Table 1. Cluster size distribution for the three compared *D.rerio* cluster datasets. Only the sequences common to all three datasets were used in the comparisons.

Size of Cluster (number of sequences)	***UniGene Build 21 24/08/01***	***TIGR ZGI V.7 07/08/01***	***CLOBB***
1	4169	9914	6848
2	1824	2231	2655
3–4	1953	1956	2321
5–8	1288	1155	1407
9–16	638	506	574
17–32	270	214	230
33–64	123	103	112
65–128	37	41	33
129–256	16	24	24
257–512	12	8	9
513–1024	5	3	2
1025–2048	1	1	0
Total Clusters (from 58,888 sequences)	**10336**	**16156**	**14215**

Table 2 Distribution of cluster events

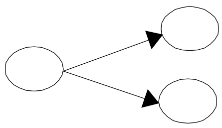
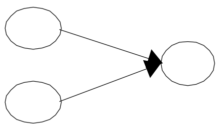
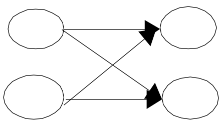
Type of Rearrangement	***CLOBB−UniGene***	***CLOBB−TIGR***	***TIGR−UniGene***
 Equivalent Clusters	6125	10440	6241
 Simple Split	246 –> 521	1638 –> 4190	107 –> 223
 Simple Merge	5838 –> 2467	998 –> 406	8876 –> 3331
 Complex	2006 – > 1223	1139 –> 1120	932 –> 541

Table 3. Detailed analysis of UniGene cluster ug.2984

	UniGene	***TIGR***	***CLOBB*−α**	***CLOBB*−β**
Total sequences in clusters containing at least one sequence derived from ug.2984	1873	1880	1900	1900
Total clusters	1	38	66	74
Clusters with > 100 seqs (sizes)	1	4 (1075, 146, 145, 136)	3 (711,482,108)	5 (425, 219, 214, 203, 143)
Clusters with only one sequence (singletons)	0	23	33	31

Table 4. Post cluster consensus assembly using CAP3 of CLOBB clusters derived from ug.2984.

Cluster(s) assembled	***Total number of clusters used in assembly (sequences)***	***Number of contigs built from > 1 sequence (A)***	***Number of singlets produced in assembly (B)***	***Total predicted tentative consensuses (A) + (B) + number of singletons***
CLOBB pooled clusters	1 (1900)	20	23	43
CLOBB−α individual clusters	33 (1867)	46	11	90
CLOBB−β individual clusters	43 (1869)	56	14	101
TIGR predictions for the same 1900 sequences	N/A	30	19	49

Table 5. Summary of features of the three cluster methods examined.

Feature	***UniGene***	***TIGR***	***CLOBB***
Underlying Clustering Method	megaBLAST	WU–BLAST & CAP3	NCBI BLAST
Stringency	Dependent on stage of clustering	Very High >= 95% identity over > 40 bp	High >= 95% identity over 30bp
Overlap allowed	N/A	< 20bp	< 10% of sequence length Those with > 10% of sequence length are allowed if they contain > 10% unassigned bases
Clusters are always contiguous ?	No	Yes	Yes
Dealing with potential chimeric clusters	Initial clustering performed with gene sequences – merging of these initial distinct clusters rejected	CAP3 does not include identified chimeric sequences	Definition of type III matches and ‘superclusters’ prevents chimeric sequences from merging unsuitable clusters.
Continuity (addition of new sequences)	New builds are compared with previous builds	Post processing	Incremental within algorithm
Historical information	Availability of previous builds	Notes showing retirement of clusters	‘superclusters’ and merge events can be tagged
Portability and adapatibility	Low	Low	High
Ease of retention of manual curation	Medium	Medium	High