Searching for Circular RNAs in Archaeal RNA-Seq Data

August 28, 2011

Kamil Slowikowski, SURF-IT 2011
Advisor: Dr. Todd Lowe
Graduate Student Advisor: Andrew Uzilov

Summary

I wrote Perl scripts to analyze 36 samples of RNA-Seq data representing 12 species of archaea: *Aeropyrum pernix*, *Haloferax volcanii* DS2, *Methanocaldococcus jannaschii*, *Pyrobaculum aerophilum*, *P. arsenaticum*, *P. calidifontis*, *P. islandicum*, *P. oguniense*, *Pyrococcus furiosus*, *Staphylothermus marinus* F1, and *Thermococcus kodakaraensis*. Thousands of reads were found to wrap around when they map to the genome, providing direct evidence of RNA circularization. A read wraps around when it maps to a locus in two halves, the first half downstream of the second half. All 6 snoRNAs flagged in a previous study (see Starostina 2004 [2]) were found to have wraparound reads. In the samples from the RNA-Seq procedure done on March 26, 2011, nontemplate nucleotide addition by reverse transcriptase SuperScript III was investigated. Results suggest that SuperScript III adds one nonspecific nucleotide beyond the 5' end of the RNA molecule. In the same samples, special linkers that were designed not to ligate to each other were found to appear in tandem (2-4 consecutive copies) in a small fraction (0-9%) of reads. Also, many reads map exactly twice to their respective genomes: one part of each read maps to one particular locus and the other part maps to a distant locus. Additionally, I reviewed and rewrote Dave Bernick's script `pairParser.pl` to throw nonsilent warnings and to output statistics about the processed FASTQ read files. The rewritten script is incomplete.

This document is a brief summary of my work and findings. All of my work is documented in more detail in Markdown files located in `CircularRNA/doc` in my home directory. Those files have also been converted into wiki pages viewable here:

- http://lowelabwiki.cse.ucsc.edu/index.php/User:Kslowikowski

Wraparound Reads

I wrote a Perl script called `wraparounds.pl` to find wraparound reads (see `CircularRna/bin` in my home directory). A wraparound read is one where the read maps to the genome in two halves and the first half of the read is downstream of the second half. The script examines only those reads that map with a CIGAR (see SAM format in Li 2009 [1]) string that matches this regular expression: `/^\d+[HM]\d+[HM]$/`. That is, the mapping must be in two halves, where one half maps to the genome and the other half does not. When this is the case, the script searches for the complementary mapping within 1000 nucleotides of the original mapping: when the mapping half does not map and the nonmapping half does. With option `--spliced`, the script will output reads where the first half of the read is upstream of the second half.

I created and added precomputed tracks for all 36 samples. I have not committed any of the tracks to the CVS repository, so they should only be visible on my private genome browser. Each read is displayed on a single row in which exons are connected by a line. The arrows on the line indicate the strandedness of the read. Each exon represents a half of the read that mapped to the genome. The vast majority of reads wraparound with both halves mapping to the same strand of the genome. There is a small number of cases

where two halves of a single read map to opposite strands of the genome. In this case, each half of the read is on its own row.

Count	Sample File
423	001.Smar_tRNA-5S.wraparounds.bed
10	002.Ape_tRNA-5S.wraparounds.bed
6	003.Pae_tRNA-5S.wraparounds.bed
71	004.Pfu_tRNA-5S.wraparounds.bed
16	005.Hvol_tRNA-5S.wraparounds.bed
26	006.Mja_E_18-70_Titanium.wraparounds.bed
169	007.Mja_S_18-70_Titanium.wraparounds.bed
2460	008.Mja_5S-500.wraparounds.bed
6106	009.Pfu_5S-500.wraparounds.bed
13483	010.Hvol_5S-500.wraparounds.bed
12	011.Mja_E_18-70_TaqGold.wraparounds.bed
5	012.Mja_S_18-70_TaqGold.wraparounds.bed
21	Pae_001.wraparounds.bed
10	Pae_002.wraparounds.bed
1	Pae_003.wraparounds.bed
10	Pae_004.wraparounds.bed
4	Pae_005.wraparounds.bed
3	Pae_006.wraparounds.bed
20	Pae_007.wraparounds.bed
8	Pae_008.wraparounds.bed
6	Pae_009.wraparounds.bed
21	Pae_010.wraparounds.bed
5	Pae_011.wraparounds.bed
14	Pae_012.wraparounds.bed
4	Pae_E_3.wraparounds.bed
3	Pae_S_3.wraparounds.bed
1	Par_E_3.wraparounds.bed
1	Par_S_3.wraparounds.bed
0	Pca_E_3.wraparounds.bed
1	Pca_S_3.wraparounds.bed
694	Pfu_007_E.wraparounds.bed
247	Pfu_008_S.wraparounds.bed
3	Pis_E_3.wraparounds.bed
0	Pis_S_3.wraparounds.bed
0	Pog_E.wraparounds.bed
0	Pog_S.wraparounds.bed

Table 1: This table shows the number of wraparound reads in each RNA-Seq sample.

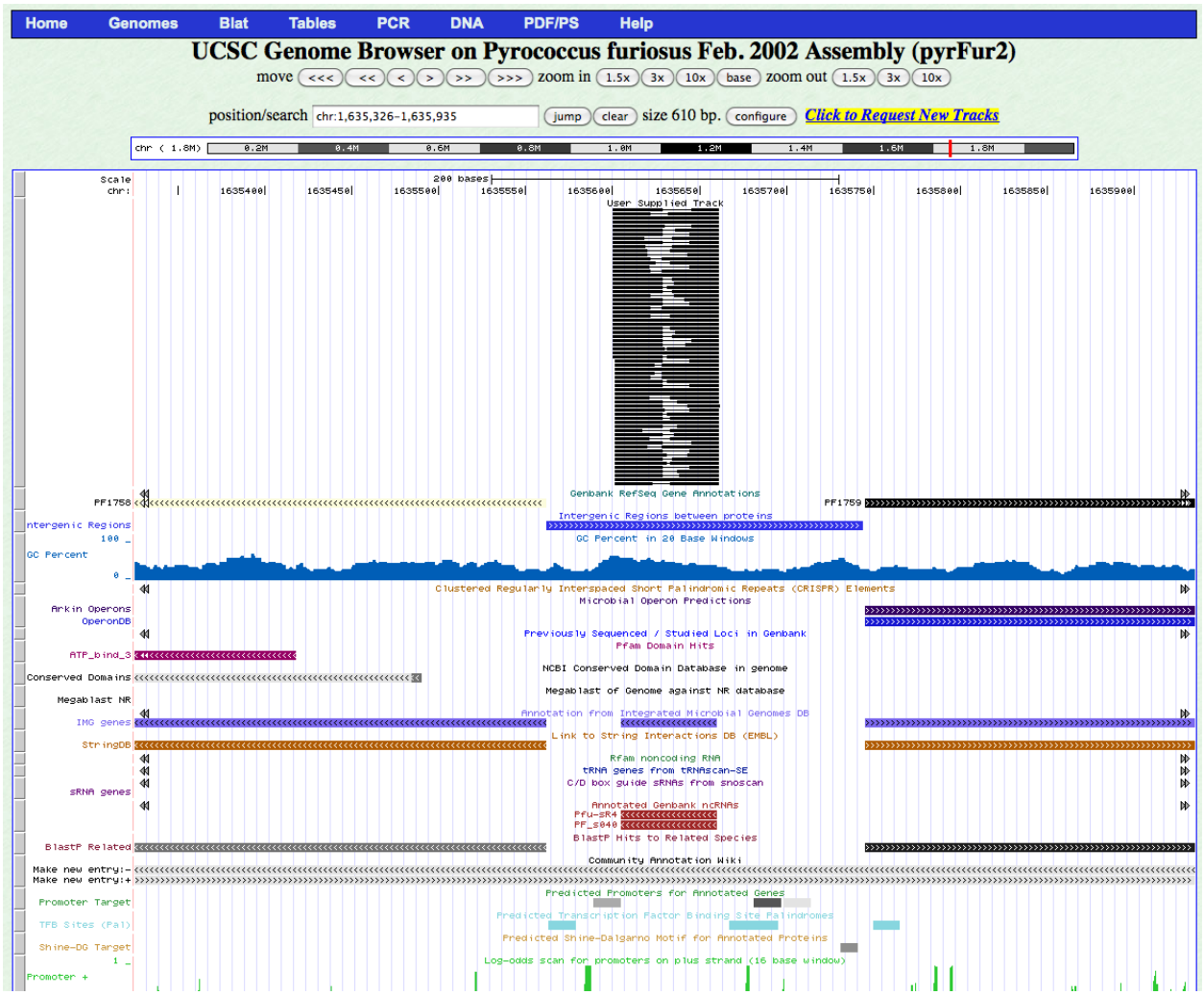

Figure 1: This is a screenshot of the genome browser showing the wraparounds (black) in *Pyrococcus furiosus* that map to snoRNA sR4 (red).

Nontemplate Nucleotide Addition by SuperScriptIII

Nontemplate addition is the addition of one or more nucleotides beyond the 5' end of the RNA molecule being reverse-transcribed. It is known that predecessors to the reverse transcriptase SuperScriptIII (MMLV, SuperScript II) perform nontemplate addition because some experimental protocols depend on the addition of nontemplate nucleotides. The commonly held belief among practicing scientists seems to be that SuperScript III does not add any nucleotides beyond the 5' end. My analysis suggests that this belief is not true.

I wrote a Perl script called `nontemplate_addition.pl` and a helper Bash script called `nontemplate_addition.sh` to examine nontemplate addition by SuperScript III in RNA-Seq samples from March 26, 2011 (see `CircularRna/bin` in my home directory). Only paired-end 1 reads that map uniquely to their respective genomes were analyzed for nontemplate addition. Their mappings were checked for CIGAR strings that match the following regular expression: `/^[1-4]H/`. That is, if a read maps to the genome and its first 1-4 nucleotides are not mapped, then it is assumed that this is a case of nontemplate addition. Further, I examined the nucleotide distribution for reads whose first nucleotide did not map. See figure 2 and figure 3.

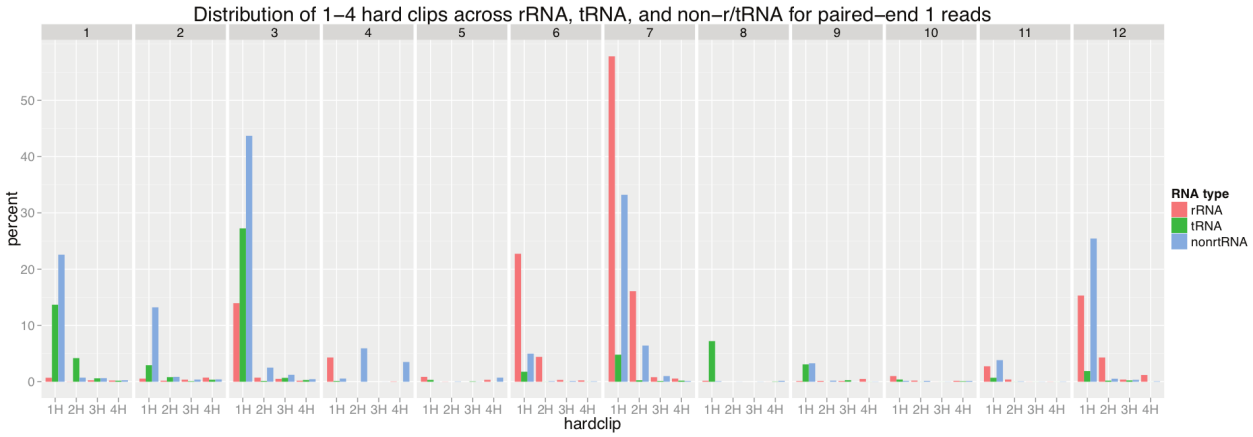

Figure 2: The percent of uniquely mapping PE1 reads that map without the first 1, 2, 3, or 4 nucleotides. The bars are split into three categories: reads that map to rRNA, tRNA, and neither rRNA nor tRNA.

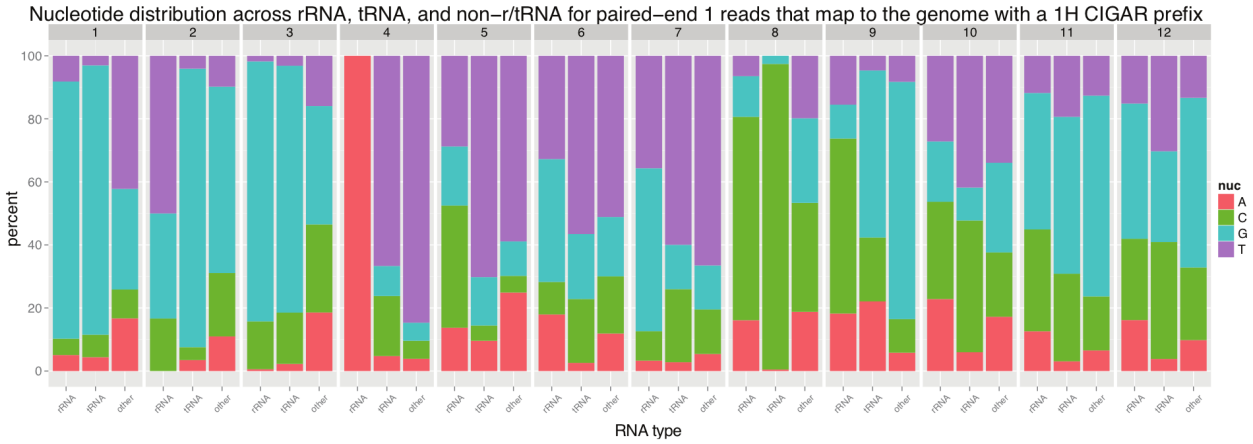

Figure 3: The nucleotide distribution for the first nucleotide for uniquely mapping PE1 reads that map without the first nucleotide. The bars are split into three categories: reads that map to rRNA, tRNA, and neither rRNA nor tRNA.

These results are not conclusive. Further work should include:

- Examination of all uniquely mapping PE1 reads that correspond to a single RNA molecule and map without the first nucleotide. Is this omitted nucleotide consistent across all of these reads?
- Examination of read ends that correspond to the 3' end of the original RNAs. Nontemplate addition happens only at the 5' end of the RNAs, so the mapping of the 3' end can serve as a control for how frequently 1, 2, 3, or 4 nucleotides are omitted from the end of a mapping.
- Examination of PE1 and PE2 reads when PE1 and PE2 reads both read through the entire original RNA. The nontemplate nucleotide(s) should be identical in PE1 and PE2.

Tandem Linkers

RNA-Seq samples from March 26, 2011 were found to contain multiple tandem copies of linker sequences. This is despite the fact that the linkers were designed not to ligate to each other. I used `grep` to search for the linker sequences (see figure 4) in the FASTA reads that are output from David Bernick's `pairParser.pl`. His script clips only one linker sequence from the start of each FASTQ read.

```
>linker1
CTGTAGGCACCATCAAT
>linker1-revcom
ATTGATGGTGCCTACAG
>linker2
CACTCGGGCACCAAGGA
>linker2-revcom
TCCTTGGTGCCCGAGTG
```

Figure 4: Linker sequences.

```
@HWI-ST611_0172:1:1:12110:2202#0/1
ACTAGCTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTAGGCACCATCAAT AGCTAGT
--
@HWI-ST611_0172:1:1:13232:2344#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG ATCCCACCCGGTCCACCACTGTAGGCACCATCAATACTATG
--
@HWI-ST611_0172:1:1:14436:2622#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG ATCCCACCGGGTCCACCACTGTAGGCACCATCAATACTATG
--
@HWI-ST611_0172:1:1:12398:3106#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTGGCACCATCAATACTATGCAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGAC
--
@HWI-ST611_0172:1:1:1332:3369#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTAGGCACCATCAATACTATGCAGATCGGAAGAGCGGTT
--
@HWI-ST611_0172:1:1:6366:3482#0/1
ACTAGCTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGCG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTAGG
--
@HWI-ST611_0172:1:1:14081:3483#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG ATGGAAACTCTGATTTTAACTACCGAAATTTTTATATATGG
--
@HWI-ST611_0172:1:1:20568:3256#0/1
GCATAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTGATCCCGGCCAGGGCGCCACTGTAG
--
@HWI-ST611_0172:1:1:18379:3768#0/1
ACTAGCTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTAGGCACCATCAATAGCTAGTAGATCGGAAGAGCGGTT
--
@HWI-ST611_0172:1:1:16689:4102#0/1
ACTAGCTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG TCCTTGGTGCCCGAGTG CTGTAGGCACCATCAATAGCTAGTAGATCGGAAGAGCGGTT
```

Figure 5: Raw RNA-Seq reads in FASTQ format. Instances of linker2-revcom are marked by inserting spaces into the nucleotide sequences.

Sample	Linker Sequence	Reads with Linker	Total Reads	Percent of Reads with Linker
001	linker1	1559	8347166	0.02
001	linker1-revcom	0	8347166	0.00
001	linker2	7	8347166	0.00
001	linker2-revcom	715508	8347166	8.57
002	linker1	35	503076	0.01
002	linker1-revcom	0	503076	0.00
002	linker2	0	503076	0.00
002	linker2-revcom	11958	503076	2.38
003	linker1	6349	3131580	0.20
003	linker1-revcom	0	3131580	0.00
003	linker2	1	3131580	0.00
003	linker2-revcom	197442	3131580	6.30
004	linker1	865	1831294	0.05
004	linker1-revcom	0	1831294	0.00
004	linker2	0	1831294	0.00
004	linker2-revcom	125787	1831294	6.87
005	linker1	1596	5473198	0.03
005	linker1-revcom	0	5473198	0.00
005	linker2	0	5473198	0.00
005	linker2-revcom	227141	5473198	4.15
006	linker1	24541	19511556	0.13
006	linker1-revcom	0	19511556	0.00
006	linker2	4	19511556	0.00
006	linker2-revcom	397090	19511556	2.04
007	linker1	381710	16142912	2.36
007	linker1-revcom	0	16142912	0.00
007	linker2	0	16142912	0.00
007	linker2-revcom	482817	16142912	2.99
008	linker1	350	1901188	0.02
008	linker1-revcom	0	1901188	0.00
008	linker2	0	1901188	0.00
008	linker2-revcom	98032	1901188	5.16
009	linker1	3	5338690	0.00
009	linker1-revcom	0	5338690	0.00
009	linker2	0	5338690	0.00
009	linker2-revcom	35018	5338690	0.66
010	linker1	7	3365288	0.00
010	linker1-revcom	0	3365288	0.00
010	linker2	0	3365288	0.00
010	linker2-revcom	64142	3365288	1.91
011	linker1	91034	16554734	0.55
011	linker1-revcom	0	16554734	0.00
011	linker2	1	16554734	0.00
011	linker2-revcom	257898	16554734	1.56
012	linker1	325935	3825142	8.52
012	linker1-revcom	0	3825142	0.00
012	linker2	0	3825142	0.00
012	linker2-revcom	51212	3825142	1.34

Table 2: Count and percent of reads that contain linker sequences. These are reads that were output by `pairParser.pl`.

Double Mappings

I wrote a Perl script called `multiple_mappings.pl` to find reads that map to the genome a specific number of times (see `CircularRna/bin` in my home directory). RNA-Seq samples from March 26, 2011 were found to contain many reads that map exactly twice. That is, in a group of reads, all of the reads in the group map to one specific locus in the genome and they also map to other places all over the genome. At least some of the time, this is because part of each read maps to the one specific locus and the remainder of each read maps elsewhere. This may be consistent with RNA splicing like trans-splicing. See position 536648 in figure 6 and figure 7.

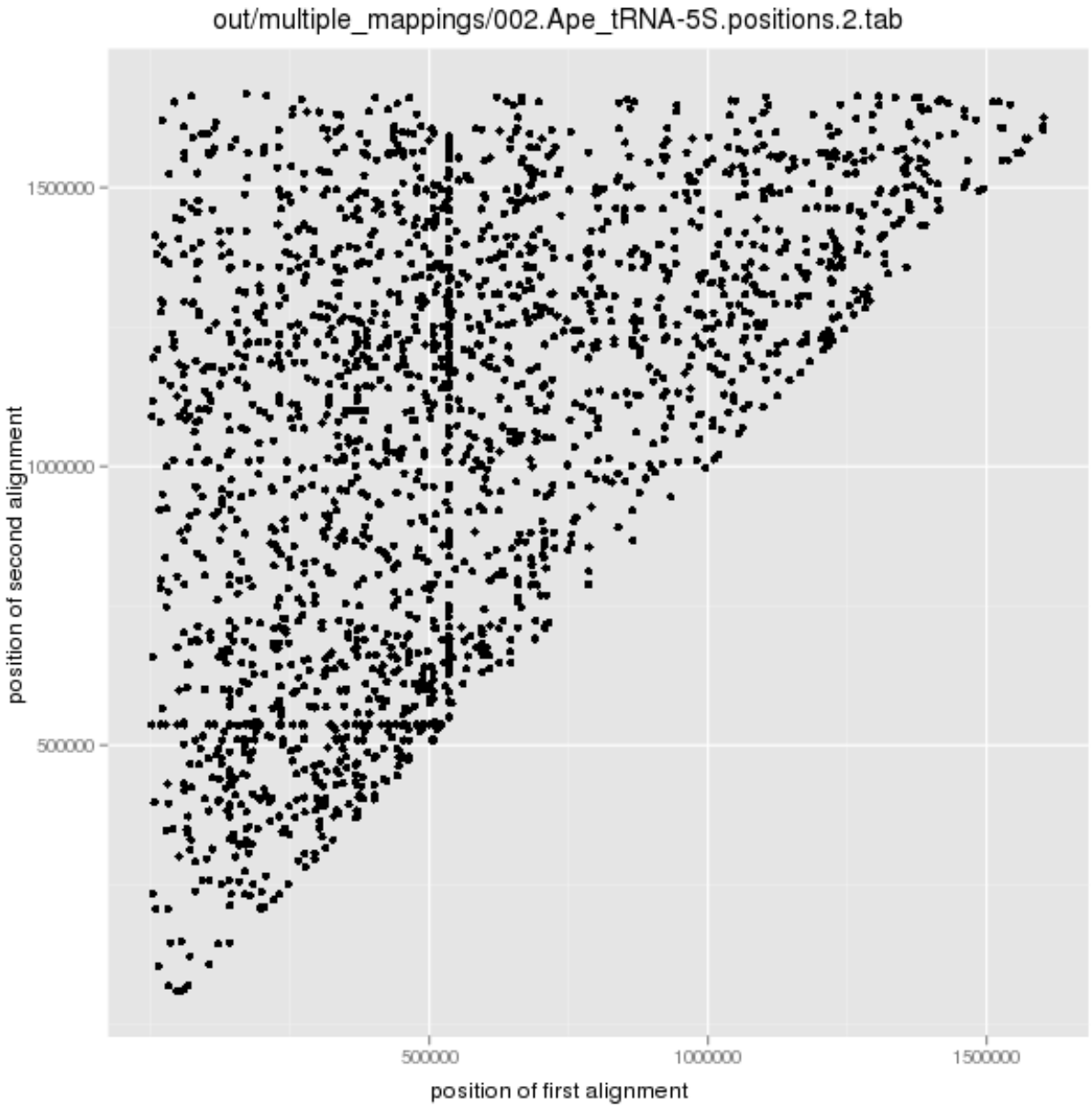

Figure 6: Reads that map exactly twice from a sample of *Aeropyrum pernix*. The position of the first mapping is on the X axis and the second mapping is on the Y axis.

```
>25:20328:31703:2
CACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCTTGTCCTTGGTGCCCGAGTGGGC
GGGATGAAGACTAGCT
25:20328:31703:2 16 chr 501759 0 19M57H * 0 0 * * AS:i:19
25:20328:31703:2 16 chr 536648 0 22H18M36H * 0 0 * * AS:i:14
--
>4:19303:166867:2
TGGCTTGTCCTTGGTGCCCGAGTGGGGTACATGGCCAACCCTCACTCACCCCTCTTACGC
CTGTGCAGCCTCCTCC
4:19303:166867:2 16 chr 536648 0 55H21M * 0 0 * * AS:i:13
4:19303:166867:2 16 chr 1100691 0 52M24H * 0 0 * * AS:i:52
--
>41:10721:166220:2
CGATCTATGGCTTGTCCTTGGTGCCCGAGTGGGCCCTGTGTGAAGAGATTACCGCGGACT
CCAGACGCCCTC
41:10721:166220:2 16 chr 509097 0 38M34H * 0 0 * * AS:i:38
41:10721:166220:2 16 chr 536648 0 44H18M10H * 0 0 * * AS:i:14
--
>63:4269:186315:2
CACTCTTTCCCTACACGACGCTCTTCCGATCTTTGGCTTGTCCTTGGTGCCCGAGTGGGC
GGGATGAAGACTAGCT
63:4269:186315:2 16 chr 501759 0 19M57H * 0 0 * * AS:i:19
63:4269:186315:2 16 chr 536648 0 22H18M36H * 0 0 * * AS:i:14
```

Figure 7: This figure corresponds with figure 6. A few of the reads that map to position 536648 also map to one other locus. For example, read `4:19303:166867:2` seems to have two parts: a ~53 nt part and a ~23 nt part. The ~53 nt part maps to 536648, but the other part of the read maps elsewhere.

References

[1] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[2] Natalia G. Starostina, Sarah Marshburn, L. Steven Johnson, Sean R. Eddy, Rebecca M. Terns, and Michael P. Terns. Circular box c/d rnas in pyrococcus furiosus. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):14097–14101, 2004.