RACE-SEQ and population-wide polymorphism susceptibility testing for endonucleotically active, RNA-targeting therapeutics.

Running title: RACE-SEQing RNA cleaving drug mechanism of action

Louise Usher[1], Pantazis I. Theotokis[2,3] and Sterghios A. Moschos[4]

1. Department of Biomedical Sciences, Faculty of Science and Technology, University of Westmnister, 115 New Cavendish Str., London W1W 6UW, UK.
2. National Heart and Lung Institute, Imperial College, South Kensington, London, SW3 6LY, UK.
3. Cardiovascular Research Centre, Royal Brompton and Harefield NHS Foundation Trust and Imperial College London, London SW3 6NP, UK.
4. Department of Applied Sciences, Faculty of Health and Life Sciences, Northumbria University, Ellison Building, Ellison Place, Newcastle Upon Tyne, Tyne and Wear, NE1 8ST, UK.

Corresponding author: Dr. Sterghios A. Moschos, Sterghios.moschos@northumbria.ac.uk, +44 (0)191 227 3013.

i. Abstract

High throughput sequencing of the products of 5' RNA ligase-mediated rapid amplification of cDNA ends (5' RLM-RACE) reactions (RACE-SEQ) enables the mapping and digital enumeration of expected and novel 5' ends in RNA molecules. The resulting data are essential in documenting the mechanism of action and precision of endonucleotically active, RNA-targeting drugs such as RNase H-active antisense or small interfering RNA. When applied to error-prone replication systems such as RNA viruses or *in vitro* RNA replicon systems, the method can additionally report the relative susceptibility of known and unknown polymorphisms to a prospective sequence-specific drug, making it a powerful tool in patient selection and stratification, as well as resistance prediction.

We describe the preparation of sequencing libraries for ultra-high depth 5' RLM-RACE analysis on two popular second-generation high throughput sequencing platforms (Illumina, Ion Torrent) and supply a detailed bioinformatic analysis pipeline for target site activity definition and enumeration. We further illustrate how the pipeline can be simply modified to generate polymorphism-specific drug susceptibility data from *in vitro* replicon experiments (RACE-SEQ-MM), in a patient-free manner, to cover both known and unknown target site variants in the population.

ii. Keywords

RACE, 5' RLM-RACE, RACE-SEQ, RACE-SEQ-MM, RNAi, siRNA, antisense, mechanism of action, pipeline.

1. Introduction

It is now well-accepted that one of the critical requirements for realizing the translational potential of molecular therapeutics is the availability of conclusive evidence of an on-target mechanism of action (*1*). Experimental data must therefore evidence effective target engagement and, where an enzymatic

function is involved, the appropriate reaction products (*2*). For sequence-specific, endonucleotically active, RNA-targeting molecular therapeutics such as RNase H-active antisense and small interfering RNA (siRNA) mediators of RNA interference (RNAi), these reaction products are the novel 5' ends in the targeted RNA, at the anticipated site of action (*3*).

Traditionally, 5' RNA ligase-mediated rapid amplification of cDNA ends (5' RLM-RACE) reactions were used to document such novel 5' ends. Briefly, this involved the ligation of a known, synthetic RNA adapter to free 5' RNA ends, production of cDNA using gene-specific reverse transcription primers (GSP), and the generation of PCR / nested PCR amplicons using primers specific to the ligated RNA adapter (upstream) and the gene specific primer (downstream) surrounding the drug active site. The resulting amplicon would then be sized by gel electrophoresis, excised, purified, and Sanger-sequenced to determine the drug site of action as the point at which the known RNA adapter sequence would stop and the targeted RNA sequence would start (Fig. 1). However, this qualitative method routinely generated unexpected electrophoresis gel bands, and/or minor Sanger sequencing traces; these were typically overlooked as non-specific amplification artefacts or contaminants, respectively. Using high throughput, second generation sequencing on such 5' RLM-RACE products (RACE-SEQ) revealed that the minor Sanger sequencing traces were due to the imprecise Slicer activity of siRNA-induced silencing complexes (siRISC) (*4*). Furthermore, as additional data emerged on the method from multiple laboratories, it became clear that the extent of information returned from this analytical approach was a function of the biology of the experimental setup, and the data analysis pipeline implemented. Thus, parameters such as the mutation rate of the RNA target (tissue RNA (*5*–*7*), endogenous targets in cell culture (*8*), viral RNA or replicon RNA in cell culture (*4*, *9*, *10*)), depth of sequencing and breadth of sequencing (*4*) orchestrate the observed outcomes, including predictive pharmacogenomic assessment of drug efficacy against targets mismatched to the drug (RACE-SEQ-MM: (*9*)).

The RACE-SEQ method has been applied on the Illumina and Ion Torrent platforms, two of the most popular 2nd generation sequencing systems available to, or accessible by most laboratories. Despite their radically different chemistries and physics, RACE-SEQ workflows on both platforms can be summarized into three steps: 5' RLM RACE-based sequencing library preparation, sequencing, and data analysis. In this chapter we focus on the library preparation (RNA adapter ligation, reverse transcription, sequencing amplicon generation) and data interpretation aspects, and highlight opportunities to reduce process complexity and cost where possible. For example, the following protocol covers RACE-SEQ or RACE-SEQ-MM library preparation without sample multiplexing, or sequencing adapter-tagging at the PCR / nested PCR step, approaches that can reduce the number of handling steps. Likewise, the data analysis components indicate how datasets can be processed either for cleavage point characterization (RACE-SEQ) or polymorphism tolerance testing (RACE-SEQ-MM).

2. Materials

Standard good laboratory practice procedures for RNA work apply throughout these methods, i.e. RNA quality between $OD_{260/280}$ 1.8-2.1 or Agilent Bioanalyzer RNA integrity number (RIN) > 6.0, use of certified RNase-free barrier tips and plasticware (including during sequencing library preparation steps), use of precision (electronic by preference) pipettors, pre-treatement of any in-house containers with 0.1% v/v diethylpyrocarbonate (DEPC), and workspace/glove treatment with RNase ZAP™ or equivalent.

2.1 Preparation of 5' RLM RACE-based sequencing amplicons.

A wide variety of 5' RLM RACE kits are available from a number of commercial reagent suppliers, including Merck (previously known as Sigma Aldrich), Thermo Fisher Scientific (Life Technologies,

Invitrogen), and Clontech, amongst others. The kits may cover all steps up to and including library purification allowing them to be ready for (Sanger) sequencing. The reader is invited to use any of these (RACE-SEQ was independently developed by two separate groups using the Invitrogen GeneRacer kit (*4*, *5*)) provided attention is paid to key parameters as detailed in the steps below. Alternatively, the required kit components such as RNA adapter oligonucleotides, primers, enzymes, buffers, and purification components can be sourced independently (e.g. Integrated DNA Technologies, Eurogentec, Guangzhou Ribobio Company, New England Biolabs) for a fraction of the cost.

2.1.1 RNA Adapter ligation

1. Total RNA in excess of 0.4 µg from a biological experiment (in a maximum volume of 12.5 µl for a 20µl adapter ligation reaction).
2. Nuclease-free water, either manufactured in house using >18MOhm water treated with 0.1% DEPC overnight and autoclaved to degrade DEPC, or commercially procured.
3. A 5' RACE RNA adapter *without* a 5' phosphate; we routinely use for experiments in *Homo sapiens* 5'-GGA CAC UGA CAU GGA CUG AAG GAG UAG AAA-3', desalted purity level, resuspended at 0.1 mM concentration in nuclease-free water (notes 1-6). Store in aliquots at -80^{o}C.
4. T4 RNA ligase I, 10,000 units/ml, and associated 2x reaction buffer (New England Biolabs, Hitchin, UK).
5. RNase Inhibitor, murine, 40,000 units/ml (New England Biolabs). Store at -20^{o}C.
6. GlycoBlue coprecipitant, RNase and DNase free, 15 mg/ml (ThermoFisher Scientific, Loughborough, UK).
7. 10 mM Adenosine 5' Triphosphate (ATP).
8. Phenol-chloroform for RNA precipitation (5:1 ratio; molecular biology grade).

9. Absolute ethanol (EtOH), molecular biology grade.
10. 3M Sodium acetate, pH 5.2 at 25 ^{o}C.
11. Nuclease-free PCR microcentrifuge tubes (typically 0.2 ml) and 1.5 ml microcentrifuge tubes.
12. Thermal cycler.
13. Wet ice.
14. Dry ice (or accessible -80^{o}C freezer shelf).
15. Refrigerated microtube centrifuge.
16. Vortex mixer.

2.1.2 Reverse transcription

1. A strand-selective, gene specific primer (GSP) for reverse transcription, designed to anneal downstream from the anticipated drug cleavage point on the target RNA, stored at 0.1 mM stock concentration in Tris-EDTA or nuclease-free water, at -20^{o}C (notes 7-13).
2. Moloney Murine Leukemia Virus Reverse Transcriptase (Mo-MuLV RT), 200,000 units/ml, and associated 10x reaction buffer (New England Biolabs; note 14).
3. RNase Inhibitor, murine, 40,000 units/ml (New England Biolabs).
4. 10 mM each deoxynucleotide triphosphate (dNTP) mix.
5. Nuclease-free PCR microcentrifuge tubes.
6. Nuclease-free water.
7. Thermal cycler.
8. Wet ice.

2.1.3 Sequencing amplicon generation, size selection, and purification.

PCR amplification offers several opportunities to streamline library preparation ahead of 2nd generation sequencing. The approach to be taken may be influenced by the choice of sequencing platform, RACE-SEQ application, and/or sequencing kit selection (notes 15-16). For downstream second generation sequencing, kits tailored to the Illumina and Ion Torrent platforms can be purchased from the instrument manufacturers or reliable 3rd party suppliers such as New England Biolabs, Roche (KAPA Biosystems), BIOO Scientific, Labgene Scientific S.A., Bioline, etc. These kits usually come with comprehensive instructions relevant to each platform, and are sold in versions for specific applications such as RNA-SEQ, genomic DNA sequencing, amplicon sequencing etc. The reader is advised to carefully review the sequencing kit protocol and sequencing adapter approach used per platform / kit supplier to select an appropriate solution for their experimental setup (notes 17-18). The reader may also prefer to outsource 2nd generation sequencing; performing library preparation and validation in house can substantially reduce oursourced sequencing costs.

1. A forward PCR primer corresponding to the 5' RNA adapter, such as 5'-GGA CAC TGA CAT GGA CTG AAG GAG TA-3' (GeneRacer Forward), stored at 0.1 mM stock concentration in Tris-EDTA or nuclease-free water (note 19).
2. The GSP used during reverse transcription, or a nested GSP primer (note 20).
3. Q5 Hot Start High Fidelity 2x Master Mix (New England Biolabs; note 21).
4. Paramagnetic PCR cleanup beads and neodymium magnet e.g. Agencourt AMPure XP beads (Beckman Coulter, High Wycombe, UK; note 22).
5. Column-based nucleic acid gel extraction kit (e.g. New England Biolabs Monarch DNA Gel Extraction Kit).
6. Nuclease-free PCR microcentrifuge tubes (typically 0.2 ml).
7. Nuclease-free 1.5 ml microcentrifuge tubes.

8. Nuclease-free water.
9. 70% ethanol (molecular biology grade) in nuclease-free water.
10. Agarose, low melting point.
11. Gel electrophoresis equipment.
12. 100 bp DNA ladder.
13. DNA loading dye (e.g. ThermoFisher 6X Orange DNA loading Dye).
14. DNA staining dye (e.g. ThermoFisher SYBR Safe).
15. 1x Tris-borate EDTA (TBE) buffer (89 mM Tris-borate, 2 mM EDTA, pH 8.3 at 25^{o}C).
16. Tris-EDTA (TE) buffer (10 mM Tris-HCl pH 8.0 at 25^{o}C, 0.1 mM EDTA).
17. Thermal cycler.
18. Isopropanol.
19. UV transilluminator compatible with gel band excision (e.g. ThermoFisher Safe Imager 2.0 Blue Light Transilluminator).
20. Wet ice.
21. High sensitivity spectrophotometric nucleic acid quantification system (ThermoFisher Nanodrop or QuBit).

2.2 Bioinformatics.

You will need to use a Linux operating system for the RACE-SEQ-lite script to run the necessary command line packages. If you are using a Bio-Linux distribution (Bio-Linux 8), then you will have most of the packages pre-installed system-wide. If you are setting up a new workstation with a Linux based operating system, such as a virtual operating system hosted within a Windows or MacOS environment, we strongly recommend that you opt for a Bio-Linux distribution, as it will save you a substantial amount of time on installing, setting up and troubleshooting your machine and your environment. If you are

using a Virtual Machine (VM) environment we recommend using a Bio-Linux distribution as the same benefits apply here as well; make sure that data storage is handled in the host operating system to optimize disk drive allocation to the VM.

2.2.1 Software packages

1. To trim the RACE adapter, the default package is Cutadapt (>= v. 1.2).
2. For the alignment of the reads to the reference sequence you will need either Bowtie (v. 1.0.0) for reads generated with Illumina sequencer (eg. NextSeq, MiSeq, NovaSeq, HiSeq, etc.) or Tmap (>= v. 3.4.1) which is the preferred aligner for reads generated with an Ion Torrent sequencer (eg. PGM, PROTON). If the sequencing has been performed using Illumina technology, then the Tmap aligner does not need to be installed as in that case the aligner of preference is Bowtie.
3. You will need the Samtools (>= v. 1.3.1) and Bedtools (>= v. 2.17.0) software packages installed.
4. To run the RACE-SEQ-lite script you will need to have a version of R (>= v. 3.4) installed on your system. All the necessary R packages will get installed and loaded automatically.
5. All command line packages need to be installed system-wide.

3. Methods

3.1 Preparation of 5' RLM RACE-based sequencing amplicons.

3.1.1 RNA Adapter ligation

1. Working on wet ice, prepare a 20µl adapter ligation reaction by adding <12.5µl of total RNA extract (>0.4 ug; volume can be made up to 12.5 µl with nuclease-free water) to 2µl of RNA adapter (0.1 mM) in a nuclease-free thermal cycling tube (tube 1; note 23).

2. Heat tube 1 in a thermal cycler to 65°C for 5 min and immediately transfer the tube 1 to wet ice. Cool on wet ice for 2 min.
3. Working on wet ice, prepare the RNA ligase mix in a separate nuclease free tube (tube 2) consisting of 2µl of 2x T4 RNA ligase I buffer, 2µl of 10 mM ATP, 0.5µl of RNase Inhibitor, and 1µl of T4 RNA ligase I (note 24).
4. Pipette mix tube 2 gently and briefly centrifuge for 10 sec at 10,000 xg to collect the tube contents to the bottom (note 25).
5. Working on wet ice, transfer 5.5µl of the ligase mix (tube 2) to the annealed RNA (tube 1); discard tube 2.
6. Using a thermal cycler, perform RNA adapter ligation in tube 1 at 37 °C for 1 hr, with the heated lid (if available) switched off (note 26).
7. To cleanup the RNA adapter-ligated RNA extract, first bring the RNA adapter ligation reaction volume in tube 1 up to 0.1 ml using nuclease-free water.
8. Add 0.1 ml phenol/chloroform to the RNA adapter ligation reaction (tube 1) and secure the tube lid shut.
9. Vortex mix at full speed for 30 sec.
10. Centrifuge at 12,000 xg for 5 minutes at room temperature.
11. *Carefully* remove tube 1 from microcentrifuge *without* disturbing the oil-water interphase (note 27).
12. *Carefully* transfer the upper aqueous phase (~0.1 ml) into a new 1.5 nuclease-free microcentrifuge tube (tube 3) using nuclease-free barrier tips. Record the volume transferred.
13. Add to tube 3 the following materials in the specified order: 1.5µl GlycoBlue, 10µl of 3M sodium acetate, and 2x volumes 96-100% v/v EtOH relative to the amount transferred in step 12 (note 28).
14. Vortex tube 3 for 30 sec.

15. To precipitate the RNA, incubate tube 3 on dry ice for 30 min (note 29).
16. Centrifuge tube 3 at 16,000 xg for 20 min at 4°C.
17. *Carefully* remove tube 3 from the centrifuge, and aspire the supernatant without disturbing the pellet: discard the supernatant.
18. Wash the pellet without disturbing it by adding 0.5 ml of 70% v/v EtOH to tube 3. Mix by inversion x3.
19. Vortex tube 3 for 10 sec.
20. Centrifuge tube 3 at 16,000 xg for 2 min at 4°C and remove the supernatant as per step 17.
21. Centrifuge tube 3 for a further 10 sec to collect and remove residual supernatant EtOH as per step 17.
22. Air dry pellets for up to 5 min.
23. Resuspend the adapter-ligated RNA pellet in 11-20µl of RNase-free water and mix well by pipetting.
24. Keep on wet ice or store at -20°C (up to 2 weeks) or -80°C (long term).

3.1.2 Reverse transcription

1. Transfer the 11µl of adapter ligated RNA produced in the previous procedure into a nuclease-free PCR microcentrifuge tube (tube 4; note 30).
2. To the 11µl of adapter ligated RNA (tube 4), add 1µl of GSP diluted to a 1 uM woking concentration using nuclease-free water, and 1µl of dNTPs (10 mM each; note 31).
3. Transfer tube 4 onto a thermal cycler and heat for 5 min at 65°C to remove secondary RNA structure. Immediately transfer tube 3 onto wet ice for 2 min.

4. In a separate nuclease-free tube (tube 5), create a reverse transcription master mix consisting of 2 µl 10X Mo-MuLV RT reaction buffer, 0.2 µl RNase Inhibitor, 1 µl of Mo-MuLV RT enzyme, and 3.8 µl of nuclease-free water (7µl final volume; note 32).
5. Transfer the 7µl contents of the reverse transcription master mix (tube 5) to the adapter-ligated RNA (tube 4) to give a total volume of 20µl and keep it on wet ice; discard tube 5.
6. Transfer reverse transcription mix (tube 4) onto a thermal cycler and perform reverse transcription at 55°C for 30 min followed by enzyme inactivation at 70°C for 15 min.
7. Store resulting cDNA at -20°C (2 weeks), -80°C (indefinitely), or proceed to sequencing amplicon generation.

3.1.3 Sequencing amplicon generation, size selection, and purification

1. Create a PCR reaction in a nuclease-free PCR microcentrifuge tube (tube 6) consisting of the following materials: For a 40µl reaction, add 20µl Q5 Hot Start High-fidelity 2X Master Mix, 2µl of 10 µM GeneRacerF1 primer, 2 µl of 10 µM GSP, 4 µl of cDNA template from tube 4 (cDNA volume should be up to 10% v/v of the PCR reaction volume) and 12 µl of nuclease-free water (note 33).
2. Transfer the PCR mix (tube 6) onto a thermal cycler and perform PCR as follows: 98°C for 30 sec; 35 amplification cycles (95°C for 15 sec, 60°C for 10 sec, and 72°C for 15 sec), 72°C for 2 min and 4°C hold (notes 34-35).
3. Analyse PCR amplicons by preparative agarose gel electrophoresis. This also allows for gel extraction-based size selection and cleanup. Prepare a 2% w/v low melting point agarose gel in TBE buffer and supplement with appropriate DNA staining dye (e.g. 2µl of SYBR Safe per 50 ml of gel; notes 36-37).
4. Load 5-20µl of PCR product diluted 1:6 with 6X Orange loading Dye alongside appropriate DNA ladder and run at 100V for 40-60 min (note 38).

5. Visualise bands under a blue light transilluminator, and excise target bands using a sterile scalpel into nuclease-free 1.5 ml microcentrifuge tubes; do not use UV transilluminators if the excised DNA band is to be extracted and used for sequencing.
6. Purify the DNA from the agarose using a gel extraction kit; steps 6-10 describe the steps relevant to the New England Biolabs Monarch DNA Gel Extraction kit. Place gel in 4x volumes of gel dissolving buffer and incubate at 50^{o}C for 5-10 min.
7. Transfer entire volume onto the Monarch DNA clean up column in batches of 0.8 ml.
8. Centrifuge column at 13,000 xg for 1 min and discard flow through (repeat steps 7 and 8 to process entire volume).
9. Wash column 2x with 0.2 ml DNA wash buffer by centrifuging at 13,000 xg for 1 min, discarding the flow through.
10. Elute in 20µl NEB elution buffer, TE, or nuclease free water.
11. To further purify and concentrate DNA if necessary, e.g. for the removal of primer dimers, follow steps 12-25 (for >100 bp amplicons) *OR* 26-36 (for <100 bp amplicons), or proceed with step 36.
12. Add 1.6x volumes (32 ul) of room temperature Agencourt AMPure XP beads.
13. Mix by pipetting at least 5x.
14. Incubate on bench for 8 min.
15. Place on neodymium magnet (or rack) for 3 min. Beads should form a visible pellet collect close to the magnet.
16. Remove supernatant by pipetting taking care not to disturb the pellet.
17. Wash beads 2x with 0.3 ml of 70% v/v EtOH without disturbing the pellet.
18. Remove residual ethanol by pipetting.
19. Air dry for 3-5 min.
20. Resuspend in 25µl of TE by pipetting.

21. Vortex at full speed for 10 sec.
22. Centrifuge at 0.3 xg for 5 sec.
23. Place on neodymium magnet for 2 min.
24. Collect supernatant (~23 ul) into a new nuclease-free tube.
25. Quantify using high sensitivity spectrophotometric approach and proceed to step 37.
26. Prepare a mix of 0.07 ml isopropanol and 0.18 ml AxyPrep beads.
27. Transfer 25µl of the bead-isopropanol beads to 10µl of PCR product.
28. Mix by pipetting 5x.
29. Incubate at room temperature for 5 min.
30. Tramnsfer tubes to neodymium magnet (or rack) for 3 min.
31. Remove supernatant by pipetting without disturbing the bead pellet proximal to the magnet.
32. Wash 2x with 0.2 ml 70% EtOH without disturbing the pellet.
33. Remove residual EtOH by 0.3 xg 5 sec centrifugation and pipetting.
34. Air dry for <5 min.
35. Resuspend in 40µl of TE buffer.
36. Quantify using high sensitivity spectrophotometric approach.
37. Proceed to library preparation in accordance to the sequencing library kit protocol suited to amplicon size and sequencing platform of choice.
38. Retrieve FASTQ data at sequencing run end for bioinformatic analysis.

3.2 Bioinformatic software installation.

1. To install the required software system-wide you will need to have administrator privileges on your LINUX system. Administrator privileges are necessary if you want to download and install new packages in your system or edit and get access to protected files created by other users.

2. If you have python already installed on your system, the easiest way to download and install the Cutadapt package is by using the native python-pip command. You can check in you have python installed by typing

```
python --version  OR
python3 --version
```

3. Once you know you have either python or python3 working you can download and install python-pip from the command line using apt-get. Once you have python-pip, download and install the latest Cutadapt version using the pip trusted repository.

```
sudo apt-get install python-pip
sudo pip install cutadapt
```

   For more information and troubleshooting please read the Cutadapt documentation. (https://cutadapt.readthedocs.io/)

4. You can install Bowtie, Samtools and Bedtools from the command line using apt-get. If you are working on a Bio-Linux machine you have Bowtie, Samtools and Bedtools already preinstalled. Otherwise use the following command to download the latest versions.

```
sudo apt-get install bowtie samtools bedtools
```

5. To download and install the latest version of the R environment system-wide use the command.

```
sudo apt-get install r-base
```

6. For Ion Torrent data, use the Tmap aligner for optimal alignment results. To download and install system-wide the latest version of the Tmap aligner as a separate command line package please follow the instructions on their official repository on GitHub. (https://github.com/iontorrent/TS/tree/master/Analysis/TMAP)

3.3 Data pre-processing

1. Depending the sequencing platform (eg. Illumina, IonTorrent) and whether it has been carried in-house or externally you will need to follow a slightly different data pre-process step.

   If you are using IonTorrent you will have a fastq file for every sample with single-end reads. If you are using an Illumina sequencer you will have paired end reads in two different files. For pair end reads concatenate the files using the cat command to perform a duplex-like sequencing dataset.

```
cat your_reads1.fastq your_reads2.fastq > your_reads.fastq
```

2. During the initial adapter trimming stage, cutadapt filters for the reads that contain the exact adapter sequence on the 5-prime end of the reads. It then trims the RACE adapter sequence from these reads and outputs only the processed sequence. Any reads that do not contain the exact adapter sequence on their 5-prime end will be discarded (notes 39-44).
3. You will need to have your reference sequence in a fasta format file and make sure there is only one header written in the fasta file. Check if your reference file has only one header using the following command, which should return “1” if true.

```
grep '^>' your_reference.fasta | wc -l
```

4. Download the RACE-SEQ-lite script (https://github.com/pantastheo/RACE-SEQ-lite) from the GitHub repository and place it in your working directory along with your reference fasta file and only one sample fastq file that you want to process.

3.4 Run the RACE-SEQ pipeline

1. Run the RACEseqMM.r script using the Rscript command and the - -*help* option to bring up a short description and the help page.

```
Rscript RACE-SEQ.r -h
```

Options	Description
-s, --start	Input the first nucleotide genomic location
-e, --end	Input the last nucleotide genomic location
-a, --adapter	Input the RACE adapter sequence [default: NO]
-m, --mismatch	Input number of mismatches during alignment [default: 0]
-p, --plot	Print output graph [default: NO]
-t, --tmap	Use the tmap aligner instead of bowtie [default: NO]
--notsv	Do not print output TSV file [default: YES]
-i, --iterate	Create the alternative references to cover all the possible SNPs of the reference between the genomics locations specified by *-s* and *-e* and then perform global alignment and generate graph [default: YES]
-h, --help	Print this help page

2. While you run the script, it automatically identifies and imports the reference file in fasta format and the reads file in fastq format as standard inputs. If there are more than one fasta or fastq files in your working directory the script will exit with an error.
3. The minimum parameters you need to set for the script to run are the - -start and - -end options, which will set the genomic locations for the default .tsv (tab separated values) output

file. The following example command will perform alignment using Bowtie with 0 mismatches and output a .tsv formatted table with the number of the cleavage incidents in the span of the genomic region specified, in a linear scale (%) and in a logarithmic scale (log10; note 45).

```
Rscript RACEseqMM.r -s 9478 -e 9498
```

4. Using the following command, the script will use the Tmap aligner instead of the default Bowtie with 0 mismatches during alignment. Then it will output a graph of the region of interest but no table with values, on account of the - -*notsv* switch.

```
Rscript RACEseqMM.r -s 9478 -e 9498 -t -p --notsv
```

5. The following example command will perform the RACE adapter trimming on the 5' end of the reads using the specific nucleotide sequence provided with 0 mismatches in the trimming process. Then it will perform alignment with the reference sequence using the default Bowtie aligner and allowing for 2 mismatches during alignment. Then it will output a graph plotted between the specified genomic locations but not a .tsv table (notes 43, 44, 46).

```
Rscript RACEseqMM.r -s 9478 -e 9498 -a GGACACTGACATGGACTGAAGGAGTAGAAA -m 2 - p --notsv
```

4. Notes

1. The RACE-SEQ method was originally developed using the Life Technologies Generacer 5' RLM-RACE Kit, which was designed to analyse full length coding RNAs with intact 5' caps. Thus, in the first step of 5' RLM RACE, a phosphatase is used to remove 5' phosphates from mRNA fragments (such as those generated by endonucleotically active antisense and Slicer-active RNAi drugs) and non-mRNA transcripts. In the second step, a pyrophosphatase is used to remove 5' caps from full-length, capped mRNAs, leaving them as the only RNA species in a sample with a 5' phosphate

group. The third step involves the ligation of a synthetic RNA adapter of known sequence to these free 5’ phosphorylated ends using an RNA ligase. In RACE-SEQ the molecular species of interest are truncated RNAs, making the phosphatase step redundant, as a minimum. The redundancy of the pyrophosphatase step is a function of experimental design, e.g. drug cleavage site distance from the canonical 5’ end of the target, presence of 5’ cap structure(s), need for relative quantification of intact/cleaved species, etc.

2. Kit-derived or custom RNA adapters must be confirmed to be absent from the host genome by BLASTn before use to avoid spurious amplicon artefacts.
3. RNA adapter length can be adjusted at either the 5’ or 3’ end to conform with amplicon size needs. For example, the RNA adapter described herein is a shortened version of the RNA Adapter originally described in the Invitrogen GeneRacer kit: 5’-CGA CUG GAG CAC GAG GAC ACU GAC AUG GAC UGA AGG AGU AGA AA-3’.
4. Review the RNA ligase documentation for any specific enzymatic preferences regarding end nucleotide composition on any custom RNA adapters.
5. When purchasing custom synthesized RNA adapters, ensure that these are not 5’ phosphorylated so that adapter concatamerisation is avoided during the RNA adapter ligation step.
6. Custom-manufactured RNA adapters supplied at standard desalted purity (lowest manufacturing cost) are adequate for 5’ RLM-RACE.
7. Guidance for the identification of GSP binding sites is available in the manuals of 5’ RLM RACE kits from ThermoFisher, Clontech, Roche, Merck. In our experience, use of standard nearest neighbor-based primer design software such as PrimerQuest, Primer 3, VisualOMP, etc. are fit for purpose with additional attention placed on primer length (22-28 nt), relatively-high GC content (50-70%) for high melting temperature-based selectivity, and a maximum of two 3’ terminal guanines or cytosines. Standard PCR primer exclusion criteria such as primer dimer and hairpin formation

apply. It is recommended that the GSP melting temperature is similar to the RNA Adapter-specific forward primer.

8. Critical to experimental success is the size of the 5' RLM RACE sequencing library amplicon to be generated by PCR or nested PCR from the RNA adapter-tagged, GSP reverse-transcribed cDNA. The amplicon must be compatible with the 2nd generation sequencing platform operating parameters. Both Illumina and Ion Torrent are well-suited to small (<400 bp) DNA library sequencing, although practical limitations usually restrict library sizes to <300 bp, which is usually adequate for appropriate GSP binding site identification.
9. Both technologies can perform bidirectional sequence reading which markedly improves the reliability of resulting reads (*10*, *11*), although only Illumina can perform true paired end sequencing on libraries exceeding nominal read quality limitations (i.e. >300 bp). This is because of the physical limitations of Ion Torrent sequencing methods (bead-based emulsion extension reactions that do not physically accommodate libraries >400 bp), which do not apply on Illumina.
10. In the unlikely scenario that a suitable GSP site cannot be defined within the typical library size limits of Illumina, paired end read fidelity can be sacrificed to generate longer cDNAs; crucially, this does not impact on read quality over the definition of the novel 5' ends as these universally fall within the high quality read length region of the platform. Alternatively, 3rd generation sequencing platforms can be used such as Oxford Nanopore or Pacific Biosciences. However, the end user must note the currently poor accuracy of Oxford Nanopore compared to Illumina and Ion Torrent, and the relatively high cost per base of Pacific Biosciences for a given read depth, which would limit use to restriction site definition but render mismatched target cleavage tolerance (RACE-SEQ-MM) pharmacogenomics unreliable.
11. For A/T rich target sequences, the use of high melting temperature nucleoside modifications such as locked nucleic acid (LNA) can contain GSP length to <28 nt. When using LNA ensure these are

located at the 5’ end of the GSP binding site, with at least one non-LNA spacer nucleotide between modifications.

12. LNA-modified GSPs are not suitable for RACE-SEQ-MM as these will introduce specificity bias towards the ‘locked’ nucleotides in the target sequence queried.
13. Standard desalted primers are adequate for all steps; typically stored at 0.1 mM in TE and diluted to working concentrations in nuclease-free water.
14. Alternatively, the ThermoFisher Scientific Superscript III reverse transcriptase can be used to comparable outcomes; the reaction would need to be supplemented with RNase Inhibitor to a final concentration of 4,000 units/ml, and 5 µM final concentration of DTT.
15. The extent of amplicon processing following this step is a function of the sequencing platform and kit to be used; here we provide the minimum necessary processing for maximal usable read return after sequencing.
16. The Ion Torrent PGM and Illumina MiSeq can only analyse 1-2 samples by RACE-SEQ (1 million read minimum requirement). Large output sequencers (e.g. Illumina HiSeq, NovaSeq; Ion Torrent Proton) permit sample multiplexing through library barcoding, but only for RACE-SEQ; RACE-SEQ-MM requires 100 million reads as a minimum. Barcoded RACE-SEQ-MM is currently possible only on the Illumina NovaSEQ platform.
17. In our experience, standard library kits for sequencing DNA amplicons results in less troubleshooting around library preparation and validation.
18. All platforms have unique chemistry specifications with regards to sequencing adapter ligation to the sequencing library. The user can eliminate the need for such additional ligation steps by 5’ tagging the GSP/RNA Adapter primer, or the nested PCR primers if nested PCR is used, with the relevant sequencing adapter nucleotides (and nucleotide barcodes, when multiplexing). For more information regarding these sequences please consult the user manuals of your sequencing

platform. If applying such extensions to your primers it is necessary to re-confirm lack of primer dimers/hairpins for the extended primer sequence, and to evaluate reaction cleanup protocols for nucleic acid size limits based on the new sequencing library length.

19. A nested PCR primer, or shorter PCR primer can also be used; in the case of the 5' RLM RACE RNA adapter proposed, the nested PCR primer can be 5'-GGA CAC TGA CAT GGA CTG AAG G-3'. When performing RACE-SEQ-MM, data fidelity can be improved by modifying primers according to Kennedy *et al.* (*11*) for bidirectional sequencing; this does not apply for Illumina paired end sequencing, unless the library size is within the read length limits of the platform/sequencing kit used. Comparable fidelity can be achieved through RNA adapter sequence filtering at 100% identity in paired end sequencing.
20. The user may also reduce the need for 5' RLM RACE amplicon ligation to sequencing adapters and barcodes by 5' modifying the forward primer or nester PCR primer with sequences appropriate to the sequencing platform to be used.
21. A high-fidelity polymerase is essential for next generation sequencing, and particularly so for RACE-SEQ-MM; it is advised to use best-in-class enzymes in these experiments and regularly review supplier offerings.
22. For efficient size selection of <100 bp amplicons use AxyPrep Mag PCR beads. For >100 bp amplicons use Agencourt AMPure XP beads.
23. The reaction can be scaled down to 10µl with no impact to data, but do not reduce RNA mass in the reaction. This step is not suited to multiplexing but is suited to parallel sample processing to increase throughput. When performing the reaction on multiple samples in parallel, equilibrate RNA concentration across samples by adjusting RNA sample volume with nuclease-free water to match the lowest concentration in your sample set.

24. A master mix can be generated to perform multiple sample reactions in parallel, by scaling components appropriately; if doing so, always adjust volumes up by 15% to avoid any liquid loss due to evaporation and use separate tips for each dispensing step.
25. 4°C centrifugation is not necessary, but preferred.
26. A 4°C final hold step is useful for maintaining sample integrity.
27. Centrifugation separates aqueous from solvent phase, with the RNA in the aqueous phase. If the interface is disturbed, simply repeat the centrifugation step.
28. i.e. if 0.1 ml was transferred at step 12, add 0.2 ml of 100% v/v EtOH).
29. If no dry ice is available, place tubes directly onto the shelf of a -80°C freezer.
30. A brief centrifugation step of tube 3 at 10,000 xg for 1 min at 4°C ahead of the transfer is advised to ensure maximal material collection and transfer.
31. This step is not suited to the preparation and use of master mixes due to the substantial cross-contamination risk: Single pipetting step dispensing and use of fresh tips per dispension is strongly advised.
32. A master mix can be generated to perform multiple sample reactions in parallel, by scaling components appropriately; if doing so, always adjust volumes up by 15% to avoid any liquid loss due to evaporation, and use separate tips for dispensing. If using ThermoFisher's SuperScript III reverse transcriptase prepare a master mix consisting of 4µl 5X First strand buffer, 1µl of 0.1 mM DTT, 0.2 µl RNase Inhibitor, 1µl of SuperScript III reverse transcriptase enzyme, and 0.8µl of nuclease-free water (7µl final volume).
33. The reaction volume can be scaled down for economy provided the product is forward compatible with the required library handling steps applicable to your sequencing platform of choice. It is recommended that you briefly centrifuge tube 4 at 10,000 xg 4°C for 1 min prior to use.

34. A touchdown protocol can be used in the case of non-specific amplification, or due to low primer melting temperatures compared to 5' RLM RACE kit provider recommendations due to e.g. high A/T content on the target of interest. A reliable starting touchdown cycling protocol involves 98°C for 30 sec; 20 cycles of i) 95°C for 15 sec, ii) 70°C for 10 sec reduced by 0.5°C per cycle, and iii) 72°C for 15 sec; 20 cycles of i) 95°C for 15 sec, ii) 60°C for 10 sec, and iii) 72°C for 15 sec; 72°C for 2 min; 4°C hold.
35. If performing nested PCR, scale down the first reaction to 15µl and use the New England Biolabs Rapid PCR Cleanup Enzyme Set to degrade residual primers and dephosphorylate dNTPs (per 5µl of first PCR amplicon add 1µl of exonuclease I and 1µl of shrimp alkaline phosphatase included in this kit; using a thermal cycler, heat to 37°C for 5 min and inactivate enzymes at 80°C for 10 min). After cleanup of the first reaction, transfer 4µl to a new nuclease free PCR microcentrifuge tube supplemented with materials as per step 1 and 2 in this protocol to 40µl total volume (use nested PCR primers instead of GeneRacer Forward and GSP), and use the same cycling conditions (assumes nested PCR primer melting temperatures of comparable (+/- 2.0°C melting temperature).
36. Adjust agarose content to required concentration for anticipated amplicon size.
37. Alternatives to agarose electrophoresis include the Perkin Elmer Labchip XT platform (set extraction window to +/- 10% amplicon size and proceed to cleanup product by ethanol precipitation or PCR cleanup columns) or the ThermoFisher Size Select E-gel system.
38. Adjust voltage to low melting point agarose tolerance levels to avoid gel melting in the tank.
39. The nucleic acid precipitation and purification steps during 5' RLM RACE have been proven in our experience, necessary, albeit insufficient for adequate depletion of excess adapter. This results in RNA adapter carryover into the sequencing adapter ligation step.

40. Size selection approaches that remove small amplicons such as gel excision, bead-based size selection, etc. can significantly reduce data due to adapter ligation amplification. The efficiency of this process depends on the size difference between the adapter ligation by-products and the 5' RLM-RACE amplicon. It is therefore important to aim for a library size that allows for efficiency when a bead cleanup step is implemented in removing adapter and primer dimers. In our hands, we aim for amplicons in excess of 150 bp allow for efficient separation using appropriate cutoff range paramagnetic beads. Remainder contaminating adapter sequences can be easily eliminated from the dataset computationally.

41. Running the Cutadapt trimming software on your fastq dataset independently and using the verbose option will give you a very detailed initial summary statistics report. Based on that report you can further adjust Cutadapt's huge array of options to filter, quality control and process the raw sequencing reads based on various parameters to reflect your experiment.

42. Based on the sequencing technology used the length of the sequencing reads may vary, thus setting the minimum and maximum length threshold at a specific cutoff level with the *-m* and -M parameters respectively as an initial filtering step based on your experiment design. In conjunction with the cut-off length parameters you can choose the *--too-short-output=FILE* and *--too-long-output=FILE* options to output the filtered reads individually for further investigation and troubleshooting. Another useful filtering parameter is the *--untrimmed-output=FILE* which will output all the reads that do not contain the RACE adapter and can be used in combination with --discard-untrimmed.

43. Looking in the filtered reads output files can give you valuable insights in the sequenced reads and will help you troubleshoot any mistakes that might have occurred during the RACE adapter hybridization, library preparation, or PCR steps. To achieve all this initial filtering steps, it is

advised to run the Cutadapt trimming step prior to running your dataset through the RACE-SEQ(-MM) pipeline.

44. The RACE-SEQ pipeline has been built with the option to skip the initial adapter trimming step which has been pre-configured to run with the minimum default options. After manually generating an optimally filtered and quality controlled trimmed dataset based on your experiment specifications you can simply run it through the RACE-SEQ pipeline. This way you will generate the final outputs optimally, saving on alignment steps, which is arguably the most computationally intensive procedure.

45. Bowtie v.2.0.0 or later is not suited to RACE-SEQ due to its mismatch-tolerating read alignment methodology around alignment seed length (see Theotokis *et al.* 2017).

46. Most importantly, the RACE-Seq-MM pipeline introduces the *–iterate* option. Using this option, the script will generate an index from the reference sequence with all the possible nucleotide combinations spanning the region of interest, based on your input parameters. Then it will generate and write all the possible sequences containing the single nucleotide variations in the region of interest. Following that, it will try to perform alignment using bowtie and utilizing all the new sequences as reference iteratively, meaning the whole script will be run three time multiplied by the length of the region of interest, plus the wildtype sequence eg. (3x21)+1=64. The output graph generated is a combination plot with the cleavage incident generated by the wildtype reference sequence as a histogram superimposed by line plots of the single variations per nucleotide for every genomic position (Fig. 3).

5. References

1. P. Morgan *et al.*, Impact of a five-dimensional framework on R&D productivity at AstraZeneca. *Nat. Rev. Drug Discov.* **17**, 167–181 (2018).
2. D. Cook *et al.*, Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nat. Rev. Drug Discov.* **13**, 419–31 (2014).

3. S. A. Moschos, L. Usher, M. A. Lindsay, Clinical potential of oligonucleotide-based therapeutics in the respiratory system. *Pharmacol. Ther.* (2016), doi:10.1016/j.pharmthera.2016.10.009.

4. H. Denise *et al.*, Deep Sequencing Insights in Therapeutic shRNA Processing and siRNA Target Cleavage Precision. *Mol. Ther. Nucleic Acids*. **3**, e145 (2014).

5. J. Tabernero *et al.*, First-in-humans trial of an RNA interference therapeutic targeting VEGF and KSP in cancer patients with liver involvement. *Cancer Discov.* **3**, 406–17 (2013).

6. M. Barve *et al.*, Phase 1 Trial of Bi-shRNA STMN1 BIV in Refractory Cancer. *Mol. Ther.* **23**, 1123–1130 (2015).

7. H. Dudek *et al.*, Knockdown of β-catenin with Dicer-Substrate siRNAs Reduces Liver Tumor Burden In vivo. *Mol. Ther.* **22**, 92–101 (2014).

8. S. Ganesh *et al.*, Direct Pharmacological Inhibition of -Catenin by RNA Interference in Tumors of Diverse Origin. *Mol. Cancer Ther.* **15**, 2143–2154 (2016).

9. P. I. Theotokis, L. Usher, C. K. Kortschak, E. Schwalbe, S. A. Moschos, Profiling the Mismatch Tolerance of Argonaute 2 through Deep Sequencing of Sliced Polymorphic Viral RNAs. *Mol. Ther. - Nucleic Acids*. **9** (2017), doi:10.1016/j.omtn.2017.08.010.

10. K. Thys *et al.*, Performance assessment of the Illumina massively parallel sequencing platform for deep sequencing analysis of viral minority variants. *J. Virol. Methods*. **221**, 29–38 (2015).

11. S. R. Kennedy *et al.*, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*. **9**, 2586–2606 (2014).

Figure Captions

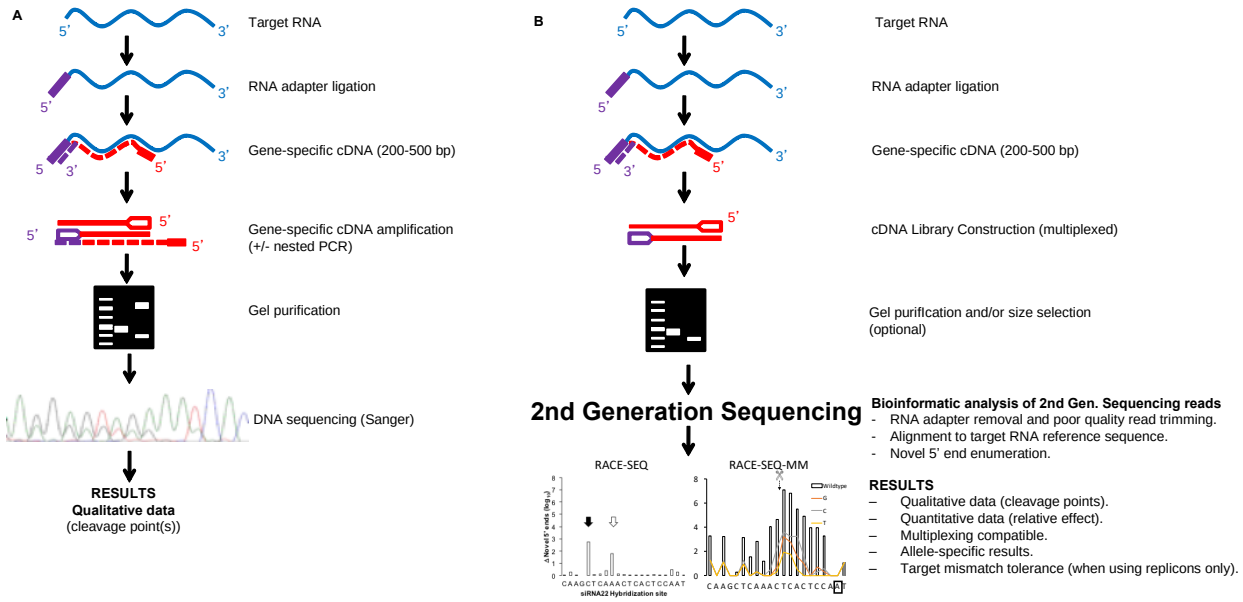

Figure 1. Comparative features of 5’ RLM-RACE and RACE-SEQ, or RACE-SEQ-MM. The gold-standard method for defining novel 5’ ends in target RNAs is a manual labor-intensive qualitative approach that can deal with single samples at a time, returning sometimes inconclusive data (A) such as mixed Sanger traces. RACE-SEQ on the other hand (B) can be optimized to parallel process multiple samples generating a wealth of quantitative and qualitative data, including the anticipated impact of unknown target polymorphisms on drug efficacy (RACE-SEQ-MM).

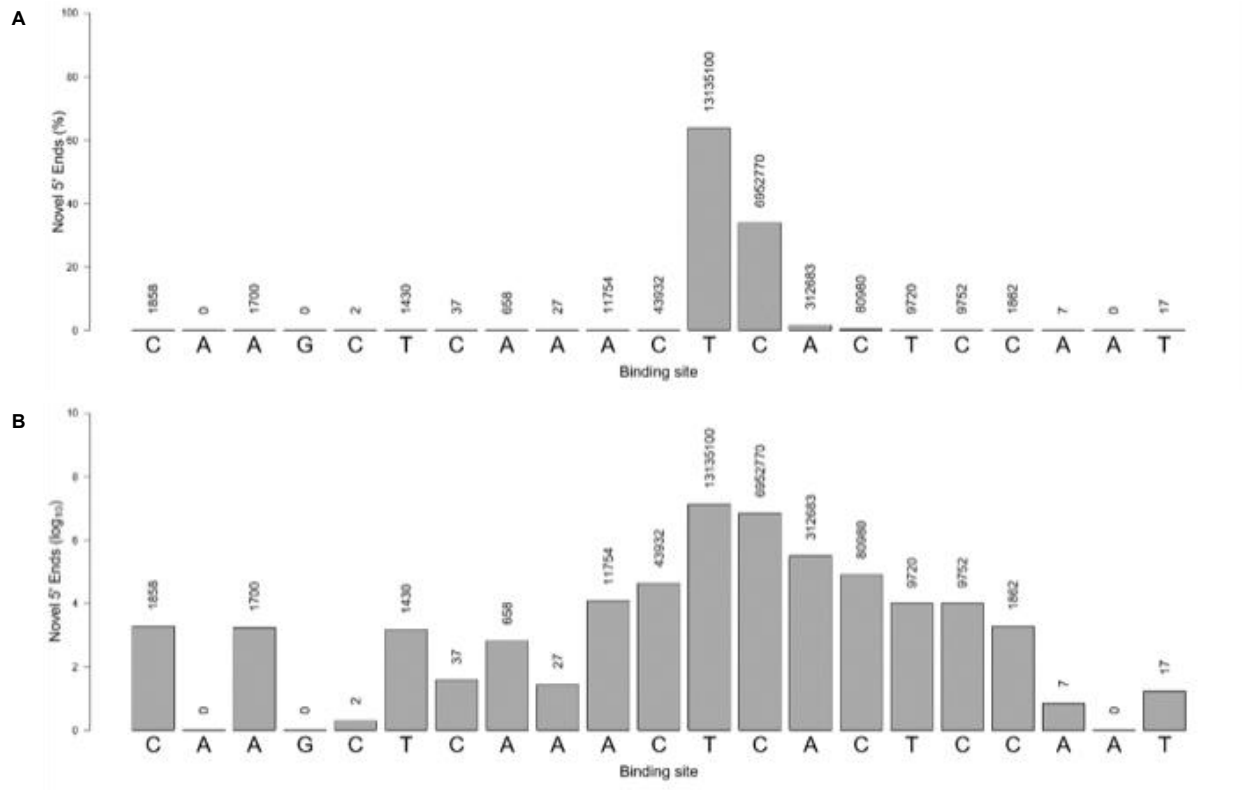

Figure 2. An example of the graph generated using the -p option while running RACE-SEQ. The two graphs representing the novel 5' cleavage incidents recorded on the same dataset between the specified genomic locations, in a linear (a) and a logarithmic (b) scale. Absolute counts of novel 5’ ends at each position are also returned above each barThe logarithmic scale unveils important details about the cleavage products that otherwise would have been overlooked due to the huge number of reads

generated during RACE-SEQ. Users are strongly advised to favour logarithmic data representation in line with data normal distribution principles.

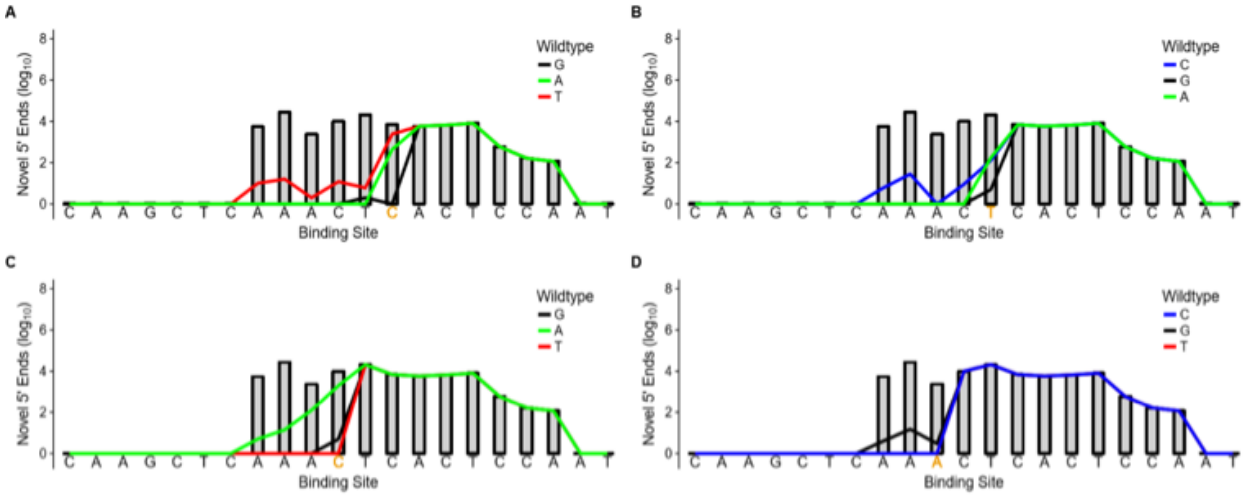

Figure 3. A combined graph generated using the *--iterate* parameter to generate RACE-SEQ-MM data. The bar plot depicts the cleavage incidents generated on the wildtype reference genome of a replicon targeted by a short interfering RNA. The three line plots are generated by substituting the nucleotide in the specified position on the drug target site reference genome (marked as a gold nucleotide on the X axis) with each of the 3 alternative ones, and aligning against the “mutated” reference sequence; these lines represent novel 5’ ends identified on replicon quasi-species genomes. This analytical approach is valid only when applied on highly mutable target RNAs, such as viral RNAs or *in vitro* replicon RNA virus systems encoding the target site of interest, and strongly depends on high read quality.