Bioinformatics, 2015, 1–9
doi: 10.1093/bioinformatics/btv632
Advance Access Publication Date: 2 November 2015
Original Paper

Sequence analysis

Specific small-RNA signatures in the amygdala at premotor and motor stages of Parkinson's disease revealed by deep sequencing analysis

Lorena Pantano[1,2,3,4,5], Marc R. Friedländer[1,2,3,4], Georgia Escaramís[1,2,3,4], Esther Lizano[1,2,3,4], Joan Pallarès-Albanell[1,2,3,4], Isidre Ferrer[6,7], Xavier Estivill[1,2,3,4,*] and Eulàlia Martí[1,2,3,4,*]

[1]Genomics and Disease Group, Bioinformatics and Genomics Programme, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona 08003, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain, [3]IMIM, Hospital del Mar Medical Research Institute, Barcelona 08003, Spain, [4]CIBER de Epidemiología y Salud Pública (CIBERESP), CRG, Instituto Carlos III Barcelona 08003, Spain, [5]Universitat Autonoma de Barcelona, Institut de Biotecnologia i de Biomedicina, Bellaterra (Cerdanyola del Valles), Barcelona, Spain and [6]Institut Neuropatologia, Servei Anatomia Patològica, IDIBELL-Hospital Universitari de Bellvitge, Universitat de Barcelona, Spain and [7]CIBER de Enfermedades Neurodegenerativas (CIBERNED), Instituto Carlos III, Barcelona, Spain

[*]To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on April 23, 2015; revised on October 9, 2015; accepted on October 23, 2015

Abstract

Motivation: Most computational tools for small non-coding RNAs (sRNA) sequencing data analysis focus in microRNAs (miRNAs), overlooking other types of sRNAs that show multi-mapping hits. Here, we have developed a pipeline to non-redundantly quantify all types of sRNAs, and extract patterns of expression in biologically defined groups. We have used our tool to characterize and profile sRNAs in post-mortem brain samples of control individuals and Parkinson's disease (PD) cases at early-premotor and late-symptomatic stages.

Results: Clusters of co-expressed sRNAs mapping onto tRNAs significantly separated premotor and motor cases from controls. A similar result was obtained using a matrix of miRNAs slightly varying in sequence (isomiRs). The present framework revealed sRNA alterations at premotor stages of PD, which might reflect initial pathogenic perturbations. This tool may be useful to discover sRNA expression patterns linked to different biological conditions.

Availability and Implementation: The full code is available at http://github.com/lpantano/seqbuster.

Contact: lpantano@hsph.harvard.edu or eulalia.marti@crg.eu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA high-throughput sequencing strategies have revealed a plethora of small non-coding RNAs (sRNAs) with diverse functions as regulators of gene expression (Esteller, 2011). While micro RNAs (miRNAs) are the best-known class of sRNAs, for many others the biogenesis, regulation and cellular roles are largely unknown. Accumulating evidence suggests that RNA fragments derived from small nucleolar RNA (snoRNA) and transfer RNA (tRNA) are not just random degradation products but rather stable elements, which may have functional activity in physiological and pathological conditions, influencing gene expression and alternative splicing events (Martens-Uzunova *et al.*, 2013). In neuronal cells, tRNA fragments

sensitize cells to oxidative-stress-induced p53 activation and p53-dependent cell death indicating that these sRNAs may participate in neurodegenerative processes (Hanada *et al.*, 2013).

In the central nervous system (CNS), miRNA are essential in cell-type specification and differentiation, and post-mitotic long-term neuronal maintenance. Perturbations of miRNA pathways have emerged as effectors of CNS damage, contributing to impaired cell homeostasis and neuronal death. However, the relevance of other types of sRNAs in analogous processes has been little explored, constituting an untapped source of bioactive compounds. Several miRNAs pathways are altered in neurodegenerative disorders, including Parkinson's disease (PD), the most common movement disorder (de Rijk *et al.*, 2000; Shulman *et al.*, 2011). We have previously shown that the expression of several miRNAs is altered in brains of patients at early/premotor and late/motor stages of PD (Minones-Moyano *et al.*, 2011, 2013). These miRNAs modulate mitochondrial function and neuronal viability, suggesting a contribution of their deregulation in early pathogenic events. In addition, miRNA profiling in peripheral blood suggests specific expression signatures in PD (Burgos *et al.*, 2014; Fernandez-Santiago *et al.*, 2015). Furthermore, in leukocytes of PD patients, sRNA deep sequencing reveals splicing changes that classify brain region transcriptomes (Soreq *et al.*, 2013). Characterization of sRNA species alteration in PD may provide the basis to understand pathogenesis of PD and to target new non-invasive diagnosis biomarkers. However, a full characterization of the sRNA transcriptome in PD (including species other than miRNAs), is still lacking.

Current bioinformatics resources for the analysis of sRNA sequencing data are mainly focused in miRNA detection and prediction. The characterization of other types of sRNA is not deeply addressed in these tools and only a few of them produce outputs for downstream analysis (Hoogstrate *et al.*, 2014; Huang *et al.*, 2010). A major drawback in the analysis and quantification of the non-miRNA sRNA is the presence of multi-mapping reads that derive from tRNAs or non-coding RNA genes with duplication events on the genome. The majority of the current bioinformatic tools apply inaccurate strategies to handle these types of sequences and: (i) directly discard them, resulting in an under-estimation of these elements or (ii) count them everywhere they map, causing an over-estimation of their expression and making incorrect the use of count-based differential expression analysis, since these methods assume that reads are counted once.

Here, we have developed a framework to (i) characterize the full set of sRNAs using an improved version of the SeqCluster tool (Pantano *et al.*, 2011) that deals with multi-mapping reads and, (ii) extract patterns of expression through data-mining analyses. We used this framework to quantify all types of sRNAs from high-throughput sequencing data of post-mortem brain samples at premotor- and motor stages of PD and age-matched controls. Subsequent data-mining analyses using the SeqCluster output uncovered sRNA signatures at premotor stages of the disease, involving several types of sRNAs. These results suggest that general sRNA perturbations occur early in PD and further indicate that our pipeline is a sensitive tool to profile all types of sRNA.

2 Methods

2.1 General characterization of the sRNA sequencing dataset

Brain samples were obtained from the Institute of Neuropathology and the University of Barcelona Brain Bank. PD-related Braak staging, RNA extraction and sRNA library preparation are detailed in the Supplementary methods (Supplementary Table S1). Reads were trimmed to 36 nt and ligation adapters removed using the adrec.jar program from seqBuster suite (Pantano *et al.*, 2010). Sequences were mapped to the hg19 genome. sRNA processing, mapping and annotation details are provided in Supplementary Methods.

2.2 Definition of sRNA clusters

After adapter removal, sequences were collapsed among samples, resulting in a set of unique sequences with the corresponding counts. In this analysis, only sequences with more than 10 counts were considered. The pipeline detects hotspots after mapping sequences onto the human genome (hg19 release). A hotspot is defined as a group of at least 10 overlapping sequences, mapping onto a specific genomic site.

The pipeline defines initial clusters, each consisting in several hotspots sharing any number of sequences. We call these sequences that map multiple times ambiguous sequences. We modified SeqCluster (Pantano *et al.*, 2011) to get better summarization, annotation and reports of clusters, and be able to use multiple mapping sequences. We assume that all hotspots with ambiguous sequences belong to the same sRNA cluster that may have one or multiple copies on the genome. In some cases, two hotspots may share a very small proportion of ambiguous sequences, maybe due to gene divergence or spurious alignments. It would be incorrect to consider this as the same sRNA cluster since the two copies are essentially different. To solve this, we apply a heuristic algorithm to end up with a sRNA cluster that can be considered as a unique unit of transcription (Fig. 1). The algorithm is based on two steps: (i) reduction and (ii) cluster correction. The reduction step joins all hotspots that share more than N sequences, (60% by default). We implemented a proportion test to determine whether the percentage of common sequences among hotspots is above that value. If the P value of the proportion test is < 0.05, the hotspots will be considered from the same sRNA cluster.

After that, if there are multiple hotspots in the same sRNA cluster with lower similarity (<60%), that cluster goes through the 'cluster correction' step. We applied the-most-voting strategy, where common sequences are assigned to the sRNA cluster with more sequences. After this step, a new sRNA cluster is created and will contain sequences uniquely to this new cluster.

This strategy generates unique/final sRNA clusters, each identifying a type of sRNA that, although it may contain sequences mapping onto multiple genomic locations (hotspots), these are only considered once. As a result, a cluster defines the expression pattern of a type of sRNA, in which groups of sequences are consistently co-expressed, irrespective of their genomic origin. The full code is available at GitHub repository (http://github.com/lpantano/seqbuster), and as a package at pypi (python package manager) (https://pypi.python.org/pypi/seqcluster).

In the present study, we detected a total of 2162 precursors (initial sRNA clusters) that were organized in 652 final sRNA clusters. Only 127 initial clusters (with hotspots having <60% of similarity) were split into three or more unique clusters (without ambiguous sequences), where ambiguous sequences contributed with less than 10% of sequences to these new clusters (Supplementary Fig. S1A). This indicates that the ambiguous sequences among precursors of different sRNA clusters were low expressed sequences compared to the rest of sequences in each sRNA cluster. In many cases, all hotspots of an sRNA cluster shared all sequences, with hotspots

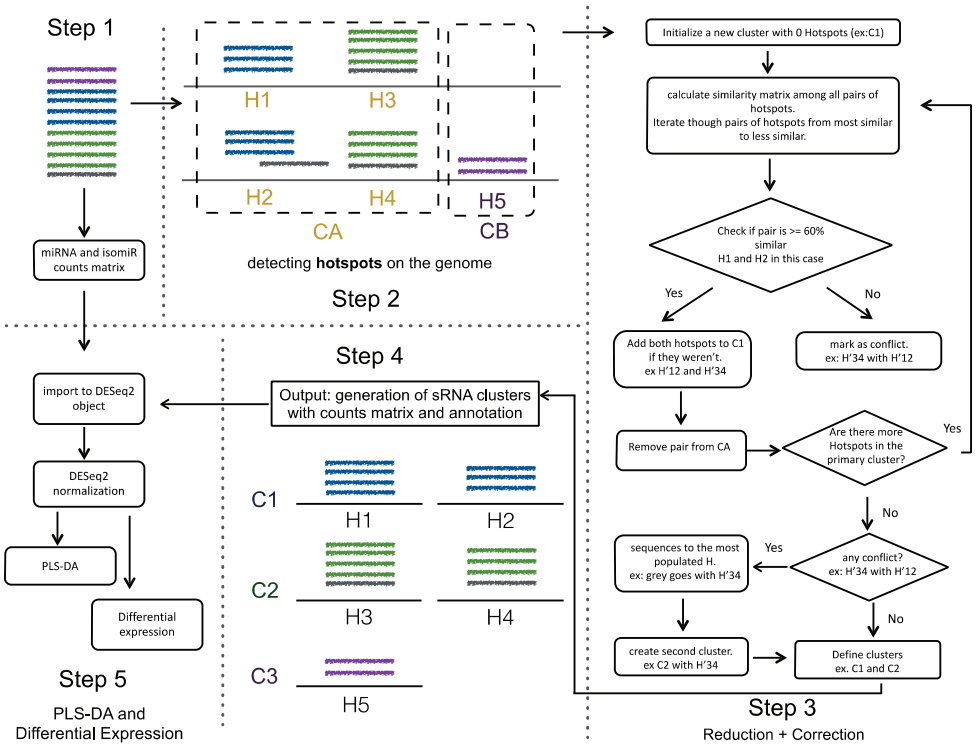

Fig. 1. General pipeline to define clusters of sRNAs (sRNA clusters), annotate miRNAs/isomiRs and perform downstream analyses to separate groups. Unique sequences are mapped onto the genome (steps 1 and 2). Miraligner detects miRNAs and isomiRs and generates a count matrix for all samples. For the rest, hotspots are defined (in this example H1–H5) as sets of overlapping sequences, according to their genomic positions (step 2). Colours show different types of sRNA, derived from different classes of RNA precursors. The grey sequence has an uncertain origin because it maps on multiple sites. Primary clusters (CA and CB, in the scheme) are subsequently defined as hotspots sharing any sequence. Then the pipeline goes through the reduction and the correction modules, based on recursively heuristic steps that reorganize the hotspots in meaningful clusters of expression (step 3). To generate a final cluster (for instance, C1 and C2 in step 4) the algorithm starts considering the most similar hotspots (H1–H2) in CA and joints them into a new hotspot (H'12). Then, the next similar hotspots within CA (H3–H4) are taken and joined into another hotspot (H'34). All new hotspots that have similarity $\leq$ 60% (or any other cut-off), will be labelled as conflict events. In this case, the common sequences go to the bigger hotspot, and each of them becomes a different cluster (in this example H'12 ends as C1 and H'34 as C2). When all initial clusters go through this step, SeqCluster annotates them with an optional GTF file and generates the count matrix for downstream analysis (step 4). Finally, both count matrixes (miRNA and cluster) are normalized with DESeq2 and used for PLS-DA and differential expression analysis (step 5)

differing only in the size (Supplementary Fig. S1B). In these cases, as the pipeline runs through progressively more similar hotspots, all of them are included in the same final sRNA cluster (without repeating the sequences) using as hotspot/s those that contain the largest number of sequences.

For instance, an initial sRNA cluster formed by 50 hotspots was reduced to three sRNA clusters with three hotspots. One of them corresponds to sRNAs derived from tRNA-ARG-CCG that was reduced from 20 precursors to one. The other two clusters are sRNAs derived from the tRNA-ARG-CCT and another tRNA-ARG-CCG. These three clusters are very similar, although they contained some region-specific differences (Supplementary Fig. S22). Therefore our strategy simplified annotation and interpretation of the data and, at the same time kept enough specificity to detect variability.

2.3 Partial least squares discriminant analysis

Partial least squares discriminant analysis (PLS-DA) is a technique specifically appropriate for analysis of high dimensionality data sets and multicollineality (Perez-Enciso and Tenenhaus, 2003). PLS-DA is a supervised method (i.e. makes use of class labels) with the aim to provide a dimension reduction strategy in a situation where we want to relate a binary response variable (in our case control or diseased status) to a set of predictor variables (in our study, sRNA clusters) (Perez-Enciso and Tenenhaus, 2003) Dimensionality reduction procedure is based on orthogonal transformations of the original variables (clusters) into a set of linearly uncorrelated latent variables (usually termed as components) such that maximizes the separation between the different classes in the first few components (Xia and Wishart, 2011). We used sum of squares captured by the model (R^2) as a goodness of fit measure.

Clusters with more than 10 counts in more than two samples were taken into account and additionally we considered only clusters expressed in at least 5 samples out of 14 being analysed (a total of 621 clusters in controls versus premotor cases, or controls

versus motor cases). To avoid false separation caused by picking up random noise rather than real signal, we conducted a permutation test (Xia and Wishart, 2011) involving 1000 data sets constructed by randomly reassigning class labels at each individual and further performing PLS-DA on the new randomized data sets. We further performed a PLS-DA for each of the main classes of sRNA clusters and for the miscellaneous cluster list. We ensured for each randomized data set that each group had a balanced number of correct and incorrect samples.

We conducted a refinement strategy to elude over parameterized models with rather poor discriminant properties (Perez-Enciso and Tenenhaus, 2003). In this sense, we obtained the most important

discriminant clusters from the four PLS-DA models based on the analysis of the three main sRNA-clusters lists (miRNA-, tRNA- and snoRNA-clusters) and the miscellaneous one, and conducted a second PLS-DA analysis including the *important* clusters associated independently to the four different models. We used variable importance for the projection (VIP) criterion that takes into account the contribution of a specific predictor for both the explained variability on the response and the explained variability on the predictors. As a rule of thumb, it is customary to retain variables with VIP > 0.8 (Perez-Enciso and Tenenhaus, 2003), however we used a more strict criteria, VIP > 1.2, to ensure variable importance into the whole model.

To evaluate significance of the refined strategy, we conducted two different approaches to respond to two different concepts. The first approach evaluates the robustness of the whole procedure, which consisted in: (i) randomize the class labels of the individuals, (ii) perform a PLS-DA analysis per each of the four different cluster lists and (iii) conduct a PLS-DA analysis including VIP clusters associated to the four different models. The second approach evaluates whether the selected clusters (according to the VIP score) are only useful for discriminating the target groups. For this latter purpose, we conducted similar permutation analysis to that explained for the general PLS-DA model, but including only the original VIP clusters in each randomized data set.

Permutation-based *P* values were calculated following this equation:

$$p - \text{value} = \frac{\sum R_o^2 < R_e^2}{n+1}$$

where R_o^2 and R_e^2 are the sum of squares captured by the model in the real and randomized data sets respectively, and n is the number of permutations.

2.4 Differential expression

We used DESeq2 for differential expression analysis and log2 transformation of the count data. We used the count matrix generated by Seqbuster and SeqCluster. Following the same rationale as in the PLS-DA analysis, only isomiRs, or clusters with more than 10 counts were taken into account and also sRNA-clusters or isomiRs consistently expressed (counts > 10) in at least 5 samples out of the 14 included in each analysis (controls versus premotor cases and controls versus motor cases). We performed permutation analysis to compare the results with the background noise of the data (See Supplementary material).

3 Results

3.1 Amygdala sRNA composition is complex, but dominated by few abundant RNAs

To elucidate the role of sRNAs in PD, we subjected 21 human brain samples to small RNA sequencing (sRNA-seq). We specifically focused in the amygdala, a brain area presenting Lewy bodies (LB) at pre-motor stages, the characteristic neuropathological hallmark of PD (Braak *et al.*, 2003, 2004). The samples comprised seven premotor cases, seven motor cases and seven control individuals (Supplementary Table S1, Supplementary methods). In addition, technical replicates were made of the seven controls, for a total of 28 distinct sequencing libraries constructed. Following sequencing on an Illumina HiSeq2000 instrument, the libraries each yielded between 10 and 20 million reads. After quality filtering, between 85 and 93% of the reads could be traced to genomic loci with high

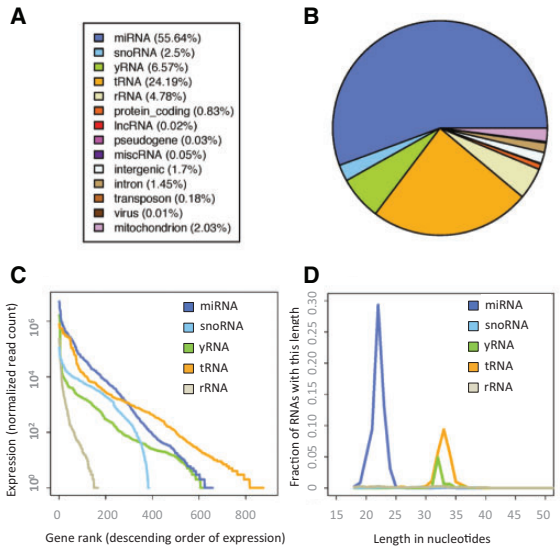

Fig. 2. Amygdala sRNA composition (**A**) sRNA composition summed over seven control individuals. (**B**) Pie chart representation. (**C**) Abundances of highly and lowly expressed sRNAs summed over seven control individuals. Colour code as above. (**D**) Length distributions of sRNAs summed over seven control individuals. The abundances are relative to all sequenced RNAs

confidence (Supplementary Methods). As expected, most of the sequenced RNAs originated from miRNAs (mean 56%), but there were also substantial contributions from tRNAs (24%), rRNAs (5%), snoRNAs (3%) and yRNAs (7%) (Fig. 2A and B). Although hundreds of genes in each of these sRNA classes were expressed, relatively few genes dominate the pool of sequenced RNA (Fig. 2C). The biogenesis of sRNAs is often reflected in the length of the molecules. As expected, the miRNAs in our samples clearly tended towards a length of 22 nucleotides (Fig. 2D).

Interestingly, the yRNA fragments clearly peaked at 32 nucleotides and the tRNA fragments at 33 nucleotides. The fragments from snoRNAs and rRNAs, did not exhibit any clear length peaks (Supplementary Fig. S3A). Investigating the relation between the amygdala sRNAs and their annotated host transcripts, we found that the most expressed miRNA, tRNA and yRNA genes: mir-181a, tRNA-Val-GTY and Y4, accounting for 18, 21 and 78% of the respective classes, all yield fragments from specific positions in their transcripts (Supplementary Fig. S4). We next investigated if these three genes, mir-181a, Y4 and tRNA-Val-GTY, are representative of their respective species. While miRNAs and tRNA fragments tend to be 22 nt 33 nt long, the yRNAs tend to be more variable in length (Supplementary Fig. S5).

Hierarchical clustering (Supplementary Methods) to group the 28 libraries based on their sRNA compositions reflects the pathology of the brain samples, since controls grouped together although they were prepared in independent batches (left of the sequenced dendrogram, Supplementary Fig. S3B). Finally, plotting the lengths RNAs showed sRNA length profile is enough to define the library compositions: libraries enriched in miRNAs tend to comprise RNAs that are ~22 nucleotides in length, while those enriched in tRNAs tend to contain RNAs that are ~33 nucleotides in length.

3.2 Improvements to the SeqCluster tool

SeqCluster organizes sRNAs in units or clusters of co-expressed molecules, consistently mapping to a host transcript (Pantano *et al.*,

Table 1. PLS-DA of sRNA expression data. PLS-DA using the total list of clusters (All), the clusters annotating onto several functional clases of non coding RNA (miRNAs, snoRNA and tRNAs) or the rest of clusters annotating onto a variety of precursors (Rest)

Type of sRNA clusters	Control versus pre-motor					Control versus motor				
	All	miRNA	snoRNA	tRNA	Rest	All	miRNA	snoRNA	tRNA	Rest
Number of sRNA clusters (%)	621 (100%)	230 (37%)	190 (31%)	104 (17%)	98 (16%)	621 (100%)	230 (37%)	188 (30%)	105 (17%)	97 (16%)
Number components	4	2	4	5	4	3	2	4	4	4
R^2 (sRNA-clusters variability)										
First comp.	26.40%	18.60%	16.40%	27.50%	17.10%	14.30%	16.80%	16.50%	28.30%	20.90%
Second comp.	13.60%	6.60%	22.20%	37.70%	20.60%	13.70%	26.60%	19.40%	27.90%	20.70%
Third comp.	16.70%		7.60%	9.90%	9.90%	16.70%		15.10%	5.70%	5.50%
Fourth comp.	6.10%		2.80%	5.80%	8.70%			5.50%	4.80%	5.70%
Fifth comp.				2.30%						
R^2 (outcome variability)	28.00%	58.40%	17.90%	99.29%	79.90%	86.10%	44.90%	73.50%	88.40%	78.70%
P value	0.608	0.345	0.719	0.001	0.074	0.049	0.317	0.055	0.014	0.077

In each PLS-DA, the number of variables (clusters), the number of components, R2 value providing clusters variability in each component and R2 value providing the outcome variability are shown. Finally, a *P* value indicates the significance of the separation between controls and affected individuals (1000 permutations).

2011). A distinctive characteristic of this tool is that it objectively assigns each sequence to a cluster, at which the correspondent counts are assigned (Supplementary Methods). We have upgraded the original pipeline, which in its current version generates a count matrix for all samples that can be used downstream profiling analyses (Fig. 1).

The new version of SeqCluster has important improvements that allow correct quantification of all types of sncRNA and a better exploratory analysis: (i) While the first version was java-based interface for exploratory analysis, this version, is a python based command line tool that can be integrated into any bioinformatic framework focused in the quantification, annotation and visualization; (ii) It is totally integrated in python allowing easy installation and usage of already published packages for common bioinformatic tasks; (iii) It handles all samples at the same time, reducing the time and number of commands needed to process a full project; (iv) It generates a count matrix that can be used for differential expression or clustering analyses; (v) It uses pysam and bedtools for the alignment and annotation that improves reproducibility, ensuring correct results since these tools are highly tested by the community; (vi) It works with known formats removing any custom format from the previous version; (vii) The algorithm to detect clusters has been modified in order to remove some dependencies that complicate installation and improve memory and time resources, taking less than 20 min to run a total of 28 samples—20 million reads/sample—in a single machine with 8GB of RAM); (viii) The whole analysis is wrapped in bcbio-nextgen framework that produces isomiRs/clusters results in html format with a single command line [http://github.com/chapmanb/bcbio-nextgen]; (ix) isomiRs count matrix generation, PLS-DA and DE analysis are integrated in a R package: isomiRs (http://github.com/lpantano/isomiRs).

To validate the quantification potential of SeqCluster, we analyzed the publicly available miRQC samples (Mestdah *et al.*, 2014) with SeqCluster and used the generated sRNA clusters count matrix for relative expression analysis. Four samples were considered: A, containing 100% Universal Human miRNA Reference RNA; B, 100% human brain RNA; and two titrations thereof (C = 0.75A + 0.25B and D = 0.25A + 0.75B). For clusters more abundant in A versus B (A > B) or in B versus A (B > A), the vast majority were correctly ordered when comparing all groups (A > C > D > B) or (B > D > C > A), respectively (see Supplementary Table S3); suggesting that SeqCluster detects titration and therefore it is an appropriate tool to quantify sRNAs (http://seqcluster.readthedocs.org/example_pipeline.html#mirqc-data).

3.3 Amygdala host transcripts give rise to specific sRNAs that are organized in hundreds of sRNA clusters

To have a qualitative and quantitative estimation of expressed sRNAs in each sample, we structured overlapping sequences in clusters, using the improved version of SeqCluster.

A total of 635 sRNA clusters were identified in the 28 samples (Supplementary Table S2), the majority of which mapped onto miRNA precursors (36.2%); onto C/D box small nucleolar RNAs (C/D box snoRNAs), H/ACA box snoRNAs and small Cajal body-specific RNAs (scaRNAs) (30%); and tRNAs (16.8%). The remaining 108 clusters (17%) contained a variety of sRNA mapping onto different types of precursors, including small cytoplasmic RNAs (scRNAs) genes, and a variety of non-characterized transcripts, that are also detected by the ENCODE consortium. The composition, abundance and mapping positions of all sequences within a cluster can be visualized in a html generated by the pipeline.

3.4 Partial least square discriminant analysis identifies sRNA expression patterns that distinguish control individuals from PD patients at premotor and motor stages.

Partial least square discriminant analysis (PLS-DA) is a supervised multivariate method that was used to explore if the expression pattern of sRNAs could classify controls versus affected individuals (Table 1). A general PLS-DA was first performed separately in controls versus premotor cases and controls versus motor cases to avoid experimental bias related with independent sequencing experiments. Using the list of sRNA clusters PLS-DA build a model that could significantly explain 86% of the response variability (control or PD-motor stages) ($P = 0.049$). A similar analysis could not discriminate controls from premotor cases ($R^2 = 28\%$, $P = 0.6$). To evaluate if different types of sRNAs contributed specifically to the model, we performed a PLS-DA for each of the main classes of sRNAs clusters and the list containing miscellaneous sRNA clusters (Table 1).

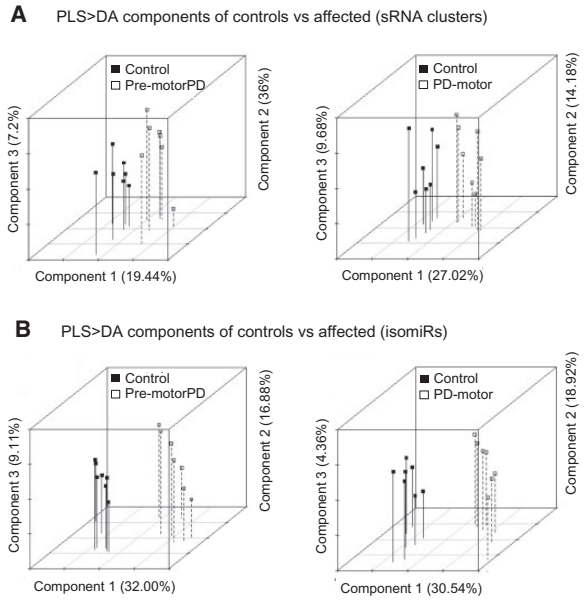

Fig. 3. Refined PLS-DA of sRNA clusters (**A**) and isomiRs (**B**) expression data. PLS-DA score scatter plot of the three components for control and affected individuals, showing separation of controls from premotor cases and controls from motor cases. Refined PLS-DA was performed with the more relevant sRNA clusters (VIP > 1.2) of each class (miRNA, snoRNAs, tRNAs, and the rest) (**A**) or the more relevant isomiRs (VIP > 1.2) (**B**) separating controls from affected individuals of a primary PLS-DA with consistently expressed sequences

A PLS-DA using the list of clusters that mapped onto tRNAs better separated controls and affected patients both at motor (R^2 = 88.4%, $P = 0.014$) and premotor stages ($R^2 = 99.3$ $P = 0.001$), suggesting that this class of sRNAs undergo early perturbation in PD.

PLS-DA provides the variable importance in projection (VIP) score that measures the relevance that each cluster provides to the model (Supplementary Fig. S6). Interestingly, 13 of the highly important tRNA-clusters (VIP score > 1.2; Supplementary Table S4 and Fig. S7) were common in PLS-DA of controls versus premotor cases and controls versus motor cases, suggesting that their combined expression pattern may define initial events in PD.

To improve the classification between control individuals and diseased patients, we performed a new PLS-DA, using a re-defined shorter list of clusters with a high (> 1.2) VIP score (Fig. 3A, Supplementary Fig. S8, Tables S5 and S6). This list could discriminate controls from PD-motor cases (R^2 = 88.4%) or PD-premotor cases (R^2 = 99.1%). This refined PLS-DA confirmed the former PLS-DA (Table 1), showing that among the top contributors, tRNA clusters were significantly enriched (Supplementary Fig. S6 and Table S7). A total of 12 clusters with high VIP scores (>1.2) were common in PLS-DA of controls versus motor cases and controls versus premotor cases, suggesting their early perturbation in PD.

PLS-DA using miRNA clusters could not separate controls and affected individuals (Table 1). However, miRNA clusters contain all sequences mapping onto each miRNA-precursor and therefore miRNA-3p or miRNA-5p forms are not distinguished. Similarly, miRNA clusters definition does not distinguish miRNA isoforms (isomiRs). IsomiRs are miRNAs that vary slightly in sequence, which result from variations in the cleavage site during miRNA biogenesis (5′-trimming and 3′-trimming variants), nucleotide additions to the 3′-end of the mature miRNA (3′ addition variants) and nucleotide modifications (substitution variants) (Marti *et al.*, 2010; Pantano *et al.*, 2010). Because miRNAs and IsomiRs have important roles in the CNS maintenance and function (Cloonan *et al.*, 2011; Fernandez-Valverde *et al.*, 2010), we performed a PLS-DA using an isomiR expression dataset, defined by the Seqbuster tool (Pantano *et al.*, 2010). More than 5000 isomiRs were consistently detected in the human amygdala samples, corresponding to 539 miRNAs. Using this list, we performed a first PLS-DA (Supplementary Table S8), and a refined version with the isomiRs with the highest VIP scores (>1.2) from the first PLS-DA could significantly discriminate controls from premotor or motor cases (Fig. 3B, Supplementary Table S8), suggesting that miRNAs are early altered in PD.

To validate these results, we applied a non-supervised method principal component analysis (PCA) using clusters or isomiRs with high VIP scores (>1.2). As shown in Supplementary Figure S9, the first two components from PCA could clearly separate controls from motor cases, explaining around 50% of clusters variability. This was not so clear at early stages of the disease, where only the second component (with 16% of clusters variability) generally separated controls from premotor cases. When analyzing isomiRs the two first components from PCA discriminated controls versus patients at both motor and premotor stages explaining as well 50% of the total variability.

3.5 Differential expression analysis identified selective sRNA clusters and isomiRs deregulated in the amygdala at premotor and motor stages

In addition to the identification of global patterns of sRNA clusters that distinguished controls from diseased samples using PLS-DA, we evaluated whether single clusters were differently expressed between control individuals and patients at premotor stages or motor stages. Differential expression was evaluated using DESeq2 with sRNA clusters. Among the differently expressed sRNA clusters (nominal P value < 0.05), hierarchical clustering analysis, which is blind to sample type, showed that the 20 with the highest significance separated all premotor cases and a control sample in the same cluster (Fig. 4A, Supplementary Table S9 and Fig. S10). Considering these 20 sRNA clusters, shuffling the identity of the patients or randomly taking groups of 10 clusters from the full set resulted in the significant loss of separation (Supplementary Fig. S11A). Of these 20 sRNA-clusters, 7 mapped onto tRNA, 7 onto C/D Box snoRNA, 5 onto miRNAs precursors and an additional sRNA-cluster onto an uncharacterized transcript. Half of these sRNA clusters presented high VIP scores (>1.2) in the PLS-DA analysis (Fig. 4A, Supplementary Table S9). Overall, these data suggest that these differently expressed clusters may define PD-premotor cases. A similar analysis identified 10 differently expressed clusters that optimally separated the group of controls from motor cases (Fig. 4B, Supplementary Table S9, Figs S10B and S12). Six of these top sRNA-clusters mapped onto tRNAs and 4 onto C/D Box snoRNAs. In accordance with their possible relevance, these clusters presented a high VIP score (VIP > 1.2) in the refined PLS-DA.

Differential expression analysis was also performed for isomiRs. A total of 85 differently expressed isomiRs (nominal P value < 0.05) corresponding to 42 miRNAs significantly separated all control individuals and premotor cases in two main groups (Fig. 4C, Supplementary Table S10). In addition, 230 differently expressed isomiRs corresponding to 104 miRNAs significantly separated controls and premotor cases. Importantly, the vast majority of isomiRs presented a significant VIP score (>1.2) according to the PLS-DA (Supplementary Table S10). In each of the comparisons, clustering

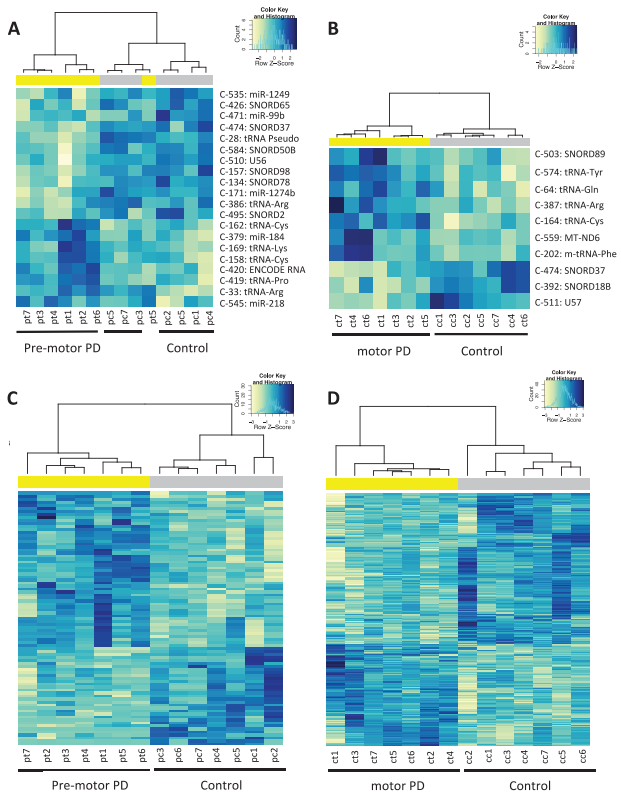

Fig. 4. Ward hierarchical clustering analysis of differentially expressed sRNA-clusters (**A**, **B**) or isomiRs (**C**, **D**) in the amygdala of controls versus pre-motor cases of PD (A, C) and controls versus motor cases of PD (B, D). Only the top significant clusters (C) or isomiRs ($P < 0.05$) optimally separating controls from affected individuals are shown (see Supplementary Table S9 and S10). 1 – correlation matrix was used as the distance matrix in the clustering. The expression counts were z-score normalized for visualization purposes and expression levels are coloured light yellow for low intensities and dark blue for high intensities

of the two groups of samples was lost if shuffling the identity of the patients or randomly choosing different groups of isomiRs among the total expressed (Supplementary Fig. S11C, D). Twenty-one miRNAs and more specifically, seven isomiRs were commonly deregulated in premotor and motor cases, suggesting their early affectation in PD.

3.6 Benchmarking with sRNAbench

To our knowledge, there are only two additional tools that, similarly to the present pipeline, can annotate different types of small RNAs and handle multi-mapped reads: sRNAbench (Hackenberg *et al.*, 2011) and flaimapper (Hoogstrate *et al.*, 2014). We therefore used these tools to repeat the analyses with our sequencing dataset, focusing in the sRNAs that map onto tRNAs because this class is highly enriched in multi-mapping sequences and is the type that optimally separated affected and control individuals.

Flaimapper failed to successfully process the individual BAM files due to RAM memory issues (it failed in a 64GB RAM memory machine). Regarding sRNAbench, it produces outputs with uniquely mapped reads and multi-mapped reads, with these last being counted repeatedly in each mapping site. We separately used both types of count matrix for PLS-DA, to determine the influence of multi-mapped reads in the analysis. While PLS-DA using uniquely mapping tRNAs could not separate controls and affected individuals, multi-mapping tRNAs tended to better distinguish groups (Supplementary Table S11). However, when comparing controls with premotor cases, differences were not significant, indicating that SeqCluster outperformed this tool. Furthermore, differential expression analysis of tRNA and isomiRs with sRNAbench output showed that the signal of differences among groups was reduced or disappeared (Supplementary Table S12). Overall these data suggest that using multi-mapped hits is important and further indicates that processing the data to unique clusters helps to separate affected individuals from controls.

4 Discussion

Our study provides the first deep characterization of the sRNA transcriptome at different stages of LB pathology and parkinsonism in post-mortem brain samples. We have used a strategy that organizes the transcriptome in unique units of consistently co-expressed sRNAs (sRNA clusters) using an improved version of the SeqCluster tool (Pantano *et al.*, 2011). This approach simplifies the complexity of the sRNA transcriptome, arranging millions of sequences into hundreds of sRNA clusters, which permits a downstream comprehensive profiling.

Using the sRNA-cluster count matrix, we could significantly separate control individuals from clinical/motor cases. However, expression profiles of tRNA clusters better separated controls from affected individuals at different stages of LB pathology, compared with other functional classes of co-expressed sRNAs. tRNA-derived sRNAs correct annotation and quantification is specially challenging, given that the majority map onto multiple locations. Our strategy successfully overcomes this problem through a progressive and heuristic allocation of each multi-mapped sequence to a specific (tRNA) gene with one or multiple precursors (see Methods). Using the expression matrix of tRNA-genes, we demonstrate that multi-mapping sequences are important to classify diseased status.

It has been recently shown that non-random 20–35 nt tRNA fragments (tRFs) guide mRNA cleavage, control translation and show cross talk with canonical sRNA silencing pathways, through a variety of mechanisms (Durdevic *et al.*, 2013; Gebetsberger and Polacek, 2013; Selitsky *et al.*, 2015; Sobala and Hutvagner, 2011; Sobala and Hutvagner, 2013). In addition, accumulation of 5′ halves of different types of tRNAs are produced in response to oxidative stress (OS) in a wide variety of eukaryotes (Saikia *et al.*, 2012; Thompson *et al.*, 2008). The majority of the top-deregulated tRNA fragments presented highly abundant 5′-tRFs, with most showing up-regulation at premotor and motor stages (Supplementary Table S2). Because OS is a major hallmark in PD brains (Ferrer *et al.*, 2011), it is tempting to speculate about the participation of this pathway in tRFs accumulation.

miRNA clusters that contained all types of co-expressed sequences mapping onto each miRNA could not discriminate controls and affected individuals. However, we captured disease stage-specific profiles using a list of isomiRs, which are slightly varying miRNA sequences with the potential to influence silencing dynamics (Cloonan *et al.*, 2011; Fernandez-Valverde *et al.*, 2010; Llorens *et al.*, 2013).

The proportion of the different classes of IsomiRs (5′ trimming, 3′-trimming, 3′-addition and nt-substitution) was similar in the group of deregulated isomiRs and the total number of isomiRs, indicating that no general alterations occur in the mechanisms generating the main types of isomiRs. Instead, disease-IsomiR profiles may

reflect changes in the biogenesis and/or stability of selective miRNAs variants; raising the question of whether mechanisms finely modulating the biogenesis of each miRNA may reflect an early dysfunction of the brain. In line with this, differential changes in the expression of specific isomiRs relative to those of the consensus (reference) counterpart have been found in several neurodegenerative disorders (Hebert *et al.*, 2013; Marti *et al.*, 2010).

In the present study, we found many deregulated miRNAs previously associated with PD and other neurodegenerative disorders, including miR-9, miR-181c, miR-146a, miR-16 and miR-124 (Supplementary Table S13). We confirmed the deregulation pattern of 26 miRNAs (out of 34) (Supplementary Table S13), whose precursors were altered in PD brains (Kim *et al.*, 2007). We also confirmed the deregulation pattern of 10 out of 18 miRNAs in total blood of PD patients (Martins *et al.*, 2011) and 3 out of 17 in leukocytes of PD patients (Soreq *et al.*, 2013) (Supplementary Table S13). Other miRNAs associated with PD (Alvarez-Erviti *et al.*, 2013; Cho *et al.*, 2013; Kim *et al.*, 2007; Minones-Moyano *et al.*, 2011; Soreq *et al.*, 2103) were not confirmed in our analysis either because deep sequencing approach could not reliably detect them or showed a high variability between samples.

In summary, the present workflow revealed an overall alteration of sRNA profiles in the amygdala of PD brains that occurs early, at premotor stages. A similar analysis in peripheral blood samples of patients at diverse disease stages and compared with control samples will answer whether our strategy is useful as a diagnosis tool in PD. tRFs and miRNAs (isomiRs) are the more relevant sRNAs types classifying affected individuals at different stages. We propose tRFs as a new class of early-stage biomarkers that may reflect OS in the brain of neurodegenerative disease patients. Other sRNAs species were deregulated in PD, mapping onto snoRNAs, repeated element sequences and other uncharacterized RNAs of uncertain function. An important future task is to go beyond the identification of disease-classifying sRNAs and associate them with functions. This will likely likely shed light onto the involvement of sRNAs pathways in PD patients as compared with controls, and between varying disease progression stages.

Acknowledgements

We thank Birgit Kagerbauer and Elena Miñones-Moyano for the preparation of sRNA libraries and the staff of the Genomics Unit at the CRG for sequencing of the sRNA libraries.

Funding

This work was supported by the Spanish Government, Instituto Carlos III – ISCIII and co-financed by the European Regional Development Fund (ERDF): PN de I+D+I 2012-2015 PI11/02036 (E. Martí), PI1100968 (I. Ferrer); Subdirección General de Evaluación y Fomento de la Investigación, SAF2008-00357: NOVADIS (X. Estivill) Ministerio de Economia y competitividad, SAF2014-60551-R: iRPaD (E. Martí) Ministerio de Economia y competitividad; Generalitat de Catalunya funding AGAUR 2009 SGR-1502 and the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements 278486: DEVELAGE (I. Ferrer), 62055: ESGI (X. Estivill) and 261123: GEUVADIS (X. Estivill); Lilly Foundation award (X. Estivill). Support for M.R. Friedländer was provided by an EMBO long-term fellowship. We acknowledge support of the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208.

Conflict of Interest: none declared.

References

Alvarez-Erviti,L. *et al.* (2013) Influence of microRNA deregulation on chaperone-mediated autophagy and alpha-synuclein pathology in Parkinson's disease. *Cell Death Dis.*, **4**, e545.

Braak,H. *et al.* (2003) Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiol. Aging*, **24**, 197–211.

Braak,H. *et al.* (2004) Stages in the development of Parkinson's disease-related pathology. *Cell Tissue Res.*, **318**, 121–134.

Burgos,K. *et al.* (2014) Profiles of extracellular miRNA in cerebrospinal fluid and serum from patients with Alzheimer's and Parkinson's diseases correlate with disease status and features of pathology. *PloS One*, **9**, e94839.

Cho,H.J. *et al.* (2013) MicroRNA-205 regulates the expression of Parkinson's disease-related leucine-rich repeat kinase 2 protein. *Hum. Mol. Genet.*, **22**, 608–620.

Cloonan,N. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.

de Rijk,M.C. *et al.* (2000) Prevalence of Parkinson's disease in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. *Neurology*, **54**, S21–S23.

Durdevic,Z. *et al.* (2013) The RNA methyltransferase Dnmt2 is required for efficient Dicer-2-dependent siRNA pathway activity in Drosophila. *Cell Reports*, **4**, 931–937.

Esteller,M. (2011) Non-coding RNAs in human disease. *Nat. Rev. Genet.*, **12**, 861–874.

Fernandez-Santiago,R. *et al.* (2015) MicroRNA association with synucleinopathy conversion in rapid eye movement behavior disorder. *Ann. Neurol.*, **77**, 895–901.

Fernandez-Valverde,S.L. *et al.* (2010) Dynamic isomiR regulation in Drosophila development. *RNA*, **16**, 1881–1888.

Ferrer,I. *et al.* (2011) Neuropathology of sporadic Parkinson disease before the appearance of parkinsonism: preclinical Parkinson disease. *J. Neural. Transm.*, **118**, 821–839.

Gebetsberger,J. and Polacek,N. (2013) Slicing tRNAs to boost functional ncRNA diversity. *RNA Biol.*, **10**, 1798–1806.

Hackenberg,M. *et al.* (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.

Hanada,T. *et al.* (2013) CLP1 links tRNA metabolism to progressive motor-neuron loss. *Nature*, **495**, 474–480.

Hebert,S.S. *et al.* (2013) A study of small RNAs from cerebral neocortex of pathology-verified Alzheimer's disease, dementia with lewy bodies, hippocampal sclerosis, frontotemporal lobar dementia, and non-demented human controls. *J. Alzheimer's Dis.*, **35**, 335–348.

Hoogstrate,Y. *et al.* (2014) FlaiMapper: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data, *Bioinformatics.*, **31**, 665–673

Huang,P.J. *et al.* (2010) DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.*, **38**, W385–W391.

Kim,J. *et al.* (2007) A MicroRNA feedback circuit in midbrain dopamine neurons. *Science*, **317**, 1220–1224.

Llorens,F. *et al.* (2013) A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genomics*, **14**, 104.

Martens-Uzunova,E.S. *et al.* (2013) Beyond microRNA–novel RNAs derived from small non-coding RNA and their implication in cancer. *Cancer Lett.*, **340**, 201–211.

Marti,E. *et al.* (2010) A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res.*, **38**, 7219–7235.

Martins,M. *et al.* (2011) Convergence of miRNA expression profiling, α-synuclein interacton and GWAS in Parkinson's disease. *PLoS One*, **6**, e25443.

Mestdah,P. *et al.* (2014) Evaluation of quantitative expression platforms in the miRNA quality control (miRQC) study. *Nat. Methods*, **11**, 809–815.

Minones-Moyano,E. *et al.* (2013) Upregulation of a small vault RNA (svtRNA2-1a) is an early event in Parkinson disease and induces neuronal dysfunction. *RNA Biol.*, **10**, 1093–1106.

Minones-Moyano,E. *et al.* (2011) MicroRNA profiling of Parkinson's disease brains identifies early downregulation of miR-34b/c which modulate mitochondrial function. *Hum. Mol. Genet.*, **20**, 3067–3078.

Pantano,L. *et al*. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.

Pantano,L. *et al*. (2011) A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome. *Bioinformatics*, **27**, 3202–3203.

Perez-Enciso,M. and Tenenhaus,M. (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.*, **112**, 581–592.

Saikia,M. *et al*. (2012) Genome-wide identification and quantitative analysis of cleaved tRNA fragments induced by cellular stress. *J. Biol. Chem.*, **287**, 42708–42725.

Selitsky,S.R. *et al*. (2015) Small tRNA-derived RNAs are increased and more abundant than microRNAs in chronic hepatitis B and C. *Scientific Reports*, **5**, 7675.

Shulman,J.M. *et al*. (2011) Parkinson's disease: genetics and pathogenesis. *Annu. Rev. Pathol.*, **6**, 193–222.

Sobala,A. and Hutvagner,G. (2011) Transfer RNA-derived fragments: origins, processing, and functions, Wiley interdisciplinary reviews. *RNA*, **2**, 853–862.

Sobala,A. and Hutvagner,G. (2013) Small RNAs derived from the 5' end of tRNA can inhibit protein translation in human cells. *RNA Biol.*, **10**, 553–563.

Soreq,L. *et al*. (2013) Small RNA sequencing-microarray analyses in Parkinson leukocytes reveal deep brain stimulation-induced splicing changes that classify brain region transcriptomes. *Front Mol. Neurosci.*, **7**, 55.

Thompson,D.M. *et al*. (2008) tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA*, **14**, 2095–2103.

Xia,J. and Wishart,D.S. (2011) Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.*, **6**, 743–760.