Amplicon library (short insert) sequenced on 5500XL, then taken through NEB Illumina gDNA kit and sequenced on HiSeq:

Insertions seem to be pretty uniformly bi-directional:

```
scott@fourierseq:~/data/tony_solid_illumina$ grep -c '^CTGCCCCGGGTTCC' Ill_SA12001/L-N19_L007_R1.fastq
541483
scott@fourierseq:~/data/tony_solid_illumina$ grep -c '^CCACTACGCCTCCG' Ill_SA12001/L-N19_L007_R1.fastq
503916
scott@fourierseq:~/data/tony_solid_illumina$ grep -c '^CTGCCCCGGGTTCC' Ill_SA12001/L-N19_L007_R2.fastq
486259
scott@fourierseq:~/data/tony_solid_illumina$ grep -c '^CCACTACGCCTCCG' Ill_SA12001/L-N19_L007_R2.fastq
425104
```

Sampling 25000 sequences from the S-N19 pool, we see about 35% of the sequences are properly formed 5' and 3' ends of the SOLiD library:

```
scott@fourierseq:~/data/tony_solid_illumina/Ill_SA12001$ head -100000 S-N19_L007_R1.fastq | grep 'CCCCGGGTT' | sort | uniq -c -w 10 | awk '{sum+=$1} END {print sum}'
8571
scott@fourierseq:~/data/tony_solid_illumina/Ill_SA12001$ head -100000 S-N19_L007_R1.fastq | grep 'ACTACGCCT' | sort | uniq -c -w 10 | awk '{sum+=$1} END {print sum}'
8894
The rest?
scott@fourierseq:~/data/tony_solid_illumina/Ill_SA12001$ head -100000 S-N19_L007_R1.fastq | grep -A 1 '^@HWI' | grep -v '^--' | grep -v '^@HWI' | grep -v 'CCCCGGGTT' | grep -v 'ACTACGCCT' | sort | uniq -c -w 10 | awk '{sum+=$1} END {print sum}'
7536
```

CTGCTGTACGGCCAAGGCGCAGTCCCTTTCAATGGACTCGGCACAACCTGGAGGACAACTAACGCGCTACGATA TAGAACTAGCTGTACTATACACCTTAT
The first sequence aligns with the short form of the bottom strand of the P2 adaptor:

```
CTGCTGTACGGCCAAGGCGCAGTCCCTTTTC - first 30 bp of the sequence occurring 2142 times
CTGCTGTACGGCCAAGGCG - the RC of the bottom strand of the P2 adaptor
```

Likely abundant b/c we did minimal PCR cycles on the original SOLiD library. Note that the part after the P2 bottom strand is Tony's construct (CAGTCCCTTTTC...).

And similarly the second sequence aligns with the short form of the top strand of the P1 adaptor:

```
GCTTTCCTCTCTATGGGCAGTCGGTGATCA - first 30 bp of the sequence occurring 1481 times
CCTCTATGGGCAGTCGGTGAT - top strand of the P1 multiplex adaptor (though the GCTTT which precede it are coded in the mux P1 primer).
```

Expanding uniqueness to 40 bp reveals Tony's construct off the P1 mux adaptor (the CATAAGCTGG):

```
scott@fourierseq:~/data/tony_solid_illumina/Ill_SA12001$ head -100000 S-N19_L007_R1.fastq | grep -A 1 '^@HWI' | grep -v '^--' | grep -v '^@HWI' | grep -v 'CCCCGGGTT' | grep -v 'ACTACGCCT' | sort | uniq -c -w 40 | sort -n -r | head
1451 GCTTTCCTCTCTATGGGCAGTCGGTGATCATAAGCTGG
659 CTGCTGTACGGCCAAGGCGCAGTCCCTTTCCATGGACTAC
631 CTGCTGTACGGCCAAGGCGCAGTCCCTTTCCATGGACTG
362 CTGCTGTACGGCCAAGGCGCAGTCCCTTTCCATGGACTG
```

as well as the fact that there is some variability at the point Tony's canonical construct intersects with what we're actually seeing in seq data.

The ones that start with CTGCTG may be particularly useful though b/c they are reverse reads much closer to the variable region (which would otherwise be missed in those reverses). The ones starting with GCTTT should be similarly useful.

**RNA-seq using NEB small RNA kit for mRNA expression**

This is probably going into a paper...