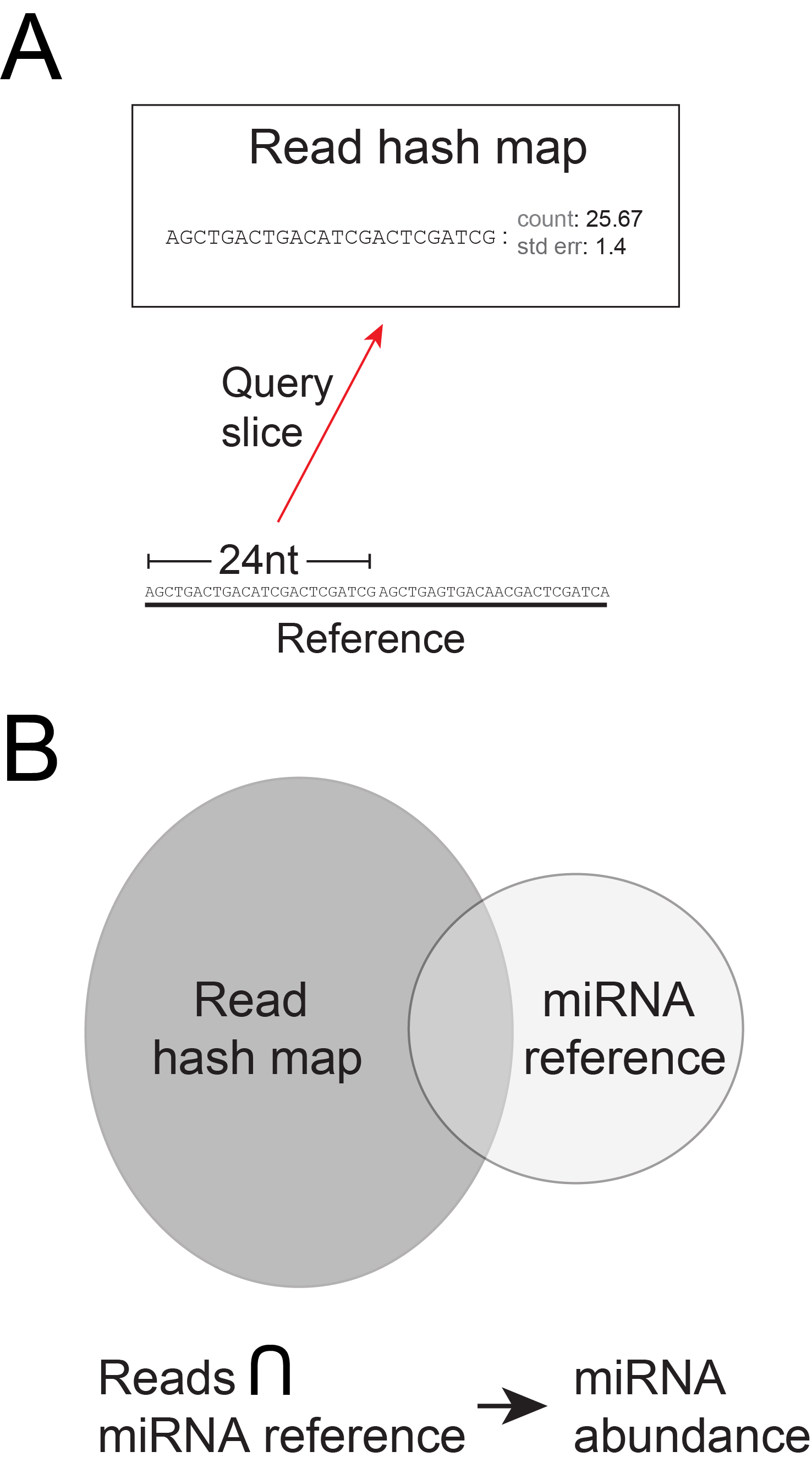
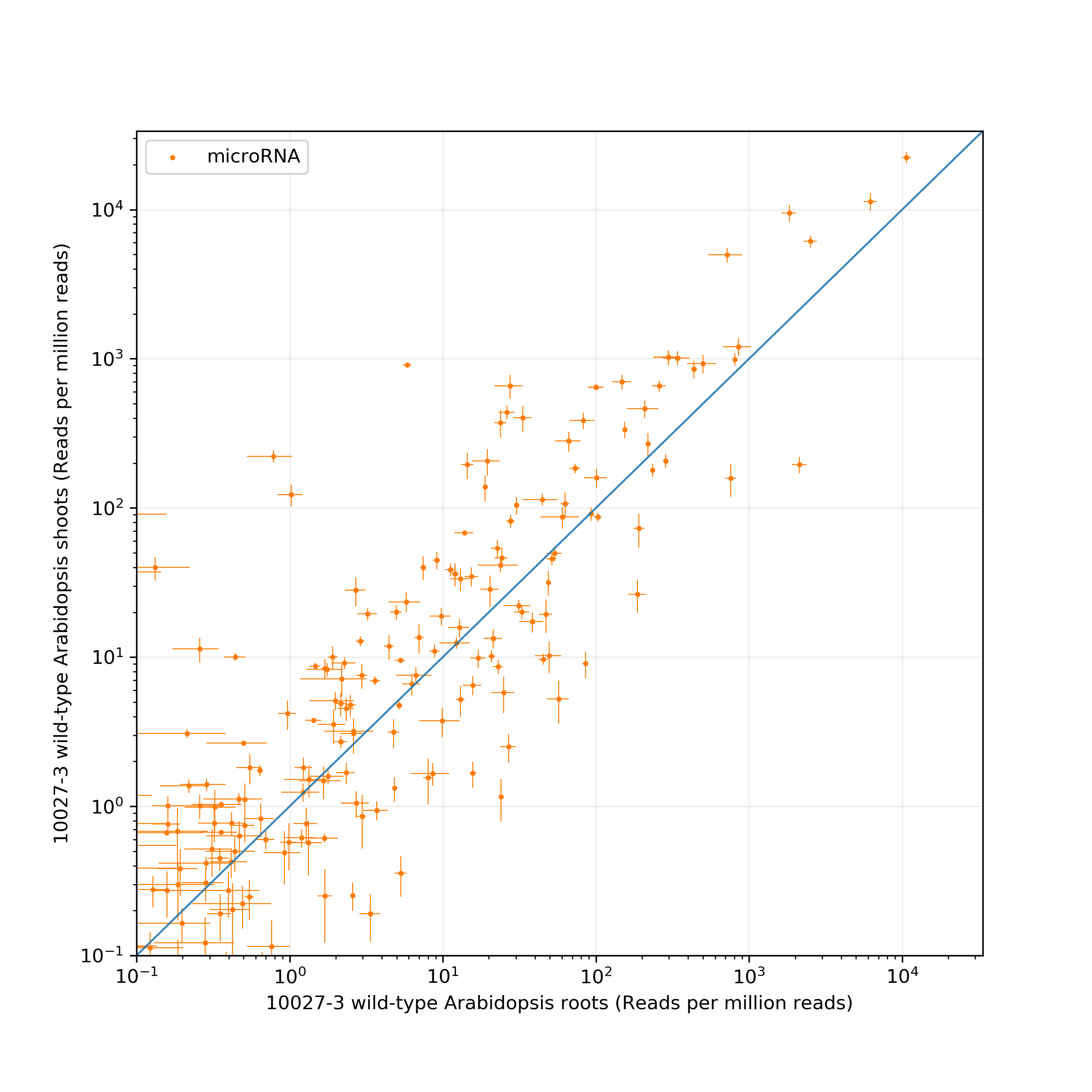
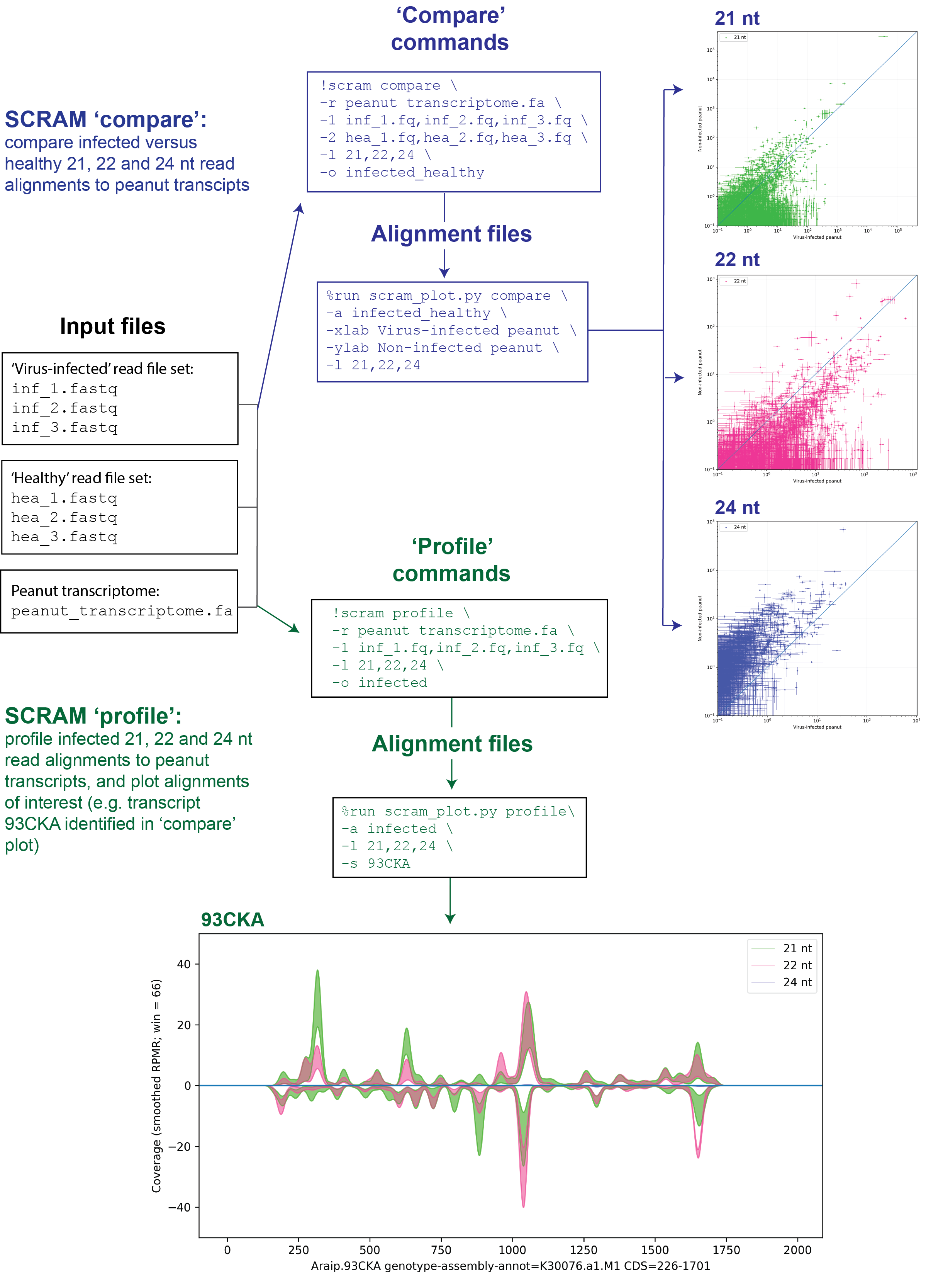
****

**Figure S1: SCRAM maps short reads with no mismatches. A.** Reads are exactly mapped to longer references (reference length >= read length) using an incrementing query slice of a set nucleotide (nt) length to interrogate a hash map. **B.** The intersection of the read hash map set and the miRNA reference set reveals the abundance of any mature miRNAs present (reference length = read length). The read hash map is generated by calculating the mean and standard error for all unique reads from a set of biological replicates (subject to parameters such as minimum and maximum read lengths, and minimum raw read count).

****

**Figure S2: A ‘compare’ plot showing microRNA abundance for root and shoot tissues of one-week-old *Arabidopsis* 10027-3 wild-type seedlings.** The *x* and *y* values for each point represent the respective mean microRNA-matching read counts for the two tissues, with standard error bars indicating the variance among these replicate reads. The plot is shown as generated, with no additional manipulation. Experimental conditions and input read data for the figure are described in Taochy et al. (2017). Mature *Arabidopsis* miRNA reference sequences were sourced from miRBase (<http://www.mirbase.org/>).



**Figure S3: An example workflow using the SCRAM pipeline.** Two sets of small RNA read files (TSWV-infected and healthy) and a reference peanut transcriptome serve as inputs. ‘Compare’ plots demonstrate a divergent impact of virus infection on the abundance of transcript-aligned 21, 22 and 24 nt siRNAs. Selected transcripts of interest can be further investigated with a ‘profile’ alignment and plot, which in this example demonstrates 21 and 22 nt, but not 24 nt, siRNA hotspots span the 93CKA transcript in virus-infected plants, suggesting it may be a target for degradation via RNAi. Experimental conditions and input read data for the figure are described in Fletcher et al. (2016).

**Table S1: SCRAM ‘compare’ and ‘profile’ data fields written to file in CSV format.** A ‘compare’ alignment outputs combined alignments to a single reference sequence per line, with five data fields. A ‘profile’ alignment shows a single discrete read alignment per line, with eight data fields.

|  |  |
| --- | --- |
| **Alignment type** | **Generated data columns (CSV)** |
| Compare | 1. Reference header 2. Mean aligned count (biological replicate set 1) 3. Standard error of the mean (biological replicate set 1) 4. Mean aligned count (biological replicate set 2) 5. Standard error of the mean (biological replicate set 2) |
| Profile | 1. Reference header 2. Reference sequence length (nt) 3. Read sequence 4. Alignment position (+ strand: 5’ ref. to 5’ read, - strand: 5’ ref to 3’ read) 5. Alignment strand 6. Mean count of the aligned read (biological replicate set 1) 7. Standard error of the mean of the aligned read (biological replicate set 1) 8. The number of times the read has aligned to all sequences in the reference file |

**Table S2: Example SCRAM alignment time benchmarks, indicating many small RNA analyses can be rapidly carried out on desktop PCs.** Alignments were carried out on a Desktop PC with 3.5GHz Xeon E5-1650 processor, 32GB RAM, 7200RPM HDD, Ubuntu Gnome 17.04 Linux OS.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Alignment type** | **Read files** | **Mean read count per file**  **(reads)** | **Reference File** | **Read lengths aligned (nt)** | **Number of reference sequences** | **Size of reference sequences**  **(nt)** | **Peak RAM usage**  **(GB)** | **Time to alignment completion**  **(seconds)** |
| Compare | 2 sets of 3 collapsed FASTA files | 8,911,819 | Arabidopsis transposable elements | 21  22  23  24 | 31,189 | 23,315,940 | 2.2 | 28 |
| Compare | 2 sets of 3 compressed FASTQ.gz files | 29,607,690 | Tomato transcriptome | 21  22  23  24 | 34,761 | 53,854,797 | 9.1 | 131 |
| Profile | 3 compressed FASTQ.gz files | 29,269,922 | Tomato transcriptome | 21  22  23  24 | 34,761 | 53,854,797 | 7.2 | 72 |
| Profile | 3 compressed FASTQ.gz files | 29,269,922 | Tomato genome | 21  22  23  24 | 13 | 823,944,041 | 13.1 | 321 |

\* Though their outputs are not directly comparable, Bowtie aligns a single uncompressed FASTQ file from the set of three read files listed to the tomato transcriptome reference in 98 seconds (71 seconds indexing and 27 seconds alignment; -v 0 and --threads 12 options).

**Table S3: Comparison of features of the SCRAM and Bowtie aligners.** Bowtie was selected for comparison as it is commonly used in many publications for siRNA read alignment1,2,3,4.

|  |  |  |
| --- | --- | --- |
| **Aligner Features** | **SCRAM** | **Bowtie** |
| Indexing of reference sequences required prior to alignment | **No** | **Yes** |
| Optional on-the-fly 3’ adapter trimming | **Yes** | **No** |
| Optional normalisation by library size, excluding reads above and below set lengths | **Yes** | **No** |
| Optional exclusion of reads below a set abundance from alignment and normalisation | **Yes** | **No** |
| Discrete read length alignments separately reported | **Yes** | **No** |
| Guaranteed optimal alignment of each read | **Yes** | **Yes** 1 |
| All valid alignments reported | **Yes** | **Yes** 1 |
| Mismatches allowed | **No** | **Yes** 2 |
| Optional output to SAM format | **No** | **Yes** |
| Optional output of alignment counts per reference sequence (‘compare’ alignment), rather than on a position by position basis (‘profile’ alignment) | **Yes** | **No** |
| Optional output of mean and standard error of aligned read counts from biological replicates, or individual aligned read counts for each input file | **Yes** | **No** |
| Supplied with an integrated plotting package | **Yes** | **No** |

1Using the –a option, all valid alignments can be reported by Bowtie, however there are significant impacts on performance and disk usage when not using collapsed FASTA files as inputs (<http://bowtie-bio.sourceforge.net/manual.shtml>).

2When using the Bowtie aligner in –v mode, the number of allowable mismatches can be set. The SCRAM aligner does not allow for mismatches (as siRNAs are derived from their target RNA, mismatches should not be present; miRNAs should exactly match their reference counterpart for identification).

**References**

1Niu, X. *et al.* (2017) Using Small RNA-seq Data to Detect siRNA Duplexes Induced by Plant Viruses. *Genes*. 6, 163

2Shamandi, N. *et al* (2015) Plants Encode a General siRNA Suppressor That Is Induced and Suppressed by Viruses. *PLOS Biol*. 13, e1002326

3Shi, B. et al (2015) Identification and regulation of host genes related to Rice stripe virus symptom production*. New Phytol*. 209, 1106–1119

4Li, Y. et al (2016) Induction and suppression of antiviral RNA interference by influenza A virus in mammalian cells. *Nat Microbiol*. 5, 16250