DraculR: A web based application for \textit{in silico} haemolysis detection in high throughput small RNA sequencing data

Melanie D. Smith$^{1,2,*}$, Shalem Y. Leemaqz$^{1}$, Tanja Jankovic-Karasoulos$^{1,2}$, Dylan McCullough$^{1}$, Dale McAninch$^{2}$, James Breen$^{3,4}$, Claire T. Roberts$^{1,2}$, Katherine A. Pillman$^{5,*}$

$^1$ Flinders Health and Medical Research Institute, Flinders University, Bedford Park SA 5042, Australia
$^{2}$ Adelaide Medical School, University of Adelaide, Adelaide SA 5005, Australia
$^3$ Indigenous Genomics, Telethon Kids Institute, Adelaide SA 5000, Australia
$^4$ College of Health & Medicine, Australian National University, Canberra ACT 2600, Australia
$^5$ Centre for Cancer Biology, University of South Australia/SA Pathology, Adelaide SA 5000, Australia

* Correspondence: melanie.smith@flinders.edu.au & Katherine.Pillman@unisa.edu.au

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Motivation: The search for novel microRNA (miRNA) biomarkers in plasma is hampered by haemolysis, the lysis and subsequent release of red blood cell (RBC) contents, including miRNAs, into surrounding fluid. The biomarker potential of miRNAs comes in part from their multi-compartment origin, and the long-lived nature of miRNA transcripts in plasma, giving researchers a functional window for tissues that are otherwise difficult or disadvantageous to sample. The inclusion of RBC derived miRNA transcripts in downstream analysis introduces a source of error that is difficult to identify post hoc and may lead to spurious results. Where access to a physical specimen is not possible, our tool will provide an in silico approach to haemolysis prediction.

Results: We present DraculR, an interactive Shiny/R application that enables a user to upload microRNA expression data from short read sequencing of human plasma as a raw read counts table and interactively calculate a metric that indicates the degree of haemolysis contamination.

Introduction

Circulating miRNAs have long been identified in human plasma and, given their stability in this medium, have strong potential as biomarkers. Whilst there are multiple techniques for quantifying the abundance of miRNAs in plasma, high throughput sequencing (HTS) detects both known and novel (i.e. putative) miRNAs with single base resolution. The fine resolution provided by HTS allows distinction between variants differing by a single nucleotide, as well as isomiRs of differing length (1) with many researchers now leveraging this technology (2–11). The biomarker potential of plasma miRNAs is in part because plasma derived transcripts commonly originate from varied endogenous compartments. Accurate profiling of plasma miRNAs is impeded when transcripts are derived from another blood source such as RBCs. When haemolysis occurs due to shearing of RBCs during blood sampling, miRNAs are released into the volume of blood drawn (12–16). The presence of the RBC-associated miRNAs alters the plasma expression profile affecting the global normalisation of sequence counts (1).

The increase in relative abundance of RBC-associated miRNAs, and the aberrant normalisation of libraries, have potential to impact the profile analysis of miRNAs (12,13,15) yet assessment of haemolysis is rarely reported. Data quality checks prior to analysis of HTS data should include an assessment of haemolysis in the plasma sample from which the sequencing library was produced. There are currently two gold standard approaches: 1. Delta quantification cycle (ΔCq), where expression levels of a known blood cell associated miRNA (miR-451) and a control miRNA (miR-23a) are determined based on the difference between the two raw Cq values; and 2. Spectrophotometry, based on absorbance maximum of free haemoglobin measured at 414 nm (12,17–19). However, both gold standard methods rely on access to the original plasma sample. In this work we present DraculR, a data driven approach for the assessment of haemolysis confounding in silico. The DraculR tool enables the user to upload self-generated or publicly available high throughput miRNA sequencing data for assessment and returns both visual and tabular recommendations for downstream analysis of read count data.

Materials and methods

DraculR is an interactive, Shiny/R web-based tool, for the in silico assessment of haemolysis contributions to small RNA sequencing libraries prepared from human plasma. DraculR utilises the Haemolysis metric (20), a measure analogous to the ΔCq (miR-23a-miR-451) method, which determines the difference between the abundance of two miRNAs, one known to vary, and one known to be invariant in the presence of haemolysis. The Haemolysis metric is calculated as the sample specific difference in geometric means of the normalised gene expression values between
two sets of microRNAs: 1) 20 miRNAs identified as indicative of haemolysis (‘signature set’), and
2) all other microRNAs (‘background’). In this case, the geometric mean of the reduced signature
set will be calculated, as defined in (1).

Let

\[ Z_x = \log_2 CPM \text{ counts of Reduced signature set} \]

and

\[ Z_y = \log_2 CPM \text{ counts of Background set} \]

where \( x = 1, 2, 3, \ldots, p_1 \) with \( p_1 \) is the number of miRNA in Reduced signature set
and \( y = 1, 2, 3, \ldots, p_2 \) where \( p_2 \) is the number of miRNA in Background
and \( i = 1, 2, 3, \ldots, n \) where \( n \) is the sample size after filtering

\[ \text{Haemolysis Metric}_i = \left( \frac{1}{p_1} \prod_{x=1}^{p_1} Z_x^i \right) - \left( \frac{1}{p_2} \prod_{y=1}^{p_2} Z_y^i \right) \]

(1)

Application

By using the Haemolysis metric calculation, researchers/clinicians can assess their samples for
evidence of haemolysis and obtain recommendations for their own individual samples as clear for
use (‘Clear’) or use with caution (‘Caution’). We have shown that when RBC-associated miRNA
transcripts are retained in a plasma sample post centrifugation, with subsequent incorporation into
the sequencing library, the relative abundance of these miRNAs is increased. This increase is
evidenced by the increase in the geometric mean of signature haemolysis miRNAs away from that
of the background miRNA giving the background miRNA a smaller relative expression than would
be expected from a pure plasma sample taken from the same individual. Using miRNA
sequencing data, DraculR is designed to analyse and visualise the distribution of miRNA counts
from our haemolysis signature set and compare this with the distribution of counts from
background miRNAs. Once calculated, the sample specific Haemolysis metric for user defined
data is returned in tabular and graphical format for download and assessment. Samples with a
Haemolysis metric \( \geq 1.9 \), our recommended threshold, indicate haemolysis as evidenced by the
RBC-associated miRNA retention. Haemolysed samples are identified by the word ‘Caution’ in the
‘Haemolysis Result’ column of the ‘Results Summary’ tab. We recommend removal, or at a
minimum further investigation, of any samples that return a Haemolysis metric above the threshold
set here prior to use in any downstream analysis.

Public data example

To illustrate the utility of the application we downloaded four publicly available human plasma HTS
miRNA datasets from NCBI GEO (21). The datasets used here were GSE153813, GSE118038,
GSE105052, GSE151341 (22–24). Where the publication associated with the given dataset included miRNA differentially expressed between the conditions being considered, and these miRNA correspond with our haemolysis signature set, these miRNA were dropped from the Haemolysis metric calculation. In each example, we were able to detect evidence of haemolysis in multiple samples (Table 1).

Table 1. Publicly available human plasma miRNA data were assessed for haemolysis using the DraculR method identifying multiple samples to use with caution in each dataset. No haemolysis information was included with the original dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Experimental Context</th>
<th>Total Samples</th>
<th>Caution</th>
<th>Differentially abundant miRNA</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE153813</td>
<td>Case:Control Profile miRNA expression at each stage of menstrual cycle; endometriosis</td>
<td>9</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>GSE118038</td>
<td>Case:Control Prostate cancer biomarker</td>
<td>70</td>
<td>32</td>
<td>miR-4732-3p, let-7a, miR-26b-5p, miR-98-5p, miR-30c-5p*, miR-21-5p</td>
<td>(24)</td>
</tr>
<tr>
<td>GSE105052</td>
<td>Case:Control Friedreich's ataxia</td>
<td>42</td>
<td>3</td>
<td>miR-128-3p, miR-625-3p, miR-130b-5p, miR-151a-5p, miR-330-3p, miR-323a-3p, miR-142-3p</td>
<td>(22)</td>
</tr>
<tr>
<td>GSE151341</td>
<td>Case:Control Early radiographic knee osteoarthritis biomarker</td>
<td>91</td>
<td>4</td>
<td>miR-335-3p, miR-199a-5p, miR-671-3p, miR-1260b, miR-191-3p, miR-335-5p, miR-543</td>
<td>(23)</td>
</tr>
</tbody>
</table>

* miRNA associated with Haemolysis Metric signature

DraculR provides a visual representation of the results in the form of a histogram (Figure 1). In this histogram, the difference between the geometric means can be seen against a background of data with known haemolysis quantification based on the ΔCq (miR-23a-miR-451) method. Samples are coloured to indicate either ‘Clear’ (light blue) or ‘Caution’ (scarlet) with background data coloured to indicate the ΔCq (miR-23a-miR-451) result of either ‘Clear (ΔCq)’ (grey with a light blue highlight) for samples with ΔCq < 7 or ‘Haemolysed (ΔCq)’ (grey with a scarlet highlight) for samples with ΔCq ≥ 7. Background data is taken from Smith et al. (20) and represents individuals for which we have both HTS and ΔCq (miR-23a-miR-451) analyses. The false positives reported...
here (ie where DraculR reports haemolysis but ΔCq < 7) may not be accurate as ΔCq values may be affected by pregnancy status.

![Figure 1](image)

**Figure 1.** DraculR uses public data to illustrate the potential for unidentified haemolysis with potential to confound biomarker analysis. Here, data pertaining to GSE105052 was retrieved from NCBI Geo for analysis. The screen shot above shows an example where five samples were identified to be used with caution. All data are presented against a background of haemolysed samples assessed using the ΔCq method.

**Conclusion**

We have developed a Shiny/R web-based application that allows users to detect and address the issue of haemolysis in plasma miRNA HTS data. DraculR addresses the need for quality control where, either through use of public data, exhaustion of sample, or exhaustion of funds, it is not possible to assess haemolysis using one of the current gold standard approaches (being delta quantification cycle (Cq) values for miR-23a-miR-451 or Spectrophotometry for haemoglobin estimation). The application is easy to use and applicable to small RNA sequencing data from human plasma. The method is robust for cases where a case-control style analysis is undertaken, and the user has *a priori* knowledge of miRNA that are anticipated to be differentially abundant between groups. Whilst a probabilistic quantification of contamination risk is not possible based on the dataset used here, we plan future work drawing on the methods used by Shah *et al.* (2016) that will include serial dilution and miRNA quantification of haemolysed plasma samples to validate and further refine our method. DraculR adds value to the growing resource of public data shared by plasma researchers by enabling *in silico* analysis of haemolysis confounding post sequencing. The detection of haemolysis using our Haemolysis metric enables the user to identify and potentially discard low quality samples which are otherwise unknown to be affected by haemolysis. This enables an additional quality metric and the subsequent increased confidence in the use of high throughput miRNA sequencing data for which no haemolysis information is available (for more details see Supplementary Material).
Acknowledgements

We wish to acknowledge the generosity of the women who donated their blood for our research. Without them, this research would not be possible. We also acknowledge valuable input from QIAGEN Genomic Services.

Funding information

This research is supported by NIH NICHD R01 [grant number HD089685-01] Maternal molecular profiles reflect placental function and development across gestation PI Roberts, an Australian Government Research Training Program (RTP) Scholarship awarded to MDS, a National Health and Medical Research Council Investigator Grant [grant number GNT1174971] awarded to CTR and a Matthew Flinders Professorial Fellowship awarded to CTR and funded by Flinders University. JB is supported by the James & Diana Ramsay Foundation. KAP is supported by the Florey Fellowship funded by the Adelaide Hospital Research Committee.
References


