Inserts Selection Report

GAO-1, GAO-2

Genome Analyzer run:

Instrument name:	*EAS269*
Instrument version:	*"GAII"*
Number of sequencing cycles:	*36*
Sequencing kit:	*Scanning buffer v 2.0*
Tiles:	*100*
Data Analysis Pipeline:	*GAPipeline-0.3.0*
Run code:	*080616*

Data:

Sample preparation protocol: small RNA

Fasteris code	Sample name	File name	Reads
GAO-1	WT SF8	080616_s_5_GAO-1.zip	2'986'692
GAO-2	TG SF8	080616_s_6_GAO-2.zip	3'361'969

Method:

First the data sets are screened for the sequence of the 5' adapter (We sometime observed that parts of the 5' adapter can be cloned instead of a small RNA insert) with the same method than described below using the last 10 bases of this adapter. Clean reads (without a match) are used for the next step.

The adapter sequences were trimmed in 4 steps:

1) The 21nt adapter sequence is used, which permits to identify "inserts" of 14nt or less.
2) If no adapter sequence was found, in successive steps the last base of the adapter was removed and the sequence was searched at the end of the reads. The minimum adapter size of 5 bases permits identifying inserts of up to 30 bases.
3) The remaining reads are search for not-exact matches of the adapter. The first 4 bases of the adapter were searched within the full reads sequences and at least 75% of the following bases must be identical to the adapter sequence.
4) Finally the sequence is search at position 20 to 25 for matches with at least 75% of the bases identical to the adapter sequence.

After trimming of the adapter sequences, the inserts were sorted in separate files according to their lengths. The reads in which no match of the adapter sequences were found were placed in the “No_Adapter_” prefix file.

Results:

Sample	Clean reads	5' adapter artifact
GAO-1	2'854'636	132'056
GAO-2	2'993'268	368'701

Insert selection:

Length	Count	Percent total reads
0	7'607	0.27%
1	449	0.02%
2	3'701	0.13%
3	126	0.00%
4	350	0.01%
5	765	0.03%
6	447	0.02%
7	821	0.03%
8	871	0.03%
9	187	0.01%
10	151	0.01%
11	335	0.01%
12	401	0.01%
13	150	0.01%
14	265	0.01%
15	474	0.02%
16	973	0.03%
17	1'634	0.06%
18	2'058	0.07%
19	5'275	0.18%
20	24'955	0.87%
21	219'216	7.68%

22	1'310'030	45.89%
23	547'664	19.19%
24	129'249	4.53%
25	31'574	1.11%
26	9'043	0.32%
27	3'049	0.11%
28	2'227	0.08%
29	1'951	0.07%
30	1'411	0.05%
Total reads with adapter:	2'307'409	80.9%

A total of 2'307'409 inserts of have been identified.
They represent 51'123'526 bases.

Length	**Count**	**Percent total reads**
0	22'121	0.74%
1	1'016	0.03%
2	9'194	0.31%
3	550	0.02%
4	1'143	0.04%
5	2'423	0.08%
6	2'133	0.07%
7	4'772	0.16%
8	2'842	0.09%
9	1'681	0.06%
10	1'144	0.04%
11	1'850	0.06%
12	3'804	0.13%
13	6'903	0.23%
14	11'160	0.37%
15	13'291	0.44%
16	14'086	0.47%
17	14'883	0.50%

18	17'516	0.59%
19	17'875	0.60%
20	48'471	1.62%
21	265'605	8.87%
22	1'248'490	41.71%
23	447'998	14.97%
24	135'928	4.54%
25	36'347	1.21%
26	8'673	0.29%
27	3'351	0.11%
28	2'575	0.09%
29	2'097	0.07%
30	1'461	0.05%
Total reads with adapter:	2'351'383	78.6%

A total of 2'351'383 inserts of have been identified.
They represent 50'759'414 bases.

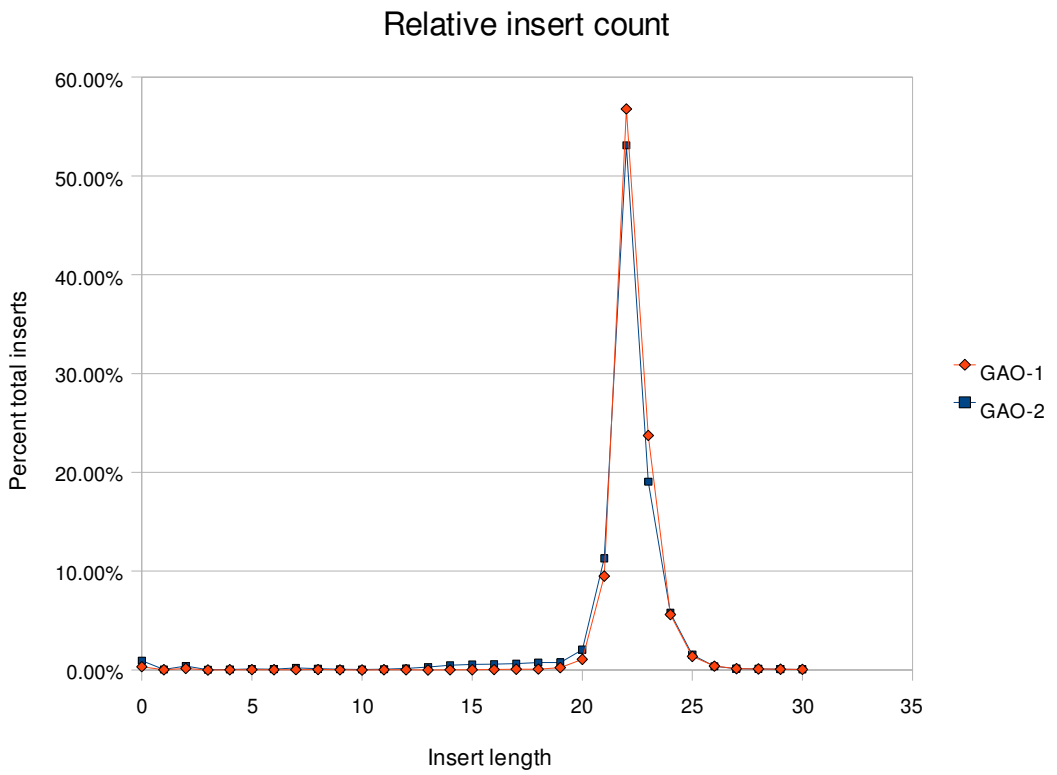

Figure 1:Relative count in regards of the insert length

80.9 and 78.6% of the clean reads from GAO-1 and GAO-2, respectively, have been identified as inserts. The remaining reads can be considered as 31 or more bases inserts.

Both samples show a single peak for the relative amount of inserts of a given length at 22 bases.

Attachment:

Raw data:
080616_s_1_seq_GAO-1.zip
080616_s_2_seq_GAO-2.zip

All inserts files in a single zip file:
2008-06-25_080616_GAO-1_inserts.zip
2008-06-25_080616_GAO-2_inserts.zip