Supplementary Figures, Tables, and Text

TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the *S. pombe* siRNA pathway

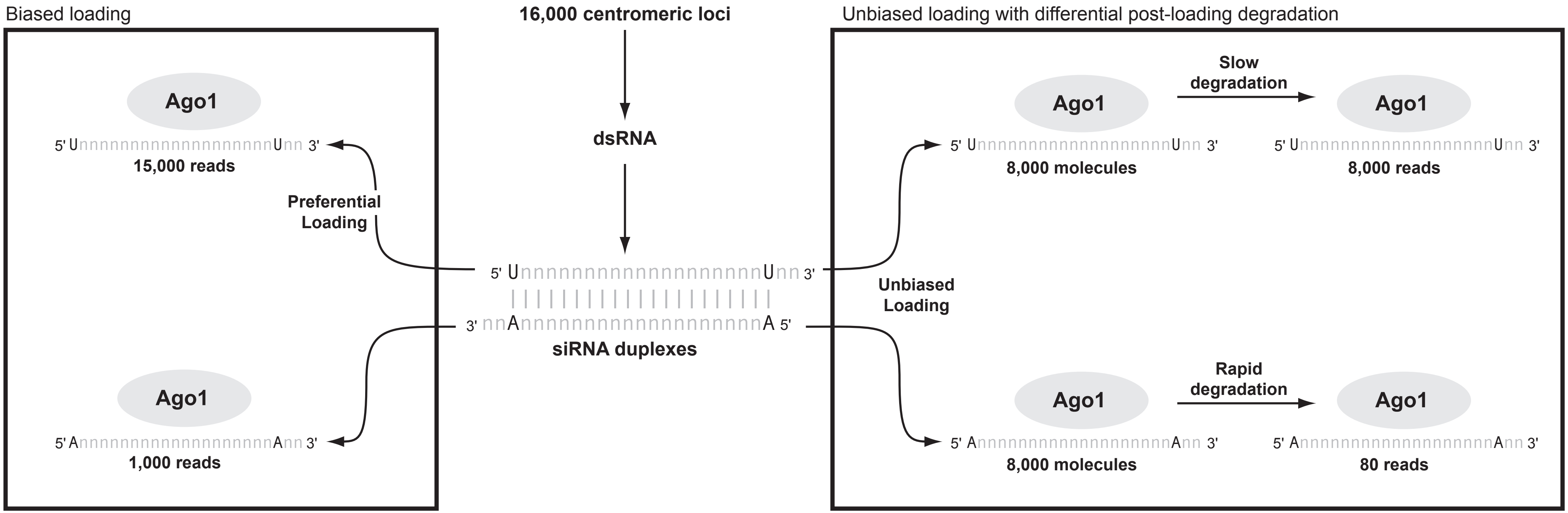
Marc Bühler[1,3,4], Noah Spies[2,4], David P. Bartel[2] and Danesh Moazed[1]

[1] Department of Cell Biology, 240 Longwood Avenue, Harvard Medical School, Boston, MA 02115 USA

[2] Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, and Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142 USA

[3] present address: Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland

[4] these authors contributed equally to this work



Supplementary Figure 1.

dinuc at 23-24	22mer preferred	23mer preferred	Ratio
AA	856	540	1.585
AC	392	278	1.410
AG	316	348	0.908
AT	718	666	1.078
CA	604	410	1.473
CC	177	178	0.994
CG	184	151	1.219
CT	367	257	1.428
GA	253	464	0.545
GC	141	344	0.410
GG	64	293	0.218
GT	214	525	0.408
TA	515	234	2.201
TC	475	325	1.462
TG	542	334	1.623
TT	805	664	1.212

Supplementary Table 1. There is an up to 2-fold preference for 22nt species over 23nt species when the base at position 23 is a U. For each locus where both the 22nt and 23nt sequences were present with identical 5' ends, the species with the greater number of normalized reads was indicated as preferred, and these counts were recorded according to the dinucleotide at position 22–23 from the common 5' end.

3 terminal nucs (2 nearest neighbors)			
Sample:	cid14	cid14	cid14
Size:	22	23	24
Total sames:	718	581	212
Total opposites:	781	560	133
Chi-sq p-value:	0.104	0.534	0.000
Sample:	wt	wt	wt
Size:	22	23	24
Total sames:	1638	1611	714
Total opposites:	1772	1599	751
Chi-sq p-value:	0.022	0.832	0.334

Supplementary Table 2. No evidence for strand preference based on terminal stability. Terminal three base pairs (2 nearest neighbors) were analyzed using a nearest neighbors algorithm and the 5' end with the highest stability was counted as same if it had more reads than the inferred duplex, or opposite if the opposite strand had more reads. Duplexes with identical stability at both ends were ignored. Because of the strong 5' U bias, only those duplexes with the same 5' nucleotide on both strands were counted.

Supplementary Table 3. Genes with highest number of antisense reads (sorted by number of reads antisense to gene in wild-type). The number of reads and sequences is counting only those reads that map uniquely to the genome, with the exception of those mapping to tlh1 and tlh2. Most gene-matching reads either map to several genomic loci (not shown here) or map to the sense strand of the gene and are presumably degradation products.

Wild-type				cid14					
Antisense		Sense		Antisense		Sense			
Reads	Seqs	Reads	Seqs	Reads	Seqs	Reads	Seqs	Systematic Name	Gene product
3,124*	2,172	1,196*	1,133	413*	597	122*	213	SPBCPT2R1.08c	RecQ type DNA helicase Tlh1
3,123*	2,171	1,228*	1,138	412*	596	121*	212	SPAC212.11	RecQ type DNA helicase
54	46	53	36	16	13	67	60	SPCC13B11.01	alcohol dehydrogenase Adh1
43	34	206	108	4	4	34	28	SPCC330.05c	orotidine 5'-phosphate decarboxylase Ura4
33	23	0	0	38	29	0	0	SPAC27E2.13	dubious
8	6	0	0	0	0	0	0	SPBC3D6.16	sequence orphan
7	7	0	0	4	4	5	5	SPBC317.01	MADS-box transcription factor Pvg4
7	5	0	0	5	2	0	0	SPBC725.06c	serine/threonine protein kinase Ppk31
6	4	0	0	2	1	0	0	SPAC212.06c	pseudogene/pseudo
6	6	0	0	2	2	0	0	SPAC22F3.04	AMP binding enzyme
4	4	26	22	0	0	25	24	SPAPB8E5.03	malic acid transport protein Mae1
4	4	0	0	7	5	0	0	SPCC285.14	TRAPP complex subunit Trs130
4	4	0	0	4	3	0	0	SPBC1861.06c	S. pombe specific UPF0300 family protein 4
4	4	0	0	1	1	0	0	SPAC25H1.09	alpha-amylase homolog Mde5
4	3	1	1	6	4	2	2	SPBC776.10c	Golgi transport complex peripheral subunit
4	3	2	1	1	1	1	1	SPBC1215.02c	NatB N-acetyltransferase complex non catalyticsubunit Arm1
4	1	0	0	0	0	1	1	SPAC25H1.05	sequence orphan
4	1	0	0	0	0	1	1	SPCC16C4.02c	DUF1941 family protein
3	3	0	0	5	5	0	0	SPBC19G7.01c	MutS protein homolog 2
3	3	1	1	3	3	4	3	SPAC22F8.11	phosphoinositide phospholipase C Plc1
3	3	1	1	1	1	6	5	SPBC947.02	AP-1 adaptor complex subunit Apl2
3	3	5	4	1	1	9	7	SPBC418.01c	imidazoleglycerol-phosphate synthase
3	3	3	3	0	0	2	2	SPAC10F6.02c	ATP-dependent RNA helicase Prp22
3	3	0	0	0	0	1	1	SPAC11H11.04	pheromone p-factor receptor
3	3	0	0	0	0	2	2	SPAC167.03c	U4/U6 x U5 tri-snRNP complex subunit Snu66
3	3	0	0	0	0	0	0	SPAC3H8.09c	poly(A) binding protein Nab3
3	3	0	0	0	0	0	0	SPAC6C3.07	sequence orphan
3	3	0	0	0	0	0	0	SPAC4F8.08	sequence orphan
3	3	0	0	0	0	0	0	SPBC3D6.10	AP-endonuclease Apn2
3	2	5	5	6	5	2	2	SPAC57A7.05	conserved protein (fungal and plant)
3	2	0	0	1	1	0	0	SPCC126.01c	conserved fungal protein
3	2	0	0	1	1	0	0	SPBC36B7.07	SNARE Tgl1
3	2	3	3	0	0	2	1	SPBC1539.07c	glutathione-dependent formaldehyde dehydrogenase
3	1	7	6	6	3	4	4	SPAC23C11.06c	hydrolase
3	1	0	0	1	1	0	0	SPBC32H8.06	TPR repeat protein, meiotically spliced
3	1	2	2	1	1	0	0	SPCC126.02c	Ku domain protein Pku70
3	1	1	1	0	0	2	2	SPBC21.05c	Ras guanyl-nucleotide exchange factor Ral2
3	1	2	2	0	0	0	0	SPBC776.06c	spindle pole body interacting protein
3	1	0	0	0	0	2	1	SPAC5H10.09c	3-methyl-2-oxobutanoatehydroxymethyltransferase
3	1	0	0	0	0	1	1	SPAC3C7.03c	RecA family ATPase Rhp55
3	1	1	1	0	0	0	0	SPBP8B7.28c	sequence orphan

* Nearly all of these reads map to the genome exactly twice: once to tlh1 and once to tlh2.

Table S4. List of strains used in this study.

Strain	Genotype
SPY137	SPY137 *h$^+$ leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4$^+$ oriA*
SPY1220	*h$^+$ leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4$^+$ oriA cid14Δ::nat^R*
SPY815	*h$^+$ leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4$^+$ oriA clr4Δ::kan^R*
SPY1220	*h$^+$ leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4$^+$ oriA cid14Δ::nat^R*
SPY28	*h$^+$ leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4$^+$ oriI*
SPY86	*h$^+$ leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4$^+$ oriI dcr1Δ::TAP-kan^R*
SPY87	*h$^+$ leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4$^+$ oriI rdp1Δ::TAP-kan^R*
SPY399	*h$^+$ leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4$^+$ oriI clr4Δ::nat^R*
SPY787	*h$^+$ leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4$^+$ oriI cid14Δ::nat^R*
SPY139	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4$^+$*
SPY1313	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4$^+$ rrp6Δ:: nat^R*
SPY1408	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4$^+$ mtr4-1(mtr4$^+$::TAP-nat^R)*
SPY1284	*h$^-$ leu1Δ ura4Δ dis3-54*
SPB45	*h? cid14Δ:: nat^R rdp1Δ:: kan^R ura4+::5BoxB/HPH leu1-32*
SPB46	*h? cid14Δ:: nat^R dcr1Δ:: kan^R ura4+::5BoxB/HPH leu1-32*

Table S5. List of oligonucleotides used in this study.

Name	Sequence
mb86	5’-AACCCTCAGCTTTGGGTCTT-3’
mb87	5’-TTTGCATACGATCGGCAATA-3’
mb90	5’-CAACCCTCAGCTTTGGGTCTTG-3’
mb91	5’-TCCTTTTGCATACGATCGGCAATAC-3’
mb510	5’-AAAATGTTTCTATGCTACTTTAACAATTCGCACAAAG-3’
mb511	5’-AAAGTGCACGCTCTAATTTTAATTTTAACAGTCTATAAAGTTTAG-3’
mb512	5’-CCTAGCCACTGGACCATGACGGA-3’
mb521	5’-AGGCCAGCTACGCTACTC-3’
mb522	5’-CGACTTACTATTAAGCATTGATTGCAAATTACATTTTG-3’
mb523	5’-AAATAGTGTCTGAACAATAATCATAAAACTTTCTATGCTAAC-3’
mb524	5’-CATAGTATCTTAGAAAAATGTGAAAAGTGTTAGTTTACTATTCTC-3’
mb525	5’-TTAAGCATAATAAAAAGATTCTTTGAAAGTGGAAGAAATCATG-3’
mb526	5’-CACTAAAAATTTGAGAAAATAATAAAACGTGTCAAGCTCTTTC-3’
mb527	5’-TTAAACGTAACCGATACATAATTTAGGCAAAAATTGTTG-3’
mb528	5’-GTTCATCTAAAAGCTTCAAAAAATATTAATATTGAGTCTAAAATCAAGT-3’
rsi1	5’-CAAGTTTGTCCAACTTCTCGGCA-3’
rsi2	5’-AGCCAATCCAGAGGCCTCACTAA-3’
rsi3	5’-TAATGATCCTTCCGCAGGTTCACC-3’
rsi4	5’-AGGTAGTGGTATTTCACCGGCGTA-3’
rsi5	5’-AAGCCAATCCAGAGGCCTCACTAA-3’
rsi6	5’-GGCGAGAAAAGACATCGGTCCAC-3’
rsi7	5’-ATTTTTTGCCTACCAACAAGA-3’
rsi8	5’-GACCAGTAAACACGCCTTGCG-3’
rsi9	5’-CCAAGTTTGTCCAACTTCTCGGCA-3’
rsi10	5’-ACCAGTAAACACGCCTTGCG-3’
rsi11	5’-GGTATTGTAAGCAGTAGAGTA-3’
rsi12	5’-CAATGGTAATTCAACTTAGTA-3’
rsi13	5’-CAGAATTCGGTAAGCGTTGGA-3’
rsi14	5’-GCAATGGTAATTCAACTTAGTA-3’
rsi15	5’-TTGGACAAACTTGGTCATTTA-3’
rsi16	5’-GTATTGTAAGCAGTAGAGTA-3’
rsi17	5’-ACTTGTTCCTACTCTCCTGTA-3’
rsi18	5’-TTCCTACTCTCCTGTATCGTA-3’
DM566	5’-TTATTGATGGCGAAGCTAGATCCG-3’
DM567	5’-AACTCCATAACCACCACCATGCTC-3’

Supplementary Text

Strand selection

Duplexes of miRNAs and synthetic siRNAs tend to bind the loading machinery asymmetrically, such that the strand least stably paired at its 5' end is preferentially loaded as the guide strand within the silencing complex[34]. We examined whether the same was true for heterochromatic siRNAs, focusing on the inferred duplexes that match sequenced 23mers deriving from the centromeric *dg*/*dh* repeats. To ensure that the identity of the 5' nucleotides did not influence the result, we considered only the 3210 duplexes in which the two 5' nucleotides were identical (**Supplementary Table 2**).

Ago1 preferentially loads siRNAs with 5'-uracil

Since preferential siRNA processing and pairing asymmetry contribute little to the enrichment for species with a 5'-U, this strong bias must arise either from preferential loading of siRNAs with a 5'-U into Ago1, or preferential stability of species with 5'-U already in Ago1. To discern between these two possibilities, we looked at centromeric reads with 5'-U whose inferred duplex partner also has a 5'-U (U...A.. species, **Table 1**). There are approximately 8,000 centromeric loci that could form such duplexes. We compared reads from these duplexes to 5'-U reads whose inferred duplex partner has a 5' A (U...U.., **Table 1**). Because U...U.. can occur independently on either genomic strand, there are approximately twice as many such loci, and indeed, we counted approximately 16,000 centromeric occurrences of the 23mer sequence U...U.. .

For each U...A.. duplex, either the + or – strand can be loaded into Ago1, presumably with identical affinity, given the lack of pairing asymmetry and that the opposite strand also has a 5'-U. The number of reads from these duplexes would be proportional to the depth of sequencing and the number of genomic loci, assuming each locus produces dsRNA at approximately the same rate. We

observed an average of slightly less than one read per centromeric locus (7,654 reads).

Suppose Ago1 loaded siRNAs with equal efficiency, regardless of 5' nucleotide, but that loaded siRNAs beginning with G, C and A were degraded much more rapidly than those with a 5' U (**Supplementary Figure 1**, right box). In this case, we would expect half of all U...U.. duplexes to load the U...U.. strand, and half to load the paired A...A.. strand instead. Once loaded, the A...A.. species would degrade quickly, resulting in the substantial 5' nucleotide bias in our sequencing data. Given the approximately 1 read to 1 genomic locus ratio from above, and 16,000 loci, we would expect approximately 8,000 U...U.. reads. However, we observed 9,245 reads, significantly more than expected by this model.

We turn instead to a model whereby Ago1 loads the strand with 5' U preferentially compared to duplex partners with 5' G, C or A (**Supplementary Figure 1**, left box). Under this model, we expect some majority of reads from U...U../A...A.. duplexes to be loaded from the U...U.. strand. At a rate of 1 read to 1 genomic locus, this would mean observing more than 8,000 reads, which is consistent with the 9,245 reads observed. The number of U...U.. reads plus the number of A...A.. reads is not twice the number of U...A.. reads, despite the fact that there are twice as many centromeric loci. We attribute this observation to a limited number of encounters of the U...U../A...A.. duplex with the Ago1-loading machinery because some duplex molecules that are released after nonproductive encounters in the suboptimal orientation are presumably degraded before they have another opportunity to encounter the loading machinery.

Because the centromeric regions are particularly AT-rich, the number of potential duplexes with 5'-U and an inferred duplex species with a 5'-G or 5'-C is much lower than that with an inferred duplex species with a 5'-A. There are approximately 8,000 U...C.. loci, and a similar number of U...G.. loci, in the centromeric regions. The model in which siRNAs are loaded with equal efficiency and then differential post-loading degradation explains the sequencing bias predicted approximately 4,000 U...C.. reads and approximately 4,000 U...G..

reads. However, the observed number of reads from U...C.. species (7,423 reads) and U...G.. species (7,835 reads) were nearly 1 per locus, consistent with the second model, and further supporting preferential loading as a major factor in the 5'-U bias.