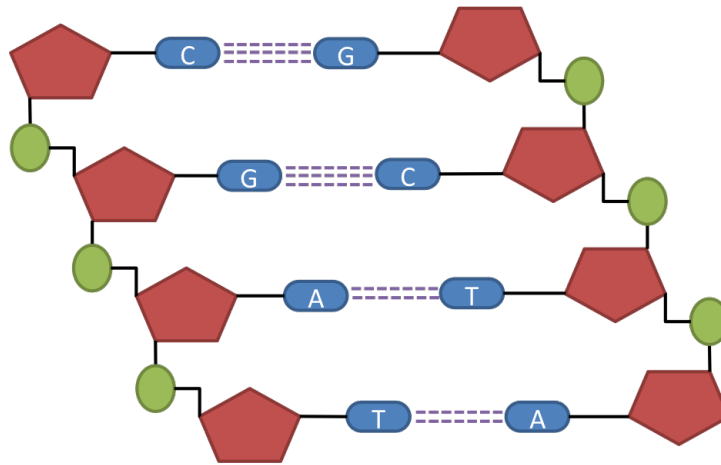


FIGURE 1.1: DNA consists of a deoxyribose backbone (light red) connected by phosphate groups (green circles). Strands are bound together by hydrogen bonds between nucleotides A and T requires two hydrogen bonds where as C and G has three hydrogen bonds between them

Figure 1.2. Transcription is a complex process and it involves many factors such as transcription start site (TSS), RNA polymerase (Pol-II), promoter region, transcription factors (TFs) and enhancers.



FIGURE 1.2: DNA is transcribed into mRNA and proteins are translated from mRNA molecules

### 1.2.1 Transcription start site

As the name suggests transcription start site (TSS) is the location where transcription of the gene into RNA begins [4]. TSS is the location where a molecule of RNA polymerase II (pol II) binds.

### 1.2.2 RNA polymerase II

RNA polymerase II (Pol II), also called as RNAP II, is an enzyme that acts as a catalyst for the transcription of DNA to synthesize precursors of mRNA, microRNA and most snRNA [5]. A variety of different transcription factors are required for Pol II to bind to upstream gene promoters and initiate transcription.

### 1.2.3 Promoter

A promoter region can be found upstream of every gene and contains particular regions where a protein complex can bind to initiate transcription. As shown in Figure 1.3 a promoter is a part of DNA that helps in the initiation of transcription of a particular gene. Promoters are located on the same strand and upstream on the DNA of genes they transcribe. They have binding sites for proteins known as transcription factors that engage RNA polymerase.

### 1.2.4 Enhancers

Enhancers play an important role in the transcription of a gene. Enhancers can be located either upstream or downstream of the transcription initiation site. Enhancers can be distal to TSS, which means they can interact from a distance of thousands of base pairs away from the initiation site [6] as shown in Figure 1.3. Some other protein complexes binds to enhancers in order to make the enhancer complex and bring it close to the promoters and increase transcription.

### 1.2.5 Transcription factors

Transcription factors (TFs) also plays an important role in the regulation of transcription. They bind to short DNA sequences 5-20 bp in length called as transcription factor binding sites (TFBSs) and plays an important role in controlling the flow of genetic information from DNA to mRNA [7]. Some TFs bind to promoter sequences near the TSS and form the transcription initiation complex, while others TFs can bind to regulatory sequences, such as enhancer sequences, either encouraging or repressing transcription of a particular gene as shown in Figure 1.3. TFs are one of the main reasons for cell and tissue specific expression of genes.

## 1.3 Alternative Splicing

Many organisms' DNA has introns and exons. Exons are the coding regions of a gene and contains information for producing proteins whereas introns are the noncoding part of the DNA, and are therefore spliced out of the primary RNA. Having functional blocks of DNA (exons) enables a single gene to be spliced differently to generate various isoforms (different mRNA from same gene), which can be translated into proteins with different structures and functions. This mechanism that enables a single gene to code for multiple

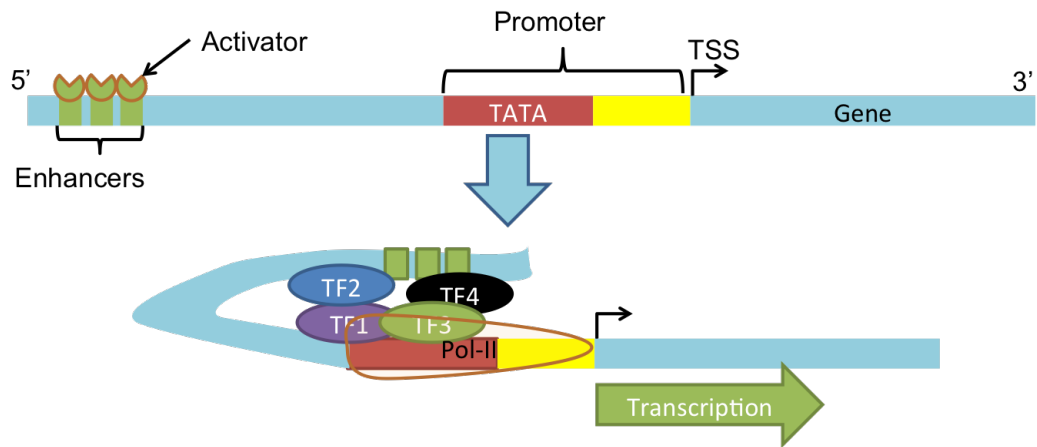


FIGURE 1.3: Interaction between enhancers, promoters along with transcription factors is shown. Promoter and enhancer regions are recognized (bound) by specific TFs. An enhancer promotes transcription and they could be distal to the gene. Activators bound to the distal elements interact with TFs. As soon as all required TFs, activators and Pol-II come together, transcription of DNA to RNA starts.

proteins is called as alternative splicing. Gene splicing occurs prior to mRNA translation, by the differential exclusion or inclusion of different exons. During the splicing event, a pre-mRNA transcribed from one gene can form different mature mRNA molecules that produce different proteins. The different forms of alternative splicing are exon skipping or inclusion, intron retention, alternative splice-site selection and mutually exclusive exons as shown in Figure 1.4.

- **Exon skipping**

In this form of gene splicing, exon(s) are excluded in the final gene transcript that leads to different mRNA isoforms.

- **Intron retention**

In this form of gene splicing, an intron is retained in the final transcript. As the non-coding (intron) portions of the gene is retained, deformity in the protein structure and function can occur.

- **Alternative 3' and 5' splice site**

In this form of gene splicing different 5' and 3' splice site are joined together. In this type of gene splicing, two or more alternative 5' splice site compete for joining to two or more alternate 3' splice site.

- **Mutually exclusive exons**

Mutually exclusive splicing makes alternative isoforms by retaining only one exon of a cluster of neighbouring internal exons in the mature transcript and is one of the ways to modulate protein function.



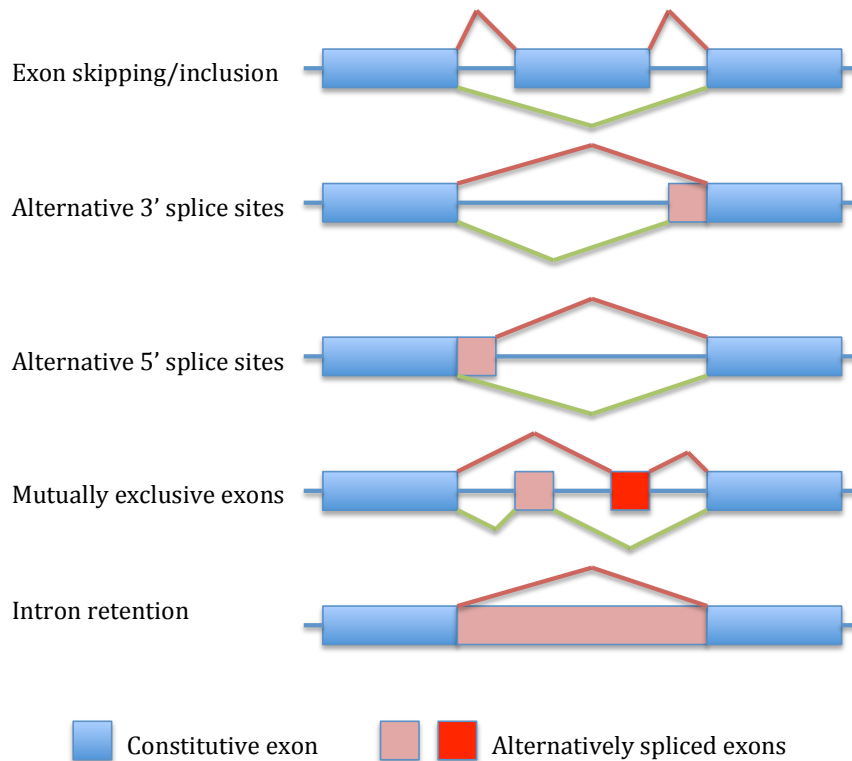


FIGURE 1.4: Different forms of alternative splicing are shown, exon skipping or inclusion, intron retention, alternative splice-site selection and mutually exclusive exons. Different types of alternative-splicing patterns of exons exist for each individual pre-mRNA.

## 1.4 Small RNA (sRNA)

As explained in Section 1.3, coding region of DNA is transcribed into mRNA, which results in proteins after being translated. However the non-coding region of the genome may also be transcribed into non-coding RNAs (ncRNAs) which are never translated into proteins. Based on their length, these ncRNAs are categorized into small ncRNAs (sRNAs) and long ncRNAs (lncRNAs). sRNAs are the type of ncRNAs whose length is less than 200 nucleotides (nt). Based on their biogenesis and biological functions major types of sRNAs include: micro-RNA (miRNA), PIWI-interacting RNAs (piRNAs), small interfering RNA (siRNAs), small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs).

### 1.4.1 MicroRNAs

MicroRNAs (miRNAs) are around 22 nt in length and play an important role in gene regulation by targeting mRNAs for cleavage or translational repression. miRNAs are the most abundant class of sRNAs and they effect the regulation of many protein-coding

genes. miRNAs inhibits the translation of mRNA into protein by binding to complementary sequences in mRNA. There are two mode of action: either the miRNA cleaves the mRNA strand into pieces or it destabilizes the mRNA through shortening of its poly (A) tail. A mature miRNA is produced through the following mechanism as shown in Figure 1.5. First RNA pol II produces pri-miRNAs which is then immediately processed by an enzyme called Drosha in the nucleus to generate pre-miRNAs. These pre-miRNAs are exported to the cytoplasm by Exportin 5. In the cytoplasm pre-miRNAs are processed by Dicer to form the mature miRNA/miRNA\* duplex. Once the mature miRNAs are produced they get assembled into the RNA-induced silencing complex (RISC complex). These mature miRNA inhibits the mRNA translation by complementarily pairing to mRNA.

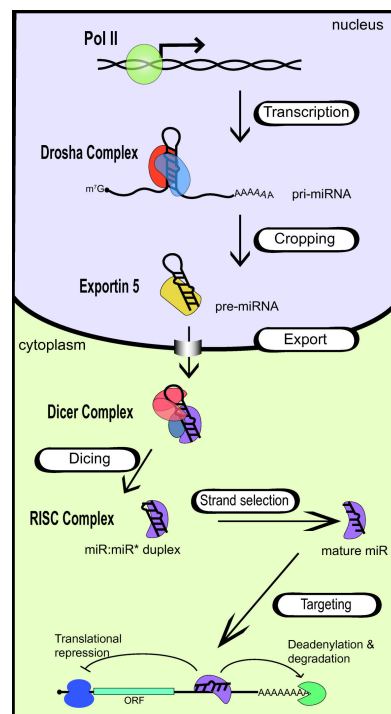


FIGURE 1.5: RNA pol II produces pri-miRNAs which is then immediately processed by an enzyme called Drosha in the nucleus to generate pre-miRNAs. These pre-miRNAs are exported to the cytoplasm by Exportin 5. In the cytoplasm pre-miRNAs are processed by Dicer to form the mature miRNA/miRNA\* duplex. Once the mature miRNAs are produced they get assembled into the RNA-induced silencing complex (RISC complex). These mature miRNA inhibits the mRNA translation by complementarily pairing to mRNA. Figure taken from [8]

### 1.4.2 PIWI-interacting RNAs

PIWI-interacting RNAs (piRNAs) are small noncoding RNAs that function as guardians of the genome. piRNAs protect the genome from the invasive transposable elements (DNA sequences in the genome, which can change their position) in the germline [9].

Intergenic repetitive (elements) regions, from which piRNAs are produced, are called piRNA clusters [10]. piRNAs, around 24-32 nt long, are mostly expressed in the germline [11]. They bind to the PIWI proteins which play a major roles in the maintenance of the genome stability in germline cells. piRNAs have an antisense complementarity to the transposon transcripts and can therefore silence them by hybridizing with them [12]. Recent evidence suggests that piRNAs are not only involved in the germline but also plays roles in the stability of somatic cells as well as in multigenerational inheritance [9]. However to date, very little is known about piRNA diversity and its target specificity in human, nearly all piRNA studies have been conducted in model organisms [13] such as mouse and drosophila. piRNAs are derived from mono or bi-directional clusters and are mainly expressed as mainly as ssRNAs [11]. In order to enforce the high expression of piRNAs in the germline primary piRNAs are subjected to an amplification system (loop) called the ping-pong cycle [9]. To this end, additional piRNAs are produced through this cycle via sense and antisense intermediates. The PIWI ribonucleoprotein (piRNP) complex functions in transposon repression, via epigenetic silencing and target degradation, as shown in Figure 1.6.

### 1.4.3 Small nucleolar RNAs

Small nucleolar RNA (snoRNA) is a class of sRNAs that are responsible for the post-transcriptional modification of ribosomal RNAs (rRNAs) [14]. They are usually 60-150 nt long. snoRNAs are known to reside inside the introns of protein coding genes as shown in Figure 1.7. They are a part of the small nucleolar ribonucleoproteins (snoRNPs), protein complexes that plays role in the pseudouridylation [15] and also in the sequence-specific 2'-O-methylation of the ribosomal RNA (rRNA) [11]. These post-transcriptional modifications of ribosomal RNAs (rRNAs) takes place in the nucleolus, which is a nuclear compartment where ribosomes are formed. The nucleolus also supports rRNA folding and stability [16].

### 1.4.4 Small interfering RNA

RNA interference is a process through which double-stranded RNA silences homologous genes [17]. Small interfering RNA (siRNAs) are around 20-25 nt double-stranded RNA molecules that can target mRNAs based on perfect complementarity as shown in Figure 1.8. In siRNAs biogenesis, two 21-nucleotide (nt) single-stranded RNAs form a 19-bp duplex with 2-nt overhangs at 3'. A Dicer and RDE-1 (RNAi deficient-1) complex processes this double-stranded RNA (dsRNA) to form siRNAs. The RNA interference (RNAi) silencing complex uses the antisense strand of the siRNA for mRNA cleavage

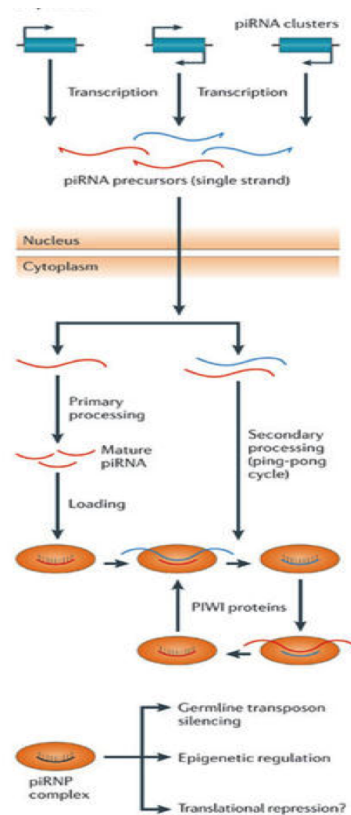


FIGURE 1.6: piRNAs are derived from mono or bi-directional clusters and are mainly expressed as ssRNAs. In order to enforce the high expression of piRNAs in the germline primary piRNAs are subjected to an amplification system (loop) called as ping-pong cycle. To this end, additional piRNAs are produced through this cycle via sense and antisense intermediates. The PIWI ribonucleoprotein (piRNP) complex functions in transposon repression via epigenetic silencing and target degradation. Figure taken from [11].

and hence promoting mRNA degradation as shown in Figure 1.8. siRNAs are more similar to miRNAs in their biogenesis and functions almost identically except: siRNAs can only bind to mRNA sequences with perfect complementarity whereas miRNAs can bind to mRNA even when it does not have perfect complementarity, secondly a siRNA can target only a single mRNA whereas a single miRNA hundreds of mRNAs. Due to the one-to-one mapping of siRNAs to mRNAs they are mostly used as a tool in molecular biology to knock down a gene in an experiment.

#### 1.4.5 Small nuclear RNAs

Small nuclear RNAs (snRNAs) are mostly found in eukaryotic cells and are also called as U-RNA. They are known to have an important role in the splicing of introns from primary genomic transcripts [18]. The average length of snRNA is around 150 nt. There are four main steps in the biogenesis of snRNPs: [19]

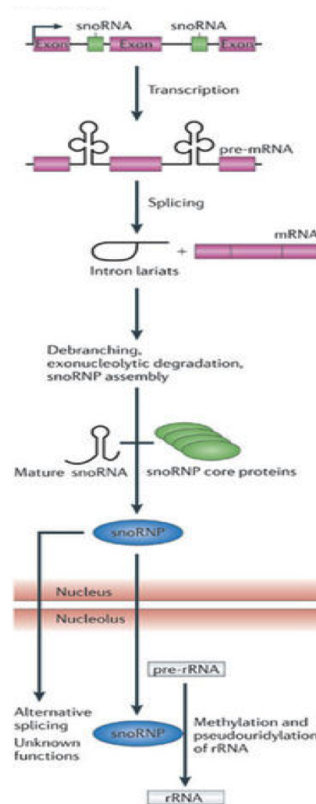


FIGURE 1.7: snoRNAs are mostly found in introns. Mature snoRNAs are formed after splicing, de-branching and trimming. In case these mature snoRNAs remain in the nucleus, they play role in alternative splicing, and if they are exported they get involved in the rRNA processing. Figure taken from [11].

- Production of a large precursor snRNA.
- Processing of the large precursor snRNA into mature snRNA.
- Introduction of site-specific covalent nucleotide modifications.
- Formation of snRNA and RNP proteins complexe.

The biogenesis of snRNPs is very complex, as different classes of snRNP follow different synthetic processing pathways; in addition, the steps are mostly dependant on the sub-cellular compartments [19]. Each snRNA has an association with a set of proteins called as ribonucleoproteins. The complex of snRNA and ribonucleoproteins is called as small nuclear ribonucleoproteins (snRNP or snurps). Prominent components of these snRNA complexes are spliceosomal RNA such as U1, U2, U4, U5 and U6, that plays a major role in the maturation of the eukaryotic precursor messenger RNA. snRNPs binds to the specific sequences on the precursor messenger RNA substrate [20] which results in two reactions: first these reactions will produce free flowing intron and secondly they will ligate the two exons in order to form a mature mRNA.

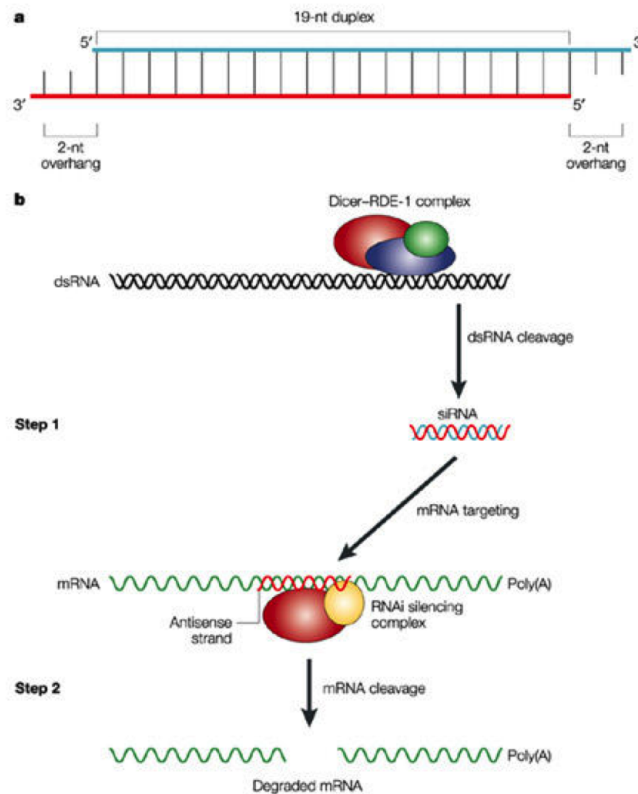


FIGURE 1.8: a) In siRNAs biogenesis, two 21-nucleotide (nt) single-stranded RNAs form a 19-bp duplex with 2-nt overhangs at 3'. b) A Dicer and RDE-1 (RNAi deficient-1) complex processes this double-stranded RNA (dsRNA) to form siRNAs. The RNA interference (RNAi) silencing complex uses the antisense strand of the siRNA for mRNA cleavage and hence promoting mRNA degradation. Figure taken from [17].

## 1.5 Next generation sequencing

The advent of next generation sequencing (NGS) technology has greatly accelerated research in life sciences. Currently, NGS is widely used for whole genome sequencing, protein-DNA interactions, methylated DNA and also for the detection and quantification of gene as well sRNA expression profiles. NGS's popularity in many research laboratories can be contributed to its low cost and high throughput, [21, 22] e.g; the entire human genome can now be sequenced in less than one day.

### 1.5.1 RNA sequencing

RNA sequencing (RNA-seq) is also called whole transcriptome shotgun sequencing. As mentioned in Section 1.5 to detect and quantify RNA in a biological sample at a given moment, NGS is widely used [21, 22]. In addition to mRNA transcripts, RNA-seq can look at different types of RNA such as total RNA, small RNA and ribosomal profiling.

It can also determine intron and exon boundaries. One can validate or update existing 5' and 3' annotated gene boundaries.

#### 1.5.1.1 Method

- **RNA Isolation:** The first step towards RNA sequencing is to isolate RNA from the samples such as tissue and mix it with deoxyribonuclease (DNase) to reduce the amount of genomic DNA.
- **RNA selection:** Depending on the biological question to be addressed, the isolated RNA can be kept- as it is or it can be depleted for ribosomal RNA (rRNA)- or in the case where the requirement is to take into account only mRNA, it can be filtered for 3' polyadenylated (poly(A)) tails. RNA's with 3' poly(A) tails are mature, processed coding sequences.
- **cDNA synthesis:** The above selected RNA is reverse transcribed to cDNA for sequencing. These cDNA fragments are then sheared, selected and amplified with adaptors attached to one or both ends [22].
- **Sequencing:** Lastly, this library is sequenced from both ends (pair-end sequencing) or one end (single-end sequencing) using next generation sequencing technology. This sequencing results in short sequences also called reads [22].

The above method can be used to sequence both mRNA and sRNA. In the case of mRNA, the isolated RNA in the first step is filtered for 3' poly(A) tails as shown in Figure 1.9. RNA's with 3' poly(A) tails are mature, processed and coding sequences. In the case of sRNA sequencing, the library preparation is modified a bit and the RNA is isolated through size selection. This can be done through different means such as size selection via magnetic beads or with a size exclusion gel. After isolation, adaptors are ligated to both ends of the small RNAs. Finally, the adaptor ligated sRNAs are converted to cDNA clones.

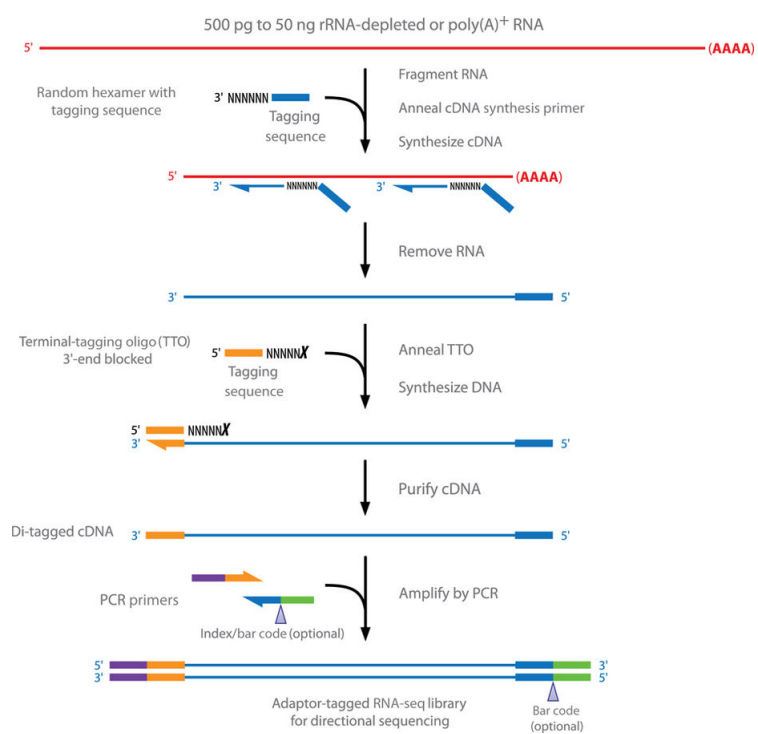


FIGURE 1.9: illustration of directional RNA-seq library preparation workflow for Illumina. Figure taken from [23].



## Chapter 2

# Bioinformatics Background Knowledge

This chapter explains the bioinformatics background knowledge required for this thesis. Main areas of focus in this chapter are:

- Principles of database management systems.
- Standard workflows for NGS data analysis (gene and small RNA expression analysis).
- Principles of supervised machine learning methods.

### 2.1 Database management systems

A database is an organized or structured collection of data [24]. When there is a need to store and process large amounts of data usually a database is applied. In general terms the word database is not specific, it can be an excel sheet storing lists of names and addresses of a company employees or a database server such as Oracle or MySQL. A database management system (DBMS) is software that allows the creation and modification of a database. There are many DBMS; some DBMS include: MySQL, PostgreSQL, Oracle, SQLite, Microsoft SQL Server, SAP, dBASE, IBM DB2, MongoDB and Neo4j. A DBMS offers the following features:

- Data definition: A DBMS allows definition, removal and modification of data structures in the database.

- A DBMS also facilitates to insert, modify, retrieve and delete data from the database.
- A DBMS is also responsible for the database administration. Administration means registering and monitoring users, enforcing data security, such as who can access what, maintaining data integrity, concurrency control and information recovery if the system fails.

A database along with its model and its database management system is collectively called as a database system [25].

### 2.1.1 DBMS Architecture

In classical DBMS architecture every user of the database has an abstract view of the data and certain details are hidden from the users such as how the data is physically stored. This feature of a DBMS enables the users to manipulate the data without worrying about where and how the data is actually stored. A database can be defined at three levels; such as internal, conceptual and external levels therefore it is named three-level DBMS architecture. Figure 2.1 shows the three levels of DBMS architecture.

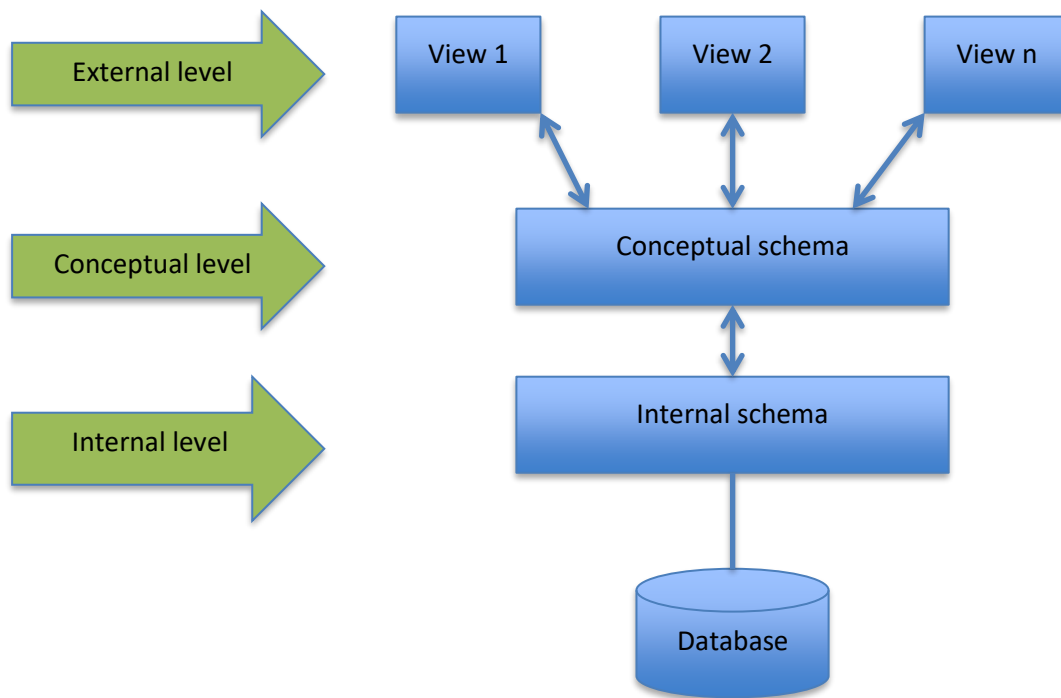


FIGURE 2.1: Three-level DBMS architecture

- **Internal level** is also called as physical level because it deals with the physical representation of the database on the machine (computer). This is the lowest level of data abstraction, which describes physical storage of the data and its organization on the storage medium.
- **Conceptual level** is also called as logical level as it deals with the logical structure of a database. It explains the data and relationships between the data, which is stored in the database. This level is not concerned with any physical organization of the data on the storage medium.
- **External level** deals with the user's view of the database, therefore it is also called as view level. As most of the users and programs do not require the whole data stored in the database. This level permits data access in a user's customized manner. In this way, it provides a powerful security mechanism by hiding some parts of the database from certain users.

There are different ways to query a database, such as web applications, web forms or even direct access to the database from a program. DBMS also offers command-line interaction for users such as programmers and database administrators. A database driver is required in order for programs to communicate to DBMS. The database drivers handle the requests and send them to the database. Once the query is send to DBMS, the query is analyzed by the query evaluation engine, then database management system applies the query and the desired data is retrieved from the physical data storage.

On the other hand a DBMS also has a concurrency control mechanism to maintain data consistency in situations such as manipulation of the same data by more than one user at the same time. Importantly a DBMS also has a recovery manager that contains several mechanisms to restore the database in case an abrupt system crash occurs. Figure 2.2 shows architecture of a DBMS and the different ways a database can be queried.

## 2.2 Types of databases

In general, databases can be categorized into relational and non-relational databases. The main differences are highlighted in the following sections.

### 2.2.1 Relational database systems

Edgar Codd first introduced a relational model for the representation of data in 1970 [26]. A relation represents the form (structure) in which the data is stored [27]. A

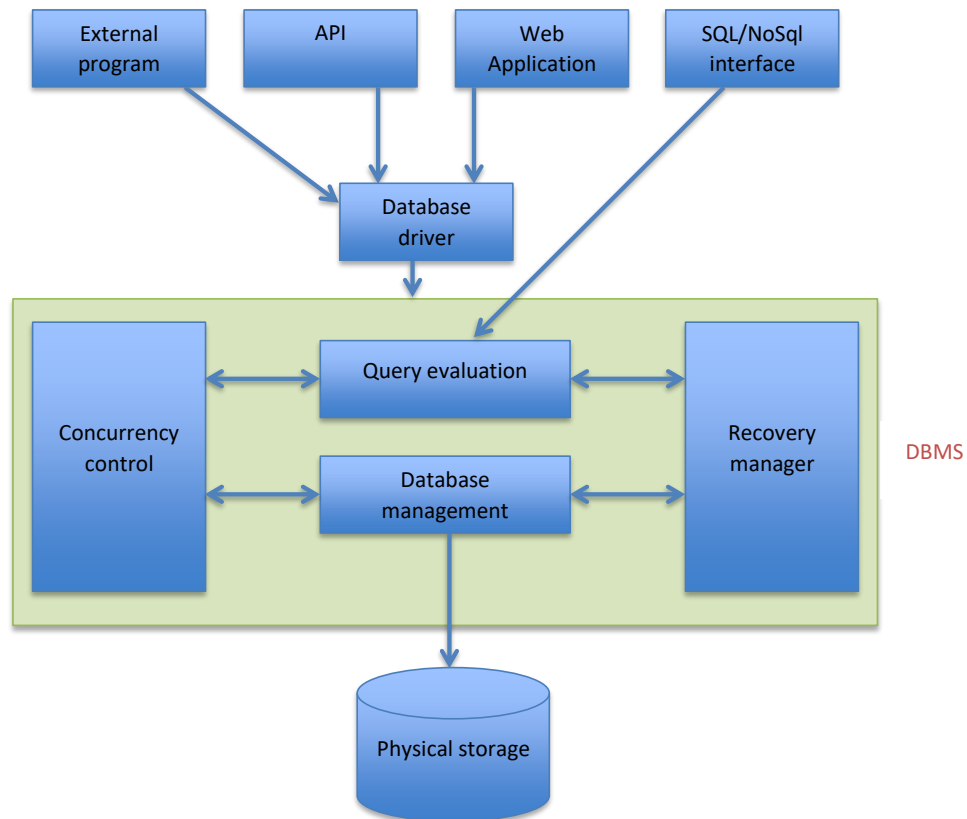


FIGURE 2.2: DBMS architecture along with different ways of querying the DBMS

relation could be an excel sheet or a mysql table. In a relational database data is usually stored in tables. Every relation has a heading and a body. A set of attributes defines the heading and the body is a set of tuples (rows) that corresponds to that heading. A heading represents the columns and a row in the table denotes a tuple. Relations follow the set theory, which means every row has to be different from each other in at least one attribute value (there must not exist identical tuples in a relation). In a standard relational database, tables also have relationships with each other. The following types of relationships exist between relations in a database:

- **One to one**

If one element from relation 1 ( $R_1$ ) is associated to at most one element from relation 2 ( $R_2$ ) and vice versa as shown in Figure 2.3.

- **One to many, many to one**

A relation is said to be one to many or many to one if an element from  $R_1$  is associated to many elements of  $R_2$  whereas one element from  $R_2$  may have a relation with at most one element from  $R_1$  as shown in Figure 2.3.

- **Many to many**

A type of relationship in which an element from R1 has zero to many relations with elements from R2. The same holds for elements from R1 as shown in Figure 2.3.

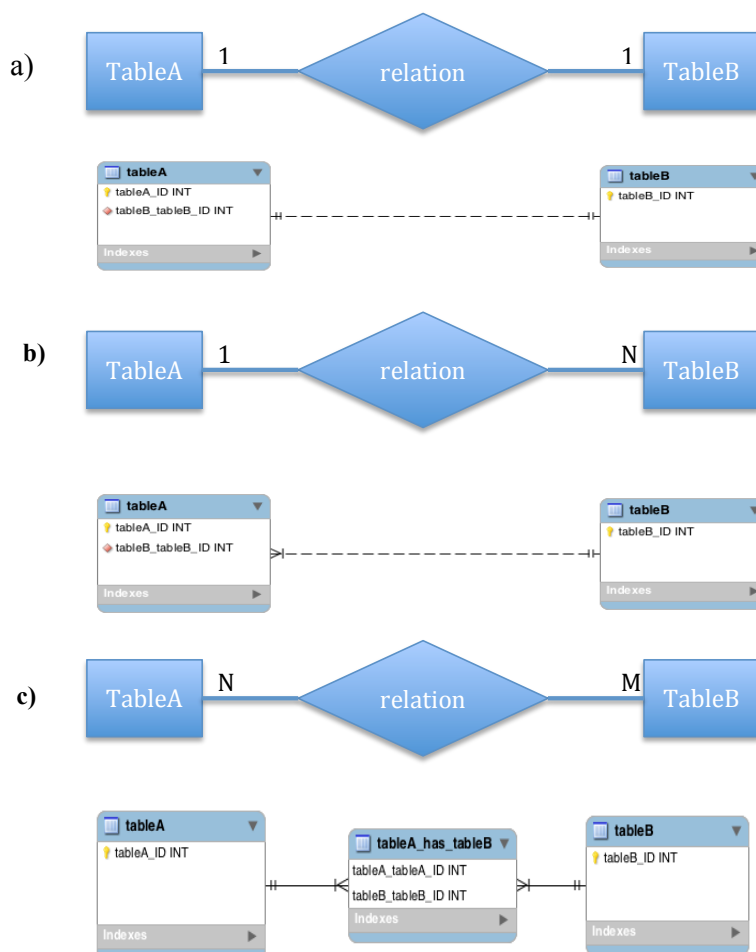


FIGURE 2.3: Shown are two representations of ERD. The figure also shows different types of relations between different tables (entities). (a) 1:1 relation (b) shows a 1:N relation between two tables and (c) represents N: M relationship between the tables. In each case the lower representation is from a DBMS where a primary key has a yellow key sign and foreign key is shown with pink diamond

### 2.2.1.1 Constraints

In relational databases constraints are used to define the domain of an attribute or a tuple. For example, a constraint on an integer attribute can restrict the integer to values between 1 and 30 only. This is one of the methods to implement business rules in the database. The two main rules for a relational model are referential integrity and entity integrity. These rules are implemented with the help of keys as explained below:

## Keys

Keys are used to identify records (tuples) in a table

- **Primary keys** A primary key is used to identify a tuple in a table uniquely. This could be a single attribute or combination of more than one attribute. This implies that no two tuples may have the same values for this attribute(s). In order to avoid duplicates a relation should always contain at least one primary key attribute. One of the constraints on the primary key is that it can never have NULL value because this leads to loss of the uniqueness. In short, a primary key is the minimal set of attributes that identifies a certain tuple in a relation. Primary keys can be used for indexing to allow faster access to the desired records.
- **Foreign key** A foreign key is used to build a relationship between different relations (tables). A foreign key is an attribute in a relation that matches the primary key attribute of the other relation. A tuple from one relation may have reference to one or several tuples in another relation with the use of a foreign key.

### 2.2.1.2 Entity relationship model (ER model)

An entity relationship model (ER model) is a data model for presenting a database in a schematic way. In case of a relational database, diagrams are created to design tables (entities) and their relationships to other tables (entities), these diagrams are called entity relationship diagrams (ERD) [28]. An example of a simple ERD is shown in the Figure 2.3.

One of the important aspects of relational databases is the minimal duplication of data, which makes them very consistent and efficient in certain transactional and concurrent update operations. However relational database schema having to be predefined and are only vertically scalable. It lacks the horizontal flexibility like NoSQL databases. Additionally they are inefficient for the storage of large and sparse data (as empty values also take space).

### 2.2.2 Non-relational database systems

As NoSQL database system was used in this thesis for the storage of unstructured data (explained later in this chapter). This section provides brief overview on the non-relational databases. They are also known as NoSQL databases. Few examples of NoSQL database are MongoDB, Neo4j, DocumentDB, Cassandra, Couchbase and HBase. Typically they can be categorized into four groups: document stores, column stores, graph

stores and key-value stores. In essence a NoSQL database is used for the storage of data without predefined explicit structures or for the storage and retrieval of data that is modeled in a non-tabular relations such as that used in relational databases. Some of the reasons for using NoSQL databases are:

- **Simple design**

Mostly no need to join many tables together for a query like relational databases.

- **Horizontal scalable**

NoSQL databases can easily scale horizontally to the clusters of machines. Data is automatically spread across servers without requiring application changes (auto-sharding).

- **Unstructured data**

It can incorporate unstructured and semi-structured data, which means it is flexible to accommodate any new type of data at any point and is not disrupted by structure changes.

- **Speed**

Due to the use of JavaScript object notation (JSON) document-like data structures, many operations are faster in NoSQL than relational databases, as it does not require joining tables (but this is achieved at the cost of space because of data duplication). In fact, joins are not supported by most NoSQL databases.

- **Cost**

Opposed to relational database systems, which rely on expensive servers, and storage systems, most of NoSQL databases usually use clusters of cheap servers. Additionally many NoSQL databases are open source and therefore free.

### 2.2.2.1 Types of NoSQL databases

NoSQL is a family of databases that are all non-relational. Broadly there are four types of NoSQL databases:

1. **Key-value database systems**

These databases stores key values as pairs. In case an update is required, the entire value of a key has to be changed, as usually there are no fields to update. It is easy to store but could limit the complexity of queries. Examples are: Redis, Dynamo, MUMPS and MemcacheDB.

## 2. Graph database systems

The concept of these databases is same as of typical graphs in computer science terminology. They consist of edges and nodes. Nodes as well as their edges can store additional properties like key-value pairs. These databases lack scalability, as generally they require all data to be on one machine. Some examples of graph based database systems are Neo4j, OrientDB and InfiniteGraph.

## 3. Column database systems

Column based databases stores all the values of a particular attribute together on-disk, which makes retrieval of a big amount of a specific attribute fast. This could be useful when analytical such as range queries over a specific field are required. Some of examples of column based NoSQL databases are HBase, Cassandra and Accumulo.

## 4. Document database systems

Records are stored as documents in these databases. A document can be a key value pair. Keys are always strings, and values can be stored as Booleans, numeric, strings, arrays, and other nested key-value pairs. Each document has its own structure; they are not required to have the same structure like rows in a relational database table. Examples of document based database systems are MongoDB, Cloudant, Apache CouchDB, and Clusterpoint.

Some of the drawbacks of NosQL databases include large amounts of data redundancies due to the lack of relationships. Additionally NoSQL databases are based on CAP theory [29], which states that it is impossible for a distributed system to provide all the three features (given below) at the same time. The three features are

1. **Consistency:** Same data is visible to all the requests at the same time.
2. **Availability:** Every request will always get a response regardless if it succeeded or failed.
3. **Partition tolerance:** The system is always functional despite failures of part of the system.

When a user meets two of the three conditions, he fails to achieve third one.



## 2.3 Standard workflows for NGS data analysis

This section explains the standard steps taken for the analysis of next generation sequencing data. Due to the scope of the thesis, we focus here mainly on mRNA and sRNA expression analysis workflows as shown in Figure 2.4.

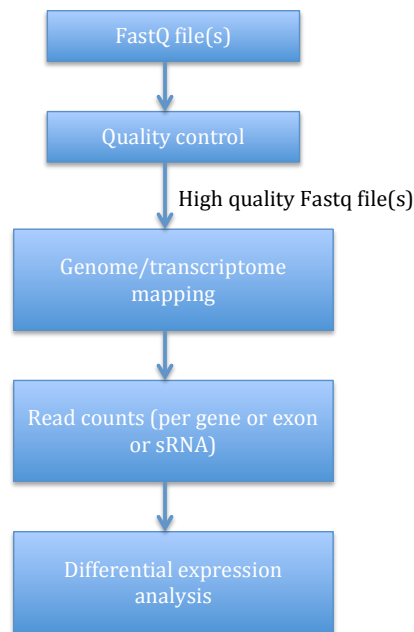


FIGURE 2.4: The main steps of the NGS (RNA-seq and sRNA-seq) analysis involves mapping of the FastQ files to the reference genome or transcriptome (sRNA-ome), followed by DE analyses for genes and/or exons or sRNAs.

### 2.3.1 Raw data (FASTQ)

Next generation sequencing data analysis starts with the raw data obtained from a sequencer, which is usually in FASTQ format. A FASTQ file stores both a biological sequence as well as its corresponding sequencing quality scores.

The FastQ format consists of 4 lines per read as shown in Figure 2.5,

- First line corresponds to the read name.
- Second line has the biological sequence represented as strings of A, C, G and T.
- Third line begins with a '+' and can be followed by the same sequence identifier as in first line or can also be used for any optional description.
- Forth and last line for the read corresponds to the sequencing quality of each base in the read.

A base quality score is the probability that the corresponding base is called incorrectly during sequencing. Phred quality score is used to represent these base qualities, and these probabilities are used to calculate overall sequencing quality.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((((***+))%%#+))(%%%).1***-+*'')**55CCF>>>>>>CCCCCCC65
```

FIGURE 2.5: Shown is the fastq format, which is output from many sequencers, first line corresponds to the read name, second line has the biological sequence, third line begins with a '+' and can be followed by any optional description and fourth line quality scores of each base in the read.

### 2.3.2 Quality control (QC)

One of the important and basic steps in NGS data analysis is quality control of the raw data. Before drawing conclusions from the data, it is important to know if the data can be trusted at all. There could be many issues with the data including both biological and technical errors such as mishandled samples, incorrectly followed protocols, sample contamination, high biological variance and sequencing errors. To this end various tools (FastQC) and methods (principal components analysis) have been developed.

#### 2.3.2.1 FastQC

FastQC [30] is a freely available tool and can be used to determine sequencing quality. As mentioned before a fastq file has the Phred quality scores that represent the probability of incorrectly calling a base. FastQC takes this file as an input and produces a basic summary that includes the quality encoding used by the sequencer, total sequences, sequences flagged as poor quality and sequence length. FastQC also provides many diagnostic plots for each input file (sample) such as per base, per tile and per sequence quality scores, per base sequence content, sequence length distribution, sequence duplication levels, overrepresented sequences (k-mers) and adapter content. All of these plots provide very detailed information on the quality of the sample file. These plots can be used to judge the overall quality of a sample. For example, per-base quality scores are shown for high quality (good) data (Figure 2.6a) and low quality (bad) data (Figure 2.6b). It is clear crystal that per-base quality drops a lot for the low quality data. One more example shown in (Figure 2.7a, 2.7b) is the 'per sequence quality scores' plots. These plots help us to see if a subset of a sample's sequences have overall low quality values. Cases where a major proportion of the sequences in a sample have overall low quality indicate a systematic problem.

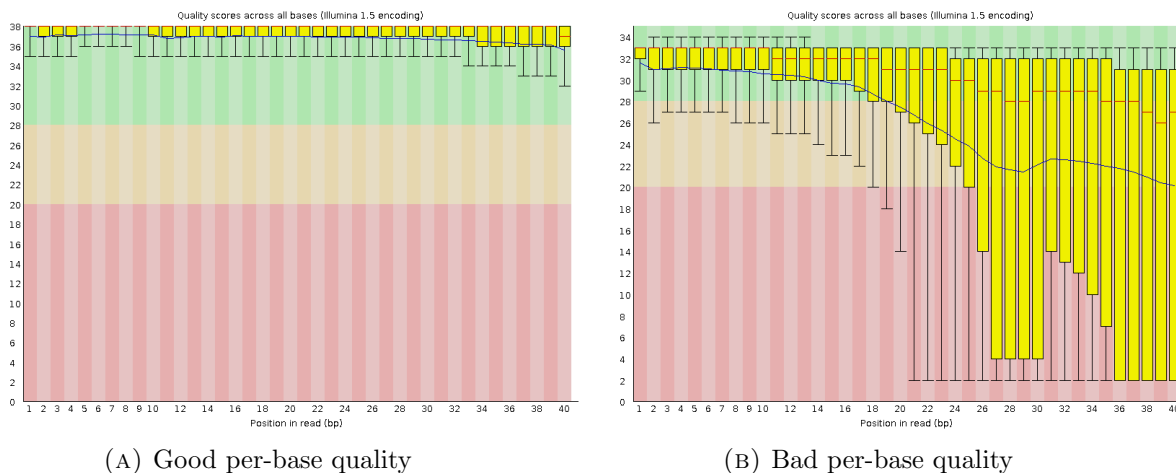


FIGURE 2.6: An overview of the range of quality values across all bases at each position. For each position a Box-and-Whisker plot is shown. Each base has a certain distribution of Phred scores from very low (red background), marginal (yellow) and high (green).

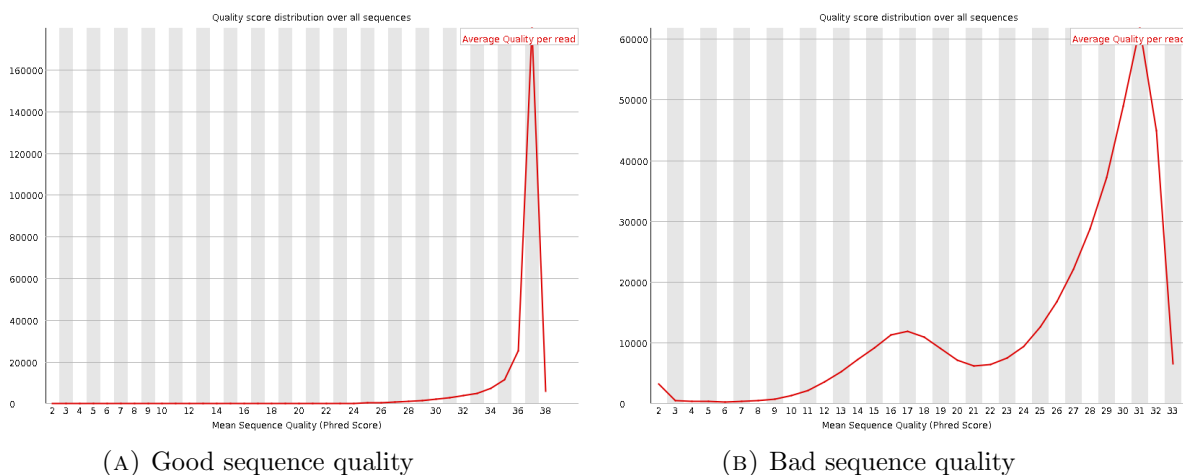


FIGURE 2.7: Shown are the average quality scores per sequence.

### 2.3.3 Adapter trimming

Once the QC is done and the data is of enough high quality to be considered for further analysis, the first step of most NGS analysis is adapter removal. As for library preparation, adapters are always ligated to every single molecule to be sequenced; therefore the adapters need to be removed before mapping to the reference genome or transcript-ome. There are several tools available for the adapter trimming including Trimmomatic [31], skewer [32], Trim Galore [33] and cutadapt [34]. All of these tools can be used for adapter trimming with different tool specific options, but mostly they vary in their speed and user friendliness.

### 2.3.4 Alignment and counting

The next common step in the analysis of mRNA, as well as sRNA, is the alignment of raw sequencing reads to the genome or transcript-ome or sRNA-ome. There are many tools available for the alignment of raw sequencing data such as GSNAP [35], MapSplice [36], RNA-Seq unified mapper (RUM) [37], Bowtie 2 [38] and STAR [39]. These aligners output SAM (sequence alignment map) [40] or BAM (the compressed BGZF format of the SAM file) file(s). The output of aligner has the information on each single read that is mapped (mapping locations, mapping quality etc) and some aligners also output the unaligned reads marking them as not aligned. Depending on the tool, the output can have much more detailed information. For example the output of STAR aligner produces a summary mapping statistics file (this file has information on the unique and multi-mapped reads, which could be very useful for quality control) and a SAM file for each sample. The SAM file has details on the genomic location of every single mapped reads along with the mapping quality.

Once the alignment is done, the next step is to summarize the mapped reads as counts for the desired features as sRNA, gene or even exon. The purpose of this is to make the further downstream analysis easy (that is to have small files with the required information only) and also many tools require these counts as input.

### 2.3.5 Differential expression (DE) analysis

Usually gene and sRNA expression sequencing experiments are performed to check quantitative changes in expression levels between different groups such as healthy versus cancer patients, wildtype (WT) versus knockout (KO) genes or even various disease or medical states. The purpose of such experiments is to identify genes or sRNAs that plays role in a particular condition such as cancer. Raw read counts can not be compared directly because there could be other factors involved in the difference of expression changed such as sequencing depth. Additionally to check if the variation is not just by chance, within group variation should also be considered. There are different methods already available that can be used to decide whether, for a given sRNA, an observed difference in read counts is significant or if this could also be seen just by chance due to random variation. Some of the most widely used methods for DE are edgeR [41] and DESeq2 [42] that are based on negative binomial (NB) distributions. These methods can be applied to test differential expression of sRNAs, genes as whole or even at the exon level.

## 2.4 Biological ontologies

It is very common in health registries to have terms that means the same or similar thing but written differently (e.g. stillbirth and fetal death) [43]. In order to be able to integrate and compare such data, one would need to know the semantic meanings of the terms. The field of computer science has established this, by using ontologies. An ontology define terms, their properties, and their relations. More formally the variables, concepts and their relationships is called an ontology. There are different ontology based systems available for biological terms such as the Ontology Lookup Service (OLS) [44] as shown in Figure 2.8. OLS provides latest biomedical ontologies at a single point of access. It can be accessed interactively via web interface as well programmatically through its API.

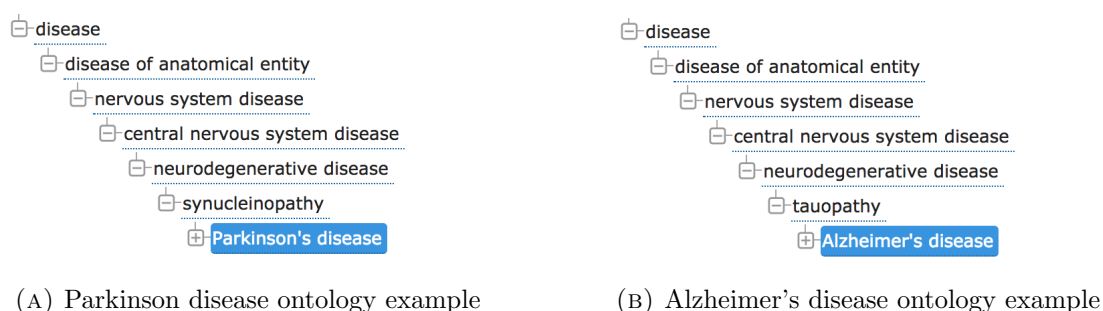


FIGURE 2.8: Shown are the ontologies for alzheimer and parkinson disease [44]. As can be seen in both A & B they share the path till neurodegenerative disease in the ontological order. It would be difficult to obtain both with single search term without the ontology association, but now one can just search for neurodegenerative disease and would get both of these diseases in the results and any other neurodegenerative diseases.

## 2.5 Principles of supervised machine learning methods

In this thesis we have used supervised machine learning methods for the biomarker detection in sRNA data and for disease prediction based on exon expression. Therefore in this section we will summarize some basic principles of supervised machine learning methods.

### Supervised machine learning

Supervised learning is the task of inferring a function from labeled data [45]. In order to train a supervised learning, a labeled dataset is partitioned into at least two sets, referred to as training and test data. The training data is a set of pairs; each pair has an input value(s) and a desired output value. A supervised learning algorithm infers a function from this training data and then this function is used for predicting output

for the unseen data also called as test data (has input value but no output value). The algorithm tries to predict an output for each of these unseen input data based on the function that was learned from the training data.

An illustration of supervised learning is shown in Figure 2.9. The supervised machine-learning problem can be either classification (categorical value dependant variables) or regression (continuous value dependant variables). We used only classification algorithms in this thesis, so we will discuss only about classification.

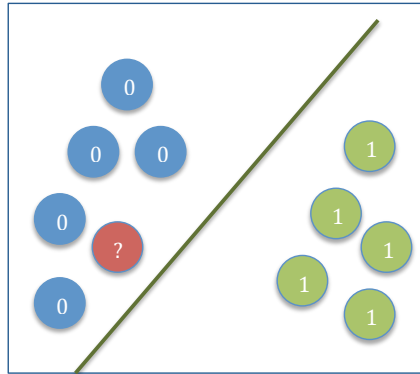


FIGURE 2.9: Supervised learning, the model (green line) is learned based on the 0 and 1 training examples and the unknown instance without a known class label (red circle) is classified as 0 according to the model

### 2.5.1 Classification

When the task of a machine-learning algorithm is to predict a category of unseen data based on the learned function from the training data, the task is termed as classification. When there are only two classes, it is called a binary classification or two-class classification, and when there are more than two classes to be predicted, this is known as multiclass classification. Handwritten digit recognition is a good example of multi-class classification, in which the objective is to assign each input vector (pixels from an image of a handwritten digit) to one of a finite number of discrete categories (0,1,...,9).

#### 2.5.1.1 Biological example

As an example, assume a set of  $N$  samples coming from healthy and individuals afflicted with a disease. Each sample has  $M$  features. The idea is to use these samples and design a system that predicts the condition of new samples (disease or healthy) that do not belong to the initial set of samples. A machine learning classifier is a type of algorithm that has been specifically designed for a task just as the one explained above, determine whether a new sample belongs to a set of mutually exclusive classes: healthy or diseased.

Machine learning algorithms are rules that need to be adjusted or trained based on the presence of evidence, which in this case could be  $N$  samples coming from previously attended healthy and patients afflicted with a disease. Once trained, the algorithm is ready to be tested on the new patients just referred; where the test consists of making a prediction: healthy or diseased. This last procedure is referred to as testing phase because the tested patient does not belong to the initial set of  $N$  patients. Some of the most widely used classification algorithms are support vector machines (SVMs) [46], K-star ( $K^*$ ) [47] and random forest [48].

### 2.5.1.2 Random forest

Random Forest is an ensemble method [48] based on the classical decision tree, where many decision trees (the forest) are produced. Each tree is given a randomly sampled subset (with replacement) of the data - hence the name random forest. As in real life the more the number of trees in a forest, the more robust is the forest. Similarly in the random forest classifier, increasing the number of trees tends to increase the accuracy. In brief random forest selects  $k$  features (randomly) from  $m$  total number of features. It constructs a decision tree on each subset of data. The above two steps are repeated in order to create  $n$  number of trees. At the end, each decision tree provides class prediction for a particular input, and random forest considers the highest voted predicted class for that instance. An illustration of random forest algorithm is shown in Figure 2.10.

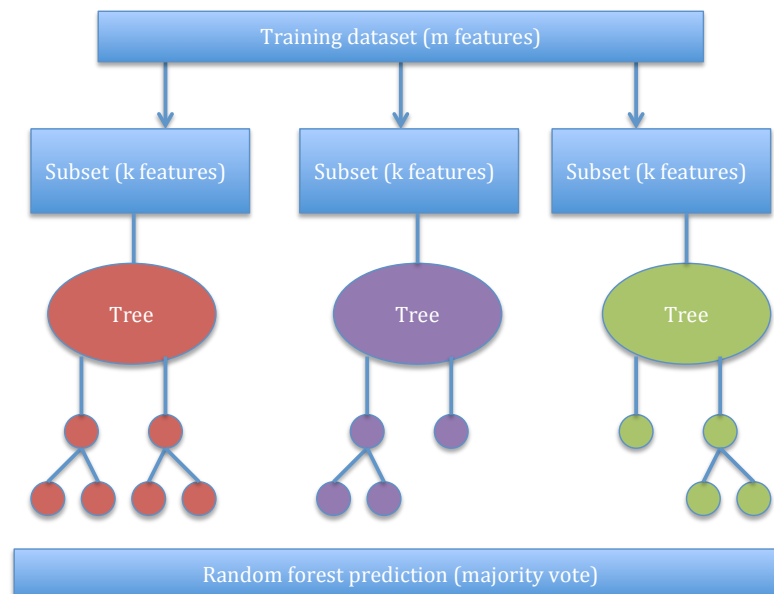


FIGURE 2.10: Illustration of random forest algorithm : Random forest selects  $k$  features (randomly) from  $m$  total number of features. It constructs a decision tree on each subset of data. The above two steps are repeated in order to create  $n$  number of trees. At the end, each decision tree provides class prediction for a particular input, and random forest considers the highest voted predicted class for that instance

### Feature importance

In many applications, it is not only important to obtain good classification performance, but also to determine the features that were relevant the most to make a prediction. Resorting once again to the example given above, a question to answer would be: what are the genes that helped the machine-learning algorithm determine whether a person (sample) is healthy or afflicted with a disease? There is extensive literature in the field of feature selection in the statistics and machine learning, but for the purposes of this thesis, a very common strategy to take is to train a classifier on several rounds using subsets of the original feature set  $m$  and evaluating an optimality function, such as the misclassification error. Then, after trying out all possible feature combinations, the selected subset of most important features is the one that optimized the misclassification error. In the case of random forest classifier the feature selection method is embedded within the training procedure, so no additional processing is required. The importance of a feature is usually estimated by computing the information gain of including an additional feature  $m_i$  into the classifier or by means of the gini index. In a real-life application, a threshold is set so that the only features kept to train the classifier are those whose information gain or gini value lies within the threshold.

## 2.6 Thesis related existing resources and research

This section describes the existing work that is related to this thesis. In addition this section briefly mentions the available resources that were used in this thesis.

### 2.6.1 sRNA-seq analysis tools

sRNA-seq is the current method of choice for the quantification of the genome-wide sRNA expression landscape. There are several local, as well as server-based, sRNA-seq analysis workflows available that differ in their analysis portfolio, performance, and user-friendliness. Some of the sRNA-seq analysis tools are described in this section.

#### 2.6.1.1 sRNA workbench

sRNA workbench [49] is an interactive pipeline for the quantification of sRNAs. This tool is able to perform quality checking and normalization of sRNA samples and to detect differentially expressed sRNAs. Additionally it can also be used for the detection of novel miRNA in the sequencing data.



### 2.6.1.2 CAP-miRSeq

CAP-miRSeq [50] is a tool that can be used for the quantification of known and novel miRNAs including variant calling and subsequent differential expression analysis. It also supports data visualization.

### 2.6.1.3 omiRas

omiRas [51] is a web server that supports the quantification, differential expression and interactive network visualization of ncRNAs. It provides users with static annotation results such as mapping statistics, quantification tables, read length distribution, differentially expressed sRNAs between differential experimental groups and also provide an interactive network of user selected miRNAs and their target genes.

### 2.6.1.4 mirTools 2.0

mirTools 2.0 [52] is a web server that can profile different ncRNAs such as snoRNA, snRNA, tRNA, rRNAs and piRNAs. It also supports functional annotation of microRNA targets genes. Additionally this tool not only supports the detection of novel microRNAs but it also detects novel piRNAs. On the other hand, mirTools 2.0 can be used for the identifying differentially expressed ncRNAs between experimental groups.

### 2.6.1.5 MAGI

MAGI [53] is another web application for the quantification and differential expression of miRNAs as well as for the prediction of miRNA target genes. MAGI provide results in an interactive web report. Additionally MAGI reports many diagnostic plots that can be used for quality control.

### 2.6.1.6 Chimira

One of the latest and widely used tool that allows for the detection of miRNA edits and modifications is Chimira [54]. It also supports differential expression of miRNAs.

### 2.6.1.7 sRNAtoolbox

Another recent addition to the sRNA-seq web applications is sRNAtoolbox [55]. sRNA-toolbox is a set of interconnected, independent modules for the analysis of sRNA-seq

data. It allows for the expression profiling, differential expression, target gene prediction and visual exploration of sRNAs. It also supports the identification of non-host organism reads by performing a blast search of all the unmapped reads (reads not mapped to the host organism). All of these modules can be used independently as well.

## 2.6.2 sRNA expression databases

sRNA annotation databases and knowledge repositories are freely and publicly available for quite sometime. The most widely used ones are miRBase (miRNA), piRNA-Bank (piRNA), snoopy (snoRNAs) and ensemble (snRNAs, snoRNAs, and rRNAs), but there are very few repositories to store sRNA expression data. Recent additions are miRmine [56], DASHR [57], miratlas [58] and YM500v3 [59].

### 2.6.2.1 miRmine

miRmine [56] is publicly available database of human miRNA expression profiles. miRmine contains the expression profiles of different publicly available miRNA-seq datasets and information about the different miRNAs expression profiles across different tissues and cell lines. Users can search for a single or multiple miRNAs across a particular tissue or a cell line. Additionally users can browse for all the expressed miRNAs in a tissue or cell line.

### 2.6.2.2 DASHR

DASHR [57] incorporates human small RNAs and their annotation. DASHR provides expression profiles of different ncRNAs (miRNAs, piRNAs, snRNAs, snoRNAs, scRNAs, tRNAs and rRNAs) across different human tissues. To date DASHR has 48000 sRNAs, 82% of them are expressed in one or more tissue types.

### 2.6.2.3 Miratlas

Miratlas [58] incorporate miRNA expression profiles and modifications from already published datasets along with its description (could be any associated term such as disease or tissue). Users can browse expression profiles by the description or by the miRNA name. Additionally users can search a dataset and download the expression of all miRNAs in that particular dataset.

#### 2.6.2.4 YM500v3

YM500v3 [59], is a database that contains expression profiles of miRNAs, snRNAs, snoRNAs, sncRNAs, piRNAs and RFs (tRNA-derived fragments) for various cancer studies. YM500v3 hosts over 11000 cancer samples. It only supports cancer datasets and no other disease types. Varieties of search and analysis options are available on YM500v3 web server including a search on sRNA for expression profile, search by cancer name and survival analysis options.

#### 2.6.3 Mutually exclusive splicing of exons

Mutually exclusive splicing of exons (MXEs) is a mechanism of functional gene and protein diversification with important roles in organismal development and diseases. The current knowledge of human mutually exclusive exons (MXEs) is very limited, that is to say that currently in human the number of MXEs range from 118 [60] to at most 167 cases [61]. Accordingly, the human genome annotation (GenBank v. 37.3) contains only 158 MXEs in 79 protein-coding genes. Although only a limited number of human MXEs is reported to date, MXEs have been shown to play a role in many essential human genes such as in SNAP-25 as part of the neuroexocytosis machinery [1]. Additionally mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the CACNA1C gene) [2, 3].

### 2.7 Goals of the Thesis

This thesis covers a very broad range of bioinformatics methods from analysis pipeline development to the data integration and development of expression atlas (database and web application development). In addition, this thesis shows very nice usage of bioinformatics analysis for the genome annotation and predicting diseases based on the expression profiles. In brief the main goals of the thesis are discussed below.

#### 2.7.1 Online analysis of small RNA deep sequencing data (Oasis)

As discussed in Section 2.6.1 many good web platforms for the analysis of sRNA-seq data exist, but some important analysis features still needs to be integrated. For example, no current web application allows for the identification of biomarkers of disease via integrated machine learning modules. Additionally, except sRNAtoolbox explained

in Section 2.6.1.7, there is currently no web application that allows detection of viral and bacterial infections in sRNA-seq samples, and detection of potential cross-species miRNAs.

Many of the aforementioned tools in Section 2.6.1 predict novel miRNAs, but none of these tools store this information (integrated into the miRNA-ome), which means this information is lost and the same miRNA might be predicted several times in different datasets. Additionally the user would have to define a local name according to their research as most of the prediction algorithms assign some random number with the genomic coordinates of the predicted miRNAs.

Finally, current sRNA-seq web services do not allow for automated analysis or batch submission of jobs via an API, a feature that could greatly facilitate analysis workflows for frequent users.

To this end, this thesis aims to develop an analysis pipeline for sRNA-seq data. The goal is to provide many different sRNA-seq analysis options over the web on a single platform, such as quantification of different sRNA species, prediction and storage of novel miRNAs with proper universally accepted nomenclature, identification of infection or contamination, differential expression analysis between different conditions of an experiment, identification of biomarkers for disease in the sRNA-seq data. User should be able to perform all these different analysis over the web application, as well as should be provided with API for automatic submission.

### 2.7.2 sRNA expression atlas (SEA)

As discussed in Section 2.6.2 there are some new and functionally well equipped additions to the sRNA expression profiles databases. Although there are still certain limitations of the sRNA expression repositories mentioned in Section 2.6.2. For example, the only database that supports more than one organism is miratlas (human and mouse), the rest focus on human sRNAs only. Two of the four mentioned databases (DASHR and YM500v3) have information on five types of sRNAs whereas the other two have information only on miRNA. DASHR and YM500v3 have information on different sRNA species but they are limited to human only or to a particular disease in the case of YM500v3. Except for YM500v3 no other databases stores expression profiles of novel predicted miRNAs. None of these databases provide search by ontology, for example to search for “neurodegenerative disease” and get all the samples in the database that are related to the term such as Alzheimer, Huntington or Parkinson’s disease as explained in Section 2.4.

Moreover, the number and scope of the currently available sRNA-seq data repositories do not reflect the recent attention that sRNAs has obtained in the recent years.

Considering these shortcomings, one of the main goals of this thesis is to build an sRNA expression repository to store sRNA expression profiles along with the experimental details such as organism, tissue, cell type, disease, age, gender and technical details like sequencer, kit and barcode etc. Additionally to build a web application that allows for the search of known and novel small RNAs across different organisms using standardized search terms and ontologies. The user should be able to query and visualize sRNA expression profiles across different tissues, cell types, and diseases in an interactive manner.

### 2.7.3 Mutually exclusive splicing of exons

Despite the important roles of mutually exclusive splicing in organismal development and diseases as mentioned in Section 2.6.3, only limited number of human MXEs is reported to date. To this end, the third major goal of this thesis is to build a method to predicted and subsequently validate mutually exclusive exons (MXEs) in the human genome.

## Chapter 3

# Results, Discussion and Outlook

In this chapter we will explain the development of Oasis 2 (online analysis of small RNA-seq data) including different modules such as sRNA detection, differential expression, classification and search modules as well as the development of small RNA expression atlas (SEA) database system and the front end for the end users. Additionally this chapter also explains the prediction of mutually exclusive splicing of exons (MXEs) and their role in disease and development.

### 3.1 Online analysis of small RNA-seq data (Oasis 2)

One of the main applications developed, as part of this thesis is Oasis (online analysis of small RNA deep sequencing data) and its second major release Oasis 2. Oasis 2 is a web application that allows for the fast and flexible online analysis of sRNA-seq data. Oasis 2 is intended for the end users in the laboratories, providing an easy-to-use web frontend including video tutorials, demo data, and best practice step-by-step guidelines on how to analyze sRNA-seq data. In this section we will highlight the main features of Oasis 2.

#### 3.1.1 Oasis 2's module

There are four main modules of Oasis 2 as shown in Figure 3.1, sRNA detection, differential expression, classification and Oasis-DB module. sRNA detection is the first analysis module of Oasis 2, and the rest of the three modules are dependant on the output of the sRNA detection module. It examines sample qualities, as well as quantifies known and novel sRNAs for each submitted sample. Oasis 2 allows for the upload of raw or

compressed FASTQ files. Once the data is submitted to Oasis 2's sRNA detection module. It performs several alignment steps and quantify sRNA molecules in the input data (sample). The sRNA detection module produces several diagnostic plots as explained in Section 3.1.3 and counts for each molecule of sRNA. These counts can be used as input to DE module to perform differential expression analysis between different experimental conditions (such as healthy and cancer patients) and to classification module to identify sRNA molecules that are distinguishing the two experimental conditions.

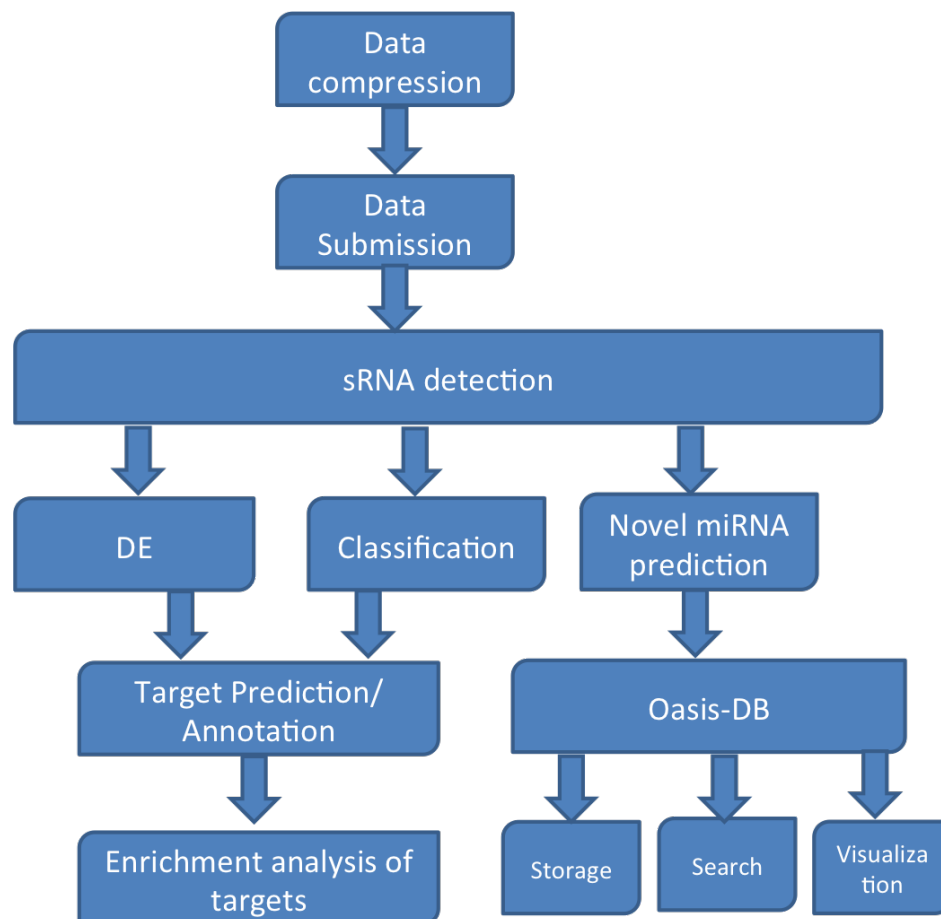


FIGURE 3.1: Oasis 2 modules and workflow : There are four main modules of Oasis 2, sRNA detection, differential expression, classification and Oasis-DB module. Oasis 2 allows for the upload of raw or compressed FASTQ files. Once the data is submitted to Oasis 2's sRNA detection module. It performs several alignment steps and quantify sRNA molecules in the input data (sample). These counts can be used for DE and classification modules. During the sRNA detection module Oasis 2 also predicts novel miRNAs which are stored in Oasis-DB and can be searched later on.

Key features of Oasis 2 are:

- **Multi-step alignment:** Alignment of the input sample's reads is split into several steps. The reason for splitting alignment into several steps such as to miRNAome, sRNAome, genome, pathogen genomes is to assure maximum annotation.

- **Speed:** The new sRNA detection workflow is faster compared to its predecessor Oasis as well as the latest additions for sRNA-seq analysis tools as shown in Table 3.1.
- **Model and non-model organism support:** It supports the analysis of any model and non-model organism. In case the user's organism of interest is not one of the 14 genomes available in Oasis 2, reads can be aligned to all novel predicted and known miRNAs.
- **Predict novel miRNAs:** It supports the prediction of novel miRNAs in the input samples using miRdeep2 [62]. It passes these novel miRNAs to Oasis 2's "Search module", which stores these in Oasis-DB with universally accepted nomenclature.
- **Infection or contamination detection:** One of the most useful features of Oasis 2 is the support to detect pathogenic contamination or may be potential pathogenic infections in the sRNA-seq samples.
- **DE module:** The second module of Oasis 2 is the DE Analysis module. It calculates differential expression of sRNAs between different experimental conditions, makes target annotation and prediction for miRNAs, and lastly provides functional analyses of miRNAs as explained in Section 3.1.4. Apart from the DE analysis between two groups, Oasis 2 also supports multi-group comparisons as well as can take into account the covariate information (such as age and gender) if available. Additionally it provides the users with downloadable, interactive web reports.
- **Classification module:** The third module of Oasis 2 is the classification module. In order to detect sRNA based biomarkers a binary classification using random forest is applied to the sRNA expression profiles. Oasis 2 classification module is augmented with sample balancing and feature pruning routines and it reports all the four possible models for a single analysis such as balanced/unbalanced with features optimized/non-optimized. The user can select any combination of the models in the interactive web report. This module also provides end user with downloadable and interactive web reports, which includes classifier's performance measures, feature importance along with predicted and known miRNAs as well as their targets that allows for the functional analyses of miRNAs as explained in Section 3.1.4.
- **Search module:** The fourth module of Oasis 2 is the search module. As explained above that Oasis 2 predicts novel miRNAs using mirDeep2 [62]. In order to enable users to search and retrieve these predicted miRNAs, Oasis-DB was developed. Oasis-DB not only stores novel predicted miRNAs by Oasis 2, but it also stores all miRBase [63] miRNA entries. In case a miRBase update (new release in future)



contains a miRNA with an identical location of a predicted miRNA in Oasis-DB, it will be automatically linked to its corresponding miRBase entry. Currently Oasis-DB contains over 700 (769) high-quality predicted miRNAs across 14 organisms. The user can search by the mature/precursor id, precursor sequence, mature sequence and reference genome or their combination. The search results shows the mature and precursor IDs of the miRNAs, their type (novel or known), the mature sequence and its genomic coordinates (chromosome, start, end and strand), the structure file (for novel miRNAs only) and the organism. Additionally the user can select a miRNA for more details. For novel miRNAs, a new page with additional details such as precursor sequence and its genomic coordinates and structure of the predicted novel miRNA (produced by mirDeep2 [62]) is shown, where as for a known miRNA, Oasis 2 redirects the user to the miRBase record for that miRNA.

- **Batch job submission (API):** Oasis 2 provide users with an API to submit multiple jobs automatically (via programs) or to submit jobs from remote systems. An API is available for the sRNA detection, DE and classification modules.

Demo Dataset	Oasis 2 <sup>1</sup>	Oasis <sup>1</sup>	MAGI	Chimira	omiRas	mirTools 2.0 <sup>7</sup>
AD (287 GB)	8 h31m50s	12h29m12s	NA <sup>2</sup>	NA <sup>4</sup>	NA <sup>5</sup>	NA
Psoriasis (48 GB)	1h35m17s	5h49m4s	48h <sup>3</sup>	3h3m12s	NA <sup>6</sup>	NA
Renal Cancer(9 GB)	31m43s	1h8m41s	8h <sup>3</sup>	47m11s	9h31m	NA

TABLE 3.1: Runtime comparison of different sRNA-seq web applications: <sup>1</sup>Run time estimate includes the data compression by OasisCompressor and decompression on Oasis 2 server side, the sRNA Detection, DE Analysis, and Classification. <sup>2</sup> MAGI failed to upload all AD files to server, may be it has a problem with the format or quality of one of the files. <sup>3</sup> These values were obtained from the MAGI website. <sup>4</sup> We could not estimate runtime for the AD dataset by Chimira as it can not analyse more than 25 files at a time. <sup>5</sup> We were not able to upload all AD samples to omiRas. <sup>6</sup> omiRas http uploading error. <sup>7</sup> As maximum file size to upload for mirTools 2.0 is limited to 30 Mb, we therefore were not able to estimate its runtime on these datasets. Table modified from Article A of this thesis.

All these features are explained extensively in the Article A of this thesis, therefore in this section we will focus on some of the secondary yet important features such as data compression, QC reports (outlier detection) and enrichment analysis of Oasis 2.

### 3.1.2 OasisCompressor

As Oasis 2 is a web application and the input for the sRNA detection module is raw sequencing data from the sequencer that usually are FASTQ file(s) and their size can vary from few megabytes (MBs) to gigabytes (GBs). Additionally some experiments have many samples (fastq files); therefore, depending on the size of this data, the upload of data to Oasis 2 server can take from minutes to hours. Extended uploads can easily

fail due to network connection issues. To this end we developed a desktop application (provided by Oasis 2) called as OasisCompressor as shown in Figure 3.2 that extracts quality metrics and compresses the read information from FASTQ files. The data compression depends on sequencing depth and sample (fastq file) entropy. For example, OasisCompressor reduced the size of Alzheimer disease dataset [64], that has 70 FASTQ files from 287 GB to 0.33 GB, achieving around 800 fold compression of the data. In brief we used a hash table to store the sequence with its frequency in the sample, where sequence is the key and its frequency (count) is the value. Depending on the entropy of the data the compression rates are different but as small RNA are short and repetitive sequences, the compression rate is mostly very high. For example psoriasis dataset [65] with 20 samples was reduced from 48 GB to 0.19 GB. Similarly renal cancer dataset [66] with 22 samples was reduced to 0.15 GB from 9 GB. Additionally, OasisCompressor also takes care of the sequencing quality information to be passed to the server, that is required by FASTQC [30]. The following quality information are extracted from the raw fastq files while compressing and then sent to the server as precomputed statistics:

- Mean and median quality score for every position in the sequence.
- Frequency of quality scores in the whole FASTQ file.
- Quantile percentiles of quality score for every position in the sequence length: Lower quartile (0.25), upper quartile (0.75), 10th quartile (0.1), and 90th quartile (0.9).

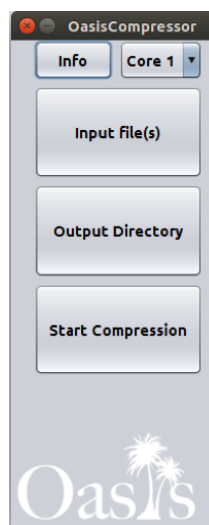


FIGURE 3.2: Two mandatory fields are the selection of FastQ files by pressing on the ‘Input file(s)’ button and selection of the output directory, an optional field is the selection of number of parallel processes for OasisCompressor

In order to produce diagnostic plots (through FASTQC) for the Oasis 2 compressed data, FASTQC was customized to be able to take the already calculated quality matrices and plot them.

### 3.1.3 Quality Control (QC)

Quality assessment of sRNA-seq samples is the most critical step for performing downstream analyses, as keeping low-quality samples in the downstream analyses (Classification and DE Analysis) might generate poor or not trustworthy results. Therefore all the analysis modules of Oasis 2 produce certain diagnostic plots and statistics such as mapping percentages, unique mapping percentages, percentages of different sRNA species in the sample such as miRNA, piRNA and snoRNA and principal component analysis (PCA) plots. All these QC plots, statistics and sRNA expression profiles are presented as interactive downloadable reports to the end user in the form of html pages as shown in Figure 3.3.

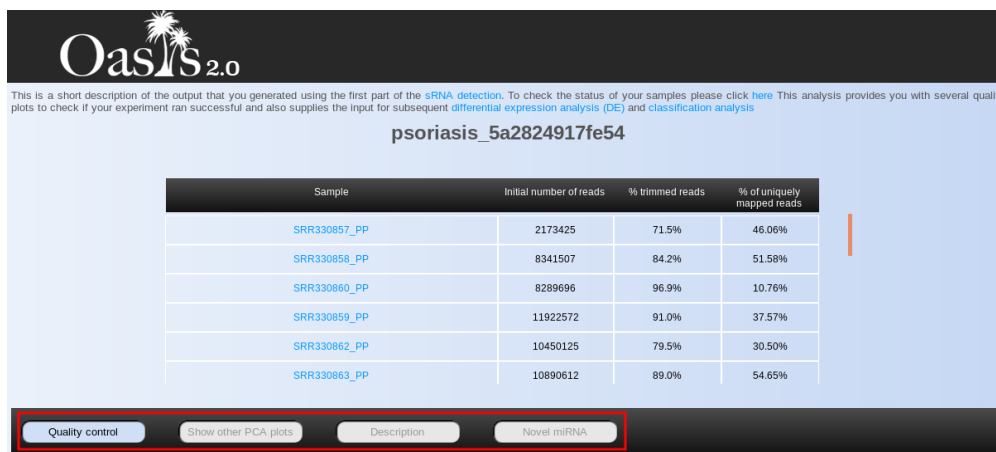


FIGURE 3.3: Browser view of the primary output of sRNA detection module. The different sub-pages with different result views can be reached via the menu marked with a red square. Shown in the table are the total number of reads, percentage of trimmed reads (adapter trimming), and percentage of uniquely aligned reads per sample. The user can click on a particular sample in the table and browse the sample specific QC plots and sRNA counts

Oasis 2 produces a comprehensive report on the QC of all the samples at different level. Some of the main diagnostic plots produced by Oasis 2 are explained below :

- Principal component analysis (PCA) plot for all samples based on sRNA expression profiles.
- Interactive bar plots for each sample to show how many reads are filtered for being too short or too long based on the minimum and maximum read length filter.

- Interactive bar plots for each sample to show percentage/number of reads mapped per sRNA species such as miRNA, piRNA, snoRNA, snRNA and rRNA.
- Interactive bar plots for each sample to show initial total number of reads, number of reads after adapter trimming and length filtering and number of uniquely mapped reads to the reference genome.
- PCA plots for different sRNA species (miRNA, piRNA, snoRNA, snRNA and rRNA) using expressions of read belonging to a particular sRNA species.
- All the plots produced by FASTQC for each sample such as per base sequence quality, per sequence quality scores, per base sequence content, sequence length distribution and k-mer content.
- PCA plots for all samples in DE and classification modules colored by groups to detect outliers in the data.

An example of Oasis 2 QC is shown in Figure 3.4. The Psoriasis dataset [65] seems to contain an outlier (SRR330860\_PP) and a mis-labelled (SRR330866\_PP) sample. The removal of these two samples from the Psoriasis dataset increased the classification accuracy from the AUC of 0.9 to 1 and increased the number of significantly DE miRNAs from 195 to 256 cases, providing strong evidence for the utility of Oasis 2' QC plots.

### 3.1.4 Functional enrichment analysis

In order to make sense from a list of sRNAs that are differentially expressed or potential biomarkers, functional enrichment analysis is required. To this end both DE and classification modules of Oasis 2 enable users to perform functional enrichment analysis for miRNAs. These analysis can be performed on the fly from the interactive web reports that the user obtained from Oasis 2's DE or classification modules. Currently Oasis 2 supports functional enrichment analysis for miRNAs based on their gene targets. This feature of Oasis 2 allows the investigation of specific biological functions for selected miRNAs. In order to perform functional analysis the user has to select a single or multiple miRNAs, choose target types (only validated targets, predicted targets or both) and lastly enrichment analysis tool(s) mentioned below. Currently Oasis 2 supports the following enrichment analysis tools:

- **gProfiler** : [67] Provides users with enriched gene ontology (GO) categories, REACTOME and KEGG pathways, TRANSFAC regulatory motifs, Human Phenotype Ontologies, BioGRID protein-protein interactions and CORUM protein complexes for the targets of selected miRNAs.

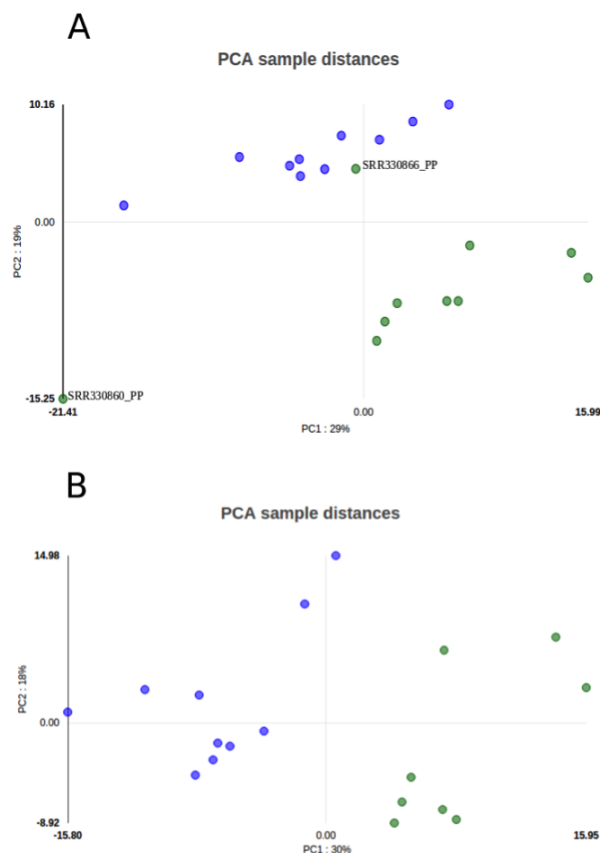


FIGURE 3.4: Shown is the PCA plot for sRNA Psoriasis data [65]. PCA sample distances of psoriasis (green) and control (blue) is shown. (A) In this PCA an outlier sample (SRR330860\_PP) and a potentially mis-annotated (SRR330866\_PP) can be seen. (B) PCA of psoriasis and control samples without misclassified/outlier samples. Removal of the aforementioned two samples increased the number of significantly (adjusted p-value <0.1) DE miRNAs from 195 to 256 cases and increased the AUC from 0.9 to 1 in the classification module, providing strong evidence for the utility of Oasis 2' QC plots. Figure taken from Article A of the thesis

- **Genemania** : [68]: Provides users with enriched GO categories and additionally it also returns a protein-protein network, showing the protein products of the selected gene targets and how they associate with each other.
- **STITCH** : [69] For the the gene targets of user selected miRNAs, STITCH computes protein-protein networks and generate a single image of all the interacting proteins. Additionally it also includes small molecules, drugs and ATPs associated with the target proteins as well.
- **STRING** : [70] Like STITCH it also computes protein-protein networks and present it as a single image for all the interacting proteins.
- **DAVID** : [71, 72] An enrichment test is performed for all target genes, using various functional annotations (GO categories, KEGG pathways, BIOCARTA pathways, protein domains etc).

In brief, we developed Oasis 2, a web based application for the fast and flexible analysis of sRNA-seq data. Its major features include sRNA detection (in any organism), multivariate DE, classification, functional/enrichment analysis, detection of pathogenic infections and contamination, search for novel and known miRNAs (14 organisms) and an API that supports the batch submission of jobs from the command line.

### 3.2 Small RNA expression atlas (SEA)

This section summarizes Article B of this thesis. This section provides an overview of the sRNA expression atlas developed in this thesis. The focus is to highlight the technical details such as architecture of SEA and the workflow for the data integration.

Small-RNA Expression Atlas (SEA) is a web application that allows for the querying, visualization, and analysis of over 2,500 published sRNA-seq expression datasets. SEA automatically downloads published sRNA-seq expression datasets and re-analyze them using Oasis 2. Additionally these sRNA-seq datasets and their respective samples are annotated with metadata; such as organism, cell line, cell type, tissue, disease, age and gender. Moreover, the available technical experimental data is also stored in SEA for each sample and dataset. SEA can be searched for sRNAs that originate from miRBase [63], ensembl [73] as well as from the repository of novel predicted miRNAs from Oasis 2 as discussed in Article A of the thesis. The major highlight of SEA is the powerful ontology-based search, and to our knowledge SEA is the only sRNA-seq database that supports ontology-based queries as explained in Section 2.4. It currently supports 10 organisms which is far more than the latest sRNA expression based repositories as shown in Table 3.2, and it is continuously updated with novel published sRNA-seq datasets and relevant sRNA information from various online resources.

Feature	SEA	miRmine	DASHR	miratlas	YM500v3
Organisms	10	1	1	2	1
sRNA types	5	1	5	1	5
Samples	>2000	304	187	461	>8000*
Novel miRNAs	+	-	-	-	-
Ontology search <sup>#</sup>	+	-	-	-	-

TABLE 3.2: SEA comparison with other sRNA-seq repositories : Comparison of SEA with latest publicly available sRNA expression databases based on a list of features we deem relevant. \*Supports mainly cancer-related datasets. # Use of ontological graphs for the annotation and querying of samples.

### 3.2.1 System design

Biological experiments vary in their experimental designs and also some experiments may have information such as tissue, cell line, disease while others may completely lack this. Due to this sparse nature of the biological experimental data we opted to use NoSQL database management systems as discussed in Section 2.2.2. MongoDB was used to store meta-information of the datasets and samples along with their expression profiles. Ontologies for organism, cell line, cell type, tissue and disease were stored in a NEO4J graph database. For sRNA genomic location, organisms and sequence, Oasis-DB was used. SEA system architecture is shown in Figure 3.5, and the workflow is as following:

- **Data acquisition:** SEA acquires raw SRA files of published sRNA-seq datasets and their primary annotation from Gene Expression Omnibus (GEO) and NCBI's Sequence Reads Archive (SRA) repository. A custom script was written in order to search and download sRNA-seq datasets.
- **Meta-information extraction:** Raw experimental annotations were obtained from GEO database in an automated manner. These annotations are textual and have therefore been parsed with certain keywords to obtain their values such as disease, tissue, cell line, age and gender etc.
- **Manual curation:** As explained above, an automated pipeline from GEO generates annotations. The sample characteristics in the GEO database are highly unstructured in terms of machine readability and the probability of false positive and true negative annotation is very high. To this end we developed an application with a graphical user interface to make annotations easier and faster for the manual curation team as explained in Section 3.2.2.
- **sRNA-seq analysis:** In order to allow for the cross study comparisons (sRNA expression across datasets), it was important to analyze all the sRNA-seq datasets with exactly the same parameters and tools. To this end we used Oasis 2 as explained Section 3.1 and Article A.
- **Data integration:** Once the datasets were downloaded, analyzed and annotated. The next step was to integrate them into SEA as shown in Figure 3.6. To this end an automated pipeline was developed. We took the advantage of Oasis 2's QC stats and plots. We only inserted high quality datasets into SEA, for example the ones that had high mapping percentages (50% of uniquely mapped reads).
- **SEA web application:** In order to enable users to search for the data stored in SEA, a web application for the end users was developed. The users can query

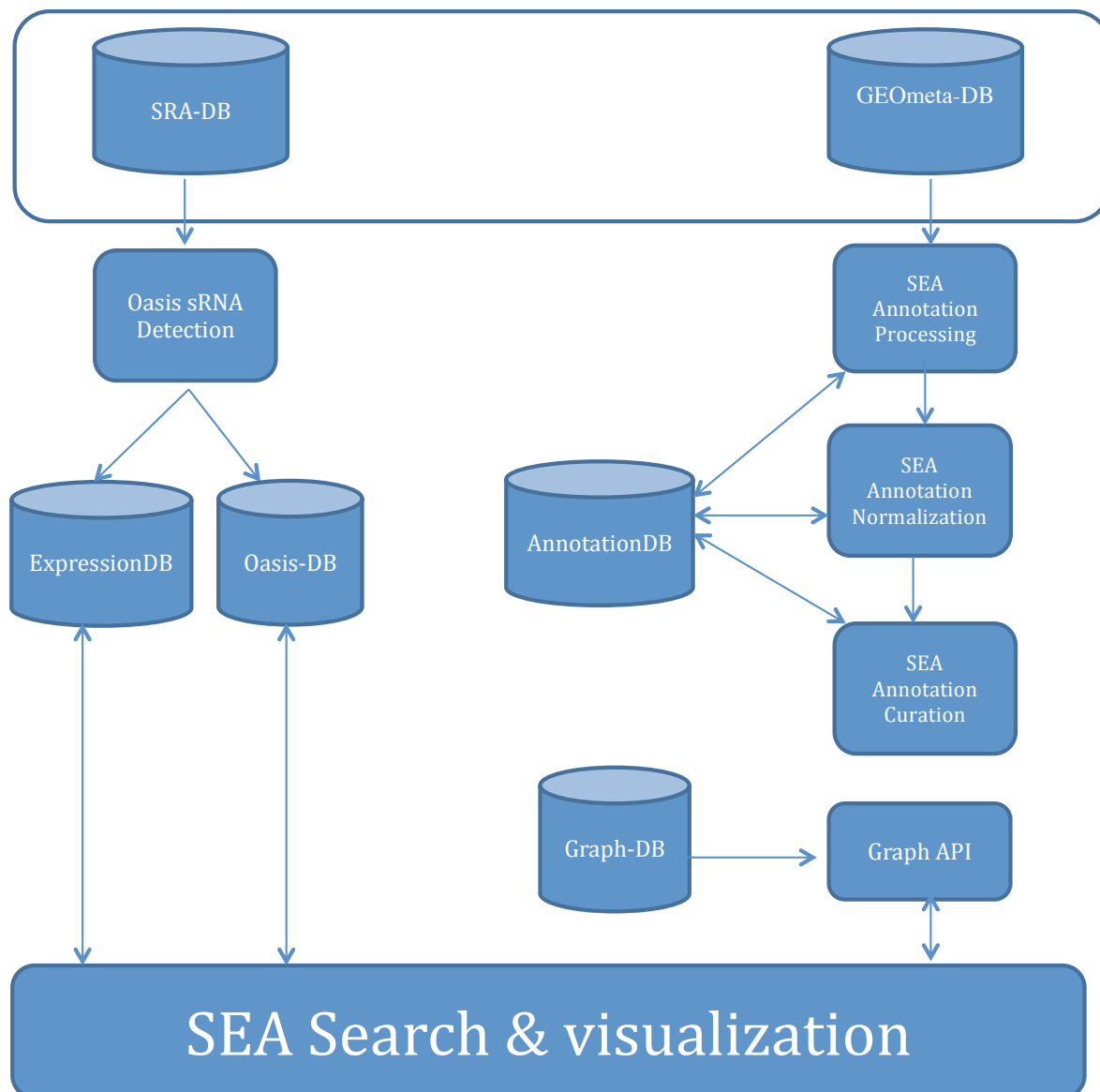


FIGURE 3.5: SEA system architecture: Raw sequencing data (from SRA) and their annotation (from GEO) are downloaded. Annotations are processed, normalized and stored to graph-DB for ontology based search. Raw sequencing data is analyzed with Oasis 2 and stored to ExpressionDB. SEA search and visualize data using these aforementioned databases.

expression of certain sRNAs across diseases, tissues, cell lines, cell types and even search for the experiments of their interests as explained in Section 3.2.3.

### 3.2.2 Annotation tool

We developed an in house tool for the curation of sRNA-seq sample's annotation. Once the curator logs in to the system, he/she can see all the datasets that are already annotated and also the ones that still needs to be annotated. The curator can click on



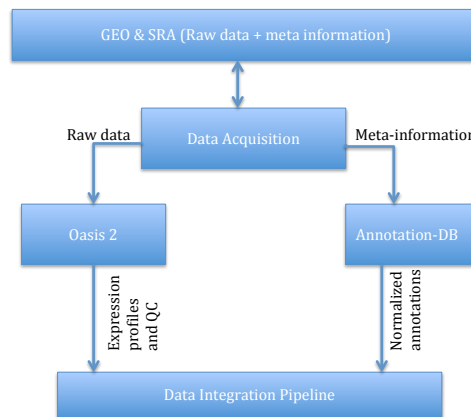


FIGURE 3.6: SEA data integration workflow : Datasets were downloaded, analyzed and annotated and then integrated into SEA. An automated pipeline was developed to check the quality of data (based on Oasis 2 QC) and check if the annotations are marked as complete by the curator. Finally the pipeline stores the datasets in SEA and notify in case there were some issues.

the dataset to annotate it and standardize/normalize its samples terms such as tissue, cell type and disease etc. There are several advantages of this annotation tool: it is easy and faster for the curator to standardize/normalize the terms by connecting them to ontologies from the drop-down menu as shown in Figure 3.7. It automatically changes the values for local annotations in case there is a change in global annotation, explained in Section 3.2.2.1. It visually shows which terms are standardized with green background and not standardized with red background as shown in Figure 3.7. Additionally this tool also keeps track of changes by each curator. Lastly the datasets can be marked as annotated, in order to make them available for search in the SEA database and application system.

### 3.2.2.1 Annotation criteria

Since ontology based annotations are at the heart of SEA, in this section we will briefly outline some basic rules for annotation of sRNA-seq samples.

#### Annotation rules

- Annotations can be defined as global-level or sample-level annotations. Global-level annotations are exactly the same for all samples of the dataset. Sample-level annotations differ among the samples of the dataset. Local annotation does not override the global one.
- In cases, where there are alternatives for a term annotation, we tried to be as specific as possible. For example, in case the sample is from breast fibrosarcoma and the term "breast fibrosarcoma" is available we will try to annotate it with the

FIGURE 3.7: Annotation tool : Shown is an example of one dataset annotation. The top panel is for global annotation and then are the per sample annotation. As shown the curator is annotating disease at the global level, the curator is suggested different ontologies with respect to what is typed in the dropdown menu and can select from it. The terms such as cell type, cell line that are normalized are green background and has an ontological key associated with them, where as the terms such as sample group has a red background color as they are not normalized

same term, although it can be annotated with breast cancer as well i.e. we choose the term deepest in the ontology tree.

- In cases where the relevant term cannot be found in an ontology, we tried to normalize the terms with synonyms or slightly less specific. The aim was to annotate as much as possible to have standard terms rather than just textual information.

### 3.2.3 SEA web application

Once the data was integrated, the next step was to provide a web interface for the end users around the world to be able to query expression of certain sRNAs across diseases or tissues and even search for the experiments of their interests. The home page of SEA is shown in Figure 3.8. SEA has one global search field that can be used to search expression of particular sRNAs across different or specific tissue(s), disease(s), cell line(s) or cell type(s). Alternatively it can be used to search for sRNA-seq datasets based on specific tissue(s), disease(s), cell line(s), cell type(s) or even organism(s). For example the user can ask questions such as: What is the expression of hsa-miR-100-5p across all human diseases? Is hsa-miR-200-5p expressed higher in alzheimer's disease as compared to cancer? Is the tissue-specific expression of hsa-miR-488-5p conserved in mouse? How many and which sRNA-seq datasets are available for cancer?

In summary, we developed SEA, a web application that allows for the search, visualiza-

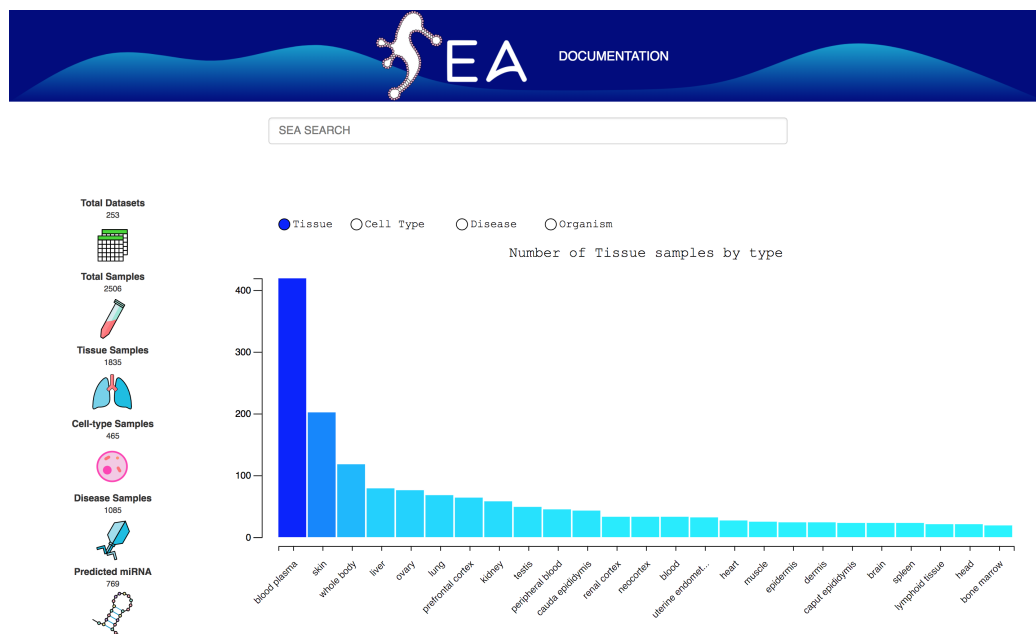


FIGURE 3.8: SEA home page: This page shows basic statistics about the SEA data repository such as total datasets, total samples, number samples with tissue, cell line or disease information. Lastly it also shows the number of predicted miRNAs (by Oasis 2) that are found to be in the published datasets that is they are expressed in the published datasets. The barplot shows number of samples per particular tissue, cell type, disease or organism.

tion and comparisons of known as well as novel small RNAs expression profiles (across ten organisms) using standardized search terms and ontologies for organism, tissue, disease, cell type, and cell line. Currently it contains expression and meta-information of over 2,500 sRNA-seq samples.

### 3.3 Mutually exclusive splicing of exons

This section summarizes Article C of this thesis. The goal of this project of the thesis was the prediction and validation of mutually exclusive splicing of exons (MXEs). Despite the aforementioned important roles of MXEs mentioned in Section 2.6.3, the current knowledge on the reported number of MXE is far from complete. To this end we build a method to predict and subsequently validate MXEs.

#### 3.3.1 Data sources

In order to predict and validate MXEs the following data sources were used:

- GenBank (v. 37.3) was used for human genome assembly and annotated proteins.

- For MXEs validation, data from 515 publically available samples comprising 31 tissues and organs, 12 cell lines and 7 developmental stages [74, 75, 76, 77, 78, 79] amounting to over 15 billion RNA-Seq reads was used.

### 3.3.2 Prediction of MXE candidates

In order to increase the current knowledge about MXEs in human, we decided to first predict a list of potential MXE candidates and then validate those using published RNA-seq data. Generally, MXEs are (cluster) characterized by splice-site compatibility, mutually exclusive presence in protein isoforms and genomic vicinity. In a first step, a set of MXE candidates from all the annotated protein-coding exons and from novel exons predicted in intronic regions was generated. Annotated exons were further filtered for those

- That appeared mutually exclusive in the transcripts.
- Neighbouring exons that have sequence similarity and are translated in the same reading frame.

In order to generate novel exon candidates, first novel exons were predicted in the existing intronic region based on sequence similarity and similar lengths [80] of the neighbouring annotated exons. Moreover, MXEs containing in-frame stop codons and exons overlapping annotated terminal exons were not included in the MXE candidates list.

As a result a set of 6,541 MXE candidates was obtained of which 1542 were protein-coding genes, including 1058 (68.6%) genes for which 1722 completely novel exons were predicted.

### 3.3.3 Validation of MXE candidates

In order to validate the predicted MXE candidates as explained in Section 3.3.2, more than 15 billion publicly available RNA-seq reads for different tissues, organs, cell lines and developmental stages were taken into account. Of 6,541, transcription of 6,466 (99%) MXE candidates were supported by RNA-seq reads mapped to the genome. However in order to be validated as true MXE, each MXE of a cluster needs to have MXE-bridging splice junction (SJ) reads to up or downstream gene regions in order to bridge the other MXE(s) of the cluster as shown in Figure 3.9. In addition, the MXEs should not have any SJ read amongst them (MXE-joining read should not exist), except for those leading to a frame shift and therefore a premature stop codon. According to these three restraints,

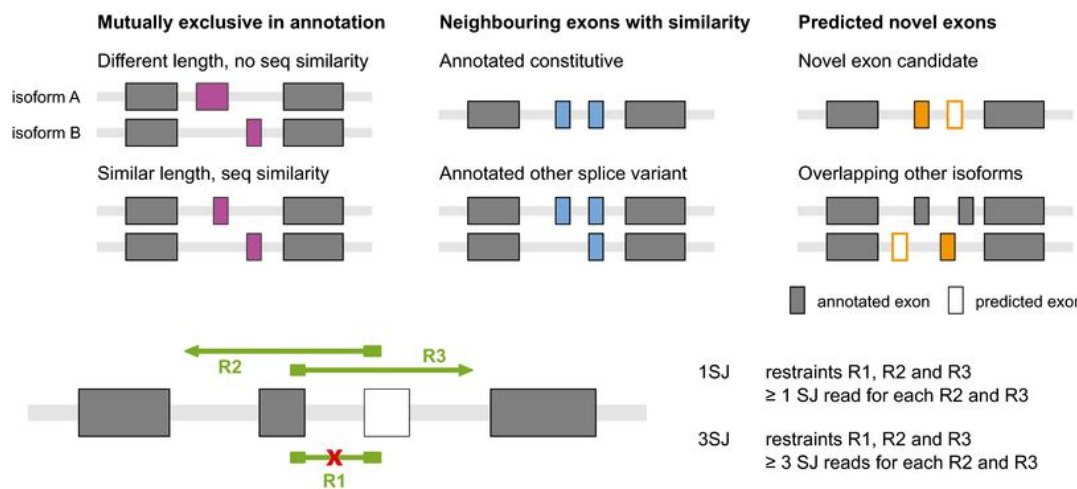


FIGURE 3.9: Illustration of different predicted and annotated exon types considered as potential MXE candidates. In order to be considered as true validated MXE, minimum of three restraints must be fulfilled: the MXEs should not be joined directly (the absence of an MXE-joining read (R1)), except for those leading to frame shift, and the MXE-bridging SJ reads (R2 and R3) must be present. Figure adapted from Article C of the thesis.

1,399 MXEs were validated with at least one SJ read per exon (1SJ), increasing the total number of human MXEs from 158 to 1,399. Moreover considering three splice junction reads per exon (3SJ) 855 MXEs were still validated. A comprehensive overview on the number of validated MXEs using different criteria is provided in the Article C of the thesis.

### 3.3.4 Spatio-temporal expression of MXEs

MXEs would need spatial and temporal splicing regulation and expression in order to modulate gene functionality. To this end, we performed a differential inclusion analysis using the above mentioned data sources 3.3.1. Of the 1,399 MXEs, 608 MXEs (345 unique genes) from Human Protein Atlas [79], 573 MXEs (389 unique genes) from Embryonic Development [78] and 552 MXEs (330 unique genes) from ENCODE datasets [75] are differentially expressed, respectively (adjusted P-value <0.05) as shown Figure 3.10. Interestingly, the differentially included MXEs comprise 43.5%, 40.9% and 39.5% of all MXEs showing that MXEs have tissue and developmental stage specific expression. As shown in Figure 3.10, many MXE clusters have one MXE which is expressed in specific tissues or at specific developmental time. This suggests spatio-temporal functional roles of MXEs in certain development and human diseases [74].

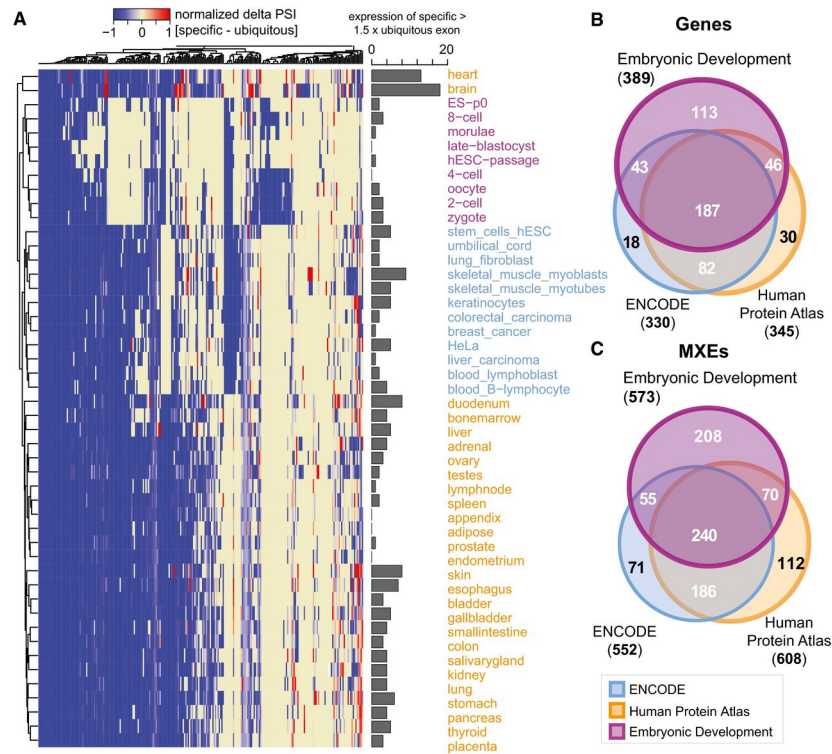


FIGURE 3.10: A. Shown in heatmap are the differentially expressed MXE clusters. Percent-spliced-in (PSI) value of the ubiquitous MXE is subtracted from the PSI value of the specific MXE and scaled between -1 (broad tissue distribution) and 1 (tissue specific). Gini coefficient is used to measure the inequality among values of a frequency distribution, where ubiquitous MXE is the minimum Gini index and specific MXE is the maximum Gini index of a cluster. Many MXE clusters have one MXE which is expressed at specific tissues or developmental time.

B & C. Overview of the differentially expressed genes/MXEs for the Human Protein Atlas, ENCODE and Embryonic Development datasets. Figure taken from Article C of the thesis.

### 3.3.5 Disease pathology prediction

As explained in Section 3.3.4 MXEs have tight developmental and tissue-specific regulation. This behavior of MXE expression can cause aberrant development and human diseases. In order to use MXE expression to predict disease pathology, all MXEs were annotated with pathogenic SNPs from ClinVar [81], resulting in 35 MXEs including 8 newly predicted exons. Of the 35 pathogenic SNP containing MXEs, 10 are associated to neurologic, 7 to neuromuscular, 6 to cardiac, 3 to cancer and 9 to other diseases based on their gene. As shown in Figure 3.11, many SNP containing MXEs are highly expressed in disease associated tissues where as the respective non-SNP-containing partner MXEs are not or hardly expressed. However, non-SNP-containing MXEs have high expression in early developmental stages. To assess this behavior of MXE specificity in terms of pathogenicity, a machine learner (random forest) was trained on the MXE expression data to predict the affected target tissue. Random Forest using leave-one-out

cross-validation strategy was used for the prediction. In order to have a minimum of 10 observations per group, we categorized diseases into two groups: cardio-neuromuscular (n = 10) and other diseases (n = 14). The classifier was able to predict Cardio-neuromuscular diseases with an accuracy of 83%, a sensitivity of 90%, a specificity of 79% and an area under the ROC curve (AUC) of 85% using MXE expression data. Despite having very few (24) observations, our data suggest that MXE expression might predict disease pathogenicity in time and space. In brief, an in silico method was de-

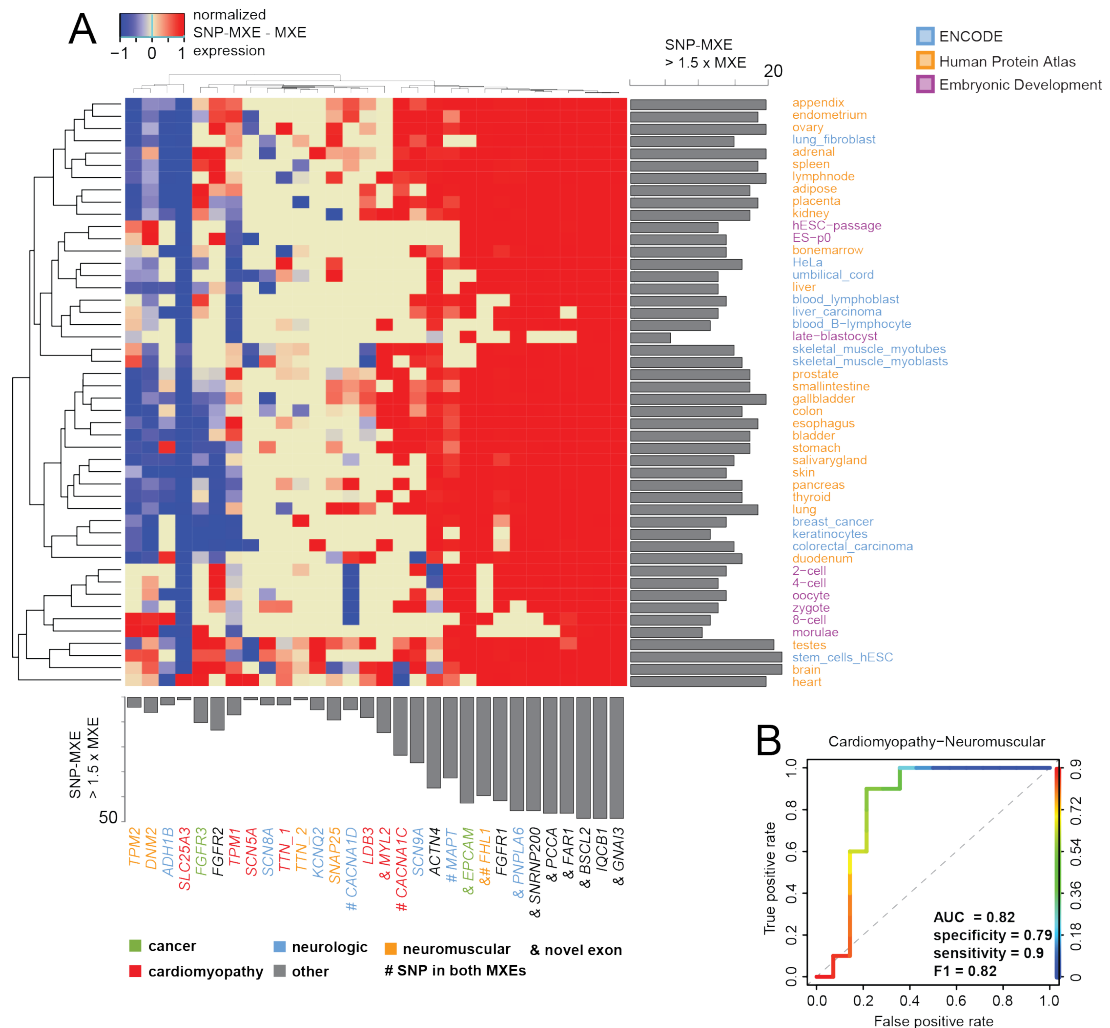


FIGURE 3.11: MXE-ratio expression predicts disease pathology :

A. Shown in heatmap are the the delta PSI values of MXE clusters. Percent-spliced-in (PSI) value of the non-SNP containing MXE is subtracted from the PSI value of the SNP containing MXE and scaled between -1 (high expression non-SNP-containing MXE, colored blue) and 1 (high expression SNP-containing MXE, colored red). Columns represent a cluster of MXE (a SNP containing and it's non-SNP containing patner). The column bars shows counts where the non-SNP containing MXE is 1.5-fold less expressed than the corresponding SNP-containing MXE. The same is shown for cell type, tissue and developmental stage by the row bar graph.

B. ROC curve showing true- and false-positive rates for the prediction of cardiomyopathy-neuromuscular disease based on MXE expression. Delta PSI values were taken in both cases. Figure modified from Article C of the thesis.

veloped to predict MXEs based on sequence similarity, similar lengths, and reading frame conservation and validate them using the publicly available billions of RNA-seq reads. Based on this method the current knowledge of human MXE is increased by almost an order of magnitude from 158 to 1,399 MXEs. These MXEs shows tissue and developmental stage specific expression and also have a potential role in diseases.

### 3.4 Conclusion and outlook

The results discussed in the above sections shows that the goals of the work have been fulfilled.

In summary, we developed **Oasis 2**, a fast and flexible web application for the analysis of sRNA-seq data. Its major functionalities include sRNA detection, multivariate differential expression (DE), and classification of small RNAs in deep sequencing data. Both DE and classification modules supports functional analyses including GO and pathway enrichment for novel and known miRNA targets. Additionally, the sRNA detection module supports the quantification of small RNAs in any organism as well supports the identification of potential cross-species miRNAs. One of the most useful features of Oasis 2 is the support to detect pathogenic contamination or potential pathogenic infections in the samples. The search module of Oasis 2 enables users to search novel (14 organisms) and known miRNAs across all miRBase supported organisms. Lastly it has an API that supports the batch submission of jobs from the command line. This will help the end users to automate the jobs submissions to Oasis 2 webserver and obtain publishable results via email. Oasis 2 generates downloadable interactive web reports for easy visualization, exploration, and analysis of data on a local system. In future, small RNA editing, modification, and mutation events can be implemented in Oasis 2. Additionally the reported output for bacterial and viral infections and contaminations can be enhanced.

In the second part of this thesis we developed **SEA**, a data repository for sRNA expression profiles that allows end users to search for ontology-based queries, supporting single or combined searches for five pre-defined terms such as organism, tissue, disease, cell type, and cell line across all datasets. However, the SEA database system contains additional (meta)-information including age, gender, developmental stage, genotype as well as technical experimental details such as the sequencing instrument and protocol details (e.g. library kit, RNA extraction procedure), which are returned as part of the user queried results. In short, SEA is a fast, flexible, and fully interactive web application for the investigation of sRNA expression and different sRNA-species. In addition it also



supports interactive result visualization at different levels, from querying and display of sRNA expression information to the mapping and quality information for each of the over 2,500 samples. As far as we are aware, SEA is the only sRNA-seq database that supports ontology-based queries. In the future, additional available meta-information such as age, gender, developmental stage, genotype as well as technical experimental details can be standardized and the search could be enhanced to allow users to query sRNAs based on them. Moreover, further sRNA-seq datasets should be incorporated into SEA. Lastly, one can store DE and classification results for all the sRNA-seq datasets having at least two groups (such as control and diseased) and make them queryable and comparable across different datasets.

Lastly, in the third project of the thesis, a high-confidence **atlas of 1,399 human MXEs** was generated based on sequence similarity, similar lengths, reading frame conservation and billions of RNA-seq reads. This high-confidence set of 1,399 MXEs extends current knowledge of human MXEs by an order of magnitude. These MXEs show tissue and developmental stage specific expression and also have a potential role in diseases. Furthermore, the data suggested that MXE expression reflects disease, which can be used to predict yet unseen diseases from published expression data. As a heuristic approach was used for the prediction of MXEs in this thesis, in the future a machine learning approach can be used for the prediction of MXEs, which may increase the predicting power of the method and could result in further novel MXEs.

The two web applications Oasis 2 and SEA are available online and can be accessed at <https://oasis.dzne.de/> and <https://sea.dzne.de/sea/sea.jsp> respectively. All the three articles Oasis 2, SEA and MXE can be accessed at <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2047-z>, <https://www.biorxiv.org/content/early/2017/08/04/133199> and <http://msb.embopress.org/content/13/12/959> respectively.

# References

- [1] Jenny U. Johansson et al. “An Ancient Duplication of Exon 5 in the Snap25 Gene Is Required for Complex Neuronal Development/Function”. In: *PLOS Genetics* 4 (Nov. 2008). DOI: [10.1371/journal.pgen.1000278](https://doi.org/10.1371/journal.pgen.1000278).
- [2] Igor Splawski et al. “Ca<sub>v</sub>1.2 Calcium Channel Dysfunction Causes a Multisystem Disorder Including Arrhythmia and Autism”. In: *Cell* 119.1 (2004), pp. 19–31. ISSN: 0092-8674. DOI: [10.1016/j.cell.2004.09.011](https://doi.org/10.1016/j.cell.2004.09.011). URL: <http://dx.doi.org/10.1016/j.cell.2004.09.011>.
- [3] Igor Splawski et al. “Severe arrhythmia disorder caused by cardiac L-type calcium channel mutations”. In: *Proceedings of the National Academy of Sciences* 102.23 (2005), pp. 8089–8096. ISSN: 0027-8424. DOI: [10.1073/pnas.0502506102](https://doi.org/10.1073/pnas.0502506102). eprint: <http://www.pnas.org/content/102/23/8089.full.pdf>. URL: <http://www.pnas.org/content/102/23/8089>.
- [4] M.J. Zvelebil and J.O. Baum. *Understanding Bioinformatics*. Garland Science, 2008. ISBN: 9780815340249. URL: [http://books.google.de/books?id=dGayL\\_tdnBMC](http://books.google.de/books?id=dGayL_tdnBMC).
- [5] Roger D Kornberg. “Eukaryotic transcriptional control”. In: *Trends in Cell Biology* (1999). DOI: [10.1016/S0962-8924\(99\)01679-7](https://doi.org/10.1016/S0962-8924(99)01679-7).
- [6] Maria D.; Town Terrence; Lee Gap Ryol; Flavell Richard A Spilianakis Charalampos G.; Lalioti. “Interchromosomal associations between alternatively expressed loci”. In: (2005).
- [7] David S. Latchman. “Transcription factors: An overview”. In: *The International Journal of Biochemistry Cell Biology* 29.12 (1997), pp. 1305–1312. ISSN: 1357-2725. DOI: [https://doi.org/10.1016/S1357-2725\(97\)00085-X](https://doi.org/10.1016/S1357-2725(97)00085-X). URL: <http://www.sciencedirect.com/science/article/pii/S135727259700085X>.
- [8] Brandi N. Davis and Akiko Hata. “Regulation of MicroRNA Biogenesis: A miRiad of mechanisms”. In: *Cell Communication and Signaling* 7.1 (2009), p. 18. ISSN: 1478-811X. DOI: [10.1186/1478-811X-7-18](https://doi.org/10.1186/1478-811X-7-18). URL: <https://doi.org/10.1186/1478-811X-7-18>.

- [9] Haruhiko Siomi Hirotsugu Ishizu and Mikiko C. Siomi. “Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines”. In: *Gene Development* (2012). DOI: <https://doi.org/10.1101/gad.203786.112>.
- [10] Alexei A.; Stark Alexander; Dus Monica; Kellis Manolis; Sachidanandam Ravi Brennecke Julius; Aravin and Gregory J. Hannon. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *jem;Drosophila;em;D*”. In: *Cell* 128.6 (2007), pp. 1089–1103. ISSN: 0092-8674. DOI: [10.1016/j.cell.2007.01.043](https://doi.org/10.1016/j.cell.2007.01.043). URL: <http://dx.doi.org/10.1016/j.cell.2007.01.043>.
- [11] Manel Esteller. “Non-coding RNAs in human disease”. In: *Nature Reviews Genetics* 12 (2011). Review Article, 861 EP –. URL: <http://dx.doi.org/10.1038/nrg3074>.
- [12] Kuniaki Saito et al. “Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in *Drosophila*”. In: *Genes Dev* 24.22 (2010). 20966047[pmid], pp. 2493–2498. ISSN: 0890-9369. DOI: [10.1101/gad.1989510](https://doi.org/10.1101/gad.1989510). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2975925/>.
- [13] Hongseok Ha et al. “A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements”. In: *BMC Genomics* 15.1 (2014), p. 545. ISSN: 1471-2164. DOI: [10.1186/1471-2164-15-545](https://doi.org/10.1186/1471-2164-15-545). URL: <https://doi.org/10.1186/1471-2164-15-545>.
- [14] Tamás Kiss. “Small Nucleolar RNAs”. In: *Cell* 109.2 (2002), pp. 145–148. ISSN: 0092-8674. DOI: [10.1016/S0092-8674\(02\)00718-3](https://doi.org/10.1016/S0092-8674(02)00718-3). URL: [http://dx.doi.org/10.1016/S0092-8674\(02\)00718-3](http://dx.doi.org/10.1016/S0092-8674(02)00718-3).
- [15] Jingwei Ni, Amy L. Tien, and Maurille J. Fournier. “Small Nucleolar RNAs Direct Site-Specific Synthesis of Pseudouridine in Ribosomal RNA”. In: *Cell* 89.4 (1997), pp. 565–573. ISSN: 0092-8674. DOI: [10.1016/S0092-8674\(00\)80238-X](https://doi.org/10.1016/S0092-8674(00)80238-X). URL: [http://dx.doi.org/10.1016/S0092-8674\(00\)80238-X](http://dx.doi.org/10.1016/S0092-8674(00)80238-X).
- [16] Thomas H. King et al. “Ribosome Structure and Activity Are Altered in Cells Lacking snoRNPs that Form Pseudouridines in the Peptidyl Transferase Center”. In: *Molecular Cell* 11.2 (2003), pp. 425–435. ISSN: 1097-2765. DOI: [10.1016/S1097-2765\(03\)00040-6](https://doi.org/10.1016/S1097-2765(03)00040-6). URL: [http://dx.doi.org/10.1016/S1097-2765\(03\)00040-6](http://dx.doi.org/10.1016/S1097-2765(03)00040-6).
- [17] Michael T. McManus and Phillip A. Sharp. “Gene silencing in mammals by small interfering RNAs”. In: *Nature Reviews Genetics* 3 (2002). Review Article, 737 EP –. URL: <http://dx.doi.org/10.1038/nrg908>.

- [18] Saba Valadkhan and Lalith S. Gunawardane. “Role of small nuclear RNAs in eukaryotic gene expression”. In: *Essays In Biochemistry* 54 (2013), pp. 79–90. ISSN: 0071-1365. DOI: [10.1042/bse0540079](https://doi.org/10.1042/bse0540079). eprint: <http://essays.biochemistry.org/content/54/79.full.pdf>. URL: <http://essays.biochemistry.org/content/54/79>.
- [19] Tamás Kiss. “Biogenesis of small nuclear RNPs”. In: *Journal of Cell Science* 117.25 (2004), pp. 5949–5951. ISSN: 0021-9533. DOI: [10.1242/jcs.01487](https://doi.org/10.1242/jcs.01487). eprint: <http://jcs.biologists.org/content/117/25/5949.full.pdf>. URL: <http://jcs.biologists.org/content/117/25/5949>.
- [20] Zhuojun Guo, Krishanthi S. Karunatilaka, and David Rueda. “Single Molecule Analysis of Protein Free U2/U6 snRNAs”. In: *Nat Struct Mol Biol* 16.11 (2009). 19881500[pmid], pp. 1154–1159. ISSN: 1545-9993. DOI: [10.1038/nsmb.1672](https://doi.org/10.1038/nsmb.1672). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784090/>.
- [21] Yongjun Chu and David R. Corey. “RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation”. In: *Nucleic Acid Ther* 22.4 (2012). 22830413[pmid], pp. 271–274. ISSN: 2159-3337. DOI: [10.1089/nat.2012.0367](https://doi.org/10.1089/nat.2012.0367) [PII]. URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426205/>.
- [22] Zhong Wang, Mark Gerstein, and Michael Snyder. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nat Rev Genet* 10.1 (2009), pp. 57–63. ISSN: 1471-0056. DOI: [10.1038/nrg2484](https://doi.org/10.1038/nrg2484). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2949280/>.
- [23] Jim Pease and Roy Sooknanan. “A rapid, directional RNA-seq library preparation workflow for Illumina sequencing”. In: *Nature Methods* 9 (2012), EP –. URL: <http://dx.doi.org/10.1038/nmeth.f.355>.
- [24] C. J. Date. *An Introduction to Database Systems, Volume I, 4th Edition*. Addison-Wesley, 1986.
- [25] Paul Beynon-Davies. *Database Systems*. Springer, 2004. ISBN: 978-1-4039-1601-3. DOI: [10.1007/978-0-230-00107-7](https://doi.org/10.1007/978-0-230-00107-7). URL: <https://doi.org/10.1007/978-0-230-00107-7>.
- [26] Edgar Frank Codd. “A Relational Model of Data for Large Shared Data Banks”. In: *Communications of the ACM* 13.6 (June 1970), pp. 377–387. URL: <http://dl.acm.org/citation.cfm?id=362685>.
- [27] C. J. Date. *Database in depth - relational theory for practitioners*. O’Reilly, 2005. ISBN: 978-0-596-10012-4. URL: <http://www.oreilly.de/catalog/databaseid/index.html>.

- [28] Peter Pin-Shan Chen. “The Entity-relationship Model—Toward a Unified View of Data”. In: *ACM Trans. Database Syst.* 1.1 (Mar. 1976), pp. 9–36. ISSN: 0362-5915. DOI: [10.1145/320434.320440](https://doi.org/10.1145/320434.320440). URL: <http://doi.acm.org/10.1145/320434.320440>.
- [29] Seth Gilbert and Nancy Lynch. “Brewer’s Conjecture and the Feasibility of Consistent, Available, Partition-tolerant Web Services”. In: *SIGACT News* 33.2 (June 2002), pp. 51–59. ISSN: 0163-5700. DOI: [10.1145/564585.564601](https://doi.org/10.1145/564585.564601). URL: <http://doi.acm.org/10.1145/564585.564601>.
- [30] Simon Andrews et al. “FastQC: A quality control tool for high throughput sequence data”. In: *Reference Source* (2010).
- [31] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. “Trimmomatic: a flexible trimmer for Illumina sequence data”. In: *Bioinformatics* 30.15 (2014). btu170[PII], pp. 2114–2120. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/>.
- [32] Hongshan Jiang et al. “Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads”. In: *BMC Bioinformatics* 15.1 (2014), p. 182. ISSN: 1471-2105. DOI: [10.1186/1471-2105-15-182](https://doi.org/10.1186/1471-2105-15-182). URL: <https://doi.org/10.1186/1471-2105-15-182>.
- [33] Felix Krueger. [http://www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/).
- [34] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet. journal* 17.1 (2011), pp–10.
- [35] Thomas D. Wu and Serban Nacu. “Fast and SNP-tolerant detection of complex variants and splicing in short reads”. In: *Bioinformatics* 26.7 (2010), pp. 873–881. DOI: [10.1093/bioinformatics/btq057](https://doi.org/10.1093/bioinformatics/btq057). eprint: [/oup/backfile/content\\_public/journal/bioinformatics/26/7/10.1093/bioinformatics/btq057/2/btq057.pdf](http://oup/backfile/content_public/journal/bioinformatics/26/7/10.1093/bioinformatics/btq057/2/btq057.pdf). URL: [+http://dx.doi.org/10.1093/bioinformatics/btq057](http://dx.doi.org/10.1093/bioinformatics/btq057).
- [36] Kai Wang et al. “MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery”. In: *Nucleic Acids Res* 38.18 (2010). gkq622[PII], e178–e178. ISSN: 0305-1048. DOI: [10.1093/nar/gkq622](https://doi.org/10.1093/nar/gkq622). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2952873/>.
- [37] Gregory R. Grant et al. “Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)”. In: *Bioinformatics* 27.18 (2011), pp. 2518–2528. DOI: [10.1093/bioinformatics/btr427](https://doi.org/10.1093/bioinformatics/btr427). eprint: [/oup/backfile/content\\_public/journal/bioinformatics/27/18/10.1093\\_bioinformatics\\_btr427/2/btr427.pdf](http://oup/backfile/content_public/journal/bioinformatics/27/18/10.1093_bioinformatics_btr427/2/btr427.pdf). URL: [+http://dx.doi.org/10.1093/bioinformatics/btr427](http://dx.doi.org/10.1093/bioinformatics/btr427).

- [38] Ben Langmead and Steven L Salzberg. “Fast gapped-read alignment with Bowtie 2”. In: *Nature methods* 9.4 (2012), pp. 357–359.
- [39] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. In: *Bioinformatics* 29.1 (2013), pp. 15–21.
- [40] Heng Li et al. “The sequence alignment/map format and SAMtools”. In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.
- [41] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* 26.1 (2010). btp616[PII], pp. 139–140. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp616](https://doi.org/10.1093/bioinformatics/btp616). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/>.
- [42] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), p. 1.
- [43] Nadine Schuurman and Agnieszka Leszczynski. “A method to map heterogeneity between near but non-equivalent semantic attributes in multiple health data registries”. In: *Health Informatics Journal* 14.1 (2008). PMID: 18258674, pp. 39–57. DOI: [10.1177/1460458207086333](https://doi.org/10.1177/1460458207086333). URL: <https://doi.org/10.1177/1460458207086333>.
- [44] Simon Jupp et al. “A New Ontology Lookup Service at EMBL-EBI”. In: (2015).
- [45] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, 2012. ISBN: 026201825X, 9780262018258.
- [46] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: [10.1007/BF00994018](https://doi.org/10.1007/BF00994018). URL: <https://doi.org/10.1007/BF00994018>.
- [47] John G. Cleary and Leonard E. Trigg. “K\*: An Instance-based Learner Using an Entropic Distance Measure”. In: *12th International Conference on Machine Learning*. 1995, pp. 108–114.
- [48] Tin Kam Ho. “Random Decision Forests”. In: *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*. ICDAR '95. Washington, DC, USA: IEEE Computer Society, 1995, pp. 278–. ISBN: 0-8186-7128-9. URL: <http://dl.acm.org/citation.cfm?id=844379.844681>.

- [49] Matthew Beckers et al. “Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench”. In: *RNA* 23.6 (2017). RA[PII], pp. 823–835. ISSN: 1355-8382. DOI: [10.1261/rna.059360.116](https://doi.org/10.1261/rna.059360.116). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435855/>.
- [50] Zhifu Sun et al. “CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data”. In: *BMC Genomics* 15.1 (2014). 6123[PII], p. 423. ISSN: 1471-2164. DOI: [10.1186/1471-2164-15-423](https://doi.org/10.1186/1471-2164-15-423). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4070549/>.
- [51] Sören Müller et al. “omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data”. In: *Bioinformatics* 29.20 (2013), pp. 2651–2652. DOI: [10.1093/bioinformatics/btt457](https://doi.org/10.1093/bioinformatics/btt457). eprint: [/oup/backfile/content\\_public/journal/bioinformatics/29/20/10.1093\\_bioinformatics\\_btt457/2/btt457.pdf](http://oup/backfile/content_public/journal/bioinformatics/29/20/10.1093_bioinformatics_btt457/2/btt457.pdf). URL: [+http://dx.doi.org/10.1093/bioinformatics/btt457](http://dx.doi.org/10.1093/bioinformatics/btt457).
- [52] Jinyu Wu et al. “mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing”. In: *RNA Biol* 10.7 (2013). 2013RNABIO0063R[PII], pp. 1087–1092. ISSN: 1547-6286. DOI: [10.4161/rna.25193](https://doi.org/10.4161/rna.25193). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3849156/>.
- [53] Jihoon Kim et al. “MAGI: a Node.js web service for fast microRNA-Seq analysis in a GPU infrastructure”. In: *Bioinformatics* 30.19 (2014). 24907367[pmid], pp. 2826–2827. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btu377](https://doi.org/10.1093/bioinformatics/btu377). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4173015/>.
- [54] Dimitrios M. Vitsios and Anton J. Enright. “Chimira: analysis of small RNA sequencing data and microRNA modifications”. In: *Bioinformatics* 31.20 (2015), pp. 3365–3367. DOI: [10.1093/bioinformatics/btv380](https://doi.org/10.1093/bioinformatics/btv380). eprint: [/oup/backfile/content\\_public/journal/bioinformatics/31/20/10.1093\\_bioinformatics\\_btv380/2/btv380.pdf](http://oup/backfile/content_public/journal/bioinformatics/31/20/10.1093_bioinformatics_btv380/2/btv380.pdf). URL: [+http://dx.doi.org/10.1093/bioinformatics/btv380](http://dx.doi.org/10.1093/bioinformatics/btv380).
- [55] Antonio Rueda et al. “sRNAtoolbox: an integrated collection of small RNA research tools”. In: *Nucleic Acids Research* 43.W1 (2015), W467–W473. DOI: [10.1093/nar/gkv555](https://doi.org/10.1093/nar/gkv555). eprint: [/oup/backfile/content\\_public/journal/nar/43/w1/10.1093\\_nar\\_gkv555/2/gkv555.pdf](http://oup/backfile/content_public/journal/nar/43/w1/10.1093_nar_gkv555/2/gkv555.pdf). URL: [+http://dx.doi.org/10.1093/nar/gkv555](http://dx.doi.org/10.1093/nar/gkv555).
- [56] Bharat Panwar, Gilbert S. Omenn, and Yuanfang Guan. “miRmine: a database of human miRNA expression profiles”. In: *Bioinformatics* 33.10 (2017), pp. 1554–1560. DOI: [10.1093/bioinformatics/btx019](https://doi.org/10.1093/bioinformatics/btx019). eprint: [/oup/backfile/content\\_](http://oup/backfile/content_)



- [public/journal/bioinformatics/33/10/10.1093\\_bioinformatics\\_btx019/2/btx019.pdf](#). URL: [+http://dx.doi.org/10.1093/bioinformatics/btx019](#).
- [57] Yuk Yee Leung et al. “DASHR: database of small human noncoding RNAs”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D216–D222. DOI: [10.1093/nar/gkv1188](#). eprint: [/oup/backfile/content\\_public/journal/nar/44/d1/10.1093\\_nar\\_gkv1188/2/gkv1188.pdf](#). URL: [+http://dx.doi.org/10.1093/nar/gkv1188](#).
- [58] Dimitrios M. Vitsios et al. “Large-scale analysis of microRNA expression, epitranscriptomic features and biogenesis”. In: *Nucleic Acids Res* 45.3 (2017). gkw1031[PII], pp. 1079–1090. ISSN: 0305-1048. DOI: [10.1093/nar/gkw1031](#). URL: [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5388392/](#).
- [59] I-Fang Chung et al. “YM500v3: a database for small RNA sequencing in human cancer research”. In: *Nucleic Acids Res* 45.Database issue (2017). 27899625[pmid], pp. D925–D931. ISSN: 0305-1048. DOI: [10.1093/nar/gkw1084](#). URL: [http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210564/](#).
- [60] Mikita Suyama. “Mechanistic insights into mutually exclusive splicing in dynamin 1”. In: *Bioinformatics* 29.17 (2013), pp. 2084–2087. DOI: [10.1093/bioinformatics/btt368](#). eprint: [/oup/backfile/content\\_public/journal/bioinformatics/29/17/10.1093\\_bioinformatics\\_btt368/2/btt368.pdf](#). URL: [+http://dx.doi.org/10.1093/bioinformatics/btt368](#).
- [61] Eric T. Wang et al. “Alternative isoform regulation in human tissue transcriptomes.” In: *Nature* 456.7221 (Nov. 2008), pp. 470–476. ISSN: 1476-4687. DOI: [10.1038/nature07509](#). URL: [http://dx.doi.org/10.1038/nature07509](#).
- [62] Marc R Friedländer et al. “miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades”. In: *Nucleic acids research* 40.1 (2012), pp. 37–52.
- [63] Ana Kozomara and Sam Griffiths-Jones. “miRBase: annotating high confidence microRNAs using deep sequencing data”. In: *Nucleic acids research* 42.D1 (2014), pp. D68–D73.
- [64] Petra Leidinger et al. “A blood based 12-miRNA signature of Alzheimer disease patients”. In: *Genome Biology* 14.7 (2013), R78. ISSN: 1474-760X. DOI: [10.1186/gb-2013-14-7-r78](#). URL: [https://doi.org/10.1186/gb-2013-14-7-r78](#).
- [65] Cailin E. Joyce et al. “Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome”. In: *Human Molecular Genetics* 20.20 (2011), pp. 4025–4040. DOI: [10.1093/hmg/ddr331](#). eprint: [/oup/backfile/content\\_public/journal/hmg/20/20/10.1093\\_hmg\\_ddr331/2/ddr331.pdf](#). URL: [+http://dx.doi.org/10.1093/hmg/ddr331](#).



- [66] Susanne Osanto et al. “Genome-Wide MicroRNA Expression Analysis of Clear Cell Renal Cell Carcinoma by Next Generation Deep Sequencing”. In: *PLoS One* 7.6 (2012). Ed. by Chad Creighton. 22745662[pmid], e38298. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0038298](https://doi.org/10.1371/journal.pone.0038298). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3380046/>.
- [67] Jüri Reimand, Tambet Arak, and Jaak Vilo. “g:Profiler—a web server for functional interpretation of gene lists (2011 update)”. In: *Nucleic Acids Res* 39.Web Server issue (2011). 21646343[pmid], W307–W315. ISSN: 0305-1048. DOI: [10.1093/nar/gkr378](https://doi.org/10.1093/nar/gkr378). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125778/>.
- [68] David Warde-Farley et al. “The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function”. In: *Nucleic acids research* 38.suppl 2 (2010), W214–W220.
- [69] Michael Kuhn et al. “STITCH 4: integration of protein-chemical interactions with user data”. In: *Nucleic Acids Res* 42.Database issue (2014). 24293645[pmid], pp. D401–D407. ISSN: 0305-1048. DOI: [10.1093/nar/gkt1207](https://doi.org/10.1093/nar/gkt1207). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3964996/>.
- [70] Andrea Franceschini et al. “STRING v9.1: protein-protein interaction networks, with increased coverage and integration”. In: *Nucleic Acids Res* 41.Database issue (2013). 23203871[pmid], pp. D808–D815. ISSN: 0305-1048. DOI: [10.1093/nar/gks1094](https://doi.org/10.1093/nar/gks1094). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531103/>.
- [71] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. In: *Nature Protocols* 4 (2008), 44 EP –. URL: <http://dx.doi.org/10.1038/nprot.2008.211>.
- [72] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists”. In: *Nucleic Acids Research* 37.1 (2009), pp. 1–13. DOI: [10.1093/nar/gkn923](https://doi.org/10.1093/nar/gkn923). eprint: [/oup/backfile/content\\_public/journal/nar/37/1/10.1093\\_nar\\_gkn923/2/gkn923.pdf](http://oup/backfile/content_public/journal/nar/37/1/10.1093_nar_gkn923/2/gkn923.pdf). URL: [+http://dx.doi.org/10.1093/nar/gkn923](http://dx.doi.org/10.1093/nar/gkn923).
- [73] Andrew Yates et al. “Ensembl 2016”. In: *Nucleic acids research* 44.D1 (2016), pp. D710–D716.
- [74] Nuno L. Barbosa-Morais et al. “The Evolutionary Landscape of Alternative Splicing in Vertebrate Species”. In: *Science* 338.6114 (2012), pp. 1587–1593. ISSN: 0036-8075. DOI: [10.1126/science.1230612](https://doi.org/10.1126/science.1230612). eprint: <http://science.sciencemag.org/content/338/6114/1587.full.pdf>. URL: <http://science.sciencemag.org/content/338/6114/1587>.

- [75] Sarah Djebali et al. “Landscape of transcription in human cells”. In: *Nature* 489.7414 (2012). 22955620[pmid], pp. 101–108. ISSN: 0028-0836. DOI: [10.1038/nature11233](https://doi.org/10.1038/nature11233). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3684276/>.
- [76] Tilgner H et al. “Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs”. In: *Genome Res* 22 (2012), 1616–1625.
- [77] Zhigang Xue et al. “Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing”. In: *Nature* 500.7464 (2013). 23892778[pmid], pp. 593–597. ISSN: 0028-0836. DOI: [10.1038/nature12364](https://doi.org/10.1038/nature12364). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4950944/>.
- [78] Liying Yan et al. “Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells”. In: *Nature Structural & Molecular Biology* 20 (2013), 1131 EP –. URL: <http://dx.doi.org/10.1038/nsmb.2660>.
- [79] Linn Fagerberg et al. “Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics”. In: *Mol Cell Proteomics* 13.2 (2014). M113.035600[PII], pp. 397–406. ISSN: 1535-9476. DOI: [10.1074/mcp.M113.035600](https://doi.org/10.1074/mcp.M113.035600). URL: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3916642/>.
- [80] Holger Pillmann et al. “Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology”. In: *BMC Bioinformatics* 12.1 (2011), p. 270. ISSN: 1471-2105. DOI: [10.1186/1471-2105-12-270](https://doi.org/10.1186/1471-2105-12-270). URL: <https://doi.org/10.1186/1471-2105-12-270>.
- [81] Melissa J. Landrum et al. “ClinVar: public archive of interpretations of clinically relevant variants”. In: *Nucleic Acids Research* 44.D1 (2016), pp. D862–D868. DOI: [10.1093/nar/gkv1222](https://doi.org/10.1093/nar/gkv1222). eprint: [/oup/backfile/content\\_public/journal/nar/44/d1/10.1093\\_nar\\_gkv1222/3/gkv1222.pdf](http://oup/backfile/content_public/journal/nar/44/d1/10.1093_nar_gkv1222/3/gkv1222.pdf). URL: [+http://dx.doi.org/10.1093/nar/gkv1222](http://dx.doi.org/10.1093/nar/gkv1222).

# Appendices

## Appendix A

### Article 1

SOFTWARE

Open Access



# Oasis 2: improved online analysis of small RNA-seq data

Raza-Ur Rahman<sup>1,2</sup>, Abhivyakti Gautam<sup>1</sup>, Jörn Bethune<sup>1,2</sup>, Abdul Sattar<sup>1,2</sup>, Maksims Fiosins<sup>1,2</sup>, Daniel Sumner Magruder<sup>1,2</sup>, Vincenzo Capece<sup>1</sup>, Orr Shomroni<sup>1</sup> and Stefan Bonn<sup>1,2,3\*</sup> 

## Abstract

**Background:** Small RNA molecules play important roles in many biological processes and their dysregulation or dysfunction can cause disease. The current method of choice for genome-wide sRNA expression profiling is deep sequencing.

**Results:** Here we present Oasis 2, which is a new main release of the Oasis web application for the detection, differential expression, and classification of small RNAs in deep sequencing data. Compared to its predecessor Oasis, Oasis 2 features a novel and speed-optimized sRNA detection module that supports the identification of small RNAs in any organism with higher accuracy. Next to the improved detection of small RNAs in a target organism, the software now also recognizes potential cross-species miRNAs and viral and bacterial sRNAs in infected samples. In addition, novel miRNAs can now be queried and visualized interactively, providing essential information for over 700 high-quality miRNA predictions across 14 organisms. Robust biomarker signatures can now be obtained using the novel enhanced classification module.

**Conclusions:** Oasis 2 enables biologists and medical researchers to rapidly analyze and query small RNA deep sequencing data with improved precision, recall, and speed, in an interactive and user-friendly environment.

**Availability and Implementation:** Oasis 2 is implemented in Java, J2EE, mysql, Python, R, PHP and JavaScript. It is freely available at <https://oasis.dzne.de>

## Background

Small RNAs (sRNAs) are a class of short, non-coding RNAs with important biological functions in nearly all aspects of organismal development in health and disease. Especially in diagnostic and therapeutic research sRNAs, such as miRNAs and piRNAs, received recent attention [18]. The current method of choice for the quantification of the genome-wide sRNA expression landscape is deep sequencing (sRNA-seq).

To date several local as well as server-based sRNA-seq analysis workflows are available that differ in their analysis portfolio, performance, and user-friendliness. Analysis workflows that need to be installed by the end-user comprise, for example, sRNA workbench [1] for the

quantification and identification of differentially expressed sRNAs and CAP-miRSeq [16] for the quantification of known and novel miRNAs including variant calling and subsequent differential expression analysis. While workflows that are installed on a local machine offer greater data security and may provide greater flexibility, they require installation, availability of servers, software and hardware maintenance as well as regular updates.

Recent additions to sRNA analysis web applications include omiRas [11], supporting quantification, differential expression and interactive network visualization; mir-Tools 2.0 [20] that allows for differential expression and gene ontology analysis of detected sRNAs; MAGI, an all-in-one workflow with detailed interactive web reports [8]; Chimira that allows for the detection of miRNA edits and modifications [17]; sRNAtoolbox [15] performs expression profiling of sRNA-seq data, differential expression as well as target gene prediction and visualization of analysis results; and Oasis [2], which supports the detection and annotation of known and

\* Correspondence: [sbonn@uke.de](mailto:sbonn@uke.de)

<sup>1</sup>Laboratory of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany

<sup>2</sup>Institute of Medical Systems Biology, Center for Molecular Neurobiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany

Full list of author information is available at the end of the article



novel sRNAs, multivariate differential expression analysis, biomarker detection, and job automation via an advanced programming interface (API). Here we present Oasis 2, an improved major release of the Oasis web application with many new and enhanced features for Biologists and Bioinformaticians (Table 1).

At the heart of Oasis 2 lies the new sRNA detection workflow that is faster and identifies more sRNAs with higher precision. In addition, Oasis 2 now supports sRNA-seq analyses for any organism, detects potential cross-species miRNAs, and reports viral and bacterial infections in samples with high precision and recall. Oasis 2 predicts and stores novel miRNAs in Oasis-DB and allows users to search and extract information for over 700 predicted high-quality miRNAs across 14 organisms. Oasis 2 classification module is improved with the use of balanced sampling and feature pruning methods that enables robust biomarker detection. Like its predecessor Oasis, Oasis 2's differential expression module supports multiple group comparisons (e.g. control vs. treatment 1 vs. treatment 2) and differential expression using co-variables such as age, gender, and medication. The differential expression and classification modules report various quality metrics including known and predicted targets of miRNAs in a downloadable, interactive web report. This web report allows for the subsequent functional enrichment analysis of miRNAs using GeneMania (interactome and GO analysis) [21], g:Profiler (GO, pathway-Kegg, Reactome) [13], STRING (protein-protein interaction network) [4], STITCH (chemical-protein interaction network) [9], and DAVID (enrichment analysis based on many biological databases) [6]. Oasis 2 is also at

the heart of the sRNA Expression Atlas (SEA, <https://sea.dzne.de>), a web application for the interactive querying, visualization, and analysis for over 2000 published sRNA samples. Lastly Oasis 2 features many new analysis and visualization options such as support for adapter trimmed data, options to trim additional barcodes, and interactive plots for sRNA detection and classification output. It has no restrictions on the size or number of samples and has no limits on the analyses per user.

### Implementation

The following paragraphs will describe the technical details of Oasis 2's novel sRNA detection, database, and classification modules. Additional information can be found in the supplementary material.

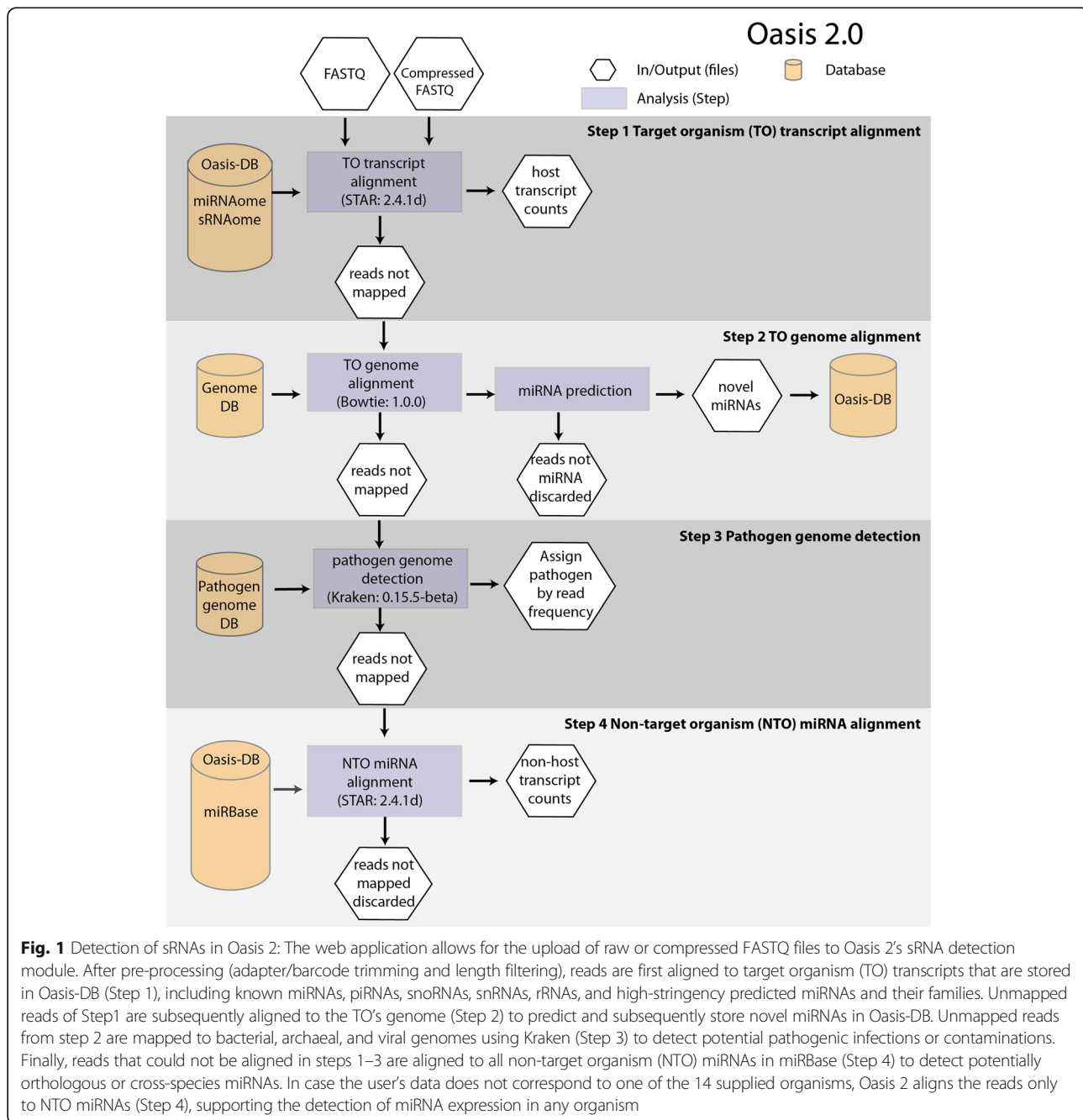
#### sRNA detection

One of the key differences between Oasis 2 and its predecessor is the fully revised detection of known and novel sRNAs. The new detection workflow increases the alignment speed, is more accurate, and supports the analysis of any model and non-model organism (Fig. 1, Additional file 1). While Oasis detected sRNAs using a single genome alignment step, Oasis 2 is based upon a four-tiered alignment strategy. Users can upload (un)-compressed data that originates from one of the 14 different organisms provided in Oasis 2 and the data will be aligned to the (i) target organism's (TO) transcripts, (ii) TO's genome, (iii) pathogen genomes, and (iv) non-target organism's (NTO) miRNA transcripts in succession (Fig. 1). In the TO Transcript alignment (step 1), reads are aligned to TO transcripts in Oasis-DB, a database that contains transcript information of miRNAs and other sRNA species (snRNA, snoRNA, rRNA and

**Table 1** sRNA-seq web application comparison

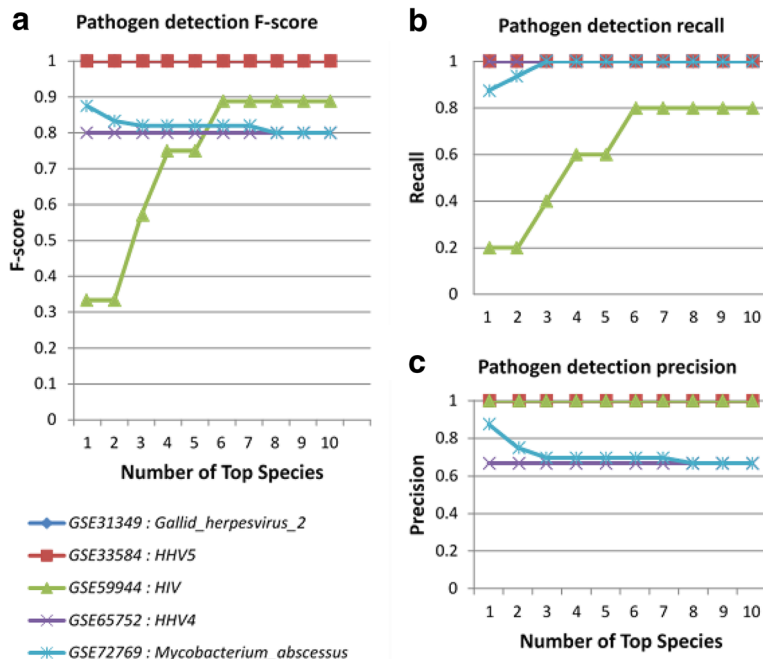
Feature	Oasis 2	Oasis	omiRas	mirTools 2.0	MAGI	Chimira	sRNAtoolbox
FASTQ compression	✓	✓			✓	✓	
miRNA prediction	✓	✓	✓	✓	✓		✓
miRNA modifications and edits						✓	✓
Novel miRNA database	✓						
Infection and cross-species analysis	✓						✓
Non-model organism	✓					✓	
Differential expression	✓	✓	✓	✓	✓	✓	✓
Multivariate differential expression	✓	✓					✓
Classification	✓	✓					
Novel miRNA target prediction	✓	✓		✓	✓		✓
Pathway/GO analysis	✓	✓	✓	✓	✓		✓
Batch job submission (API)	✓	✓					
Genome browser							✓

Of note, this comparison does not include all available sRNA analysis web applications. It only considers the most recent web applications that we deemed most competitive and we do not compare to standalone software solutions that have to be locally installed



piRNAs) from miRBase, piRNAbank, Ensembl, predicted novel miRNAs, and sRNA families. In this step reads of length 15–19 nucleotides are aligned with no mismatches whereas reads of length 20–32 nucleotides are mapped allowing for 1 mismatch (Step 2 in Fig. 1). In the TO Genome alignment (step 2), reads that do not align to TO transcripts are subsequently aligned to the reference genome allowing for 1 mismatch and no more than five potential genomic target regions to predict novel, high-quality miRNAs (Additional file 1 section 1.2

'Alignment and counting'). Predicted novel miRNAs are then added to Oasis-DB as described in section 2.2 'Detection and storage of novel miRNAs'. In the Pathogen Genome detection (step 3), reads that could not be aligned to the TO transcriptome or TO genome are used to identify pathogenic sRNA signatures from bacteria and viruses, supplying information on potentially infected samples (Fig.2 & Additional file 1). To this end, we indexed Oasis Pathogen-Genome-DB that consists of 4336 viral and 2784 bacterial/archaeal genomes with



**Fig. 2** Pathogen detection performance: To assess the performance of ‘pathogen detection module’, sRNA datasets with defined viral or bacterial infections were analyzed and the F-score (a), recall (b), and precision (c) of the pathogen predictions were measured for the top 10 reported organisms. Overall, the prediction of bacterial (*M. abscessus*) and viral (*HIV*, *HHV4*, *HHV5*, *Gallid\_herpesvirus\_2*) infections resulted in high F-scores, recall, and precision, especially when the top 5 predicted pathogen species are reported. In consequence, Oasis 2 currently reports the top five predicted pathogen species based on their read counts

Kraken [19] using a k-mer length of 18. In the Non-TO miRNA alignment (step 4), reads that could not be aligned to TO transcripts, the TO genome or pathogen genomes are aligned without any mismatches to all NTO transcripts of miRBase to detect potential orthologous or cross-species miRNAs. In cases where the data does not belong to one of the 14 supported genomes available in Oasis 2, reads can be aligned to all known and novel predicted miRNAs and miRNA families stored in Oasis-DB (Additional file 1).

In addition to the new alignment strategy, the sRNA detection module also supports data with already trimmed adapters. It also has an option for barcode removal, which is required for the analysis of libraries generated with e.g. the NEXTflex kit. In the case of barcode removal, Oasis 2 first discards the 3’ adapter sequence (in case the adapter is not already trimmed), and then removes an additional N (user defined, default is 0) bases from the adapter-clipped reads.

**Detection and storage of novel miRNAs**

Another major improvement of Oasis 2 is the ability to query and visualize detailed information for over 700 high-quality predicted miRNAs across 14 organisms (Fig. 1, Additional file 1: Figure S1). Oasis-DB comprises information on all MiRDeep2 [5] predicted miRNAs that pass stringent selection criteria during the sRNA

detection step of Oasis 2 (2.1 & Additional file 1), including the miRNA ID, organism, chromosomal location, precursor and mature sequences, structure, read counts, prediction scores, and detailed information on the software and its versions used to predict the miRNA. To assure that Oasis-DB contains only high-quality miRNA entries, novel predicted miRNAs have to pass the three criteria. The log-odds score assigned to the hairpin by miRDeep2 (miRDeep2-score) should be greater than 10, the predicted miRNA hairpin should not have sequence similarity to reference tRNAs or rRNAs, and the estimated randfold *p*-value of the excised potential miRNA hairpin should be equal to or lower than 0.05.

Novel predicted miRNAs are added to Oasis-DB using the standard nomenclature (Additional file 1 section 1.4 ‘Oasis-DB miRNA insertion and naming’).

In addition to novel miRNAs, Oasis-DB also stores information on all other sRNAs and sRNA families (Additional file 1). To provide access to Oasis-DB we created a novel web frontend, the Oasis 2 ‘Search’ module, which allows users to query miRNAs by mature/precursor ID or sequence, and the organism they come from. Information on high-confidence novel miRNAs is also shared with SEA, a web application that provides expression information of known and novel miRNAs for over 2000 samples (<https://sea.dzne.de>).



### Classification and differential expression

To allow for enhanced sRNA-based biomarker detection several profound changes to the Oasis 2 classification module were made, resulting in more robust biomarker detection with increased accuracy (Additional file 1: Figure S2, Additional file 1 section 'Oasis 2 classification module'). To increase the performance of the Random Forest-based (RF) classification module we first implemented balanced sampling (Additional file 1), making sure RF predictions would not be biased in the case of uneven class distribution. Since RFs can perform poorly on data that contains few informative and many non-informative features, the classification module was augmented with a feature pruning routine (Additional file 1), reporting prediction performance for the full and best RF models. In addition to providing information on model accuracy using the out-of-bag (OOB) error, Oasis 2 now also provides model performance information based on cross-validation. All classification results can be explored in interactive web reports, allowing for a detailed quality and performance analysis of the predicted biomarkers.

Moreover, we have improved the quality of output plots in the DE module and updated the DESeq2 version for the analysis of differential sRNA expression. Further details about DE module can be found in Additional file 1 section 1.5 'Oasis 2 differential expression module' and Additional file 1: Table S3.

### Technologies and compatibility

Oasis 2 is implemented in Java, J2EE, mysql, Python, R, PHP and JavaScript. For the usage JavaScript should be enabled in the browser. Oasis 2 functionality was tested on all major browsers (Table 2). It has no restrictions on the size or number of samples and has no limits on the analyses per user. Potential user-specific problems can arise when i) an institution or university has upload limits, ii) proxy settings that would interrupt or prohibit long uploads, or iii) JavaScript is disabled or blocked. Oasis 2 is freely available at (<https://oasis.dzne.de>).

### Results

We compared the set of analysis options and the analysis speed of Oasis 2 to six state-of-the-art sRNA analysis web applications, including Oasis, omiRas, mirTools 2.0,

MAGI, Chimira and sRNAtoolbox, and found that it compares favorably in the number of analysis options (Table 1) and the analysis speed (Table 3). When tested on four publically available datasets, Oasis 2 detected 19 out of 27 (70%) differentially expressed (DE) genes that were previously validated (true positives) and did not detect 4/4 (100%) miRNAs that showed a significant DE in deep sequencing but could not be validated with qPCR (false positives), highlighting both the sensitivity and specificity of Oasis 2. Finally, we compared the performance of the novel classification module to the one implemented in Oasis, showing that prediction accuracy as well as robustness are increased.

### Detection and differential expression of sRNAs

To estimate if the novel sRNA detection workflow of Oasis 2 identifies and quantifies sRNAs correctly we analyzed four published datasets containing validated sRNA changes using Oasis 2 with default settings. Of note, none of the above-mentioned publications looked into the DE of other small RNA classes (snRNA, snoRNA and rRNA and piRNAs), so the analyses were restricted to miRNAs.

### Alzheimer disease data

We started by analyzing an Alzheimer disease (AD) sRNA dataset that consists of 48 Alzheimer and 22 control samples [10] using Oasis 2 and default settings. The original publication uses a Wilcoxon-Mann-Whitney test detecting 125 known DE miRNAs. Oasis 2 detected 103 DE miRNAs using an adjusted  $p$ -value  $< 0.1$ , of which 62(60%) overlapped with the original analysis. The overlap of 60% seems reasonable, given the different statistical approaches and miRBase versions used for the detection and DE analysis of the miRNAs. In the original publication 8/10 known miRNAs were validated to be differentially expressed in the same direction, whereas two miRNAs (hsa-miR-1285-5p and hsa-miR-26a-5p) were not validated in the same direction (instead of up-regulation they showed downregulation in qPCR). Interestingly these two miRNAs were not detected to be differentially expressed by Oasis 2. On the other hand Oasis 2 was able to detect 3/3 upregulated miRNAs (hsa-let-7d-3p, hsa-miR-5010-3p and hsa-miR-151a-3p), 3/5 downregulated miRNAs (hsa-miR-532-5p, hsa-miR-26b-5p and hsa-let-7f-5p), and it did not detect two downregulated miRNAs (hsa-miR-103a-3p, hsa-miR-107). In summary, Oasis 2 was able to detect 6/8 (75%) validated differentially expressed known miRNAs and not detecting 2/2 false positives from the original study. Unfortunately, two novel miRNAs validated in the original study are not added to miRBase yet, therefore we were not able to compare to them.

**Table 2** Oasis 2 browser compatibility

Browser	Version
Chrome	61.0.3163.100, 62.0.3202.62
Mozilla Firefox	55.0.3, 56.0 (64-bit), 57.0 (64-bit)
Chromium	62.0.3202.75
Safari	11.0.1
Internet explorer	11

Browsers that are used to test Oasis 2 functionalities

**Table 3** Runtime comparison of different sRNA-seq web applications

Demo Dataset	Oasis 2 (total) <sup>1</sup>	Oasis (total) <sup>1</sup>	MAGI (total)	Chimira (total)	omiRas	mirTools <sup>7</sup> 2.0	sRNAtoolbox
AD (287 GB) <sup>4</sup>	8 h31m50s	12h29m12s	NA <sup>2</sup>	NA <sup>4</sup>	NA <sup>5</sup>	NA	NA
Psoriasis (48 GB)	1h35m17s	5h49m4s	48h <sup>3</sup>	3h3m12s	NA <sup>6</sup>	NA	NA
Renal Cancer (9 GB)	31m43s	1h8m41s	8h <sup>3</sup>	47m11s	9h31m	NA	NA

<sup>1</sup>Run time estimate includes the data compression and decompression, the sRNA Detection, DE Analysis, and Classification. <sup>2</sup>We could not get MAGI to upload all AD files. Most probably it has a problem with the quality or format of one of the files. <sup>3</sup>These values were obtained from the MAGI website. <sup>4</sup>Chimira does not support the analysis of more than 25 files at a time, which prohibited us from getting runtime estimates for the AD dataset. <sup>5</sup>omiRas did not finish uploading files, which prohibited us from getting runtime estimates for the AD dataset. <sup>6</sup>omiRas http uploading error. <sup>7</sup>We cannot compare the runtime of mirTools 2.0 as maximum file size to upload is limited to 30 Mb. The sRNAtoolbox web application has been non-functional since 30/05/2017, which prohibited any runtime comparison (<http://bioinfo2.ugr.es:8080/srnatoolbox/quick-start/>)

### Psoriasis data

Oasis 2's performance was next assessed using a set of 10 Psoriasis and 10 control samples [7]. The original publication uses a hypergeometric test to assess differential expression (Pearson's chi-square test) that is followed by a Bonferroni multiple-testing correction.

In accordance with the analyses performed in the original publication, we only considered non-redundant pre-miRNAs. Oasis 2 found 195 DE miRNAs (166 non-redundant known pre-miRNAs) (adjusted  $p$ -value < 0.1) whereas the original publication contains only 98 DE miRNAs (70 non-redundant known pre-miRNAs). Of the 70 DE pre-miRNAs in the original study, 51 (72.85%) could also be found in the list of Oasis 2 DE miRNAs (Table 4). In addition, 5/8 (62.5%) experimentally validated DE miRNAs (miR-21, miR-31, miR-944, miR-135band miR-675) were detected by Oasis 2, not identifying validated miRNAs miR-124, miR-431 and miR-219-2-3p that show high expression variation in the original publication. Furthermore, Oasis 2 identified 2/3 (67%) predicted novel DE miRNAs (hsa-miR-203b and hsa-miR-3613) while missing hsa-miR-4490 (miRBase v21). In addition, Oasis 2 did not detect the false positive miR-431\* (1/1, 100%) that was predicted to be DE in the original Psoriasis study [7] but could not be validated by qPCR. In summary, Oasis 2 was able to detect 7/11

(64%) validated differentially expressed known and novel miRNAs and did not detect the only available false positive miRNA from the original study.

Of note, Oasis 2' PCA analysis highlights a potentially mis-annotated Psoriasis sample and another outlier sample (Fig. 3A). Removal of these two samples (Fig. 3B) increased the number of significantly (adjusted  $p$ -value < 0.1) DE miRNAs from 195 to 256 cases. We would like to emphasize that this data was already analyzed in two publications and to our knowledge this is the first time that these 'problematic' samples were detected, providing strong evidence for the utility of Oasis 2' QC plots.

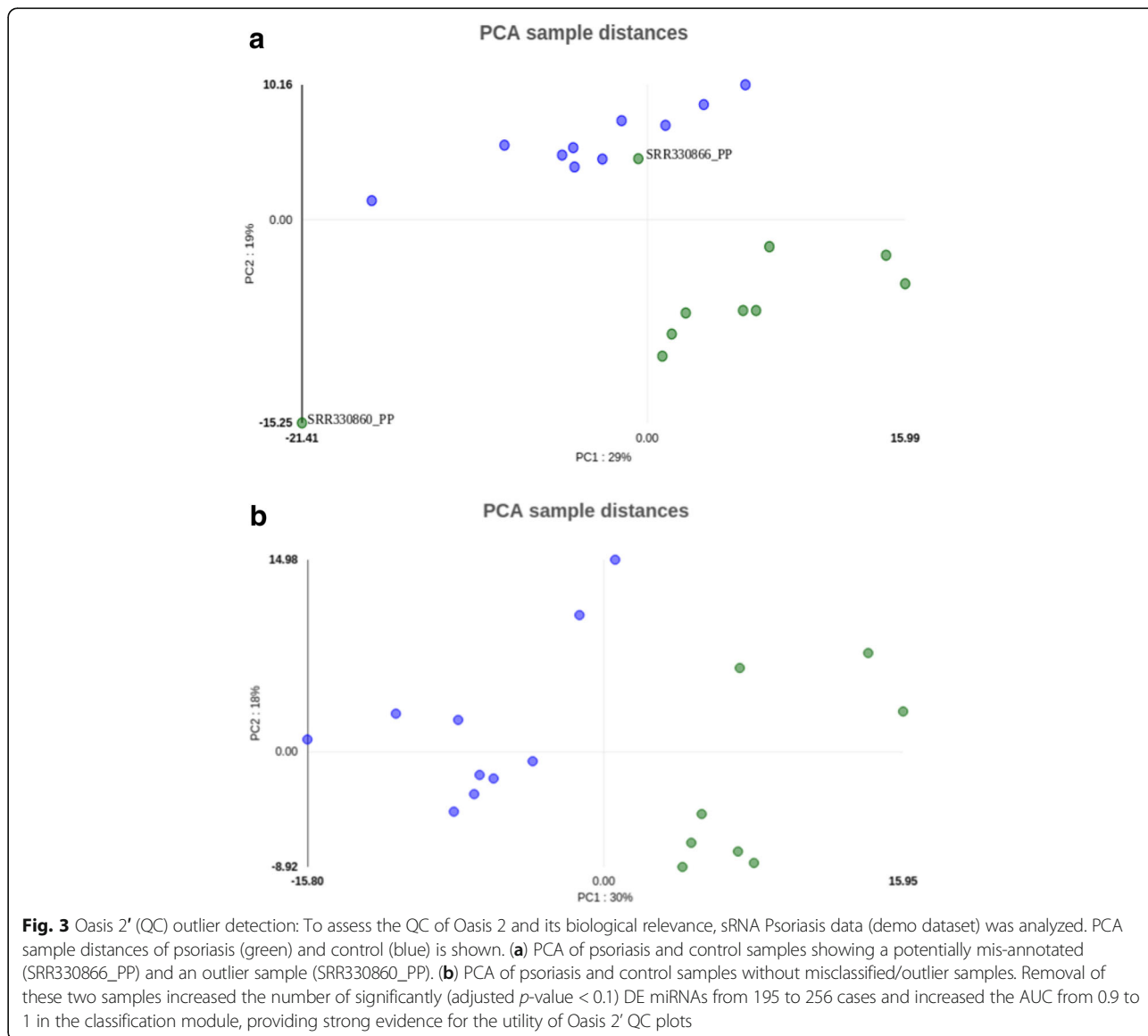
### Renal cancer data

In this work 11 renal cancer and 11 remission samples [12] were analyzed. This is longitudinal data from 11 patients and as such paired but we were unable to extract the pairing information from the GEO database annotations. Therefore the data was analyzed with Oasis 2 in un-paired mode and compared to the published, paired analysis with edgeR [14]. Despite of these technical issues the two analyses showed high overlap. Oasis 2 found 150 DE miRNAs (adjusted  $p$ -value < 0.1) whereas the original publication lists only 70 DE miRNAs. Of these 70 DE miRNAs 53 (76%) could also be found in the significant Oasis 2 miRNAs (Table 4). Of note, with

**Table 4** Overlap of differentially expressed sRNAs using three datasets

	Statistic <sup>1</sup>	Overlap <sup>2</sup>	Validated overlap <sup>3</sup>	FP overlap <sup>4</sup>
AD	Wilcoxon-Mann-Whitney	60%	75%(6/8) <sup>5</sup>	0% (0/2)
Psoriasis	Pearson's chi-squared	73%	64% (7/11)	0% (0/1)
Renal Cancer	edgeR [14]	76%	80% (4/5)	NA
Schizophrenia	DESeq2 (Dejian et al., 2015)	41%	67%(2/3)	0% (0/1)

<sup>1</sup>Oasis 2 uses a negative binomial distribution as basis for its statistical evaluation of the differential expression. A very similar approach is taken by the edgeR package that has been used in the Renal Cancer study. The Psoriasis data was analyzed using a Pearson's chi-squared test and the AD dataset was analyzed using the non-parametric Wilcoxon-Mann-Whitney test. Schizophrenia dataset used the same approach like Oasis 2. <sup>2</sup>Overlap of differentially expressed miRNAs comparing Oasis 2's results to published data. The percentage is calculated in reference to the shorter DE list. <sup>3</sup>Overlap of differentially expressed miRNAs that have been validated independently in addition to the sRNA-seq experiment. <sup>4</sup>False positive (FP) differentially expressed miRNAs detected by Oasis 2. <sup>5</sup>Only known validated DE miRNAs are considered



the exception of miR-122 all the validated miRNAs from the original work were detected using Oasis 2 (miR-21-5p, miR-210-3p, miR-199, miR-532-3p).

**Schizophrenia and schizoaffective disorder data**

In this experiment induced pluripotent stem cells were used to study neuropsychiatric disorders associated with 22q11.2 microdeletions [3]. Controls and patients with 22q11.2 microdeletions diagnosed with a psychotic disorder were compared (9 controls and 7 patients). Oasis 2 found 34 DE miRNAs (adjusted  $p$ -value < 0.1) whereas the original publication identified 45 DE miRNAs. Of these 45 DE miRNAs 14 (41%) were also detected as differentially expressed by Oasis 2 (Table 4). In the original

publication four miRNAs were validated by qPCR, two significantly up-regulated (miR-23a-5p and miR-146b-3p), one significantly down-regulated (miR-185-5p), and a miRNA that showed no difference in expression (miR-767-5p). Oasis 2 was able to confirm 2/3 (67%) validated differentially expressed miRNAs (miR-23a-5p and miR-185-5p) and did not confirm 1/1 (100%) false positive miRNAs miR-767-5p.

Overall, Oasis 2 detected 19/27 (70%) independently validated DE miRNAs in the published datasets despite of the different statistical approaches and miRBase versions used (Table 4). Detailed analysis results are accessible in Oasis 2's 'Demo Data' webpage. Our results provide strong evidence that Oasis 2 provides biologically meaningful results to the end user.

### Pathogen detection and sample classification

To assess the performance of the pathogen detection we analyzed 5 datasets with known viral or bacterial infections (Additional file 1: Table S6). We calculated the precision, recall, and F-score for the detection of the particular pathogen strain in the dataset while considering only the top ranking, first two, three, and up to the first ten reported species (Fig. 2). Species were ordered based on the number of read counts. In general, the viral or bacterial species and strains were detected with high precision and recall, reaching F-scores of  $\sim 0.8$  when the top five viral and bacterial species were considered. In consequence, Oasis 2 currently reports the top five bacterial, archaeal, and viral species found, allowing for the detection of potential infective agents or the discovery of experimental sample contaminations.

To benchmark the improved classification routine, we compared the performance of the old Oasis classification module (unbalanced sampling with all variables) to the new Oasis 2 classification module using balanced sampling and feature optimization using three demo datasets (see [Detection and Differential Expression of sRNAs](#) and Additional file 1: Figure S2). From a theoretical perspective, balanced sampling should increase prediction accuracy only in the case of class imbalances. In consequence, the novel classification module enhances the AUC for the imbalanced AD (22 controls, 48 patients) demo dataset by 2% (old AUC 0.95, new AUC 0.97), while it marginally changes classification performance for the balanced Psoriasis (10 control and 10 Psoriasis samples) (old AUC 0.90, new AUC 0.91) and Renal carcinoma (11 control and 11 cancer samples) (new and old AUC 1.00) data. Feature pruning should be crucial when a dataset contains a lot of uninformative features and very few informative features. To this end we have taken an unpublished dataset (6 controls, 6 treatments) that contains at least one feature that perfectly separates the two classes but otherwise contains mostly uninformative features. Whereas the old classification module reaches an AUC of 0 on this dataset, the new module reaches an AUC of 0.833.

Moreover, we also compared the accuracy of the new Oasis 2 classification module on the AD dataset to the published accuracy in the original manuscript [10]. Unfortunately, we were unable to obtain the primary output of the SVM and could not follow the post-processing steps of the machine learning results as performed in the original publication (e.g. removal of miRNAs that also occur in other diseases). In brief, the original publication provides a biomarker signature of 12 miRNAs (10 annotated and two novel) that reaches an average accuracy of 80%. The Oasis 2 classification reaches an accuracy of  $\sim 87\%$  (AUC of 0.97) using 320 features (no preprocessing for other diseases) and has an out-of-bag error of  $\sim$

10%. Two miRNAs in the original paper list (has-miR-151a-3p, hsa-let-7f-5p) were also found in the top 10 features (miRNAs) obtained with Oasis 2 classification.

The classification analysis of the three demo datasets (see 3.1) yielded stable and robust biomarker predictions that further corroborated the quality of the enhanced classification module.

### Runtime estimates

We next estimated the runtime of Oasis 2 using the above-mentioned AD, Psoriasis, and Renal cancer datasets and compared the results to runtime estimates for omiRas, mirTools 2.0, MAGI, Chimira and sRNAtoolbox, five recently developed web applications for the analysis of sRNA-seq data (Table 3, Additional file 1: Table S7). Performances of the sRNA Detection, DE Analysis, and Classification modules were measured on the Oasis 2 server. For benchmarking the Oasis 2 runtime we compared it to the runtime estimates of the above-mentioned web applications by submitting the AD, Psoriasis, and Renal Cancer datasets to the respective services (Table 3). Of note, runtime estimates for MAGI were taken from the MAGI webpage, which we assume constitutes a 'best case scenario' in favor of MAGI (low server analysis load). In addition, we could not compare to mirTools 2.0 as the maximum upload file size is limited to 30 Mb. Furthermore, the sRNAtoolbox web application was also not accessible during the period of testing and writing this manuscript.

Overall, Oasis 2 is significantly faster than MAGI, Chimira, and omiRas. For the smallest dataset (Renal Cancer) Oasis 2 was  $\sim 1.5$  times faster than Chimira,  $\sim 15$  times faster than MAGI, and  $\sim 18$  times faster than omiRas. While the runtime differences between Oasis 2 and Chimira were rather small when only few samples were analyzed, Oasis 2 was  $\sim 2$  times faster than Chimira,  $\sim 30$  times faster than MAGI for the 48 Gb Psoriasis dataset. Unfortunately, we were unable to estimate the runtime of omiRas for the Renal Cancer dataset since it did not finish file upload. Oasis 2 analyzed the largest dataset (AD, 287 Gb) in 8 h31m50s while none of the other tools mentioned above supported the analysis of the AD samples. In summary, Oasis 2 is the fastest of the state-of-the-art web applications we could compare to and has no restrictions on the sample number or size.

### Conclusions

Oasis 2 is fast, reliable, and offers several unique features that make it a valuable addition to the ever-growing number of sRNA-seq analysis applications. Especially the analysis support for all organisms, the detection and storage of novel miRNAs, the differential expression and classification modules, and the interactive results visualization supporting GO and pathway enrichment analyses enable

biologists and medical researchers to quickly analyze, visualize, and scrutinize their data. Oasis 2 also offers rich per experiment and per sample quality control, which might be one of the most important steps in the initial data analysis. The utility of a good quality control is exemplified in the analysis of the Psoriasis dataset, which seems to contain a mis-labelled (SRR330866\_PP) and an outlier (SRR330860\_PP) sample (Fig. 3). The removal of the outlier and mis-labelled samples in the Psoriasis dataset increased the number of significantly DE miRNAs from 195 to 256 cases and increased the classification accuracy for the same dataset from AUC of 0.9 to 1. We would like to emphasize that this data was already analyzed in two publications and to our knowledge this is the first time that these ‘problematic’ samples were detected, providing strong evidence for the utility of Oasis 2’ QC plots. Additionally the modular structure of Oasis 2 (sRNA detection, DE and classification) makes this task even easier, as the user can run only DE (without outliers) rather than going through the sRNA detection step again. In addition Oasis 2 provides PDF and video tutorials that explain its usage and details on how to interpret its results. Future developments will include the detection of small RNA editing, modification, and mutation events as well as more detailed reports on bacterial and viral infections and contaminations.

## Additional file

**Additional file 1:** Oasis2-Suppl-Material.docx: This file contains supplementary material and figures as well. (DOCX 125 kb)

## Acknowledgements

We would like to thank Ashish Rajput, Ting Sun, Vikas Bansal, Michel Edwar Mickael, the DZNE IT, and all of the Oasis users for helpful suggestions.

## Funding

This work was supported by the DFG (BO4224/4–1), the Network of Centres of Excellence in Neurodegeneration (CoEN) initiative, the Volkswagen Stiftung (Az88705), iMed – the Helmholtz Initiative on Personalized Medicine, and the BMBF grant Integrative Data Semantics in Neurodegeneration (031L0029B, IDS\_N).

## Availability of data and materials

Oasis 2 freely available at <https://oasis.dzne.de>. Oasis 2’ demo data is available at [https://oasis.dzne.de/small\\_rna\\_demo.php](https://oasis.dzne.de/small_rna_demo.php). Additional datasets mentioned and analyzed in this article can

GSE46579

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46579>

GSE31037

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31037>

GSE37616

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE37616>

GSE59944

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59944>

GSE65752

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65752>

GSE31349

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE31349>

GSE33584

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33584>

GSE72769

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72769>

## Authors’ contributions

SB initiated the study and designed the web application as well as analyses together with RR. RR and AG designed the Oasis-DB to store novel predicted miRNA. MF enhanced the classification module. JB and VC worked on the backend implementations of different modules. AS analyzed sRNA-seq data on different web servers to benchmark Oasis 2. DSM and OS worked the interactive user interface and tutorials. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

N/A

## Consent for publication

N/A

## Competing interests

The authors declare that they have no competing interests.

## Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Laboratory of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany. <sup>2</sup>Institute of Medical Systems Biology, Center for Molecular Neurobiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany. <sup>3</sup>German Center for Neurodegenerative Diseases, Tübingen, Germany.

Received: 25 August 2017 Accepted: 29 January 2018

Published online: 14 February 2018

## References

1. Beckers, et al. Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench. *RNA*. 2017;823–35.
2. Capece V, et al. Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*. 2015;31:2205–7.
3. Dejian, et al. MicroRNA Profiling of Neurons Generated Using Induced Pluripotent Stem Cells Derived from Patients with Schizophrenia and Schizoaffective Disorder, and 22q11.2 Del. *plosone*. 2015.
4. Franceschini, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 2013;D808–15.
5. Friedländer MR, et al. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res*. 2012; 40:37–52.
6. Huang, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*. 2007;8:R183.
7. Joyce CE, et al. Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Hum Mol Genet*. 2011;20: 4025–40.
8. Kim J, et al. MAGI: a node.js web service for fast microRNA-Seq analysis in a GPU infrastructure. *Bioinformatics*. 2014;30:2826–7.
9. Kuhn, et al. STITCH 4: Integration of protein-chemical interactions with user data. *Nucleic Acids Res*. 2014;D401–7.
10. Leidinger P, et al. A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biol*. 2013;14:R78.
11. Müller, et al. omiRas: a Web server for differential expression analysis of miRNAs derived from small RNASeq data. *Bioinformatics*. 2013;2651–2.
12. Osanto S, et al. Genome-wide microRNA expression analysis of clear cell renal cell carcinoma by next generation deep sequencing. *PLoS One*. 2012;7
13. Reimand, et al. G:Profiler - A web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res*. 2011;W307–15.
14. Robinson MD, et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26: 139–40.



15. Rueda, et al. sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.* 2015;W467–W473.
16. Sun, et al. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genomics.* 2014;15:423.
17. Vitsios DM, Enright AJ. Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics.* 2015;31:3365–7.
18. Witwer KW. Circulating MicroRNA biomarker studies: pitfalls and potential solutions. *Clin Chem.* 2014;000
19. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15:R46.
20. Wu, et al. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on highthroughput sequencing. *RNA Biol.* 2013;1087–92.
21. Zuberi, et al. GeneMANIA prediction server 2013 update. *Nucleic Acids Res.* 2013;W115-22.

Submit your next manuscript to BioMed Central  
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)



## Appendix B

### Article 2

## **SEA: The small RNA Expression Atlas**

Raza-Ur Rahman<sup>1,2</sup>, Abdul Sattar<sup>1,2</sup>, Maksims Fiosins<sup>1,2</sup>, Abhivyakti Gautam<sup>1</sup>, Daniel Sumner Magruder<sup>1,2</sup>, Jörn Bethune<sup>1,2</sup>, Sumit Madan<sup>3</sup>, Juliane Fluck<sup>3</sup>, and Stefan Bonn<sup>1,2,4,\*</sup>

<sup>1</sup>Laboratory of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany.

<sup>2</sup>Institute of Medical Systems Biology, Center for Molecular Neurobiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany.

<sup>3</sup>Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany.

<sup>4</sup>German Center for Neurodegenerative Diseases, Tübingen, Germany.

\*Contact: [sbonn@uke.de](mailto:sbonn@uke.de)



## Abstract

Small RNAs (sRNAs) are important biomolecules that exert vital functions in organismal health and disease, from viruses to plants, animals, and humans. Given the ever-increasing amounts of sRNA deep sequencing data in online repositories and their potential roles in disease therapy and diagnosis, it is important to enable federated sRNA expression querying across samples, organisms, tissues, cell types, and diseases. Here we present the sRNA Expression Atlas (SEA), a web application that allows for the search of known and novel small RNAs across ten organisms using standardized search terms and ontologies. SEA contains re-analyzed sRNA expression information for over 2000 published samples, including many disease datasets and over 700 novel, high-quality predicted miRNAs. We believe that SEA's simple interface and fast search in combination with its detailed interactive reports will enable researchers to better understand the potential function and diagnostic value of sRNAs across tissues, diseases, and organisms.

**Availability and Implementation:** SEA is implemented in Java, J2EE, Python, R, PHP and JavaScript. It is freely available at <http://sea.dzne.de>

---

\* To whom correspondence should be addressed.

## 1 Introduction

Small RNAs (sRNAs) are a class of short, non-coding RNAs with important biological functions in nearly all aspects of organismal development in health and disease. Especially in diagnostic and therapeutic research sRNAs such as miRNAs and piRNAs received recent attention (Witwer, 2014). Reflecting the importance of sRNAs in biological processes as well as disease diagnosis and therapy is the increasing number of deep sequencing sRNA studies (sRNA-seq). To harvest the true potential of existing data it is important to allow for the querying, visualization, and analysis of sRNA-seq data across organisms, tissues, cell types, and disease states. This would allow researchers, for example, to search for disease-specific sRNA biomarker signatures across all disease entities investigated. Data integration and interoperability require (i) a streamlined analysis workflow to reduce analysis bias between experiments and (ii) also necessitate standardized annotation using ontologies to search and retrieve relevant samples. Here we are presenting the **s**mall-RNA **E**xpression **A**tlas (SEA), a web application that allows for the querying, visualization, and analysis of over 2000 published sRNA-seq expression datasets. SEA automatically downloads and re-analyzes published data using Oasis 2, annotates relevant meta-information using standardized terms, synchronizes sRNA information with other databases, allows for the querying of terms across ontological graphs, and presents quality curated sRNA expression information as interactive web reports (Capece *et al.*, 2015). It currently supports 10 organisms and is continuously updated with novel published sRNA-seq datasets and relevant sRNA information from various online resources.

## 2 System Design

SEA stores sRNA expression information as well as deep and standardized meta-data on the samples, analysis workflows, and databases used. Data and meta-data information is normalized using ontologies to allow for standardized search and retrieval across ontological hierarchies (see section 2.3 for details). The following sections will detail the system design of SEA.

### 2.1 Acquisition of sRNA datasets

SEA acquires raw SRA files of published sRNA-seq datasets and their primary annotation from Gene Expression Omnibus (GEO) and NCBI's Sequence Reads Archive (SRA) repository. GEO makes two databases in SQLite format available for download: GEOMETADB for annotations and SRADB for SRA sequences. An automated data acquisition pipeline searches for new sRNA data bi-weekly, keeping SEA continuously updated. Novel datasets are downloaded and stored in SEA's raw data repository while corresponding annotations are stored in SEA's annotation database. Raw data is subsequently processed automatically by SEA's sRNA analysis workflow (2.2) while annotations are processed automatically with SEA's annotation workflow (2.3). Processed files and annotations are subsequently semi-automatically curated.

### 2.2 Data analysis and storage

Following the acquisition of sRNA datasets, the SEA analysis workflow automatically analyzes new files using the Oasis 2.0 API (see [biorxiv.org](http://biorxiv.org) for latest manuscript) (Capece *et al.*, 2015) (<https://oasis.dzne.de>). The SEA analysis workflow determines data quality and detects and quantifies sRNAs, including the prediction of novel, high-quality miRNAs. Low quality files are flagged automatically and subjected to manual curation. Any files not passing manual curation are removed from SEA. Subsequently, sRNA counts of high-quality samples are stored in the sRNA expression database while corresponding quality information is saved in the data quality repository. SEA also stores expression information of high-quality predicted miRNAs including the ID, organism, chromosomal location, precursor and mature sequences, structure, read counts, prediction scores, and detailed information on the software and its versions used to predict the miRNA. SEA's primary analysis results including per sample quality and expression information can be examined and downloaded as interactive web reports. Detailed information on the primary analysis of sRNAs and predicted miRNAs can be found in the Oasis 2 manuscript ([biorxiv.org](http://biorxiv.org)).

In order to reduce bias that could be introduced into the data by using different analysis routines, every sample in SEA has been analyzed by identical analysis workflows using identical databases and annotations. In case of changes in databases or analysis routines, SEA additionally stores versioning information about the software and databases used for an analysis. In addition, SEA contains information about the Geo series accession (GSE) and sample accession (GSM) identifiers along with the sample ID from the Sequence Read Archive (SRA) database (SRR) (Barrett *et al.*, 2013). Given that most meta-data is quite different between experiments we opted to store this expression data and meta-data in a Not Only SQL (NoSql) MongoDB<sup>2</sup> database management system. We optimized search and retrieval times by indexing for the most common queries and most relevant terms.

### **2.3 Standardized annotation**

To allow for the interoperability of data it is important to standardize annotations using ontologies and semantic mapping (Schuurman and Leszczynski, 2008). Ontologies define standard terms, their properties, and the relations between them and dataset terms that are connected to Ontologies are called ‘normalized’. The Ontologies and the number of normalized terms in SEA are listed in Table 1.

SEA’s sRNA annotation workflow maps free-text GEO annotations to standardized terms in three consecutive steps. In general, GEO data annotations are free text that can be parsed into key-value pairs. In a first fully automated step the annotation workflow extracts key-value relations and stores them in the annotation database. As GEO data information is unstructured and contains very different information, we opted for a NoSql annotation database with an optimized indexing for prototypical questions (see also section 2.2).

The second fully automated step normalizes the extracted keys and values using Ontologies as standard dictionaries. SEA has a list of pre-defined keys, five of which (organism, tissue, disease, cell type, and cell line) can be currently queried for in SEA. Each extracted key is compared to pre-defined keys. For values, the ontologies are used as standard terminology dictionaries. For each pre-defined key, SEA has one of several corresponding ontologies. Each extracted value is searched in the corresponding ontologies and, if the same or a similar term is found, connected with it.

---

<sup>2</sup> <https://www.mongodb.com/>

Automatic annotation is followed by semi-automatic manual curation. For that purpose, we developed an internal curation Web interface using Groovy/Grails<sup>3</sup>, which allows browsing and editing of annotations from the annotation database as well as manual normalization of keys and values in annotations, searching among pre-defined keys and corresponding ontologies. Thus, curators examine all keys and values for consistency and update missing or additional information with standardized terms where necessary (e.g. protocols, kit version, lot and batch numbers, publications). At the moment, all SEA annotations are manually curated, a quality standard that we intend to keep for every future SEA entry.

## 2.4 Querying and visualization

To enable the search across ontological hierarchies we integrated the relevant ontologies into the graph database Neo4j<sup>4</sup> (Figure 1). Graph databases are NoSQL databases which support storage of objects and connections between them, as is the case for ontologies. Following the manual curation (see section 2.3), sample annotations are uploaded to the SEA ontology graph database including all ontological parent terms (having an ‘is-a’ relation to it). This allows search by ontology terms, as well as by their parents, which are in fact groups of terms (e.g. ‘cancer’ or ‘neurodegenerative disease’). SEA accesses the ontology graph database via the Ontology Lookup Service using a REST interface, supporting complex and compound queries and query auto-completion (Côté *et al.*, 2010).

---

<sup>3</sup> <https://grails.org/>

<sup>4</sup> <https://neo4j.com/>

### 3 Results & Conclusions

SEA is designed for the biological or medical end-user that is interested to define where and when an sRNA of interest is expressed. Prototypical questions that can be addressed with SEA are: What is the expression of hsa-miR-488-5p across all human tissues? Is hsa-miR-488-5p expressed higher in adenocarcinomas as compared to other cancer types? Is the tissue-specific expression of hsa-miR-488-5p conserved in mouse? Its unique selling points are the deep and standardized annotation of meta-information, the re-analysis of published data with Oasis 2 to reduce analysis bias, a user-friendly search interface that supports complex queries, and the fast and interactive visualization of analysis results across 10 organisms (Table 2) and various sRNA-species. SEA also contains information on the expression of over 700 high-quality predicted miRNAs, across organisms and tissues. Last but not least, SEA is continuously growing and aims to eventually encompass all sRNA-seq datasets across all organisms deposited in GEO and other repositories. Genome versions will be updated with every major release of SEA. SEA will be backwards compatible in the future by allowing users to choose previous genome versions and annotations. A detailed comparison of SEA to other existing sRNA expression databases highlights that SEA is superior in terms of supported organism, annotations, diseases, and tissues. SEA contains over 2000 samples in its database, which is considerably less than YM500v3 (Chung *et al.*, 2016), which hosts over 8000 cancer samples. It is to be noted, however, that the YM500v3 database only supports cancer datasets and no other disease types (Table 3).

As far as we are aware SEA is the only sRNA-seq database that supports ontology-based queries, supporting single or combined searches for five pre-defined keys (organism, tissue, disease, cell type, and cell line) across all datasets. However, the SEA database system contains additional (meta)-information including age, gender, developmental stage, genotype as well as technical experimental details such as the sequencing instrument and protocol details (e.g. library kit, RNA extraction procedure). We plan to normalize most of this additional information in future versions of SEA. This will allow users, for example, to query and analyze sRNA expression effects that are introduced by library kit or sequencing platform differences (both of these features can introduce large biases in the detection and expression of sRNAs). Other future developments will include information on sRNA editing, modifications, and mutation events.

In summary, SEA supports interactive result visualization on all levels, from querying and display of sRNA expression information to the mapping and quality information for each of the over 2000 samples. SEA is a fast, flexible, and fully interactive web application for the investigation of sRNA expression across cell lines, tissues, diseases, organisms, and sRNA-species. As such, SEA should be a valuable addition to the landscape of sRNA expression databases.

## **ACKNOWLEDGEMENTS**

We would like to thank Mariah Snyder, Ashish Rajput, Ting Sun, Vikas Bansal, Michel Edwar Mickael, the DZNE IT, and all of the Oasis users for helpful suggestions.

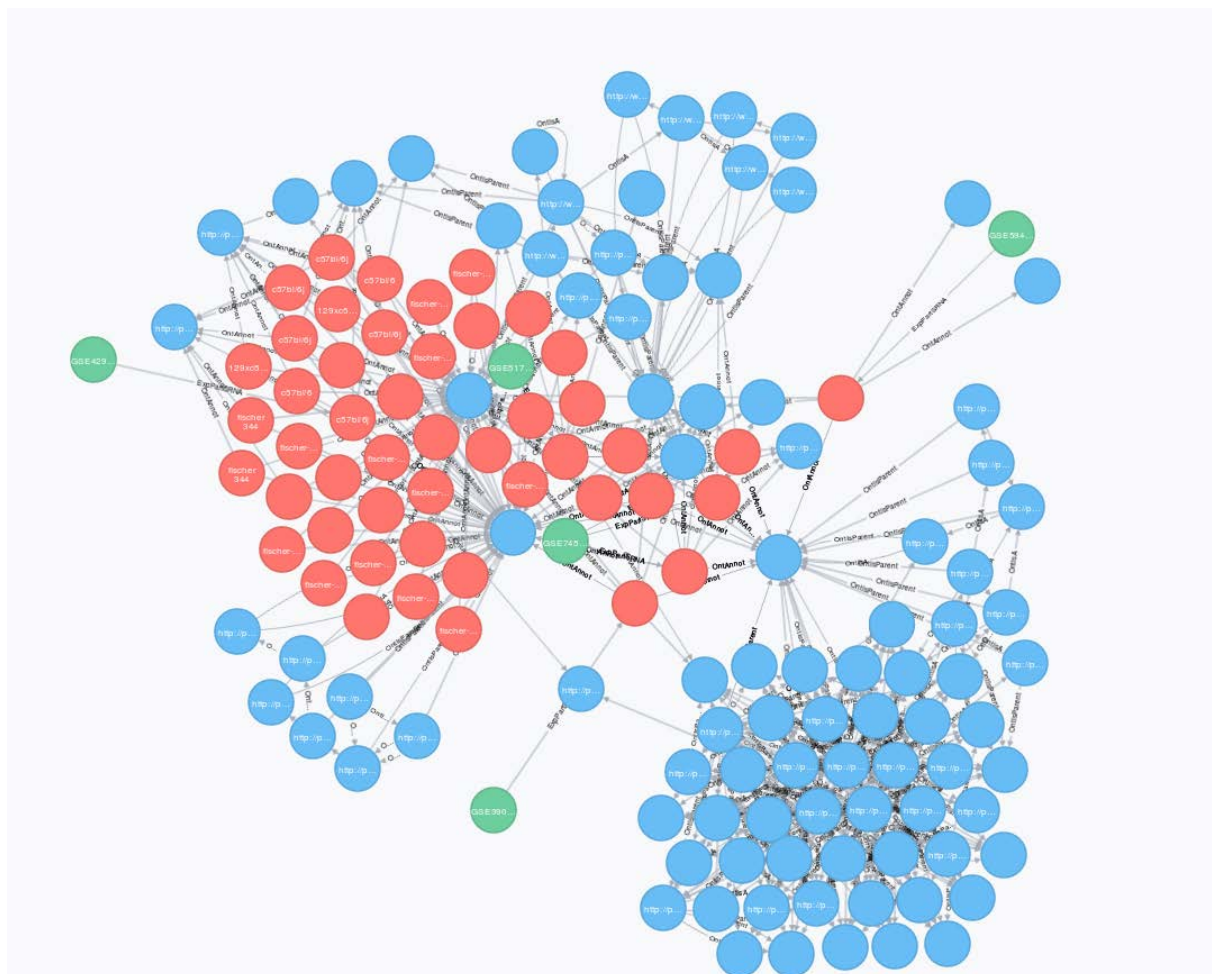
## **FUNDING**

This work was supported by the DFG (BO4224/4-1), the Network of Centres of Excellence in Neurodegeneration (CoEN) initiative, the Volkswagen Stiftung (Az88705), iMed – the Helmholtz Initiative on Personalized Medicine, and the BMBF grant Integrative Data Semantics in Neurodegeneration (031L0029B, [IDSN](#)).



## REFERENCES

- Barrett,T. *et al.* (2013) NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.*, **41**, 991–995.
- Capece,V. *et al.* (2015) Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*, **31**, 1–3.
- Chung,I.-F. *et al.* (2016) YM500v3: a database for small RNA sequencing in human cancer research. *Nucleic Acids Res.*, **45**, D925–D931.
- Côté,R. *et al.* (2010) The Ontology Lookup Service: Bigger and better. *Nucleic Acids Res.*, **38**, 155–160.
- Leung,Y.Y. *et al.* (2016) DASHR: Database of Small human noncoding RNAs. *Nucleic Acids Res.*, **44**, D216–D222.
- Panwar,B. *et al.* (2017) miRmine: A Database of Human miRNA Expression Profiles. *Bioinformatics*.
- Schuurman,N. and Leszczynski,A. (2008) Ontologies for Bioinformatics. *Bioinform. Biol. Insights*, **2**, 187–200.
- Vitsios,D.M. *et al.* (2017) Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis. *Nucleic Acids Res.*, **45**, 1079–1090.
- Witwer,K.W. (2014) Circulating MicroRNA Biomarker Studies: Pitfalls and Potential Solutions. *Clin. Chem.*, **000**.



**Fig.1.** Objects in the SEA graph database (Neo4j). A fragment of the SEA graph database is visualized, where green nodes represent datasets, red nodes represent samples and blue nodes represent ontology terms. Grey edges represent ‘is a’ relations between the different datasets, samples, and ontology terms.

**Table 1.** SEA keys and used ontologies (as of April, 21<sup>st</sup> 2017).

<b>Key</b>	<b>Ontology(s)</b>	<b># Annotations</b>	<b># Terms</b>
<b>Organism</b>	NCBI Taxonomy <sup>5</sup>	2105	10
<b>Tissue</b>	BRENDA tissue / enzyme source <sup>6</sup>	1595	86
<b>Disease</b>	Human Disease Ontology <sup>7</sup>	791	68
<b>Cell type</b>	Cell Ontology <sup>8</sup>	517	57
<b>Cell line</b>	Cell Line Ontology <sup>9</sup>	39	12
	Experimental Factor Ontology <sup>10</sup>	253	55

---

<sup>5</sup> <https://www.ncbi.nlm.nih.gov/taxonomy>

<sup>6</sup> <http://www.brenda-enzymes.info/>

<sup>7</sup> <http://www.disease-ontology.org/>

<sup>8</sup> <http://obofoundry.org/ontology/cl.html>

<sup>9</sup> <http://www.clo-ontology.org/>

<sup>10</sup> <http://www.ebi.ac.uk/efo/>

**Table 2.** Supported SEA organisms and their corresponding genome versions.

<b>Organism</b>	<b>genome-version</b>	<b>genome-date</b>
<b>Bos taurus</b>	UMD3.1	2009-11
<b>Caenorhabditis elegans</b>	WBcel235	2012-12
<b>Danio rerio</b>	GRCz10	2014-09
<b>Drosophila melanogaster</b>	BDGP6	2014-07
<b>Mus musculus</b>	GRCm38	2012-01
<b>Gallus gallus</b>	Ggal4	2011-11
<b>Rattus norvegicus</b>	Rnor_6.0	2014-07
<b>Homo sapiens</b>	GRCh38	2013-12
<b>Sus scrofa</b>	Sscrofa10.2	2011-08
<b>Anopheles gambiae</b>	AgamP4	2006-02

Feature	SEA	miRmine <sup>1</sup>	DASHR <sup>2</sup>	miratlas <sup>3</sup>	YM500v3 <sup>4</sup>
Organisms	10	1	1	2	1
sRNA types	5	1	5	1	5
Samples	>2000	304	187	461	>8000*
Novel miRNAs	+	-	-	-	-
Ontology search <sup>#</sup>	+	-	-	-	-





**Table 3.** Comparison of sRNA expression databases. This table includes recent sRNA expression databases and a list of features we deem relevant. \*Supports mainly cancer-related datasets. <sup>#</sup>Use of ontological graphs for the annotation and querying of samples. <sup>1</sup>(Panwar et al., 2017), <sup>2</sup>(Leung et al., 2016), <sup>3</sup>(Vitsios et al., 2017), <sup>4</sup>(Chung et al., 2016)

## Appendix C

### Article 3



# The landscape of human mutually exclusive splicing

Klas Hatje<sup>1,2,†</sup>, Raza-Ur Rahman<sup>2,3</sup>, Ramon O Vidal<sup>2,‡</sup>, Dominic Simm<sup>1,4</sup>, Björn Hammesfahr<sup>1,§</sup>, Vikas Bansal<sup>2,3</sup> , Ashish Rajput<sup>2,3</sup> , Michel Edwar Mickael<sup>2,3</sup>, Ting Sun<sup>2,3</sup>, Stefan Bonn<sup>2,3,5,\*</sup>  & Martin Kollmar<sup>1,\*\*</sup> 

## Abstract

Mutually exclusive splicing of exons is a mechanism of functional gene and protein diversification with pivotal roles in organismal development and diseases such as Timothy syndrome, cardiomyopathy and cancer in humans. In order to obtain a first genome-wide estimate of the extent and biological role of mutually exclusive splicing in humans, we predicted and subsequently validated mutually exclusive exons (MXEs) using 515 publically available RNA-Seq datasets. Here, we provide evidence for the expression of over 855 MXEs, 42% of which represent novel exons, increasing the annotated human mutually exclusive exome more than fivefold. The data provide strong evidence for the existence of large and multi-cluster MXEs in higher vertebrates and offer new insights into MXE evolution. More than 82% of the MXE clusters are conserved in mammals, and five clusters have homologous clusters in *Drosophila*. Finally, MXEs are significantly enriched in pathogenic mutations and their spatio-temporal expression might predict human disease pathology.

**Keywords** alternative splicing; differential expression; mutually exclusive splicing; splicing mechanisms

**Subject Categories** Chromatin, Epigenetics, Genomics & Functional Genomics; Genome-Scale & Integrative Biology; Transcription

**DOI** 10.15252/msb.20177728 | Received 2 May 2017 | Revised 4 November 2017 | Accepted 10 November 2017

**Mol Syst Biol.** (2017) **13**: 959

## Introduction

Alternative splicing of pre-messenger RNAs is a mechanism common to almost all eukaryotes to generate a plethora of protein

variants out of a limited number of genes (Matlin *et al.*, 2005; Nilsen & Graveley, 2010; Lee & Rio, 2015). High-throughput studies suggested that not only 95–100% of all multi-exon genes in human are affected (Pan *et al.*, 2008; Wang *et al.*, 2008; Gerstein *et al.*, 2014) but also that alternative splicing patterns strongly diverged between vertebrate lineages implying a pronounced role in the evolution of phenotypic complexity (Barbosa-Morais *et al.*, 2012; Merkin *et al.*, 2012). Five types of alternative splicing have been identified to contribute to most mRNA isoforms, which are differential exon inclusion (exon skipping), intron retention, alternative 5' and 3' exon splicing, and mutually exclusive splicing (Blencowe, 2006; Pan *et al.*, 2008; Wang *et al.*, 2008; Nilsen & Graveley, 2010). Mutually exclusive splicing generates alternative isoforms by retaining only one exon of a cluster of neighbouring internal exons in the mature transcript and is a sophisticated way to modulate protein function (Letunic *et al.*, 2002; Meijers *et al.*, 2007; Pohl *et al.*, 2013; Tress *et al.*, 2017a). The most extreme cases known so far are the arthropod *DSCAM* genes, for which up to 99 mutually exclusive exons (MXEs) spread into four clusters were identified (Schmucker *et al.*, 2000; Lee *et al.*, 2010; Pillmann *et al.*, 2011).

Opposed to arthropods, current evidence suggests that vertebrate MXEs only occur in pairs (Matlin *et al.*, 2005; Gerstein *et al.*, 2014; Abascal *et al.*, 2015a), and genome-wide estimates in human range from 118 (Suyama, 2013) to at most 167 cases (Wang *et al.*, 2008). Despite these relatively few reported cases, mutually exclusive splicing might be far more frequent in humans than currently anticipated, as has been recently revealed in the model organism *Drosophila melanogaster* (Hatje & Kollmar, 2013). Apart from their low number, MXEs have been described in many crucial and essential human genes such as in the  $\alpha$ -subunits of six of the 10 voltage-gated sodium channels (*SCN* genes) (Copley, 2004), in each of the glutamate receptor subunits 1–4 (*GluR1–4*) where the MXEs are called flip and flop (Sommer *et al.*, 1990), and in *SNAP-25* as part of

1 Group Systems Biology of Motor Proteins, Department of NMR-Based Structural Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

2 Group of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany

3 Center for Molecular Neurobiology, Institute of Medical Systems Biology, University Clinic Hamburg-Eppendorf, Hamburg, Germany

4 Theoretical Computer Science and Algorithmic Methods, Institute of Computer Science, Georg-August-University, Göttingen, Germany

5 German Center for Neurodegenerative Diseases, Tübingen, Germany

\*Corresponding author. Tel: +49 40 7410 55082; E-mail: sbonn@uke.de

\*\*Corresponding author. Tel: +49 551 5036960; E-mail: mako@nmr.mpibpc.mpg.de

†Present address: Roche Pharmaceutical Research and Early Development, Pharmaceutical Sciences, Roche Innovation Center Basel, Basel, Switzerland

‡Present address: Max-Delbrück-Center for Molecular Medicine, Berlin, Germany

§Present address: Research and Development—Data Management (RD-DM), KWS SAAT SE, Einbeck, Germany

the neuroexocytosis machinery (Johansson *et al.*, 2008). Although MXEs within a cluster often share high similarity at the sequence level, they are usually not functionally redundant, as their inclusion in the mRNAs is tightly regulated. Thus, mutations in MXEs have been shown to cause diseases such as Timothy syndrome (missense mutation in the *CACNA1C* gene) (Splawski *et al.*, 2004, 2005), cardiomyopathy (defect of the mitochondrial phosphate carrier *SLC25A3*) (Mayr *et al.*, 2011) or cancer (mutations in, e.g., the pyruvate kinase *PKM* and the zinc transporter *SLC39A14*) (David *et al.*, 2010).

Despite the implications of mutually exclusive splicing in organismal development and disease, current knowledge on the magnitude of MXE usage and its relevance in biological processes is far from complete. In order to obtain a genomewide, unbiased estimate of the extent and biological role of mutually exclusive splicing in humans, a set of 6,541 MXE candidates was compiled from annotated and novel predicted exons, and rigorously validated using over 15 billion reads from 515 RNA-Seq datasets.

## Results

### The human genome contains 855 high-confidence MXEs

Compared to other splicing mechanisms, mutually exclusive splicing in humans seems to be a rare event. MXEs are characterized by genomic vicinity, splice-site compatibility and mutually exclusive presence in protein isoforms. Accordingly, the human genome annotation (GenBank v. 37.3) contains only 158 MXEs in 79 protein-coding genes (Appendix Figs S1–S3). MXEs are often phrased “homologous exons” in the literature because they likely originated from the same ancestral exon. We refrain from using this term throughout our analysis, because several MXEs present in the genome annotation do not show any sequence homology and many neighbouring exons with high sequence similarity are not spliced in a mutually exclusive manner.

In a first attempt to chart an atlas of genomewide mutually exclusive splicing in humans, we decided to predict potential MXE candidates and validate those using published RNA-Seq data. In a first step, we generated a set of MXE candidates in the human genome

(v. 37.3) from all annotated protein-coding exons and from novel exons predicted in intronic regions including only internal exons in the candidate list (Fig 1A, Appendix Figs S1–S4). From the annotated exons, we selected those that appeared mutually exclusive in transcripts, and neighbouring exons that show sequence similarity and are translated in the same reading frame. To generate novel exon candidates, we predicted exonic regions in neighbouring introns of annotated exons based on sequence similarity and similar lengths (Pillmann *et al.*, 2011). We did not consider potential MXEs containing in-frame stop codons such as the neonatal-specific MXE reported for the sodium channel *SCN8A* (Zubović *et al.*, 2012), and exons overlapping annotated terminal exons (Appendix Fig S2). The reconstruction resulted in a set of 6,541 MXE candidates in 1,542 protein-coding genes, including 1,058 (68.6%) genes for which we predicted 1,722 completely novel exons in previously intronic regions (Fig 1B). Most introns in human genes are extremely long necessitating careful and strict validation of the MXE candidates to exclude false-positive predictions (Lee & Rio, 2015).

To validate the predicted MXE candidates, we made use of over 15 billion publically available RNA-Seq reads, selecting 515 samples comprising 31 tissues and organs, 12 cell lines and seven developmental stages (Barbosa-Morais *et al.*, 2012; Djebali *et al.*, 2012; Tilgner *et al.*, 2012; Xue *et al.*, 2013; Yan *et al.*, 2013; Fagerberg *et al.*, 2014; Dataset EV1). The data were chosen to encompass common and rare potential splice events in a broad range of tissues, cell types and embryonic stages. Accordingly, the transcription of 6,466 (99%) of the MXE candidates is supported by RNA-Seq reads mapped to the genome (Appendix Fig S3A). To be validated as true mutually exclusive splicing event, each MXE of a cluster needed to exhibit splice junction (SJ) reads from every MXE to up- or downstream gene regions bridging the other MXE(s) of the cluster (Fig 1A). In addition, MXEs should not exhibit any SJ reads to another MXE except when the combined inclusion causes a frame shift and therefore a premature stop codon (Fig 1A, Appendix Figs S3A and D, S5, and S6). These stringent criteria define a high-confidence set of MXEs, requiring three constraints for a cluster of two MXEs and already 18 constraints for a cluster of five MXEs (Appendix Fig S7). In case of clusters with more than two MXE candidates, the validation criteria were applied to the cluster including all MXE candidates as well as to all possible sub-clusters to

**Figure 1. The human genome contains 1,399 high-confidence MXEs.**

- Schematic representation of the various annotated and predicted exon types included in the MXE candidate list. For MXE validation, at least three restraints must be fulfilled: the absence of an MXE-joining read (R1), except for those leading to frame shift, and the presence of two MXE-bridging SJ reads (R2 and R3).
- Prediction and validation of 1,399 1SJ (855 3SJ) human MXEs. Top: Dataset of 6,541 MXE candidates from annotated and predicted exons. Bottom left: MXE candidates for which splice junction data are currently missing hindering their annotation as MXE or other splice variant. Bottom right: Validation of the MXE candidates using over 15 billion RNA-Seq reads. The outer circles represent the validation based on at least a single read for each of the validation criteria (1SJ), while the validation shown in the inner circles required at least three reads (3SJ).
- MXE saturation analysis. Whereas increasing amounts of RNA-Seq reads should lead to the confirmation of further MXE candidates, more RNA-Seq reads might also result in the rejection of previously validated MXEs. The green curves show the number of validated MXEs in relation to the percentage of total RNA-Seq reads used for validation. The orange curves indicate the number of initially “validated MXEs” that were rejected with increasing amounts of reads. Grey dashed lines indicate the point of saturation, which is defined as the point where a twofold increase in reads leads to rejection of less than 1% of the validated MXEs. Of note, whereas the rejection of validated MXEs saturates with 20% of the data, the amount of novel MXE validations is still rapidly increasing.
- Distribution of validated MXEs in two-exon and multi-exon clusters.
- Size and distribution of multi-cluster MXEs.
- The *CUX1* gene (cut-like homeobox 1) contains two interleaved clusters of MXEs (clusters 1 and 2) and two standard clusters each with two MXEs (clusters 3 and 4). The exon 3 and exon 4 variants each are orthologous exons. The exon 4 variants are mutually exclusive (cluster 2). Exon 3a is a differentially included exon and only spliced together with exon 4a. The exons 3b, 3c, 3d and 3e are part of a cluster of four MXEs (cluster 1) and are only spliced together with exon 4b (Appendix Figs S16 and S17). Novel exons are labelled with an asterisk.



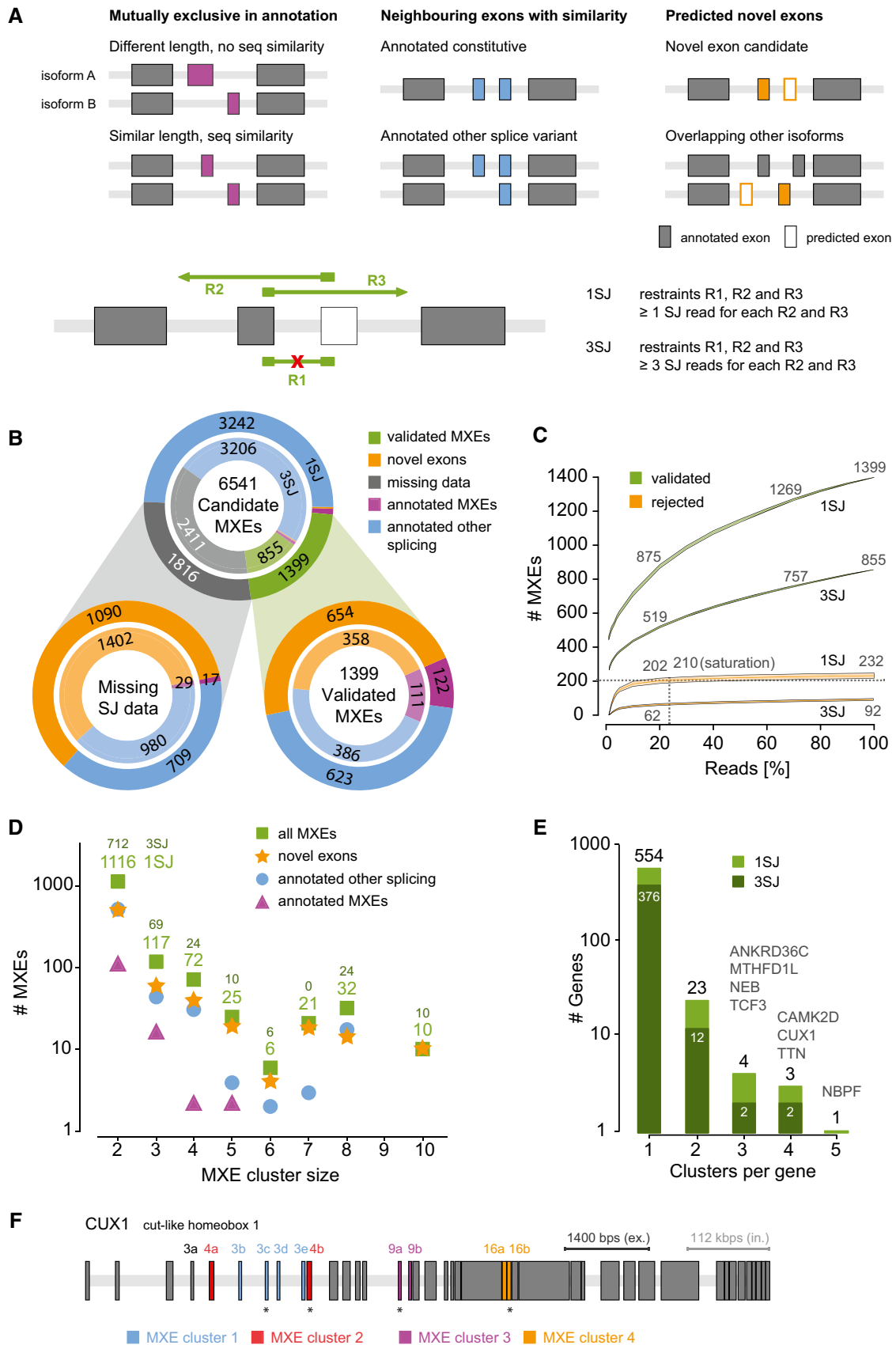


Figure 1.

identify the largest cluster fulfilling all MXE criteria. According to these criteria, 1,399 MXEs were verified with at least one SJ read per exon (1SJ), supported by 2.2 million exon mapping and 34 million SJ reads, increasing the total count of human MXEs by almost an order of magnitude (158–1,399) (Fig 1B, Dataset EV2); 855 MXEs were found to be supported by at least three splice junction reads per exon (3SJ) validated by 1.5 million exon mapping and 27 million SJ reads (Appendix Figs S3B and C, S8–S10). The 1,399 (855, numbers in brackets refer to the 3SJ validation) verified MXEs include 122 (112) annotated MXEs (Fig 1B “annotated MXE”), 623 (388) exons that were previously annotated as constitutive or differentially included (“annotated other splicing”) and 654 (358) exons newly predicted in intronic regions (“novel exon”). Our analysis also showed that 29 of the 158 annotated MXEs are in fact not mutually exclusively spliced but represent constitutively spliced exons or other types of alternative splicing (Appendix Figs S2 and S3E). Finally, 1,741 (2,336) MXE candidates including 1,090 (1,402) newly predicted exons and 17 (29) of the annotated MXEs are supported by 0.5 million exon and 13 million SJ matching reads but still have to be regarded as MXE candidates because not all annotation criteria were fulfilled (Appendix Fig S3A and E).

To estimate the dependence of MXE confirmation and rejection on data quantity, we cross-validated the MXE gain (validation) and loss (rejection) events for several subsets of the total RNA-Seq data (Fig 1C, Appendix Fig S11, Materials and Methods “Saturation analysis”). The course of the curves provides strong evidence for the validity of the MXEs because a single exon-joining read would already be sufficient to reject an MXE cluster while at least two SJ reads are needed to validate one. Whereas even 15 billion RNA-Seq reads do not achieve saturation for the amount of validated MXEs, the gain in rejected MXE candidates is virtually saturated using 25% of the data.

To further validate the list of MXEs, we compared MXE clusters that contained two “annotated other splicing” exons to splicing information from GTEx portal (<https://www.gtexportal.org/home/>). Although GTEx portal uses an alternative aligner and different alignment settings, all MXEs that we compared showed mutually exclusive behaviour in GTEx portal (Appendix Fig S12), substantiating our results. Lastly, we selected six brain-expressed novel MXEs for qPCR validation in human brain total RNA. All assayed MXEs showed perfect coherence with the alignment results, confirming mutually exclusive splicing of all assayed novel MXEs in human brain (Appendix Fig S13, Dataset EV3).

Many of the 1,399 (855) MXEs have roles in the cardiac and muscle function and development, while cassette exons are enriched for microtubule- and organelle localization-related terms (Appendix Fig S14).

In summary, the high-confidence set of 1,399 (855) MXEs extends current knowledge of human MXE usage by an order of magnitude, (re)-annotating over a thousand existing and predicted exons and isoforms, while suggesting the existence of further human MXEs.

### The human genome contains large cluster and multi-cluster MXEs

In general, mutually exclusive splicing can be quite complex. This is best demonstrated by genes in arthropods that contain both multiple MXE clusters (“multi-cluster”) and large clusters with up to 53 MXEs

such as in the *Drosophila Dscam* genes (Graveley et al, 2004; Pillmann et al, 2011). This is in strong contrast to mutually exclusive splicing in vertebrates as there is to date no evidence of multi-cluster or higher order MXE clusters (Matlin et al, 2005; Pan et al, 2008; Wang et al, 2008; Gerstein et al, 2014; Abascal et al, 2015a,b).

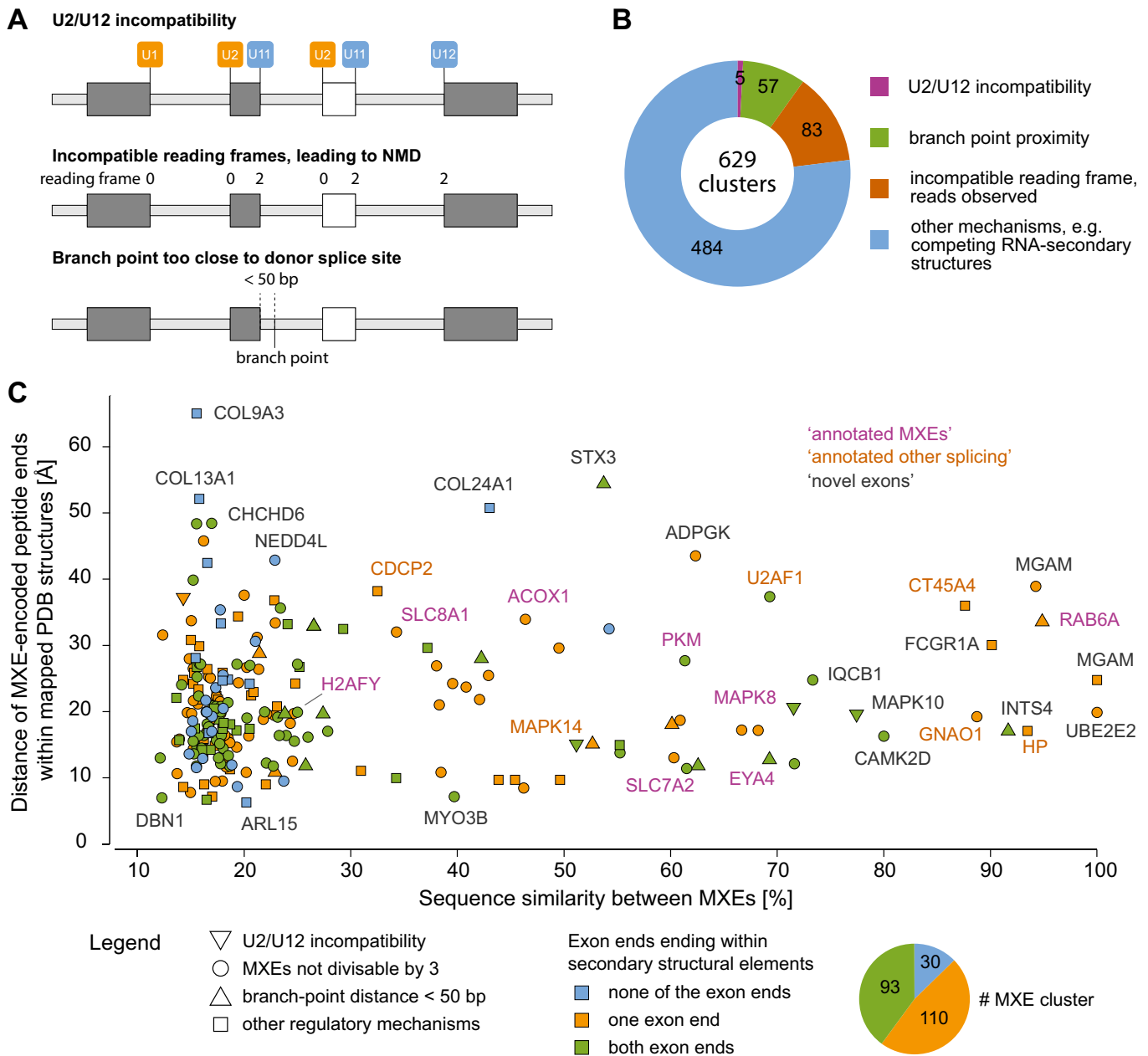
The analysis of the 1,399 validated human MXEs provides first evidence for clusters of multiple MXEs in the human genome (Fig 1D, Appendix Fig S15). While most MXEs are present in clusters of two exons (1,116 MXEs), a surprisingly high number of clusters have three to 10 MXEs (283 MXEs in 71 clusters).

Interestingly, although a large part of the verified MXEs contain a single MXE cluster (554 genes, Fig 1E), we could also provide evidence for human genes containing multiple MXE clusters. Thus, *TCF3*, *NEB*, *ANKRD36C* and *MTHFD1L* contain three clusters and *TTN*, *CAMK2D* and *CUX1* four clusters of MXEs. A very interesting case of complex interleaved mutually exclusive splicing can be seen for *CUX1*, the transcription factor cut-like homeobox 1. It contains a cluster of MXEs (exons 3b–3e) that is differentially included into a set of two exons (exon 3 and exon 4), and the two sets are themselves mutually exclusive (Fig 1F, Appendix Figs S16 and S17). The identification of large clusters with multiple MXEs and many genes with multiple clusters shows that complex mutually exclusive splicing is not restricted to arthropods (Schmucker et al, 2000; Graveley, 2005; Lee et al, 2010; Hatje & Kollmar, 2013) but might be present in all bilateria.

### Mutually exclusive presence of coding exons in functionally active transcripts

To understand which splicing mechanisms might be primarily responsible for the regulation of mutually exclusive splicing in humans, we investigated several mechanisms that were shown to act in some specific cases and were proposed to coordinate mutually exclusive splicing in general (Fig 2A; Letunic et al, 2002; Smith, 2005). We identified five cases (0.79% of all clusters) of U2 and U12 splice acceptor incompatibility (Appendix Fig S18) and 57 (9%) cases of potential steric interference, a too short distance between splice donor sites and branch points (< 50 bp; Fig 2B and Appendix Fig S19). Although 377 (60%) of the MXE clusters contain exons with exon lengths not divisible by three which would result in non-functional transcripts in case of combined inclusion, MXE-joining reads were found for only 83 (22%) of these clusters (Fig 2B; Appendix Figs S3B and D, and S20). Surprisingly, the majority of the annotated MXEs are of this type (91 of 122; 75%) as well as many exons previously annotated as other splice types (44 of 662), but only few of the novel MXEs predicted in intronic regions (25 of 615; Appendix Fig S3A and D). These numbers suggest that splicing of the remaining 484 MXE clusters is tightly regulated by other mechanisms (Fig 2B) such as RNA–protein interactions, interactions between small nuclear ribonucleoproteins and splicing factors (Lee & Rio, 2015), and competitive RNA secondary structural elements (Graveley, 2005; Yang et al, 2012; Lee & Rio, 2015). Competing RNA secondary structures are, however, usually not conserved across long evolutionary distances. A potential case of a docker site and selector sequences downstream of each exon variant was identified for the cluster of four MXEs in the *CD55* gene (Appendix Fig S21).

In contrast to cassette exons and micro-exons, which tend to be located in surface loops and intrinsically disordered regions instead



**Figure 2. MXE presence is regulated at the RNA and protein folding level.**

A Schematic representation of MXE splicing regulation via splice-site incompatibility, branch point proximity and translational frame shift leading to NMD.

B Observed usage of MXE splicing regulation in 629 MXE clusters.

C By mutually exclusive inclusion into transcripts, MXEs of a cluster are supposed to encode the same region of a protein structure. If the respective regions of the protein structures are embedded within secondary structural elements (the ends of the exon-encoded peptides are part of  $\alpha$ -helices and/or  $\beta$ -strands), it is highly unlikely that the translation of a transcript will result in a folded protein in case the respective exon is missing (skipped exon). If the MXEs have highly similar sequences and do not encode repeat regions, it seems unlikely that either could be present in tandem or absent at all in a folded protein. Here, we have combined protein structure features (colours) with splicing regulation information (symbols). Accordingly, 87% of the MXE-encoded protein regions are embedded in secondary structural elements (orange and green symbols), and most of the remaining MXEs can only be spliced mutually exclusive because splicing as differentially included exons would lead to frame shifts (blue circles). As examples, we labelled many MXE clusters distinguishing annotated MXEs (purple letters), known exons that we validated as MXEs (orange letters), and clusters containing novel exons (dark-grey letters).

of folded domains (Buljan *et al.*, 2012; Ellis *et al.*, 2012; Irimia *et al.*, 2014), all MXEs, whose protein structures have been analysed, are embedded within folded structural domains as has been shown for, for example, *DSCAM* (Meijers *et al.*, 2007), *H2AFY* (Kustatscher

*et al.*, 2005), the myosin motor domain (Kollmar & Hatje, 2014) and *SLC25A3* (Tress *et al.*, 2017a). As we have shown in the beginning, there is also a subset of 73 MXEs not showing any sequence homology (“annotated no similarity”). It is unlikely that the encoded

peptides account for identical secondary structural elements. Rather, if the MXEs of this subset are true MXEs, there is a small subset (about 5%) of MXEs whose mutual inclusion leads to considerably altered protein folds or affects surface loops and disordered regions similar to cassette exons.

Because MXEs are supposed to modulate protein functions through variations and not alterations in specific restricted parts of the structure, we thought it could be possible to distinguish MXEs from cassette exons at a protein structural level. Such an analysis could provide complementary evidence for the validation as MXE in contrast to two (or more) neighbouring cassette exons. While one and only one of the exons of a cluster of MXEs has to be included in the transcript, the defining feature of a cassette exon is that it can either be present or absent. If MXEs were mis-classified and in fact neighbouring cassette exons, it would therefore be possible that all exons of the cluster were present or absent from the transcript, and accordingly the protein structure. These differences between MXEs and cassette exons impose three restrictions on their localization within protein folds (Appendix Fig S22). Thus, (i) if one or both ends of the MXE-encoded peptide end within a secondary structural element, it seems impossible that the respective peptide could be absent from the protein because this would break up multiple spatial interactions. This suggests that respective protein regions cannot be encoded by cassette exons. (ii) High sequence similarity between MXEs suggests important conserved structural interactions even if the peptide ends are not part of secondary structural elements. For example, it seems highly unlikely that a cluster of two exons encoding transmembrane helices could be spliced as cassette exons because absence or presence of both exons would switch the membrane site of all subsequent sequence. (iii) In case of cassette exons and absence of the exons, it must be possible that the remaining sequence still folds correctly. This can be assessed if a protein structure is available with the respective exon-encoded region present. Supposing the respective region was absent, the remaining ends would need to be joined to result in a correctly folded domain, which seems extremely unlikely if the peptide ends are far apart. Such regions are also more likely encoded by MXEs. To assess this model, we mapped the validated MXEs against the PDB database (Fig 2C, Appendix Fig S22, Dataset EV4; Rose *et al.*, 2015). Of the 1,399 MXEs, 273 MXEs (20%) from 233 MXE clusters (37%) matched to human or mammalian protein structures (Appendix Fig S22). For 87% of these MXEs, at least one of the exon termini is embedded within a secondary structural element, suggesting that these exons are in fact true MXEs and not mis-classified cassette exons (Fig 2C, yellow and green coloured symbols). This high level of structural conservation also strongly supports the hypothesis that MXEs modulate but do not considerably alter protein functions (Letunic *et al.*, 2002; Yura *et al.*, 2006; Abascal *et al.*, 2015a; Tress *et al.*, 2017a). Of the remaining 13% (Fig 2C, blue coloured symbols), many MXEs would lead to frame shifts if they were spliced as cassette exons (both exons present or absent in the transcript, blue circles), and in multiple cases (e.g. *COL9A3*, *COL24A1* and *COL13A1*), the peptide ends are far apart indicating strong folding problems in case the respective exons were absent in the transcripts. In total, there are only a handful cases such as the MXE cluster in *ARL15* (Fig 2C) whose mutually exclusive presence in proteins cannot be explained by the analysed splicing restrictions, by NMD targeting, or by folding constraints.

### MXEs mainly consist of one ubiquitous exon and otherwise regulated exons

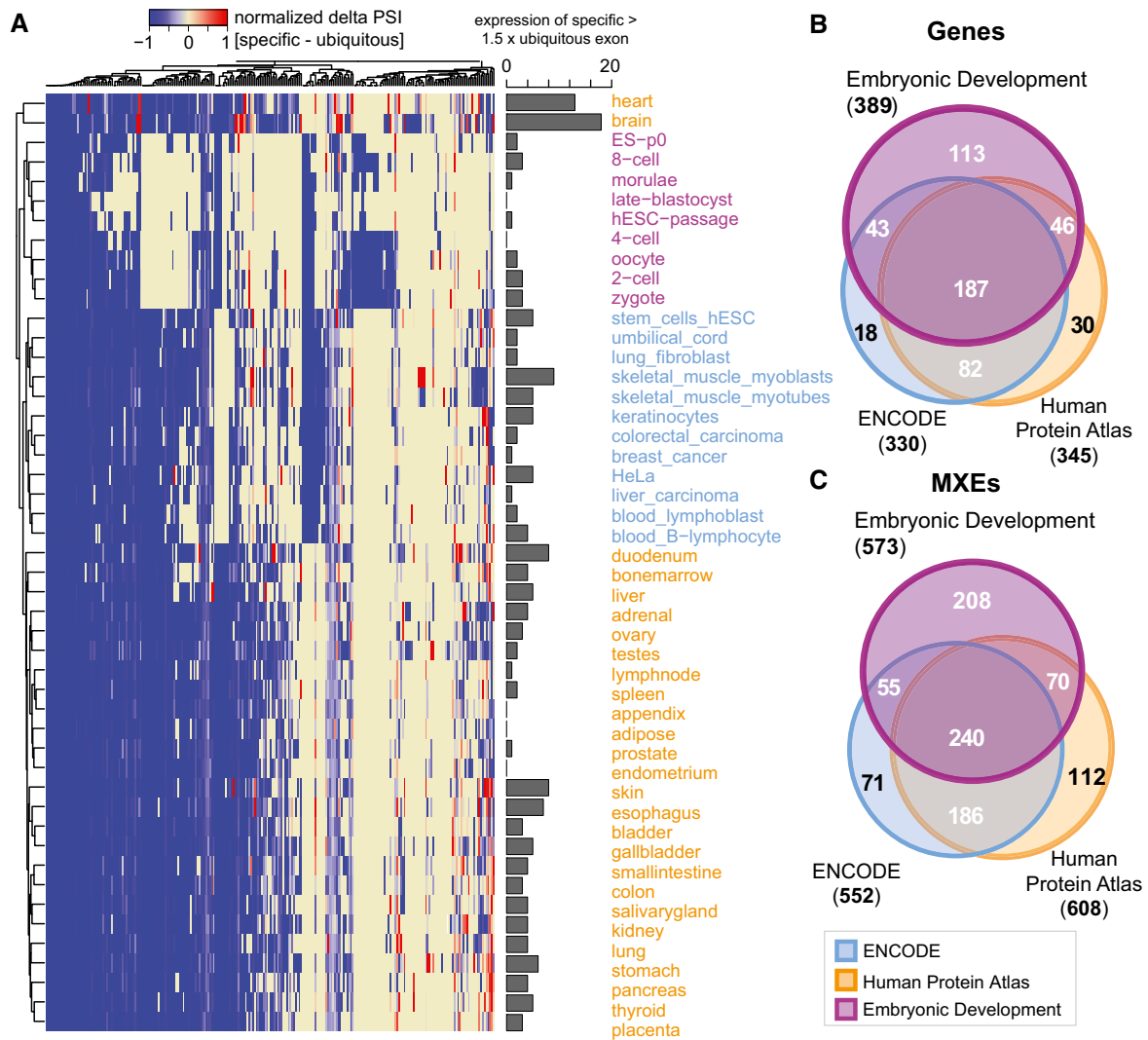
To modulate gene functionality, mutually exclusive splicing would need spatial and temporal splicing regulation and expression. To understand the expression patterns of MXEs, we conducted a differential inclusion analysis using the Human Protein Atlas (Fagerberg *et al.*, 2014), Embryonic Development (Yan *et al.*, 2013) and ENCODE datasets (Djebali *et al.*, 2012). Of the 1,399 MXEs, 608 MXEs (345 unique genes), 573 MXEs (389 unique genes) and 552 MXEs (330 unique genes) are differentially expressed, respectively (adjusted  $P$ -value < 0.05; Fig 3A, Appendix Figs S23–S26, Dataset EV5 and EV6). Most notably, the differentially included MXEs comprise 43.5, 40.9 and 39.5% of all MXEs indicating that MXEs are to a very large extent tissue- and developmental stage-specifically expressed.

The comparison of the genes containing differentially expressed MXEs from these three projects shows that 519 (88.7%) of all 585 MXE cluster containing genes have at least a single MXE differentially expressed in one of the covered tissues, cell types or developmental stages (Fig 3B). The 519 genes contain 942 differentially expressed MXEs (67% of the total 1,399 MXEs; Fig 3C). This number is in agreement with earlier analyses on small sets of MXEs (66 and 57%) (Wang *et al.*, 2008; Abascal *et al.*, 2015a). Expectedly, the expression of novel MXEs seems to be considerably more tissue specific than the expression of annotated MXEs and cassette exons (Appendix Fig S23). Lastly, 208 MXEs from 113 genes are preferentially expressed during embryonic development indicating that many MXEs are specific to certain developmental stages (Fig 3B and C).

The analysis of MXE specificity reveals that in many clusters one MXE dominates expression, whereas other MXEs are expressed at selected developmental time points and in specific tissues (Fig 3, Appendix Figs S23–S26). This modulation suggests crucial spatio-temporal functional roles for MXEs and can in many cases not be observed at the gene level, as gene counts can remain largely invariant. A well-known case for similar expression of MXEs in newborn heart but expression of only one MXE variant in adult heart is the ion channel *CACNA1C* (Diebold *et al.*, 1992), an example for the switch of expression are the MXEs of the *SLC25A3* gene (Wang *et al.*, 2008). We surmise that the observed specificity in combination with a generally lower expression could also explain the discovery of 654 (358) novel exons that have so far eluded annotation efforts (Fig 1A, Appendix Fig S23). In conclusion, the tight developmental and tissue-specific regulation of MXE expression suggests that changes in MXE function or expression might cause aberrant development and human disease (Xiong *et al.*, 2015). Pathogenic mutations in MXEs are known to cause Timothy syndrome, cardiomyopathy, cancer and kidney disease (Kaplan *et al.*, 2000; Splawski *et al.*, 2004, 2005; David *et al.*, 2010; Mayr *et al.*, 2011).

### MXEs are high-susceptibility loci for pathogenic mutations

To obtain a comprehensive overview of MXE-mediated diseases, we annotated all MXEs with pathogenic SNPs from ClinVar (Landrum *et al.*, 2016), resulting in 35 MXEs (eight newly predicted exons) with 82 pathogenic SNPs (Fig 4A, Dataset EV7). Disease-associated MXEs show tight developmental and tissue-specific expression with



**Figure 3. MXE expression is tightly regulated across tissues and development.**

- A** Heatmap showing all differentially expressed MXE clusters with at least three RPKM. Here, we used the Gini coefficient, which is a measure of the inequality among values of a frequency distribution (Ceriani & Verme, 2012) and has successfully been used to determine tissue-enriched gene sets (Zhang *et al*, 2017), to determine highly tissue-specific MXEs (maximum normalized Gini index of cluster) and MXEs with a broad tissue expression distribution (minimum Gini index). For each MXE cluster, the per cent-spliced-in (PSI) value of the ubiquitous MXE (minimum Gini index) is subtracted from the PSI value of the specific MXE (maximum Gini index of cluster) (delta PSI value) and scaled between -1 (broad tissue distribution) and 1 (highly tissue specific). Each column represents an MXE pair, and each row represents MXE expression in a tissue, cell type or at a developmental time point. The bar graph summarizes counts where the specific MXE is 1.5-fold more spliced in than the ubiquitous MXE.
- B** Overview of differentially expressed genes for the Embryonic Development, ENCODE and Human Protein Atlas datasets.
- C** Overview of differentially expressed MXEs for the Embryonic Development, ENCODE and Human Protein Atlas datasets.

prominent selective expression in heart and brain, and cancer cell lines (Fig 4B and C, Dataset EV7). Interestingly, the percentage of pathogenic SNP-carrying MXEs is twofold higher than the percentage of all pathogenic SNP-carrying exons (Fisher's exact test,  $P$ -value =  $3 \times 10^{-11}$ ). A similar enrichment can be found for cassette exons (Fisher's exact test,  $P$ -value =  $2.2 \times 10^{-16}$ ) suggesting that in general alternative splicing-associated exons are susceptibility loci for pathogenic mutations. The genes with MXEs carrying pathogenic SNPs are predominantly associated with neurological disease (10), neuromuscular disorders (7), cardiomyopathies (6) and cancer (3) and are enriched in voltage-gated cation channels

(e.g. *CACNA1C* and *CACNA1D*), muscle contractile fibre genes (e.g. *TPM1*), and transmembrane receptors (e.g. *FGFR1-3*; Fig 4, Appendix Fig S27, Dataset EV7).

Disease-associated MXEs have high amino-acid identity (average 49.1%, SD 23.1%), reaching up to 89% in *ACTN4* (Appendix Fig S28), suggesting similar functional roles and in consequence similar pathogenic potential for many MXE pairs (Fig 4C, Appendix Fig S29). Four of all SNP-containing MXE clusters contain mutations in both MXEs (*FHL1*, *MAPT*, *CACNA1C* and *CACNA1D*), whereas 31 currently have pathogenic SNPs in only one MXE. The MXE expression analysis shows that many SNP-carrying MXEs are highly

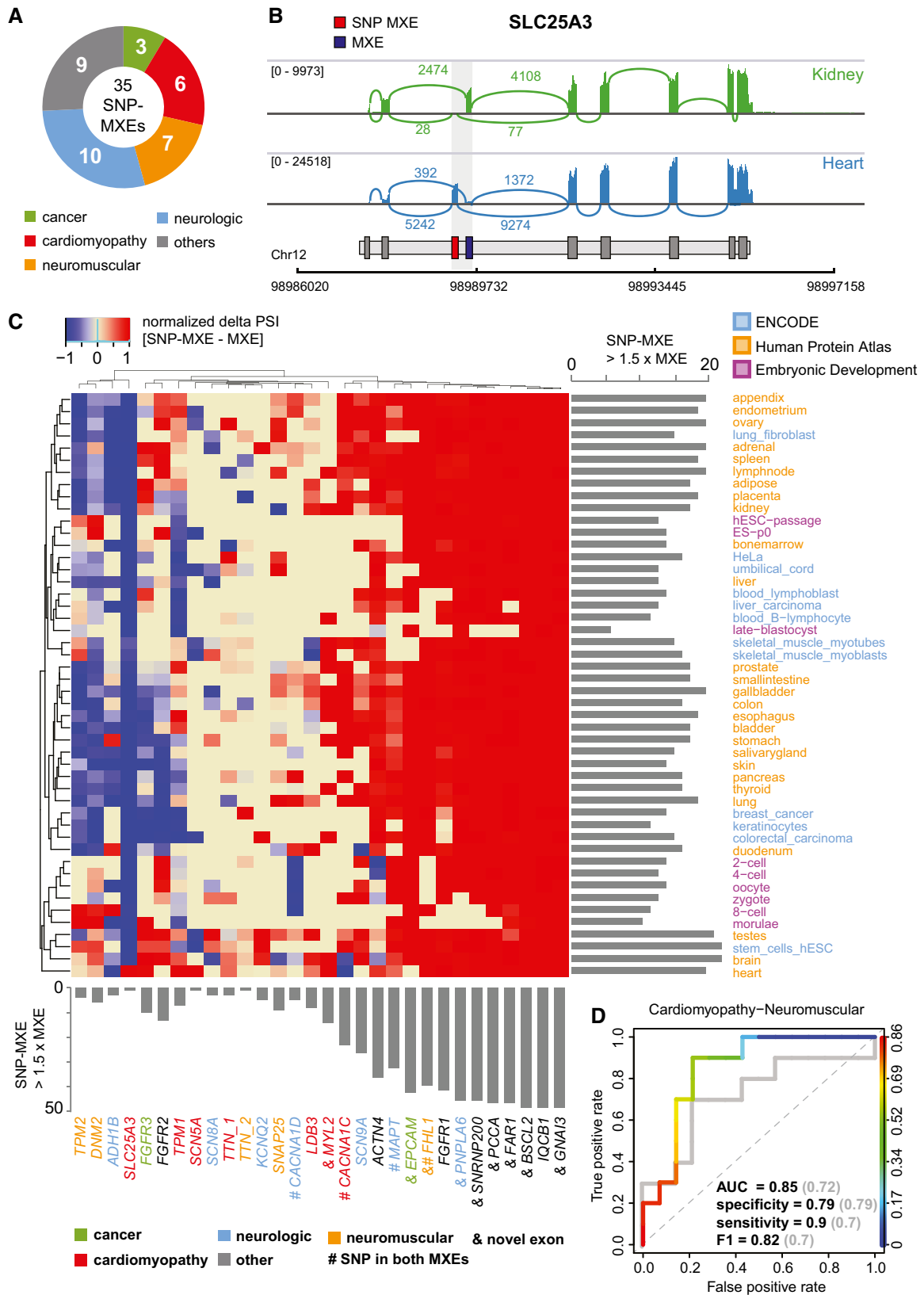


Figure 4.



**Figure 4. MXE-ratio expression predicts disease pathology.**

- A Thirty-five MXE clusters contain 82 pathogenic mutations causing neurologic (10), neuromuscular (7), cardiac (6), cancer (3) or other diseases (9).
- B Sashimi plots showing exon as well as splice junction reads (including number of reads) in kidney and heart for *SLC25A3*.
- C Heatmap showing the delta PSI values (PSI value of the non-SNP-containing MXE subtracted from the PSI value of the SNP-containing MXE) of MXE clusters containing pathogenic SNPs scaled between  $-1$  and  $1$  (blue = high expression non-SNP-containing MXE, red = high expression SNP-containing MXE). Columns represent MXE clusters and rows tissues, cell types and developmental stages. The column bar graph summarizes counts where the SNP-containing MXE is 1.5-fold more expressed than the non-SNP-containing MXE, whereas the row bar graph shows this for each tissue, cell type and developmental stage.
- D Receiver-operating characteristic (ROC) curve showing true- and false-positive rates for cardiomyopathy-neuromuscular disease prediction based on spatio-temporal MXE (coloured lines and black text) and RPKM-based gene (grey lines and text) expression (delta PSI values).

expressed, especially in disease-associated tissues where the respective non-SNP-carrying MXEs are not or barely expressed (Fig 4B and C, Appendix Fig S29). Examples include *ACTN4*, *TPM1* and *SLC25A3* (Appendix Figs S28, S30, and S31). Moreover, MXEs with pathogenic SNPs are usually not or non-exclusively expressed at early developmental stages (Appendix Fig S28–S31), while high and exclusive expression could lead to early embryonic death or severe multi-organ phenotypes (e.g. *FAR1*, Appendix Fig S32). Conversely, several non-SNP-carrying MXEs are highly expressed in early development and are otherwise mainly expressed at equal and lower levels compared to the SNP-carrying MXEs (Appendix Figs S29E–S31). The absence of pathogenic SNPs in these MXEs suggests functional compensation of the pathogenic SNP-carrying MXEs or early lethality, both of which would result in no observable phenotype.

Of the 35 MXE clusters with pathogenic mutations eight contain novel exons (Fig 4C, Dataset EV7). A mutation in exon 9a (p.Asp365Gly) of *FAR1*, a gene of the plasmalogen-biosynthesis pathway, causes rhizomelic chondrodysplasia punctata (RCDP), a disease that is characterized by severe intellectual disability with cataracts, epilepsy and growth retardation (Buchert et al, 2014). Novel MXE 9b is expressed in the same tissues but at eightfold lower levels suggesting partial functional compensation of the MXE 9a mutation, which might be responsible for the “milder” form of RCDP as compared to pathogenic mutations in other genes of the pathway (*PEX7*, *GNPAT* and *AGPS*) (Appendix Fig S32). A tissue-specific compensation mechanism had already been proposed but a reasonable explanation could not be given because *FAR2* expression shows a different tissue profile and individuals with deficits in peroxisomal  $\beta$ -oxidation, a potential alternative supply for fatty alcohols, have normal plasmalogen levels (Buchert et al, 2014). Because of the young age of the affected children, it is not known yet whether a mutation in constitutive exon 4 (p.Glu165\_Pro169delinsAsp), which could not be compensated in a similar way as the exon 9a mutation, leads to a strong RCDP-like phenotype (no survival of the first decade of life) or to a milder form such as the one caused by the exon 9a mutation.

In conclusion, it is tempting to speculate that MXE pathogenicity might be governed by high or exclusive expression in affected target tissues that is usually absent from early developmental processes, a pattern of expression that seems at least partially inversed for MXEs without pathogenic SNP annotations. To assess whether MXE pathogenicity follows observable rules, we trained a machine learner on MXE expression data and predicted the affected target tissue (Fig 4D, Dataset EV8). To obtain at least 10 observations per category with an expression  $> 3$  RPKM, diseases were grouped into cardio-neuromuscular ( $n = 10$ ) and other diseases ( $n = 14$ ) and predicted using leave-one-out cross-validation with a Random Forest. Cardio-neuromuscular diseases could be predicted with an accuracy of 83% ( $P$ -value  $< 0.01$ ), a specificity of 79%, a sensitivity of 90% and an area under the ROC curve (AUC) of 85% (Fig 4D, Dataset EV8, Appendix Fig S29). Conversely, cardio-neuromuscular disease could be predicted with an AUC of 72% using RPKM-based gene expression values (Fig 4D). Although based on only 24 observations, our data suggest that MXE expression might predict disease pathogenicity in space and potentially also in time.

**Evolutionary dynamics of MXEs in mammals and bilaterians**

While tissue-specific gene expression is conserved between birds and mammals, the alternative splicing of cassette exons is conserved only in brain, heart and muscles and is mainly lineage-specific (Barbosa-Morais et al, 2012; Merkin et al, 2012). Accordingly, a core set of only  $\sim 500$  exons was found with conserved alternative splicing in mammals and high sequence conservation, which was a small subset of the thousands of cassette exons identified in total. In contrast, although the total number was considerably smaller, most of the known human MXEs have been shown to be highly conserved throughout mammals if not even vertebrates (Letunic et al, 2002; Copley, 2004; Abascal et al, 2015b). In order to assess the conservation of human MXEs across mammals, we identified orthologous proteins in 18 representative species from all major sub-branches spanning 180 million years of evolution and predicted MXEs therein (Fig 5, Appendix Fig S33, Dataset EV9). Based on a

**Figure 5. Evolutionary dynamics of MXEs in mammalian evolution.**

Clusters of validated MXE were sorted by chromosome and chromosomal position. The names of the corresponding genes and the cluster-IDs are given in the outermost circle, and the presence of the respective MXEs (MXE clusters) in other annotations and mammals is indicated by coloured bars. Because the generation of the set of MXE candidates was based on the GenBank annotation, we analysed the presence of the validated MXEs in complementary annotations. Thus, the outer circles show whether the validated MXEs are also annotated as MXEs in Ensembl and Aceview, and whether the validated MXEs are present at all as exons in the Ensembl annotation as indicated by the legend. The lengths of the bars denote the percentage of matching exons for each cluster. For comparison, we show the annotation as MXE in two different Ensembl versions highlighting the dynamics of exon annotations over time. The comparison of the GenBank with the latest Ensembl annotation (v. 37.75) showed considerably less exons annotated as MXEs (58) in Ensembl although these include six of the “novel exons” (Appendix Fig S1). The presence of the respective validated MXEs in each of the analysed 18 mammals is shown by coloured bars. The 18 mammals, their phylogenetic relation and the total numbers of MXEs shared with human are presented at the bottom. The innermost circle represents the number of exons within each cluster of MXEs.

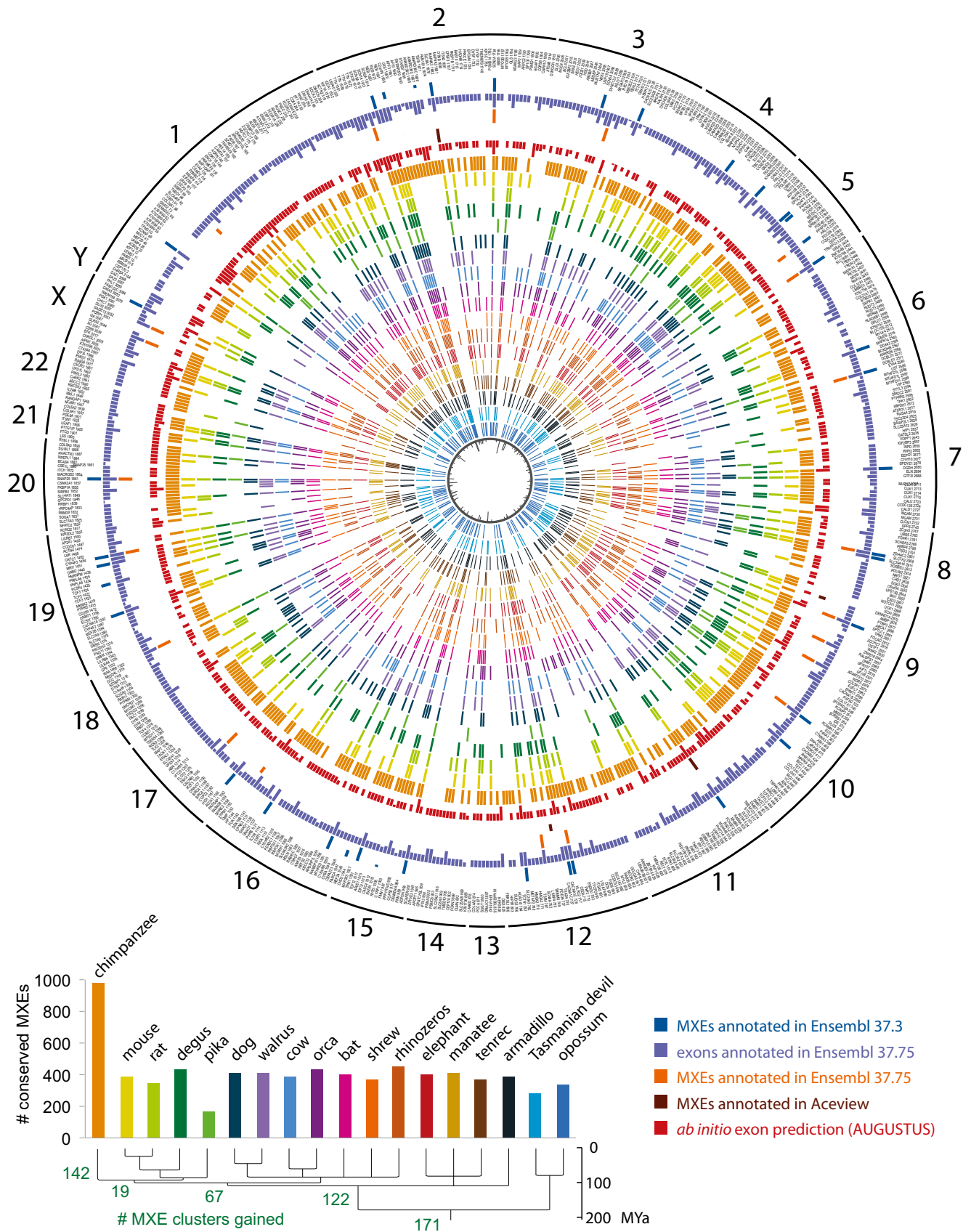


Figure 5.



simple model expecting each shared cluster to be already present in the last common ancestor of the respective species, we identified a core set of at least 173 (28%) of the human MXE clusters conserved throughout mammals (Fig 5, Appendix Fig S33). Other 122 MXE clusters were most likely present in the last common ancestor of the eutherians (16 species, placental mammals). The core set of mammalian MXE clusters includes 83 clusters shared between at least 16 of the species and 61 clusters shared between 17 species suggesting that their spurious absence in single mammals is likely due to genome assembly gaps or problems in identifying the correct orthologous genes. The remaining 29 MXE clusters of the core set have a scattered distribution across the 18 mammals indicating multiple independent branch- and species-specific cluster loss events. Such taxon-specific loss events include the MXE clusters in the *SRPK1* and *PQBP1* genes, which are absent in Glires (including mouse and pika), the cluster of 10 MXEs in *ABI3BP* that has been lost in the ancestor of mouse and rat, and the MXEs in *OSTF1* and *PTPRS*, which are absent in Afrotheria. The MXE clusters in *IKZF3*, *MBD1* and *ATP10B*, for example, are present in all Eutheria but not in Metatheria (marsupials). The MXE cluster gain rate within eutherian evolution towards human is relatively constant over time with about 23 clusters per 10 million years. Interestingly, each of the 16 eutherian species also lost a similar number of MXE clusters (127 clusters on average, Appendix Fig S33). In total, 82% of the human clusters containing validated MXEs are found in at least one further mammal (Fig 5). In summary, the large core set of mammalian MXEs and the overall conservation of MXE clusters suggest that MXEs are considerably more conserved than cassette exons. This observation supports expectations from considering the encoded protein structures where MXEs are supposed to provide alternative sequences for conserved secondary structural elements, while cassette exons are on average considerably shorter and add flexibility to surface loops (Buljan et al, 2012; Ellis et al, 2012; Irimia et al, 2014).

To get a first glimpse on mutually exclusive splicing evolution across bilaterians, we identified a set of 44 orthologous genes from genes containing MXEs in *Drosophila* (Hatje & Kollmar, 2013) and human genes containing MXE candidates (Appendix Fig S34, Dataset EV10). Of these orthologous genes, 28 contain validated MXEs in human, nine were validated to be spliced differently in human, and seven could not be validated in human because read mapping data are still missing; 20 (71%) of the genes containing validated MXEs represent cases of incompatible reading frames leading to NMD in case of joined inclusion, and for 18 of these MXE clusters multiple MXE-joining reads were found (Appendix Figs S34 and S35). We further analysed the 28 orthologous genes with validated MXEs and found five genes with homologous MXE clusters (identical position in gene, identical exon phase), 13 genes with MXE clusters in human that have homologous exons in *Drosophila* and eight genes with MXEs in human where the corresponding sequence regions in the orthologous *Drosophila* genes are part of larger exons (Appendix Figs S35 and S36). The presence of orthologous MXE clusters has been attributed to convergent evolution (Copley, 2004), although the respective analysis was in part based on the comparison of non-orthologous genes (e.g. comparing human sodium channel genes [e.g. *SCN1A*] with the *Drosophila* calcium channel *cac* gene and not the orthologous sodium channel *para* gene). At least for muscle myosin heavy chain genes it could

be demonstrated that *Drosophila* already lost several MXE clusters compared to, for example, *Daphnia pulex* (crustacean) and lophotrochozoans (Kollmar & Hatje, 2014) and that the evolutionary history of the MXEs within each cluster is remarkably complex with multiple independent exon duplications and losses (Odrionitz & Kollmar, 2008). Thus, detailed studies including more bilaterian and non-bilaterian taxa would be necessary to finally conclude convergent or divergent evolution for each of the human and *Drosophila* MXE clusters. Although the overlap of MXEs in orthologous genes of human and *Drosophila* is very low, the MXE gain and loss rates are very similar (Hatje & Kollmar, 2013) indicating a conserved role of tandem exon duplication in bilaterians. Gene structures can be highly conserved between kingdoms (Rogozin et al, 2003), and certain exons therefore seem to be predisposed to undergo duplication. In summary, these findings provide strong evidence for many MXE gain and loss events during mammalian evolution, suggesting a pronounced role of these processes in speciation and establishing phenotypic differences.

## Discussion

Using stringent criteria, including sequence similarity, reading frame conservation and similar lengths, and billions of RNA-Seq reads, we generated a strongly validated atlas of 1,399 human MXEs providing insights into mutually exclusive splicing mechanics, specific expression patterns, susceptibility for pathogenic mutations and deep evolutionary conservation across 18 mammals. The presented increase in human MXEs by an order of magnitude lifts MXEs into the present-day dimension of other human alternative splice types (Pan et al, 2008; Wang et al, 2008; Gerstein et al, 2014). Saturation analysis and the existence of 1,816 expressed but unconfirmed MXE candidates suggest a potential 27% increase in the MXE-ome with a twofold increase in data. Although alternative splice variants are abundant at the transcriptome level, recent mass spectrometry analyses suggested only small numbers of alternative transcripts to be translated (Abascal et al, 2015a; Ezkurdia et al, 2015; Blencowe, 2017; Tress et al, 2017a,b). Interestingly, MXEs were particularly enriched in the translated alternative transcripts, compared to other splice variants. However, ribosome profiling data showed high frequencies of ribosome engagement of cassette exons indicating that these isoforms are likely translated (Weatheritt et al, 2016). Similar results have been obtained through polyribosome profiling (Sterne-Weiler et al, 2013; Floor & Doudna, 2016). These observations suggest that most of the MXEs evaluated at the transcript level will also be found in the proteome.

About half (47%) of the 1,399 MXEs represent novel exons, which are often expressed at low levels and whose expression is restricted to few tissues and cell types, possibly explaining their absence from current genome annotations. Extrapolating these observations to all splice types and genes suggests the existence of thousands yet unannotated exons in introns. This estimation is in accordance with a recent analysis of more than 20,000 human RNA-Seq datasets that revealed over 55,000 junctions not present in annotations (Nellore et al, 2016). In this analysis, junctions found in at least 20 reads across all samples were termed “confidently called”. Although the total number of reads required for MXE validation in our analysis is lower ( $\geq 2$  SJ reads in the 1SJ case,  $\geq 6$  SJ

reads in the 3SJ case), the numbers seem more conservative given that we used 40 times less data for the validation.

The almost 10-fold increase in the human MXE-ome supports recent suggestions that mutually exclusive splicing might play a much more frequent role than anticipated (Pan *et al*, 2008; Wang *et al*, 2008; Ezkurdia *et al*, 2012; Abascal *et al*, 2015a). By comparing differentially expressed MXEs across cell types, tissue types and development, we could show that 14% of all genes with MXE clusters are shared between the three data sources, and 39% between any two. Most notably, however, it is almost always a different MXE from the same cluster that is differentially expressed, and only 3.3% of the MXEs are differentially expressed in all three data sources. We believe that this indicates a high spatio-temporal regulation of all MXEs in two-exon and multi-exon clusters. We rarely observed switch-like expression with only one of the MXEs of each cluster present in each cell- or tissue type or developmental stage. Rather, one of the MXEs (“default MXE”) of each cluster was present in most or all samples and the other MXEs were expressed in several selected tissues and developmental stages (“regulated MXEs”) in addition to the default MXE. Although the “regulated MXE” is usually expressed at lower level compared to the “default MXE”, there is almost always at least a single tissue or developmental stage where it is expressed at higher level. This supports previous assertions on the modulatory and compensatory effects of the regulated MXE on the enzymatic, structural or protein interaction functions of the affected protein domains (Letunic *et al*, 2002; Tress *et al*, 2017a).

The concerted annotation and splicing analysis of novel exons have deep implications for the detection and interpretation of human disease (Bamshad *et al*, 2011; Gonzaga-Jauregui *et al*, 2012; Xiong *et al*, 2015; Bowdin *et al*, 2016). For one, exome and panel sequencing remains the method of choice for the detection of genetic diseases and both methods rely on current exon annotations (Chong *et al*, 2015). Furthermore, our data suggest that MXE expression might reflect disease pathogenesis that could allow for the prediction of the affected organ(s). It is intriguing to speculate that the observed expression–disease association is a general dogma, which could be used to predict yet unseen diseases from published expression data, potentially bringing about a paradigmatic shift in (computational) disease research.

## Materials and Methods

### Data sources

The human genome assembly and annotated proteins (all isoforms) were obtained from GenBank (v. 37.3) (Benson *et al*, 2013). For MXE candidate validation, we selected data from 515 publically available samples comprising 31 tissues and organs, 12 cell lines and seven developmental stages (Barbosa-Morais *et al*, 2012; Djebali *et al*, 2012; Tilgner *et al*, 2012; Xue *et al*, 2013; Yan *et al*, 2013; Fagerberg *et al*, 2014) amounting to over 15 billion RNA-Seq reads. The data were chosen to encompass common and rare potential splice events in a broad range of tissues, cell types and embryonic stages. These RNA-Seq data were obtained from either GEO (NCBI) or ENA (EBI) databases (Dataset EV1). The description of the respective tissues and developmental stages is also listed in Dataset EV1.

### Reconstruction of gene structures

The gene structures for the annotated proteins were reconstructed with Scipio (Keller *et al*, 2008; Hatje *et al*, 2013) using standard parameters except `-max_mismatch=7`, `-region_size=20000`, `-single_target_hits`, `-max_move_exon=10`, `-gap_to_close=0`, `-blat_oneoff=false`, `-blat_score=15`, `-blat_identity=54`, `-exhaust_align_size=20000`, and `-exhaust_gap_size=50`. We let Scipio start with `blat_tilesize=7` and, if the entire gene structure could not be reconstructed, reduced the `blat_tilesize` step by step to 4. All parameters are less stringent than default parameters to increase the chance to reconstruct all genes automatically.

### Predicting mutually exclusive spliced exons

The human genome annotation does not contain specific attributes for alternative splice variants and thus does not allow extracting or obtaining lists for specific splice types. As mutually exclusive spliced exons (MXEs), we regarded those neighbouring exons of a gene locus that are present in only one of the annotated splice variants. These MXEs were termed “annotated MXEs”. However, exons appearing mutually exclusive are not necessarily spliced as MXEs. Terminal exons, for example, are included in transcripts by alternative promoter usage and by alternative cleavage and polyadenylation. MXEs were predicted in the reconstructed genes using the algorithm implemented in WebScipio (Pillmann *et al*, 2011). The minimal exon length was set to 10 aa (`-min_exon_length=10`). WebScipio determines the length of each exon (“search exon”) and generates a list of potential exonic regions with identical lengths (to preserve the reading frame) within the neighbouring up- and downstream introns. To account for potential insertions, we allowed length differences between search exon length and potential new exonic region of up to 60 nucleotides in steps of three nucleotides [`-length_difference=20` (given in aa)], thus obtaining a list of “exon candidates”. WebScipio then translates all exon candidates in the same reading frame as the search exon and removes all sequences that contain an in-frame stop codon. In case of overlapping exonic candidate regions, we modified the original WebScipio algorithm to favour exonic regions with GT–AG splice junctions over other possible splice sites (GC–AG and GG–AG). The translations of the exon candidates are then compared to the translations of the search exons, and candidates with an amino-acid similarity score of more than 10 (`-min_score=10`) are included in the final list of MXE candidates. Because the exon candidate scoring is done at the amino acid level, WebScipio expects candidates for 5′ exons of genes to start with a methionine, and candidates for 3′ exons of genes to end with a stop codon. This minor limitation is due to WebScipio’s original development as gene reconstruction software. MXE candidates for terminal exons were only searched in direction to the next/previous internal exon. The reason for looking for MXE candidates of annotated terminal exons is that we cannot exclude that further up- and downstream exons are missing in the annotation, which would turn the new MXE candidates to internal exons. Because of the described minor limitation, however, we can only propose MXE candidates if supposed additional up- and downstream exons are non-coding exons. Because terminal exons are

included in transcripts by alternative promoter usage and by alternative cleavage and polyadenylation, we treated the list of terminal exon candidates separately (Appendix Fig S4). This list might be of interest for further investigation for other researchers. Except for this Appendix Fig S4, we entirely focused on internal MXE candidates.

### Definition of criteria for RNA-Seq evaluation of the MXE candidates

While the sole mapping of RNA-Seq reads reveals the transcription of the respective genomic region, it does not prove the inclusion into functional transcripts. The mutually exclusive inclusion of the MXE candidates into functional transcripts requires at least the following splice junction (SJ) reads (Appendix Fig S5): (i) There must be SJ reads matching from every MXE to up- or downstream gene regions bridging the other MXEs of the cluster. The latter criterion takes into account that the annotated exons neighbouring the clusters of MXEs might not themselves be constitutive but alternative exons as, for example, in *NCX1* (Appendix Fig S6). (ii) SJ reads mapping from one to another MXE candidate lead to MXE candidate rejection except for those MXEs leading to a frame shift. Without this constraint, which has not been set in earlier analyses (Wang *et al*, 2008), MXEs cannot be distinguished from neighbouring differentially included exons, which are quite common in human (data not shown; see e.g. Hammesfahr & Kollmar, 2012 and Appendix Fig S6). Thus, there are three constraints for a cluster of two MXEs while clusters of three and five MXEs, for example, already require seven and 18 constraints, respectively (Appendix Figs S5 and S7). Under more stringent conditions, also SJ reads from MXEs to the neighbouring annotated exons independent of their splice type would be required giving rise to five constraints for a cluster of two MXEs (Appendix Fig S5).

Note that as a matter of principle the read coverage of MXEs and other alternative splicing events is considerably lower than that of constitutive exons due to their mutually exclusive inclusion in the transcripts. For example, each of the exons of a cluster of three MXEs is expected to only have, on average, one-third the coverage of the constitutive exons of the same gene. The number of predicted exons, of which both sites are supported by splice junction reads, is also considerably lower than the total number of supported MXE candidates (Appendix Fig S3), which we think is due to the general low coverage of the exons and not due to read mapping and exon border prediction problems (Appendix Fig S3).

### Validation of the MXE candidates by RNA-Seq mapping

SRA files were converted to FASTQ files using `fastq-dump` software (v. 2.1.18). FASTQ files were mapped onto the human reference genome (hg19) using the STAR aligner (v\_2.3.0e\_r291) (Dobin *et al*, 2013). To this end, we first generated a reference genome index with `-sjdbGTFfeatureExon`, `-sjdbGTFtagExonParentTranscript`, a splice junction overhang size of 99 (`-sjdbOverhang`) and GTF annotation files containing all transcripts and all MXE candidates. The MXE candidate GTF file was extracted from Kassiopeia database and is available for download there (Hatje & Kollmar, 2014). The mapping was done for each sample separately. We allowed a rather stringent maximum

mismatch of 2 (`-outFilterMismatchNmax 2`; STAR default is 10) and the output was forced to SAM format (`-outStd SAM`). Otherwise, default settings were used. The resulting files with the mapped reads were sorted, converted to BAM format and indexed with SAMtools (`sort -n`) for further processing (Li *et al*, 2009).

### Distinguishing MXEs from other splice variants

For the analysis of the read mapping data, we disassembled clusters with more than two MXE candidates into all possible sub-clusters. For example, a cluster with four MXE candidates [1,2,3,4] was fractionated into the following sub-cluster: [1,2], [2,3], [3,4], [1,2,3], [2,3,4], [1,2,3,4]. Each of these sub-clusters was analysed independently according to the validation criteria (splice junction reads present, exon-joining reads absent). If all criteria were satisfied for one of the sub-clusters, all MXE candidates of the respective sub-cluster were labelled “verified”. In a second analysis, each cluster of MXE candidates was analysed for exon-joining reads, which denote constitutive splicing or splicing as differentially included exons. However, MXE candidates of clusters and sub-clusters with exon-joining reads but exon lengths not divisible by three were also flagged as “verified” because their combined inclusion would lead to a frame shift in the translation of the transcript.

### Limits of the MXE dataset

Similar to every genome annotation dataset, also the current dataset of RNA-Seq validated MXEs has some limitations. Some are inherent to the still incomplete human genome annotation that was used as basis for generating the list of MXE candidates. As mentioned above and shown in Appendix Fig S2C, there are genes with mis-annotated terminal exons overlapping MXEs. Also, there are “transcripts” in the GenBank dataset that combine exons from (now) different genes. The presence of these “transcripts” in the genome annotation might be the result of mis-interpreting cDNA data as coding sequence although these might be the result of some level of mis-splicing.

Similarly, mis-splicing might be an important reason for validating true MXEs as “non-MXEs”. A single exon-joining read turns MXE candidates into non-MXEs, whose mutually exclusive splicing might otherwise be supported by thousands of MXE-bridging SJ reads. Given these limitations, we expect that many of the exons, that we currently tag as constitutive or other alternative splicing, might in fact be MXEs. On the other hand, our MXE dataset might also contain some exons that are in fact non-MXEs. This is well demonstrated in the saturation analysis (Fig 1C) showing that although more data will lead to the validation of many more exons as MXEs, for which SJ reads are currently missing, there will be clusters that will be rejected as soon as more data include exon-joining reads. In addition, some MXEs with only a few supporting SJ reads might in fact be pseudoexons. However, we also did not observe any SJ reads for about 15% of the annotated exons, which are nevertheless not regarded as pseudoexons (Fig 1B, Appendix Fig S3). Finally, some MXEs determined from transcripts showing complex splicing might in fact be mutually exclusive in transcripts, but not in the sense of a cluster of uninterrupted neighbouring exons.

### Saturation analysis

Theoretically, increasing the number of samples should also increase the number of validated MXEs, as the total increase in read number for different observed or novel tissues should increase the read evidence for the predicted MXEs. At the same time, increasing the number of reads also heighten the chance of rejecting an MXE candidate. This raises the question of what the expected number of validated and rejected MXEs for increasing numbers of samples is. Additionally, it would be interesting to obtain the theoretical point of saturation, the maximum expected number of MXEs in the human genome.

To obtain this information, sub-samples of STAR-aligned RNA-Seq splice junction (SJ) reads were used to estimate the expected recall and false-positive rate (Fig 1C, Appendix Fig S11). The number of verified MXEs was calculated using SJ reads for different percentages of the data. Similarly, the number of rejected MXEs was obtained. To reduce the bias from data sampling, datasets were chosen randomly and the saturation analysis was performed in 30 independent runs. To calculate the mean of validated and rejected MXEs at respective percentages of the total RNA-Seq data used for validation, we used the respective numbers from the 30 independent runs.

To estimate the potential increase in MXEs given more sequencing data, we fit the sub-sampling data to the number of expected MXEs  $f(x)$  using Matlab and the optimal fits were obtained for a power function

$$f(x) = a * x^b + c$$

with the linear coefficient  $a$ , the exponential coefficient  $b$  and the error term  $c$  (Appendix Fig S11B). Given a twofold increase in the number of reads, the expected number of validated MXEs (1SJ) is  $1,769 \pm 47$  (95% confidence interval), validated MXEs (3SJ) is  $1,081 \pm 12$ , rejected MXEs (1SJ) is  $227 \pm 9$ , and the number of rejected MXEs (3SJ) is  $95 \pm 5$  (Appendix Fig S11B). While the number of validated MXEs is far from saturation (a 100% increase in data results in 27% increase in the number of validations), the number of rejected MXEs seems to be saturated (a 100% increase in data results in 2% increase in the number of rejections).

### qPCR validation of MXE candidates

Total RNA was purified from healthy human brain tissue (substantia nigra) using Trizol kit (Tri Reagent, Sigma T9424) following manufacturer's instructions. RNA was further purified using the RNA Clean & Concentrator © TM -5 kit (Zymo Research, cat. R1013). The RNA quality was investigated using the 6000 nano assay on a Bio-analyzer 2100 (Agilent Technologies). Reverse transcription was carried out using the iScript © cDNA Synthesis kit (cat# 1708890, Bio-Rad) using approximately 500 ng of total RNA in a volume of 20  $\mu$ l.

Relative expression levels of the genes of interest as well as one housekeeping gene (glyceraldehyde 3-phosphate dehydrogenase [*Gapdh*]) were determined by qPCR using a LightCycler® 480. All qPCR experiments were performed in duplicates using SYBR™ Green PCR Master Mix (cat # 4309155). For each PCR, 20 ng cDNA

was used and negative controls contained no cDNA. The qPCR was run under the following conditions: pre-incubation at 95°C for 5 min, denaturation at 95°C for 10 s, annealing 60°C for 15 s, extension at 72°C for 10 s repeated for 40 cycles (Sybr green standard protocol II). Detailed information on the primers and qPCR results can be found in Dataset EV3.

### Analysis of the splice mechanism

To determine the distance between intron donor site and branch point, we analysed all introns smaller than 500 bp using the standalone version of SVM-BPfinder (beta) (Corvelo *et al*, 2010) to predict branch point locations. Longer introns harbour high numbers of branch point candidates, and the accuracy of the branch point prediction considerably decreases. Longer introns also often contain multiple branch points with different splicing kinetics (Corvelo *et al*, 2010) so that a steric hindrance criterion for splicing multiple MXEs into the same transcript might not apply anymore. Branch points are usually located in the 3' regions of the introns and it seems highly unlikely to identify only a single potential branch point within an, for example, > 2,000-bp intron, which would in addition be located within the 5' 50 bps. Thus, the highest-scoring location within the < 500-bp introns was taken as best guess for the branch point and the distance to the intron donor site determined.

In order to identify U12-type introns, we analysed all donor splice sites of the introns preceding the clusters of MXEs and those subsequent to all MXEs using the consensus pattern described by Sharp and Burge (Sharp & Burge, 1997). The acceptor splice sites of U12-type introns do not show conserved patterns and were therefore not used here for verification.

Binding windows for competing intron RNA secondary structures were predicted for all candidate clusters of MXEs using the SeqAn package (Döring *et al*, 2008). The identified binding windows of all homologous genes were aligned using MUSCLE (Edgar, 2004) and the RNA secondary structures predicted by RNAalifold (ViennaRNA package) (Lorenz *et al*, 2011).

### Mapping MXE sequences onto protein structures

To identify the best structural models for the sequences encoded by the MXEs, we mapped the protein sequences of the respective genes against available protein structure data. To this end, we made use of a recently developed database, called Allora (<http://allora.motorprotein.de>), in which genomic information is mapped onto protein structures. Allora currently contains 94,148 PDB entries (derived from the RCSB Protein Data Bank, <http://www.rcsb.org>, Rose *et al*, 2015) with 247,959 chains, of which 120,665 represent unique sequences. Based on the database references in the PDB entries, the full-length proteins were fetched from UniProt KB (UniProt Consortium, 2015) or GenBank (Benson *et al*, 2013) and the corresponding gene structures of the eukaryotic proteins reconstructed with WebScipio (Hatje *et al*, 2013). In Allora, all PDBs belonging to the same UniProt or GenBank entries are connected. BLAST+ (Camacho *et al*, 2009) was used to search for the most similar UniProt/GenBank protein sequence compared to the human proteins containing MXEs. The hit with the lowest E-value was taken, and the associated PDB chains were aligned to the human protein using m-coffee (Wallace *et al*, 2006). The MXE part of the alignment was



extracted for further analysis ( $\Rightarrow$  “MXE structure”). As “intron distances”, we determined the distances between the CA atoms of the first and the last residues of the MXE structures.

### Evaluating the differential inclusion of MXEs into transcripts

Splice junction read counts were extracted from STAR output “SJ.out.tab” files. For each MXE in a cluster, the per cent-spliced-in (PSI) value was calculated by dividing the number of junction reads of the MXE by the sum of junction reads for all MXEs in the same cluster. Differential inclusion analysis on the Human Protein Atlas, Embryonic Development and ENCODE datasets was performed using a Kruskal–Wallis rank sum test with a Benjamini–Hochberg (BH) multiple testing correction. Values were computed using the “kruskal.test” and “p.adjust” functions in R. For each project, we created a design matrix with sample name and experimental condition and replicate numbers. The results of the differential inclusion analysis are summarized in Dataset EV5.

### Differential expression of pairs of annotated and novel MXEs

For each sample (tissue, cell type and developmental stage), we calculated the median RPKM (reads per kilobase of transcript per million mapped reads) from the replicates for each MXE. To compile a set of MXEs with significant expression, only pairs of MXEs were selected of which either the annotated or the novel exon had a median expression of more than 3. The number of MXEs for this analysis would not considerably decrease if a cut-off of 30 were chosen (252 MXEs at a cut-off of 3 versus 240 MXEs at a cut-off of 30). For each pair of MXEs, we subtracted the PSI value of the ubiquitous/known/non-SNP-containing MXE from the PSI value of the respective specific/novel/SNP-containing MXE (delta PSI values) and scaled those values between  $-1$  (high PSI for ubiquitous/known/non-SNP-containing MXE) and  $1$  (high PSI for specific/novel/SNP-containing MXE) (see also Figs 3A and 4C, Appendix Fig S23). In case an MXE pair was not expressed in a certain tissues (NA or 0), the value was set to 0.

### Inequality analysis

The mean PSI values of each MXE were calculated for each tissue in the Human Protein Atlas project, each developmental stage in the embryonic development (Peking University) project, and each cell type in the ENCODE (Caltech) project. For each MXE, the Gini index (Ceriani & Verme, 2012) was calculated independently for each project based on the mean PSI values using the Gini function with standard parameters from the `ineq` R package version 0.2-13 (Achim Zeileis, Christian Kleiber, <https://CRAN.R-project.org/package=ineq>; Cowell, 2011). For the analysis of MXE clusters, only those clusters were taken into account that include at least two MXEs with an RPKM  $\geq 10$  in at least one dataset within each project. Furthermore, we excluded clusters where all MXEs have “NA” PSI values within each project (244, 96 and 225 clusters, respectively).

### Identification of pathogenic SNPs in MXEs

To identify potentially pathogenic SNPs in MXEs, the MXEs were compared to the ClinVar SNP database (ClinVar VCF file

downloaded on 11 Aug 2016, version updated at 30 Jun 2016, Landrum et al, 2016). The ClinVar variant summary file (VCF file) was converted into a BED file keeping all original information. Positions overlapping between MXEs and ClinVar-SNPs were accessed using the BEDTools feature intersection software (Quinlan & Hall, 2010). SNPs are classified as pathogenic or non-pathogenic according to ClinVar’s “ClinicalSignificance” field annotation. All entries containing “benign” and all structural variations were removed. All ClinVar-SNPs overlapping with MXEs were manually verified in order to keep only potentially pathogenic variations.

To access the statistical significance of disease enrichment in MXEs and cassette exons, we compared the amount of pathogenic SNP-containing to non-SNP-containing exons. Of 615,410 annotated exons, 21,030 (3.4%) contain pathogenic SNPs; of 1,399 MXEs, 99 (7.1%) contain pathogenic SNPs; and of 31,745 cassette exons, 2,143 (6.8%) contain pathogenic SNPs. The  $\sim 2$ -fold enrichment of alternative splicing-associated exons (MXEs and cassette exons) is highly significant (Fisher’s exact test,  $P$ -value MXE =  $3 \times 10^{-11}$ ,  $P$ -value cassette =  $2.2 \times 10^{-16}$ ).

### Disease prediction using pathogenic SNPs in MXEs

In order to predict disease from MXE expression, we first filtered for MXEs that had a minimal RPKM value of 3 and then subtracted the expression of the non-SNP-containing MXE from the SNP-containing MXE for all MXE pairs with mutations, across all developmental stages, tissues and cell types (49 features per MXE pair). Delta PSI values (PSI for SNP-containing MXE—PSI for non-SNP-containing MXE) were subsequently scaled and centred, and the MXE pairs were annotated to two disease classes, cardiomyopathy-neuromuscular disease ( $n = 10$ ) or other diseases ( $n = 14$ ). We regrouped genes into these categories to obtain relatively balanced categories while keeping a minimum of 10 observations per category.

Classification with limited observations needs careful execution, as over-fitting (high variance) and under-fitting (high bias) are common problems. To avoid high variance or bias, several crucial steps were taken. First, we did not optimize hyperparameters, using a Random Forest with 250 trees and a maximum tree depth of 16 (number of predictors/3). Second, we used leave-one-out cross-validation to avoid sampling bias and model instability. Third, diseases were grouped into two categories of relatively even size (see above). Models were built using the R packages `caret` (Kuhn, 2008) and `randomForest`, and ROC curves were generated with `ROCR` (Sing et al, 2005).

Of note, models trained on PSI values (considering only the PSI value of the SNP-containing MXE, data not shown) or RPKM values (Appendix Fig S29) obtained similar accuracies as the model trained on delta PSI values, indicating the stability of the prediction across slight variations in feature pre-processing.

### Gene ontology enrichment analysis

We used WebGestalt for Gene Ontology enrichment analyses (Wang et al, 2013). The lists of unique genes in gene symbol format were uploaded to WebGestalt and the GO Enrichment Analysis selected. The entire human genome annotation was set as background and 0.05 as threshold for the  $P$ -value for the significance test using the

default statistical method “hypergeometric”. Categorical enrichment of MXEs and cassette exons was summarized in a heatmap.

### Protein–protein interaction analysis

The protein–protein interaction network was built by using GeneMANIA webservice (Warde-Farley *et al*, 2010). The list of unique genes containing a pathogen SNP was submitted to GeneMANIA’s webservice, and we downloaded the resulting network in SVG format and manually included disease and ontology information.

### Assessing the dynamics of MXE annotations over time

MXEs might have already been annotated/described although not been included in the NCBI reference dataset. This might especially account for newer annotations based on the recently published ENCODE project data. Therefore, we obtained alternative protein sequence datasets from Aceview (Thierry-Mieg & Thierry-Mieg, 2006) and Ensembl (Yates *et al*, 2016). Further datasets like the VEGA and GENCODE annotations are continuously integrated into Ensembl and were therefore not considered separately. The Aceview database has been built in the year 2000 to represent comprehensive and non-redundant sequences of all public mRNA sequences. The human dataset has last been updated in November 2011, thus before the availability of the ENCODE data.

To assess the novelty of our MXE assignments with respect to the timely updates and changes of the human annotations, we compared our data with that of Aceview and with the latest annotation from Ensembl (Fig 5, Appendix Fig S1). As at the beginning of the project, only a few MXEs are annotated as such in other databases. Surprisingly, however, many of the previously annotated exons (independent of their splicing status) were removed from the latest Ensembl annotation, although our RNA-Seq mapping not only strongly supports their inclusion into transcripts but also their splicing as MXEs. This shows that further collaborative efforts are needed to reveal a stable and persistent human gene annotation.

### Ab initio exon prediction

Exon prediction by *ab initio* gene finding software is another means of generating a database of potential coding sequences. *Ab initio* exon prediction was done with AUGUSTUS (Stanke & Waack, 2003) using default parameters to find alternative splice forms and the feature set for *Homo sapiens*.

### Identifying orthologous proteins in 18 mammals

Cross-species searches in 18 mammals (Dataset EV9) were done with WebScipio (Hatje *et al*, 2013) with same parameters as for gene reconstructions except `-min_identity=60`, `-max_mismatch=0` (allowing any number of mismatches), `-gap_to_close=10`, `-min_intron_length=35`, `-blat_tilesize=6` and `-blat_oneoff=true`. MXE candidates in cross-species gene reconstructions were searched with `-length_difference=20`, `-min_score=15` and `-min_exon_length=15`, for all exons in all introns but not in up- and downstream regions. Reasons for not detecting clusters of MXEs might be gene and MXE loss events, sequence divergence precluding ortholog identification, and

assembly gaps. For determining the origin of a conserved MXE cluster, we used a simple model expecting each shared cluster to be already present in the last common ancestor of the respective species. This approach is equivalent to inferring ancestral character states with Dollo parsimony (Farris, 1977).

### Comparing human genes with MXEs to orthologous genes in *Drosophila melanogaster*

Orthologous genes in *D. melanogaster* for all human genes containing MXE candidates were obtained with the Ensemble BioMart service (Yates *et al*, 2016). This list of orthologous genes was filtered with the list of *D. melanogaster* genes containing MXEs, which was obtained from Hatje and Kollmar (2013), to obtain a list of genes with both types of exons, (i) MXEs in human and MXEs in *D. melanogaster*, and (ii) MXE candidates in human but validated to be spliced differently and MXEs in *D. melanogaster*. Several of the human and *D. melanogaster* genes contain multiple clusters of MXEs. Thus, we compared all genes manually to determine whether MXEs are orthologous in both species, whether MXEs in human have orthologous exons in *D. melanogaster*, and whether MXEs in human do not correspond to exons in *D. melanogaster* genes.

### Data availability

All generated data can be searched, filtered and browsed at Kassiopeia ([www.motorprotein.de/kassiopeia](http://www.motorprotein.de/kassiopeia); Hatje & Kollmar, 2014). The primary RNA-Seq datasets used in this study are available in the following databases:

<http://www.ebi.ac.uk/ena/data/view/ERP003613>  
<http://www.ebi.ac.uk/ena/data/view/ERP000546>  
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552>  
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44183>  
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33480>  
<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30567>

Expanded View for this article is available online.

### Acknowledgements

We would like to thank Prof. Paul Lingor and Lucas Araujo Caldi Gomes of the University Medicine Göttingen for providing total RNA isolated from human brain. We would like to thank Daniel Sumner Magruder from the Bonn group for critical suggestions. In particular, we would like to thank André Ahrens from the Kollmar group for his tremendous help in implementing the Allora database and performing the mapping of MXEs onto protein structures. The Kollmar group would like to thank Prof. Christian Griesinger for his continuous generous support. We would like to thank Dr. Robert P. Zinzen and Dr. Carla Margulies for critical reading of the manuscript.

### Author contributions

MK initiated the study and designed the analyses together with SB. KH and BH performed MXE predictions. KH integrated the human MXE data into the Kassiopeia database. KH implemented MXE candidate extraction and RNA-Seq analysis with help from ROV and SB. KH and MK did the cluster distribution, splicing mechanism and protein structure mapping analyses. KH, ROV, R-UR, VB, SB and MK performed the differential expression analysis. AR, MEM and TS performed the RT–PCR experiments. ROV, R-UR, VB and SB performed the SNP

mapping and prediction analysis. The comparison of different human gene annotations was done by KH and DS. KH did the prediction of MXEs in other mammals, and their comparison together with MK. MK and SB wrote the manuscript. ROV, KH, VB and R-UR contributed to manuscript text and Supplementary Materials. All authors read and approved the final manuscript.

### Conflict of interest

The authors declare that they have no conflict of interest.

## References

- Abascal F, Ezkurdia I, Rodriguez-Rivas J, Rodriguez JM, del Pozo A, Vázquez J, Valencia A, Tress ML (2015a) Alternatively spliced homologous exons have ancient origins and are highly expressed at the protein level. *PLoS Comput Biol* 11: e1004325
- Abascal F, Tress ML, Valencia A (2015b) The evolutionary fate of alternatively spliced homologous exons after gene duplication. *Genome Biol Evol* 7: 1392–1403
- Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* 12: 745–755
- Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, Slobodeniuc V, Kutter C, Watt S, Çolak R, Kim T, Misquitta-Ali CM, Wilson MD, Kim PM, Odom DT, Frey BJ, Blencowe BJ (2012) The evolutionary landscape of alternative splicing in vertebrate species. *Science* 338: 1587–1593
- Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. *Nucleic Acids Res* 41: D36–D42
- Blencowe BJ (2006) Alternative splicing: new insights from global analyses. *Cell* 126: 37–47
- Blencowe BJ (2017) The relationship between alternative splicing and proteomic complexity. *Trends Biochem Sci* 42: 407–408
- Bowdin S, Gilbert A, Bedoukian E, Carew C, Adam MP, Belmont J, Bernhardt B, Biesecker L, Bjornsson HT, Blitzer M, D'Alessandro LCA, Deardorff MA, Demmer L, Elliott A, Feldman GL, Glass IA, Herman G, Hindorff L, Hisama F, Hudgins L et al (2016) Recommendations for the integration of genomics into clinical practice. *Genet Med* 18: 1075–1084
- Buchert R, Tawamie H, Smith C, Uebe S, Innes AM, Al Hallak B, Ekici AB, Sticht H, Schwarze B, Lamont RE, Parboosingh JS, Bernier FP, Abou Jamra R (2014) A peroxisomal disorder of severe intellectual disability, epilepsy, and cataracts due to fatty acyl-CoA reductase 1 deficiency. *Am J Hum Genet* 95: 602–610
- Buljan M, Chalancon G, Eustermann S, Wagner GP, Fuxreiter M, Bateman A, Babu MM (2012) Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. *Mol Cell* 46: 871–883
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421
- Ceriani L, Verme P (2012) The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. *J Econ Inequality* 10: 421–443
- Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, Harrell TM, McMillin MJ, Wiszniewski W, Gambin T, Coban Akdemir ZH, Doheny K, Scott AF, Avramopoulos D, Chakravarti A, Hoover-Fong J, Mathews D, Witmer PD, Ling H, Hetrick K et al (2015) The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am J Hum Genet* 97: 199–215
- Copley RR (2004) Evolutionary convergence of alternative splicing in ion channels. *Trends Genet* 20: 171–176
- Corvelo A, Hallegger M, Smith CWJ, Eyraas E (2010) Genome-wide association between branch point properties and alternative splicing. *PLoS Comput Biol* 6: e1001016
- Cowell F (2011) *Measuring inequality*. Oxford: Oxford University Press
- David CJ, Chen M, Assanah M, Canoll P, Manley JL (2010) HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature* 463: 364–368
- Diebold RJ, Koch WJ, Ellinor PT, Wang JJ, Muthuchamy M, Wiecek DF, Schwartz A (1992) Mutually exclusive exon splicing of the cardiac calcium channel alpha 1 subunit gene generates developmentally regulated isoforms in the rat heart. *Proc Natl Acad Sci USA* 89: 1497–1501
- Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T et al (2012) Landscape of transcription in human cells. *Nature* 489: 101–108
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21
- Döring A, Weese D, Rausch T, Reinert K (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* 9: 11
- Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113
- Ellis JD, Barrios-Rodiles M, Colak R, Irimia M, Kim T, Calarco JA, Wang X, Pan Q, O'Hanlon D, Kim PM, Wrana JL, Blencowe BJ (2012) Tissue-specific alternative splicing remodels protein-protein interaction networks. *Mol Cell* 46: 884–892
- Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML (2012) Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Mol Biol Evol* 29: 2265–2283
- Ezkurdia I, Rodriguez JM, Carrillo-de Santa Pau E, Vázquez J, Valencia A, Tress ML (2015) Most highly expressed protein-coding genes have a single dominant isoform. *J Proteome Res* 14: 1880–1887
- Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, Asplund A, Sjöstedt E, Lundberg E, Szgyarto CA-K, Skogs M, Takanen JO, Berling H, Tegel H, Mulder J, Nilsson P et al (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13: 397–406
- Farris JS (1977) Phylogenetic analysis under Dollo's law. *Syst Zool* 26: 77–88
- Floor SN, Doudna JA (2016) Tunable protein synthesis by transcript isoforms in human cells. *Elife* 5: e10921
- Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, Pei B, Harmanci AO, Duff MO, Djebali S, Alexander RP, Alver BH, Auerbach R, Bell K, Bickel PJ, Boeck ME et al (2014) Comparative analysis of the transcriptome across distant species. *Nature* 512: 445–448
- Gonzaga-Jauregui C, Lupski JR, Gibbs RA (2012) Human genome sequencing in health and disease. *Annu Rev Med* 63: 35–61
- Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC (2004) The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. *RNA* 10: 1499–1506
- Graveley BR (2005) Mutually exclusive splicing of the insect Dscam Pre-mRNA directed by competing intronic RNA secondary structures. *Cell* 123: 65–73

- Hammesfahr B, Kollmar M (2012) Evolution of the eukaryotic dynactin complex, the activator of cytoplasmic dynein. *BMC Evol Biol* 12: 95
- Hatje K, Hammesfahr B, Kollmar M (2013) WebScipio: reconstructing alternative splice variants of eukaryotic proteins. *Nucleic Acids Res* 41: W504–W509
- Hatje K, Kollmar M (2013) Expansion of the mutually exclusive spliced exome in *Drosophila*. *Nat Commun* 4: 2460
- Hatje K, Kollmar M (2014) Kassiopia: a database and web application for the analysis of mutually exclusive exomes of eukaryotes. *BMC Genom* 15: 115
- Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallières M, Tapial J, Raj B, O'Hanlon D, Barrios-Rodiles M, Sternberg MJE, Cordes SP, Roth FP, Wrana JL, Geschwind DH, Blencowe BJ (2014) A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159: 1511–1523
- Johansson JU, Ericsson J, Janson J, Beraki S, Stanić D, Mandić SA, Wikström MA, Hökfelt T, Ögren SO, Rozell B, Berggren P-O, Bark C (2008) An ancient duplication of exon 5 in the Snap25 gene is required for complex neuronal development/function. *PLoS Genet* 4: e1000278
- Kaplan JM, Kim SH, North KN, Rennke H, Correia LA, Tong HQ, Mathis BJ, Rodríguez-Pérez JC, Allen PG, Beggs AH, Pollak MR (2000) Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis. *Nat Genet* 24: 251–256
- Keller O, Odronitz F, Stanke M, Kollmar M, Waack S (2008) Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics* 9: 278
- Kollmar M, Hatje K (2014) Shared gene structures and clusters of mutually exclusive spliced exons within the metazoan muscle myosin heavy chain genes. *PLoS One* 9: e88111
- Kuhn M (2008) Building predictive models in R using the caret package. *J Stat Softw* 28: 1–26
- Kustatscher G, Hothorn N, Pugieux C, Scheffzek K, Ladurner AG (2005) Splicing regulates NAD metabolite binding to histone macroH2A. *Nat Struct Mol Biol* 12: 624–625
- Landrum MJ, Lee JM, Benson M, Brown J, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 44: D862–D868
- Lee C, Kim N, Roy M, Graveley BR (2010) Massive expansions of Dscam splicing diversity via staggered homologous recombination during arthropod evolution. *RNA* 16: 91–105
- Lee Y, Rio DC (2015) Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem* 84: 291–323
- Letunic I, Copley RR, Bork P (2002) Common exon duplication in animals and its role in alternative splicing. *Hum Mol Genet* 11: 1561–1567
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25: 2078–2079
- Lorenz R, Bernhart SH, Höner Zu Siederdisen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL (2011) ViennaRNA package 2.0. *Algorithms Mol Biol* 6: 26
- Matlin AJ, Clark F, Smith CWJ (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol* 6: 386–398
- Mayr JA, Zimmermann FA, Horváth R, Schneider H-C, Schoser B, Holinski-Feder E, Czermin B, Freisinger P, Sperl W (2011) Deficiency of the mitochondrial phosphate carrier presenting as myopathy and cardiomyopathy in a family with three affected children. *Neuromuscul Disord* 21: 803–808
- Meijers R, Puettmann-Holgado R, Skiniotis G, Liu J, Walz T, Wang J, Schmucker D (2007) Structural basis of Dscam isoform specificity. *Nature* 449: 487–491
- Merkin J, Russell C, Chen P, Burge CB (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* 338: 1593–1599
- Nellore A, Jaffe AE, Fortin J-P, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips Iii RA, Karbhari N, Hansen KD, Langmead B, Leek JT (2016) Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the sequence read archive. *Genome Biol* 17: 266
- Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463: 457–463
- Odronitz F, Kollmar M (2008) Comparative genomic analysis of the arthropod muscle myosin heavy chain genes allows ancestral gene reconstruction and reveals a new type of 'partially' processed pseudogene. *BMC Mol Biol* 9: 21
- Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40: 1413–1415
- Pillmann H, Hatje K, Odronitz F, Hammesfahr B, Kollmar M (2011) Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics* 12: 270
- Pohl M, Bortfeldt RH, Grützmann K, Schuster S (2013) Alternative splicing of mutually exclusive exons—a review. *Biosystems* 114: 31–38
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842
- Rogozin IB, Wolf YI, Sorokin AV, Mirkin BC, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr Biol* 13: 1512–1517
- Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 43: D345–D356
- Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL (2000) *Drosophila Dscam* is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* 101: 671–684
- Sharp PA, Burge CB (1997) Classification of introns: U2-type or U12-type. *Cell* 91: 875–879
- Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCr: visualizing classifier performance in R. *Bioinformatics* 21: 3940–3941
- Smith CWJ (2005) Alternative splicing—when two's a crowd. *Cell* 123: 1–3
- Sommer B, Keinänen K, Verdoorn TA, Wisden W, Burnashev N, Herb A, Köhler M, Takagi T, Sakmann B, Seeburg PH (1990) Flip and flop: a cell-specific functional switch in glutamate-operated channels of the CNS. *Science* 249: 1580–1585
- Splawski I, Timothy KW, Sharpe LM, Decher N, Kumar P, Bloise R, Napolitano C, Schwartz PJ, Joseph RM, Condouris K, Tager-Flusberg H, Priori SG, Sanguinetti MC, Keating MT (2004) Ca(V)1.2 calcium channel dysfunction causes a multisystem disorder including arrhythmia and autism. *Cell* 119: 19–31
- Splawski I, Timothy KW, Decher N, Kumar P, Sachse FB, Beggs AH, Sanguinetti MC, Keating MT (2005) Severe arrhythmia disorder caused by



- cardiac L-type calcium channel mutations. *Proc Natl Acad Sci USA* 102: 8089–8096; discussion 8086–8088
- Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19: ii215–ii225
- Sterne-Weiler T, Martinez-Nunez RT, Howard JM, Cvitovik I, Katzman S, Tariq MA, Pourmand N, Sanford JR (2013) Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res* 23: 1615–1623
- Suyama M (2013) Mechanistic insights into mutually exclusive splicing in dynamin 1. *Bioinformatics* 29: 2084–2087
- Thierry-Mieg D, Thierry-Mieg J (2006) AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol* 7(Suppl 1): S12.1–S12.14
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* 22: 1616–1625
- Tress ML, Abascal F, Valencia A (2017a) Alternative splicing may not be the key to proteome complexity. *Trends Biochem Sci* 42: 98–110
- Tress ML, Abascal F, Valencia A (2017b) Most alternative isoforms are not functionally important. *Trends Biochem Sci* 42: 408–410
- UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204–D212
- Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res* 34: 1692–1699
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456: 470–476
- Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne Set Analysis toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 41: W77–W83
- Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* 38: W214–W220
- Weatheritt RJ, Sterne-Weiler T, Blencowe BJ (2016) The ribosome-engaged landscape of alternative splicing. *Nat Struct Mol Biol* 23: 1117–1123
- Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, Jovic N, Scherer SW, Blencowe BJ, Frey BJ (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347: 1254806
- Xue Z, Huang K, Cai C, Cai L, Jiang C, Feng Y, Liu Z, Zeng Q, Cheng L, Sun YE, Liu J, Horvath S, Fan G (2013) Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* 500: 593–597
- Yan L, Yang M, Guo H, Yang L, Wu J, Li R, Liu P, Lian Y, Zheng X, Yan J, Huang J, Li M, Wu X, Wen L, Lao K, Li R, Qiao J, Tang F (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 20: 1131–1139
- Yang Y, Sun F, Wang X, Yue Y, Wang W, Zhang W, Zhan L, Tian N, Shi F, Jin Y (2012) Conservation and regulation of alternative splicing by dynamic inter- and intra-intron base pairings in Lepidoptera 14-3-3 $\xi$  pre-mRNAs. *RNA Biol* 9: 691–700
- Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, Gordon L, Hourlier T, Hunt SE, Janacek SH, Johnson N, Juettemann T, Keenan S, Lavidas I, Martin FJ et al (2016) Ensembl 2016. *Nucleic Acids Res* 44: D710–D716
- Yura K, Shionyu M, Hagino K, Hijikata A, Hirashima Y, Nakahara T, Eguchi T, Shinoda K, Yamaguchi A, Takahashi K-I, Itoh T, Imanishi T, Gojobori T, Go M (2006) Alternative splicing in human transcriptome: functional and structural influence on proteins. *Gene* 380: 63–71
- Zhang JD, Hatje K, Sturm G, Broger C, Ebeling M, Burtin M, Terzi F, Pomposiello SI, Badi L (2017) Detect tissue heterogeneity in gene expression data with BioQC. *BMC Genom* 18: 277
- Zubović L, Baralle M, Baralle FE (2012) Mutually exclusive splicing regulates the Nav 1.6 sodium channel function through a combinatorial mechanism that involves three distinct splicing regulatory elements and their ligands. *Nucleic Acids Res* 40: 6255–6269



**License:** This is an open access article under the terms of the Creative Commons Attribution 4.0 License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## Raza Ur Rahman

Email: [raza-ur.rahman@zmnh.uni-hamburg.de](mailto:raza-ur.rahman@zmnh.uni-hamburg.de)

### Education

#### PhD, Bioinformatics

Georg-August-Universität, Göttingen , Germany.

[Since 2014]

#### Master in Bio-Informatics [MS (BI)]

Saarland University, Saarbrücken, Germany.

[Oct 2011 – Mar 2014]

#### Bachelor of Sciences Major Bio-Informatics [BS (BI)]

Mohammad Ali Jinnah University (M.A.J.U), Islamabad, Pakistan.

[August 2006 – June 2010]

### Employment Background

#### 1. IMC information multimedia communication

As a Software Engineer (Java\J2EE)

[1 Aug 2012 to 30<sup>th</sup> Nov 2012]

#### 2. Digital Processing Systems Inc (DPS)

As a junior Software Engineer (Java\J2EE)

[Jan 05, 2011 to 19<sup>th</sup> Oct 2011]

#### 3. Mohammad Ali Jinnah University, Islamabad Campus

Department of Computer Science and Bioinformatics

[Feb 02, 2009 to May 2010].

### PUBLICATIONS

#### • Published

- **Raza-Ur Rahman**, Abhivyakti Gautam, Jörn Bethune, Abdul Sattar, Maksims Fiosins, Daniel Sumner Magruder, Vincenzo Capece, Orr Shomroni and Stefan Bonn. (2018). **Oasis 2: improved online analysis of small RNA-seq data**. BMC Bioinformatics (volume19).
- **Raza-Ur Rahman**, Abdul Sattar, Maksims Fiosins, Abhivyakti Gautam , Daniel Sumner Magruder, Jörn Bethune, Sumit Madan , Juliane Fluck , and Stefan Bonn. (2017). **SEA: The small RNA Expression Atlas**. bioRxiv preprint <https://doi.org/10.1101/133199>
- Hatje, Klas and **Rahman, Raza-Ur** and Vidal, Ramon O and Simm, Dominic and Hammesfahr, Björn and Bansal, Vikas and Rajput, Ashish and Mickael, Michel Edwar and Sun, Ting and Bonn, Stefan and Kollmar, Martin (2017). **The landscape of human mutually exclusive splicing**. Molecular Systems Biology (volume 13).
- Vincenzo Capece, Julio C. Garcia Vizcaino, Ramon Vidal, **Raza-Ur Rahman**, Tonatiuh Pena Centeno, Orr Shomroni, Irantzu Suberviola, Andre Fischer and Stefan Bonn . (2015). **Oasis: online analysis of small RNA deep sequencing data**. *Bioinformatics* 31, 1–3
- Rashi Halder, Magali Hennion, Ramon O. Vidal, Orr Shomroni, **Raza-Ur Rahman**, Ashish Rajput, Frauke van Bebber, Anna-Lena Schuetz, Susanne Burkhardt, Eva Benito, Julio C. Garcia Vizcaino, Vincenzo Capece, Tonatiuh Pena Centeno, Magdalena Navarro Sala, Sanaz Bahari Javan, Christian Haass, Bettina Schmid, Andre Fischer, Stefan Bonn **DNA methylation changes in plasticity genes accompany the formation and maintenance of memory**. *Nature Neuroscience*, 19(1), 102–110.

- Tonatiuh Pena Centeno, Orr Shomroni, Magali Hennion, Rashi Halder, Ramon Vidal, **Raza-Ur Rahman**, Andre Fischer, Stefan Bonn **Genome-wide chromatin and gene expression profiling during memory formation and maintenance in adult mice**. *Scientific data*
- **In Preparation**
  - Eugenio F. Fornasiero, Sunit Mandad, **Raza-Ur Rahman**, Tonatiuh Pena Centeno, Ramon O. Vidal, Hanna Wildhagen, Burkhard Rammner, Sarva Keihani, Felipe Opazo, Inga Urban, Till Ischebeck, Koray Kirli, Eva Benito, André Fischer, Sven Dennerlein, Peter Rehling, Ivo Feussner, Henning Urlaub, Stefan Bonn, Silvio O. Rizzoli. **The codon sequences predict protein lifetimes and other parameters of the protein life cycle in the mouse brain**. *eLife*
  - Eugenio F. Fornasiero, Sunit Mandad, Hanna Wildhagen, Burkhard Rammner, Inga Urban, Till Ischebeck, Eva Benito, Koray Kirli, **Raza-Ur Rahman**, Sven Dennerlein, Peter Rehling, Ivo Feussner, André Fischer, Stefan Bonn, Henning Urlaub, Silvio O. Rizzoli. **The analysis of protein lifetimes in the mouse brain reveals basic turnover principles**. *Nature Neuroscience*

## Softwares

- **Oasis 2**: Improved online analysis of small RNA-seq data. <https://oasis.dzne.de>
- **SEA**: Small RNA Expression Atlas. <https://sea.dzne.de/sea/sea.jsp>
- **Memory-epigenome-browser**: A genome browser for the interactive visualization of (in house) NGS data. <https://oasis.dzne.de/JBrowse-1.11.4/index.html>