Michael J. Axtell - Penn State University
mja18@psu.edu
March 30, 2012

Mapping trimmed SOLiD small RNA data: Comparison of translated-trimmed read mapping to colorspace mapping.

MOTIVATION

To test how the translation to DNA-space of trimmed color-space small reads affects the percentage of mappable reads.

METHODS

Small RNA data*:* Four small RNA datasets were used, all from *Arabidopsis thaliana*. Small RNAs from wild-type Col-0 leaves, Col-0 inflorescences, as well as from *hyl1-2* leaves and *hyl1-2* inflorescenes were used (Liu et al., 2012; NCBI GEO GSE29802). Raw .csfasta and .QV.qual files were combined into .cs-fastq formatted files using csfasta2cs-fastq.pl version 0.2.

Adapter Trimming: 3' adapters were found and trimmed from .cs-fastq files using trim_SOLiD_sRNA_cs-fastq.pl version 2.1. Two different settings were used to create .fastq and .cs-fastq trimmed datasets. .fastq (i.e., translated) trimmed reads were created with options -t Y, -q Y while .cs-fastq (i.e., colorspace) trimmed reads were created with options -t N, -q Y, -h N. Thus, the 3' hybrid color was removed from the trimmed .cs-fastq output.

Mapping: The TAIR10 *Arabidopsis thaliana* genome assembly (Athal_167 via phytozome) was used, including the plastid and mitochondrial genomes. DNA-space and color-space bowtie indexes were built using bowtie-build will all default settings (v 0.12.7). bowtie version 0.12.7 was used for mapping. Only properly trimmed reads were used for mapping -- the 'pseudotrimmed' reads, where adapters were not found, were not used. Mapping settings were as follows:

1. .fastq data, 0 mismatches allowed: -v 0 -a -m 50 -S -p 6
2. .cs-fastq data, 0 mismatches allowed: -v 0 -C -a -m 50 -S -p 6 --snpfrac 0.0000000001 --col-keepends
3. .fastq data, 1 mismatch allowed: -v 1 -a --best --strata -m 50 -S -p 6
4. .cs-fastq data, 1 mismatch allowed: -v 1 -C -a --best --strata -m 50 -S -p 6 --snpfrac 0.0000000001 --col-keepends

Sam-formatted mapping output data were piped through an awk command (awk '$3!="*"') to remove information on non-mapped reads.

RESULTS:

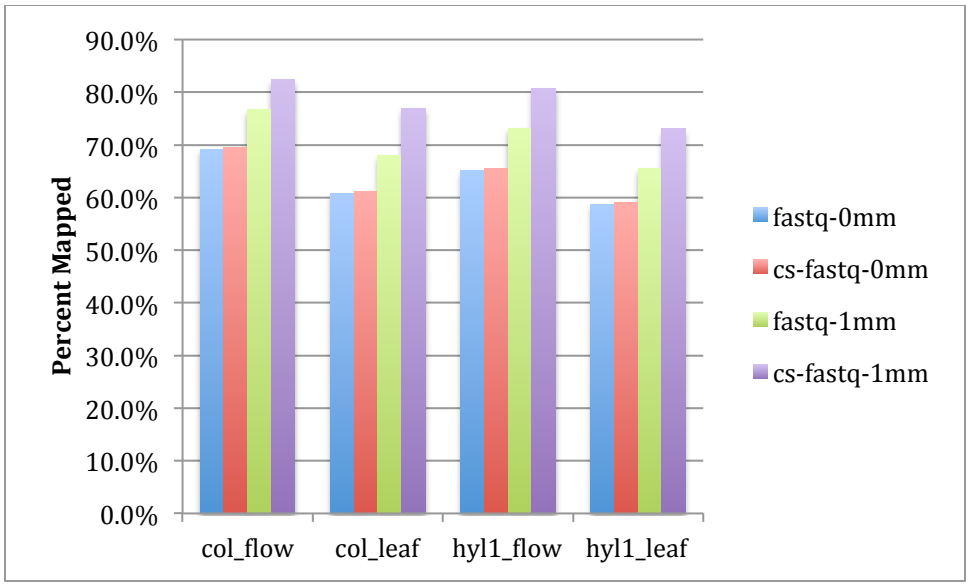

Figure 1. Percentages of mapped reads. mm= mismatches allowed for mapping.

fastq-0mm	cs-fastq-0mm	Colorspace Advantage
69.2%	69.5%	0.2%
60.8%	61.2%	0.4%
65.2%	65.4%	0.3%
58.7%	59.1%	0.4%

Table 1. Percentage mapped with zero mismatches. 'Colorspace advantage' is the percentage gain in mapped reads between mapping in colorspace and DNA-space.

fastq-1mm	cs-fastq-1mm	Colorspace Advantage
76.7%	82.4%	5.7%
67.9%	77.0%	9.1%
73.2%	80.8%	7.6%
65.6%	73.0%	7.5%

Table 2. As in table 1 except for allowing 1 mismatch.

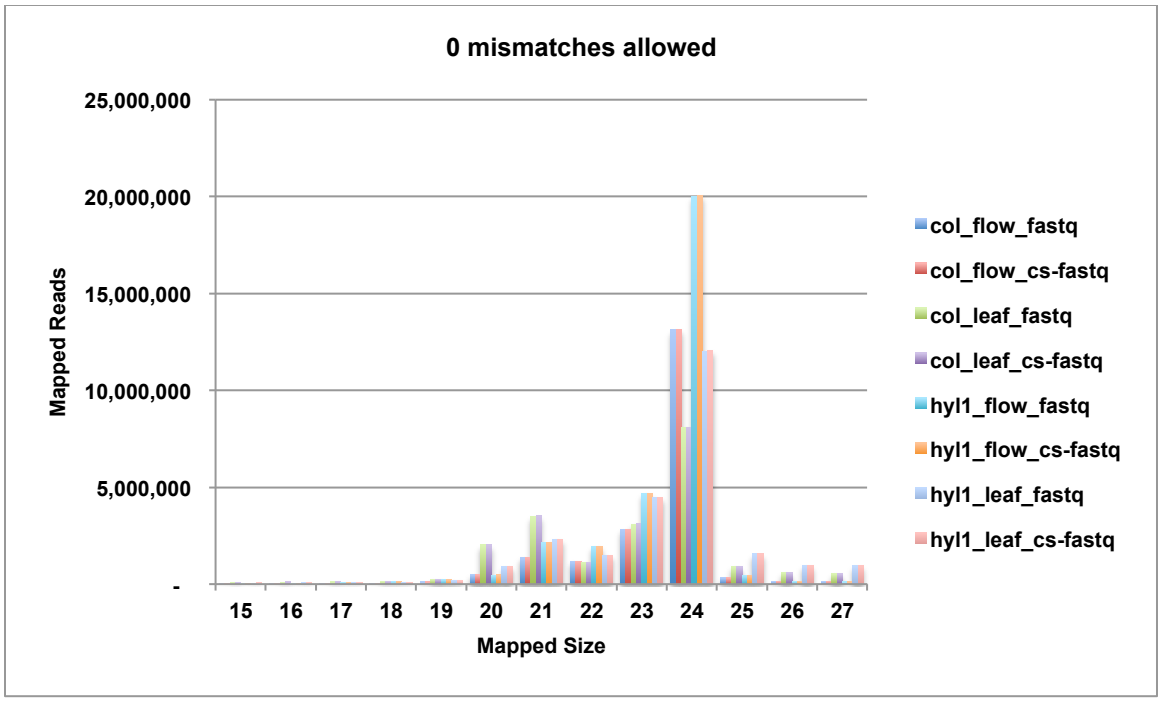

Figure 2. Size distribution of reads mapped with 0 mismatches.

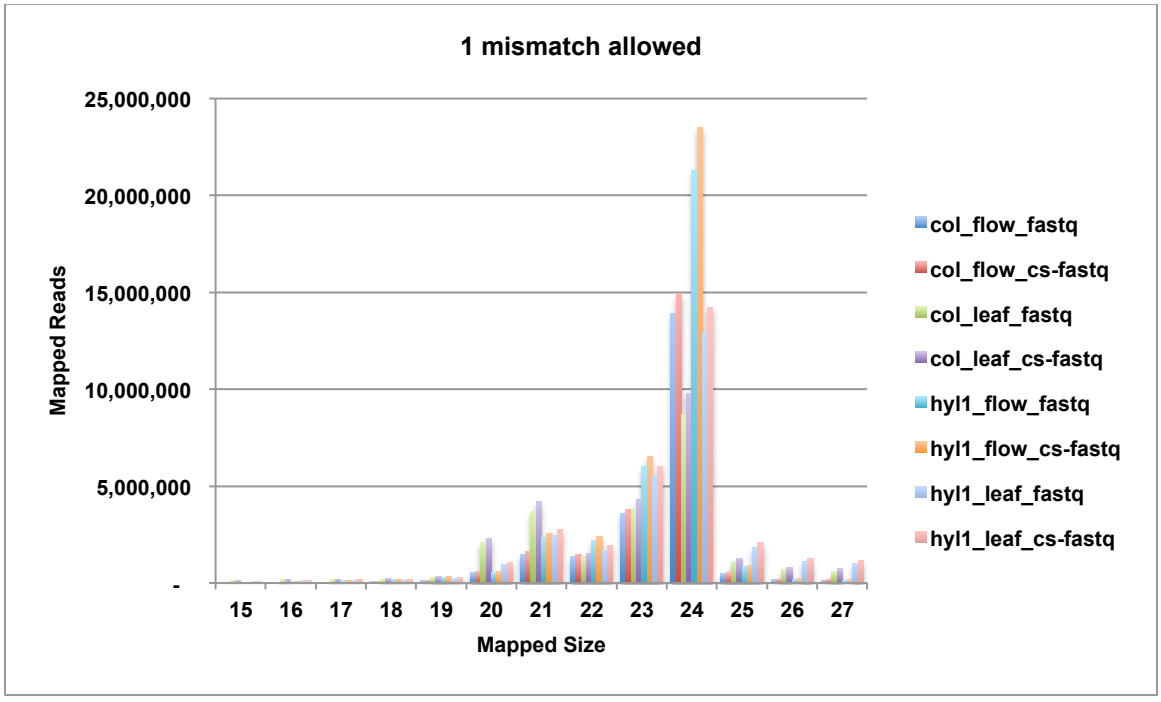

Figure 3. Size distribution of reads mapped with 1 mismatches.

CONCLUSIONS

Mapping in colorspace nets more mappable reads with an identical size distribution of mapped reads. When only allowing perfectly matched reads, the difference is negligible, between 0.2 and 0.4%. However, when trying to rescue small RNA reads with single nt errors, colorspace is giving substantially more mapped reads .. between 5.7% - 9.1% in these datasets. In general, there is no penalty for keeping reads in color-space, and if one is trying to rescue reads with sequencing errors color-space is clearly superior to DNA translation.

REFERENCE

Liu C, Axtell MJ, and Fedoroff NV. (2012). The helicase and RNaseIIIa domains of *Arabidopsis* DCL1 modulate catalytic parameters during microRNA biogenesis. (*Submitted*).