4 Results

4.1 IRS-PCR based methods for mapping the zebrafish genome

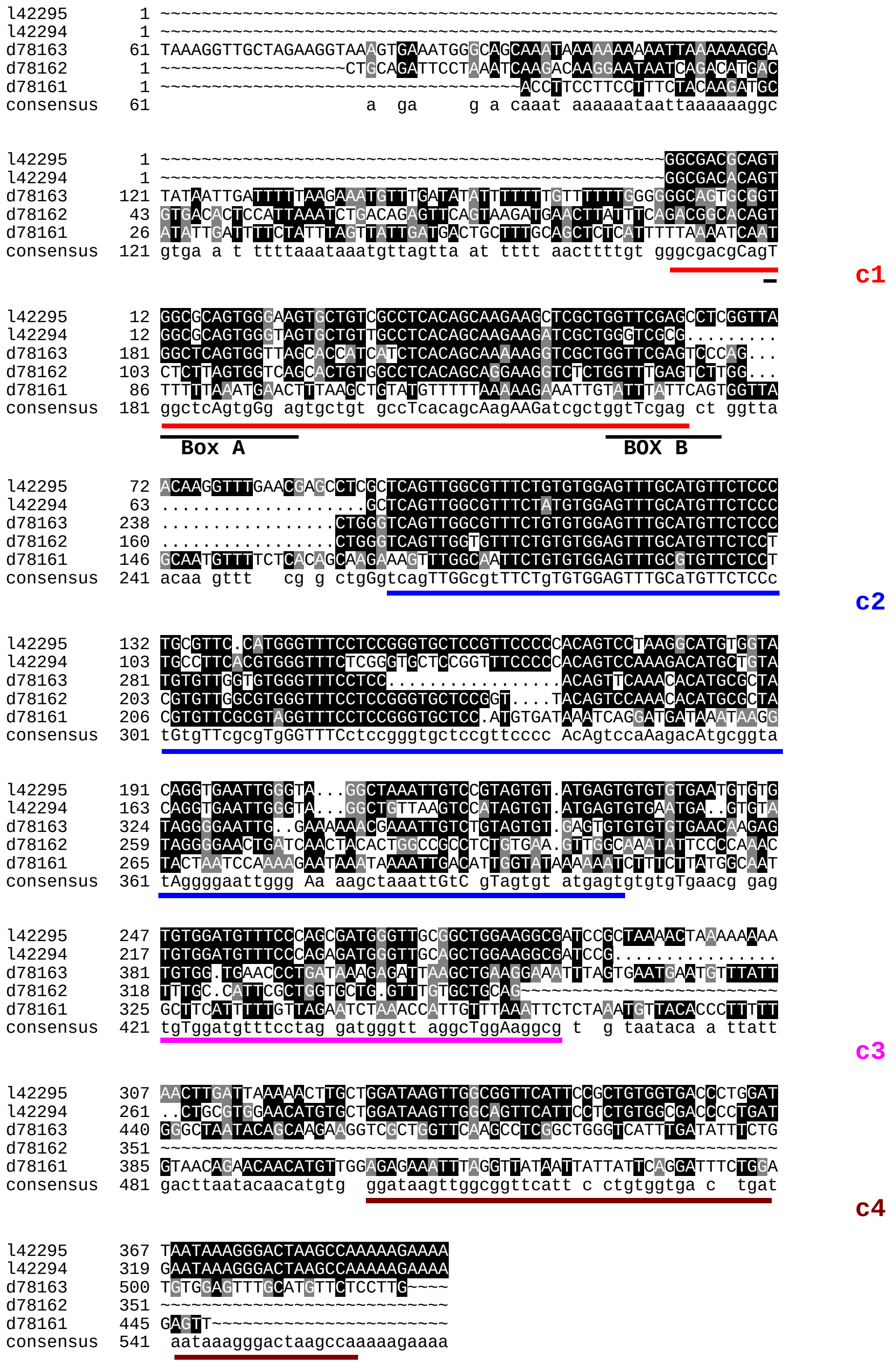
4.1.1 Analysis of the repetitive Mermaid/DANA element.

Two short interspersed repetitive sequences (SINEs) have been described in the literature independently: DANA (Izsvak et al., 1996) and mermaid (Shimoda et al., 1996). Homology searches in databases shows that both elements have a common homologous region. To further investigate their relationship, mermaid and DANA sequences were fetched from the GenEMBL nucleotide sequence database. They were aligned and edited using the multiple sequence alignment program *PileUp* and the editor *LineUp* of the GCG package (Genetics Computer Group (Devereux et al., 1984). The web-based program Boxshade http://www.ch.embnet.org/software/BOX_form.html was used to generate a consensus sequence with conserved bases highlighted (Figure 9).

The alignment shows that the mermaid element consists of the conserved regions c1-c2 of DANA, rather than being a distinct element on its own. The complete DANA element additionally contains the conserved regions c3-c4. The discrepancy between the two "classes" of elements is caused by the methods employed to identify the repeats. The probes and primers in the study of Izsvak *et al.* were suitable to find the complete DANA repeat, while Shimoda *et al.* probed for the smaller c1-c2 region. This also explains that, while DANA was found to be restricted to the genus Danio, mermaid is widespread in vertebrates and mermaid probes also bind to primate DNA.

In order to amplify a maximum of products in an IRS-PCR reaction, the oligonucleotide primers have to be optimised to match to a maximum of the (potentially divergent) target sequences. The primer should be designed to bind to a highly conserved region, and should match the consensus sequence, while variable bases in the target sequence have to be compensated by wobble bases. Because the number of sequences deposited in the databases grows rapidly and permanently, homology searches using the original DANA/mermaid sequences as query were done repeatedly during the course of this work. Homologous sequences were fetched from the databases and aligned in order to generate an optimised consensus sequence. Because this dataset is larger than that used by Shimoda *et al.* to design mermaid primers, the consensus sequence is supposed to be more reliable.

The database searched for DANA/mermaid specific sequences was a zebrafish specific subset of the GenEMBL database retrieved by sequence retrieval service (SRS http://srs.ebi.ac.uk/) and saved on a local server. A zebrafish DANA sequence (Accession number L42295) served as query sequence. Sequence homology search was done using an implementation of the *BLAST* algorithm (basic local alignment search tool, Altschul et al., 1990) in the GCG package.



```
l42295       1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
l42294       1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
d78163      61 TAAAGGTTGCTAGAAGGTAAAGTGAAATGGGCAGCAAATAAAAAAAAAATTAAAAAAGGA
d78162       1 ~~~~~~~~~~~~~~~~~~CTGCAGATTCCTAAATCAAGACAAGGAATAATCAGACATGAC
d78161       1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ACCTTCCTTCCTTTCTACAAGATGC
consensus   61                     a  ga     g a caaat aaaaaataattaaaaaaggc

l42295       1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~GGCGACGCAGT
l42294       1 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~GGCGACACAGT
d78163     121 TATAATTGATTTTTAAGAAATGTTTGATATATTTTTTTGTTTTTTGGGGGGCAGTGCGGT
d78162      43 GTGACACTCCATTAAATCTGACAGAGTTCAGTAAGATGAACTTATTTCAGACGGCACAGT
d78161      26 ATATTGATTTCTATTTAGTTATTGATGACTGCTTTGCAGCTCTCATTTTTTAAAATCAAT
consensus  121 gtga a t ttttaaataaatgttagtta at tttt aacttttgt gggcgacgCagT
                                                                                c1

l42295      12 GGCGCAGTGGGAAGTGCTGTCGCCTCACAGCAAGAAGCTCGCTGGTTCGAGCCTCGGTTA
l42294      12 GGCGCAGTGGGTAGTGCTGTTGCCTCACAGCAAGAAGATCGCTGGGTCGCG.........
d78163     181 GGCTCAGTGGTTAGCACCATCATCTCACAGCAAAAAGGTCGCTGGTTCGAGTCCCAG...
d78162     103 CTCTTAGTGGTCAGCACTGTGGCCTCACAGCAGGAAGGTCTCTGGTTTGAGTCTTGG...
d78161      86 TTTTTAAATGAACTTTAAGCTGTATGTTTTTAAAAAGAAATTGTATTTATTCAGTGGTTA
consensus  181 ggctcAgtgGg agtgctgt gccTcacagcAagAAGatcgctggtTcgag ct ggtta
                 Box A                                     BOX B

l42295      72 ACAAGGTTTGAACGAGCCTCGCTCAGTTGGCGTTTCTGTGTGGAGTTTGCATGTTCTCCC
l42294      63 ....................GCTCAGTTGGCGTTTCTATGTGGAGTTTGCATGTTCTCCC
d78163     238 .................CTGGGTCAGTTGGCGTTTCTGTGTGGAGTTTGCATGTTCTCCC
d78162     160 .................CTGGGTCAGTTGGTGTTTCTGTGTGGAGTTTGCATGTTCTCCT
d78161     146 GCAATGTTTTCTCACAGCAAGAAAGTTTGGCAATTCTGTGTGGAGTTTGCGTGTTCTCCT
consensus  241 acaa gttt   cg g ctgGgtcagTTGGcgtTTCTgTGTGGAGTTTGCaTGTTCTCCc
                                                                                c2

l42295     132 TGCGTTC.CATGGGTTTCCTCCGGGTGCTCCGTTCCCCCACAGTCCTAAGGCATGTGGTA
l42294     103 TGCCTTCACGTGGGTTTCTCGGGTGCTCCGGTTTCCCCCACAGTCCAAAGACATGCTGTA
d78163     281 TGTGTTGGTGTGGGTTTCCTCC.................ACAGTTCAAACACATGCGCTA
d78162     203 CGTGTTGGCGTGGGTTTCCTCCGGGTGCTCCGGT....TACAGTCCAAACACATGCGCTA
d78161     206 CGTGTTCGCGTAGGTTTCCTCCGGGTGCTCC.ATGTGATAAATCAGGATGATAAATAAGG
consensus  301 tGtgTTcgcgTgGGTTTCctccgggtgctccgttcccc AcAgtccaAagacAtgcggta

l42295     191 CAGGTGAATTGGGTA...GGCTAAATTGTCCGTAGTGT.ATGAGTGTGTGTGAATGTGTG
l42294     163 CAGGTGAATTGGGTA...GGCTGTTAAGTCCATAGTGT.ATGAGTGTGAATGA..GTGTA
d78163     324 TAGGGGAATTG..GAAAAAACGAAATTGTCTGTAGTGT.GAGTGTGTGTGTGAACAAGAG
d78162     259 TAGGGGAACTGATCAACTACACTGGCCGCCTCTGTGAA.GTTGGCAAATATTCCCCAAAC
d78161     265 TACTAATCCAAAAGAATAAATAAAATTGACATTGGTATAAAAAATCTTTCTTATGGCAAT
consensus  361 tAggggaattggg Aa aagctaaattGtC gTagtgt atgagtgtgtgTgaacg gag

l42295     247 TGTGGATGTTTCCCAGCGATGGGTTGCGGCTGGAAGGCGATCCGCTAAAACTAAAAAAAA
l42294     217 TGTGGATGTTTCCCAGAGATGGGTTGCAGCTGGAAGGCGATCCG................
d78163     381 TGTGG.TGAACCCTGATAAAGAGATTAAGCTGAAGGAAATTTAGTGAATGAATGTTTATT
d78162     318 TTTGC.CATTCGCTGGTGCTG.GTTTGTGCTGCAG~~~~~~~~~~~~~~~~~~~~~~~~~
d78161     325 GCTTCATTTTTGTTAGAATCTAAACCATTGTTTAAATTCTCTAAATGTTACACCCTTTTT
consensus  421 tgTggatgtttcctag gatgggtt aggcTggAaggcg t  g taataca a ttatt
                                                                                c3

l42295     307 AACTTGATTAAAAACTTGCTGGATAAGTTGGCGGTTCATTCCGCTGTGGTGACCCTGGAT
l42294     261 ..CTGCGTGGAACATGTGCTGGATAAGTTGGCAGTTCATTCCTCTGTGGCGACCCCTGAT
d78163     440 GGGCTAATACAGCAAGAAGGTCGCTGGTTCAAGCCTCGGCTGGGTCATTTGATATTTCTG
d78162     351 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
d78161     385 GTAACAGAACAACATGTTGGAGAGAAATTTAGGTTATAATTATTATTCAGGATTTCTGGA
consensus  481 gacttaatacaacatgtg  ggataagttggcggttcatt c ctgtggtga c  tgat
                                                                                c4

l42295     367 TAATAAAGGGACTAAGCCAAAAAGAAAA
l42294     319 GAATAAAGGGACTAAGCCAAAAAGAAAA
d78163     500 TGTGGAGTTTGCATGTTCTCCTTG~~~~
d78162     351 ~~~~~~~~~~~~~~~~~~~~~~~~~~~~
d78161     445 GAGTT~~~~~~~~~~~~~~~~~~~~~~~
consensus  541  aataaagggactaagccaaaaagaaaa
```

Figure 9

Alignment of DANA (Accession numbers L142294 and L142295) and mermaid repeats (D78161, D78162, D78163). The conserved regions are underlined. The RNA polymerase III specific promoter sequence (Box A - Box B) of this tRNA related repeat is highlighted. The conserved regions c1 and c2 are shared by DANA and mermaid, while c3 and c4 are DANA specific.

BLAST search results were processed using MSPcrunch and Blixem software (Sonnhammer and Durbin, 1994). MSPcrunch removes regions of low complexity (e.g. poly-A sequences) from the BLAST output, and aligns the homologous sequences along their matching regions in the query sequence. Blixem is used to view the output of MSPcrunch (Figure 10). In this manner, a good survey of conserved and variable regions is obtained. In Figure 10, the organisation of the DANA repeat, having four conserved regions interrupted by variable, low complexity regions, is obvious.

For a more detailed analysis, the MSPcrunch filtered sequences were retrieved from the database and aligned using PileUp and LineUp (GCG). The resulting alignment is shown in Figure 11. As expected for a SINE with tRNA-like features, the C1 region is recognised as a candidate tRNA by the program tRNAscan-SE (Lowe and Eddy, 1997) albeit only if relaxed search parameters are used. A BLAST search of the Fugu genome http://fugu.hgmp.mrc.ac.uk/ also identifies a sequence homologous to C1 and C2 regions of the mermaid repeat.

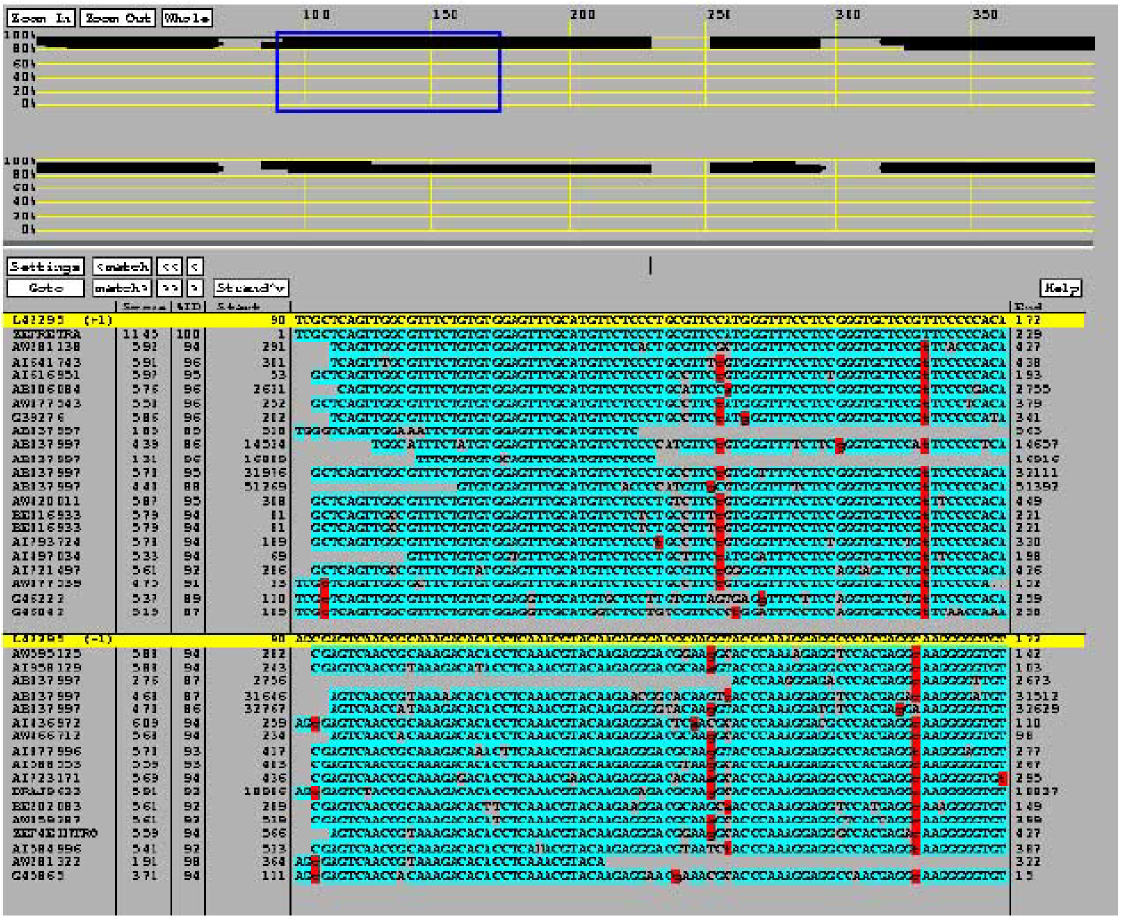


Figure 10

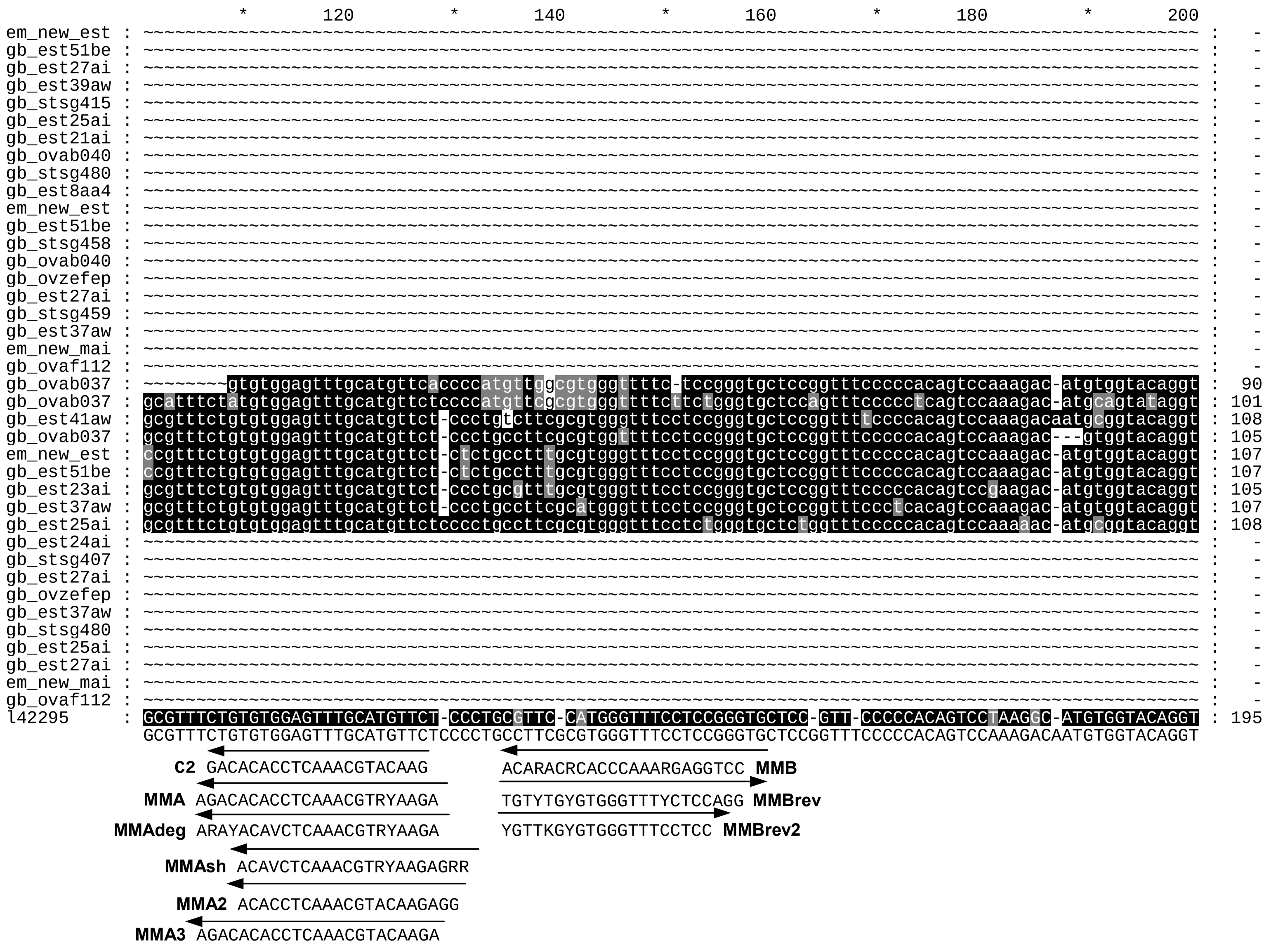
Alignment of a DANA sequence (Accession number L42295) with homologous sequences from a zebrafish specific database. The two fields on top represent the whole query sequence. The black bars show where homologous sequences align, in either direction. The four conserved regions of the DANA element are distinguishable due to the accumulation of homologous sequences. The section marked by the blue rectangle is shown in detail in the bottom part of the illustration. The query sequence is highlighted in yellow and shown in either orientation. Sequences found in the database are shown with accession numbers. Bases identical to the query sequence are highlighted in blue.

Figure 11

(Next page) Detailed alignment of the mermaid/DANA homologous regions. The variable and low-complexity regions are filtered out by MSP crunch. Different oligonucleotide primers, which were tested for IRS-PCR are shown with sequences.

```
                      *        20         *        40         *        60         *        80         *       100
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est39aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg415 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est21ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est8aa4 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg458 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg459 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~tg :   2
gb_est41aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
gb_est23ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~tcagttt :   7
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gctcagttg :   9
gb_est24ai : ~~~~~~gcagtggcgcagtaggtagtactgtcgcctcacagcaagaaggtcgctgggttgagcctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  62
gb_stsg407 : ggcgacacagttgcgcagtaggtagtgctgtcaccttacagcaagaaggtcgctggttcgagcctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  68
gb_est27ai : ggcgacgcagtggcgcagtaggtagtgctgtcacctcacagcaagaaggtcgctggtt~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  58
gb_ovzefep : ggcgacgcagtggcgcagtaggttgtgctgtcgcctcacagcaagaaggtcgctggttcgagcctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  68
gb_est37aw : ggcgaggcagtggcgcagtaggtagtgctgttgcctcacagcaagaaggtcgctgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  56
gb_stsg480 : ggcgatgcagtggcgcagtaggtagtgctgttgcctcacagcaagaaggtcgctggttcgaacctcggtt~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  70
gb_est25ai : ggcgaggcagtggcgcagtaggtagtgctgtcgtctcacagcaagaaggtcgctggttctaacctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  68
gb_est27ai : ggcgaggcagtggcacagtaggtagtgctgtcgcctaacagcaagaaggtcgctggttc~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  59
em_new_mai : ~~~gacgaagtggtgcagtaggtattggtgtcgcctcacagcaagaaggtcgctggttctagcctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  65
gb_ovaf112 : ~~~gacgaagtggtgcagtaggtattggtgtcgcctcacagcaagaaggtcgctggttctagcctcgg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :  65
l42295     : GGCGACGCAGTGGCGCAGTGGGAAGTGCTGTCGCCTCACAGCAAGAAGCTCGCTGGTTCGAGCCTCGGTTAACAAGGTTTGAACGAGCCTCGCTCAGTTG : 100
             GGCGACGCAGTGGCGCAGTAGGTAGTGCTGTCGCCTCACAGCAAGAAGGTCGCTGGTTCGAGCCTCGGTTAACAAGGTTTGAACGAGCCTCGCTCAGTTG

            ◄───────────────────
             CCGCTGYGTCACCGCGTCA C1
```



```
                      *       120         *       140         *       160         *       180         *       200
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est39aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg415 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est21ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est8aa4 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg458 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg459 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~gtgtggagtttgcatgttcaccccatgttggcgtgggttttc-tccgggtgctccggtttcccccacagtccaaagac-atgtggtacaggt :  90
gb_ovab037 : gcatttctatgtggagtttgcatgttctccccatgttcgcgtgggttttcttctgggtgctccagtttccccctcagtccaaagac-atgcagtataggt : 101
gb_est41aw : gcgtttctgtgtggagtttgcatgttct-ccctgtcttcgcgtgggtttcctccgggtgctccggttttccccacagtccaaagacaatgcggtacaggt : 108
gb_ovab037 : gcgtttctgtgtggagtttgcatgttct-ccctgccttcgcgtggttttcctccgggtgctccggtttcccccacagtccaaagac---gtggtacaggt : 105
em_new_est : ccgtttctgtgtggagtttgcatgttct-ctctgcctttgcgtgggtttcctccgggtgctccggtttcccccacagtccaaagac-atgtggtacaggt : 107
gb_est51be : ccgtttctgtgtggagtttgcatgttct-ctctgcctttgcgtgggtttcctccgggtgctccggtttcccccacagtccaaagac-atgtggtacaggt : 107
gb_est23ai : gcgtttctgtgtggagtttgcatgttct-ccctgcgtttgcgtgggtttcctccgggtgctccggtttcccccacagtccgaagac-atgtggtacaggt : 105
gb_est37aw : gcgtttctgtgtggagtttgcatgttct-ccctgccttcgcatgggtttcctccgggtgctccggtttccctcacagtccaaagac-atgtggtacaggt : 107
gb_est25ai : gcgtttctgtgtggagtttgcatgttctcccctgccttcgcgtgggtttcctctgggtgctctggtttcccccacagtccaaaaac-atgcggtacaggt : 108
gb_est24ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg407 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
l42295     : GCGTTTCTGTGTGGAGTTTGCATGTTCT-CCCTGCGTTC-CATGGGTTTCCTCCGGGTGCTCC-GTT-CCCCCACAGTCCTAAGGC-ATGTGGTACAGGT : 195
             GCGTTTCTGTGTGGAGTTTGCATGTTCTCCCCTGCCTTCGCGTGGGTTTCCTCCGGGTGCTCCGGTTTCCCCCACAGTCCAAAGACAATGTGGTACAGGT
                ←──────────────────────        ←──────────────────────────
         C2 GACACACCTCAAACGTACAAG          ACARACRCACCCAAARGAGGTCC MMB
               ←───────────────────────        ──────────────────────────→
        MMA AGACACACCTCAAACGTRYAAGA        TGTYTGYGTGGGTTTYCTCCAGG MMBrev
               ←───────────────────────        ───────────────────────→
     MMAdeg ARAYACAVCTCAAACGTRYAAGA        YGTTKGYGTGGGTTTCCTCC MMBrev2
                   ←───────────────────────
      MMAsh ACAVCTCAAACGTRYAAGAGRR
                   ←──────────────────────
       MMA2 ACACCTCAAACGTACAAGAGG
               ←───────────────────────
       MMA3 AGACACACCTCAAACGTACAAGA
```

```
                  *       220         *       240         *       260         *       280         *       300
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est39aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg415 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est21ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est8aa4 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~tttcccagagatgggttgcggctggaagggcatccgct~~~ :  38
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~tttcccagagatgggttgcggctggaagggcatccgct~~~ :  38
gb_stsg458 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ttcccagagatgggttgcggctggaagggcatccgct~~~ :  37
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccagagatgggttgcgactggaagggcatccgct~~~ :  42
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gtttcccagtgatgggttgcggctggaagggcatccgct~~~ :  39
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccatagatgggttgcggctggaaggacatccgct~~~ :  42
gb_stsg459 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccagagatgggttgcggctggaaggacatccgct~~~ :  42
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccagagatgggttgcggctagaagg~~~~~~~~~~~~ :  33
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccagagatgggttgcagctggaa~~~~~~~~~~~~~~ :  31
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~gatgtttcccagagatgggttgcagctggaa~~~~~~~~~~~~~~ :  31
gb_ovab037 : gaattgaataaattaaattggccgtagtgtatga~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 124
gb_ovab037 : gaattgaataagctaaattggccgtagtgtatg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 134
gb_est41aw : gaattgggtaggctaaattgtccgtagtgtatga~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 142
gb_ovab037 : gaattgggtaggctaaattgtccgtagtgta~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 136
em_new_est : taattggctaggctaaattgtccgtagtgtatga~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 141
gb_est51be : taattggctaggctaaattgtccgtagtgtatga~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 141
gb_est23ai : gaattgggtaggctaaattgtccatagtgtatg~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 138
gb_est37aw : gaattgggttggctaaattgt~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 128
gb_est25ai : gaattgggtaggctaaattgtccgtagtgtatga~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 142
gb_est24ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg407 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
l42295     : GAATTGGGTAGGCTAAATTGTCCGTAGTGTATGAGTGTGTGTGAATGTGTGTGTGGATGTTTCCCAGCGATGGGTTGCGGCTGGAAGGCGATCCGCTAAA : 295
             GAATTGGGTAGGCTAAATTGTCCGTAGTGTATGAGTGTGTGTGAATGTGTGTGTGGATGTTTCCCAGAGATGGGTTGCGGCTGGAAGGGCATCCGCTAAA
                                                                    ----------------------->
                                                                     ATGTTTCCCAGWGATGGGTTG C3
```

```
                      *        320         *        340         *        360         *        380         *
em_new_est : ~~~~~~~~~~~~~~~~~~~taaaaacttgctggataagttggcggttcattccgctgtggcgaccccgaattaataaagggactaagc~~~~~~~~~~~ :  69
gb_est51be : ~~~~~~~~~~~~~~~~~~~taaaaacttgctggataagttggcggttcattccgctgtggcgaccccgaattaataaagggactaagc~~~~~~~~~~~ :  69
gb_est27ai : ~~~~~~~~~~~~~~~~~~~taaaaacttgctggataagttggcggttcattctgctgtggcgaccccggattaatagagggactaagccgacaagaaaa :  80
gb_est39aw : ~~~~~~~~~~~~~~~~~~~~aaaaacttgctggataagttggcggttcattccactgtggcgaccccggattaaaaacgggactaagccgacaagaaaa :  79
gb_stsg415 : ~~~~~~~~~~~~~~~~~~~taaaancttgctggataagttggcggttcattccactgtggcgaccccggattaataaagggactaagccgacaagaaaa :  80
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~aaaaacttgctggataagttggtggttcattccgctgtggcgaccccggattaataaagggactaagcc~~~~~~~~~~ :  69
gb_est21ai : ~~~~~~~~~~~~~~~~~~~taaaaacgtgctggataagttggcagttcattccgctgtggcgaccccggattaataaagggattaagctgaaaagaaaa :  80
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~taaaaacgtgctggataagttggcggttcattccgctgtcgcgaccccggattattaaagggact~~~~~~~~~~~~~~~ :  65
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~taaaaacgtgctaaataagttggcggttcattccgctgtggcgaccccagattaataaagggactaagc~~~~~~~~~~~ :  69
gb_est8aa4 : ~~~~~~~~~~~~~~~~~~ttaaaaacgtgctggataaattggcggttcattccactgtggtgaccctggattattaaaggaactaag~~~~~~~~~~~~ :  69
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg458 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab040 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg459 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est41aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovab037 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_est : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est51be : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est23ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est24ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg407 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovzefep : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est37aw : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_stsg480 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est25ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_est27ai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
em_new_mai : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
gb_ovaf112 : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
l42295     : ACTAAAAAAAAAACTTGATTAAAAACTTGCTGGATAAGTTGGCGGTTCATTCCGCTGTGGTGACCCTGGATTAATAAAGGGACTAAGCCAAAAAGAAAA : 394
             ACTAAAAAAAAAACTTGATTAAAAACTTGCTGGATAAGTTGGCGGTTCATTCCGCTGTGGCGACCCCGGATTAATAAAGGGACTAAGCCGACAAGAAAA
                                                ----------------->
                                                CATTCCGCTGTGGYGACC C4
```

4.1.2 Identification of a new interspersed repetitive element in the zebrafish genome

During sequencing of cDNA libraries created at the Max-Planck-Institute for molecular genetics, a sequence was found, which hit a number of different genes in the database, suggesting that it might be an interspersed repetitive element (M. Clark, personal communication). For example two copies of that repetitive sequence were found in an intron of the zebrafish bone morphogenetic protein 4 gene (Hwang et al., 1997). To further investigate the properties of this repetitive sequence, homologous sequences identified by a BLAST search were retrieved from the GenEMBL database and used for creating a multiple sequence alignment, carried out by the GCG program PileUp. Corrections of the alignment had to be done by hand using the multiple sequence alignment and editing tool Seqlab (GCG). A consensus sequence was calculated and conserved nucleotides were marked by shading using the program GenDoc (http://www.psc.edu/biomed/genedoc/). The alignment is shown in Figure 13. A schematic representation of the repeat element in the *bmp*4 gene is shown in Figure 12. The element shows an interspersed distribution in the genome and its length (500 bp) is at the upper limit of the size range of SINEs. It consists of 4 conserved regions, separated by regions of low sequence complexity with varying length. The constant regions C1 and C4 are flanked by poly(A) and poly(T) stretches, respectively, as usually found at the 3' ends of SINEs. This repetitive element was tentatively called *arielle*.

In contrast to the mermaid repeat, no well conserved RNA polymerase III promoter could be found. Scanning the sequence with a tRNA identification software (tRNAscan-SE, Lowe and Eddy, 1997) did not identify a tRNA related structure. This software uses a combination of different algorithms to identify tRNAs on the basis of their promoters and their ability to form a secondary structure. A BLAST search of the Fugu genome did not identify any significant matches. Thus, while sharing some characteristics with other SINEs, it seems not to be a typical representative of this class of retroelements, and it is probably restricted to zebrafish and maybe close relatives.

Interestingly, the conserved region C1 has been found to be homologous to a genomic region close to a transposon insertion in an experimental system (Kawakami et al., 2000). This suggests that transposon insertion might be biased towards sites containing certain types of repeats. Constant region C3 contains 3 copies of a tandemly arrayed 18-bp sequence repeat. A copy of the arielle repeat in the bmp4 gene, which was analysed in more detail, was found to be delimited by a pair of direct and inverted repeats.

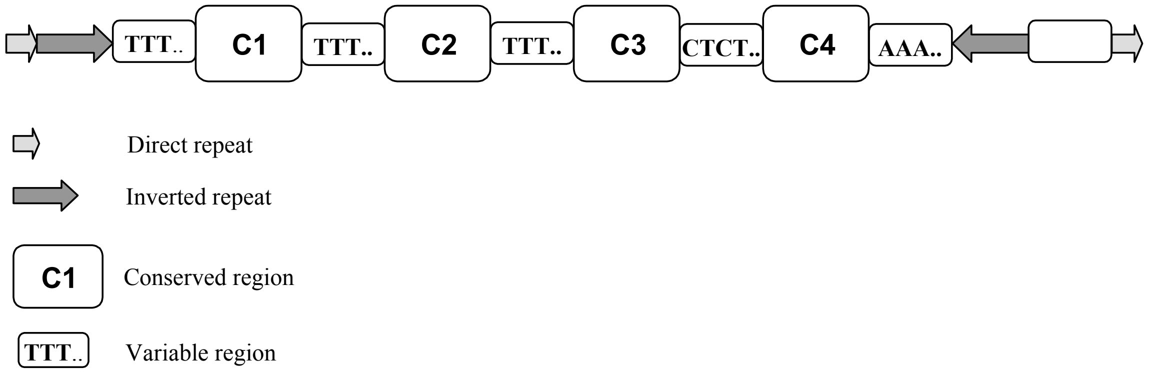
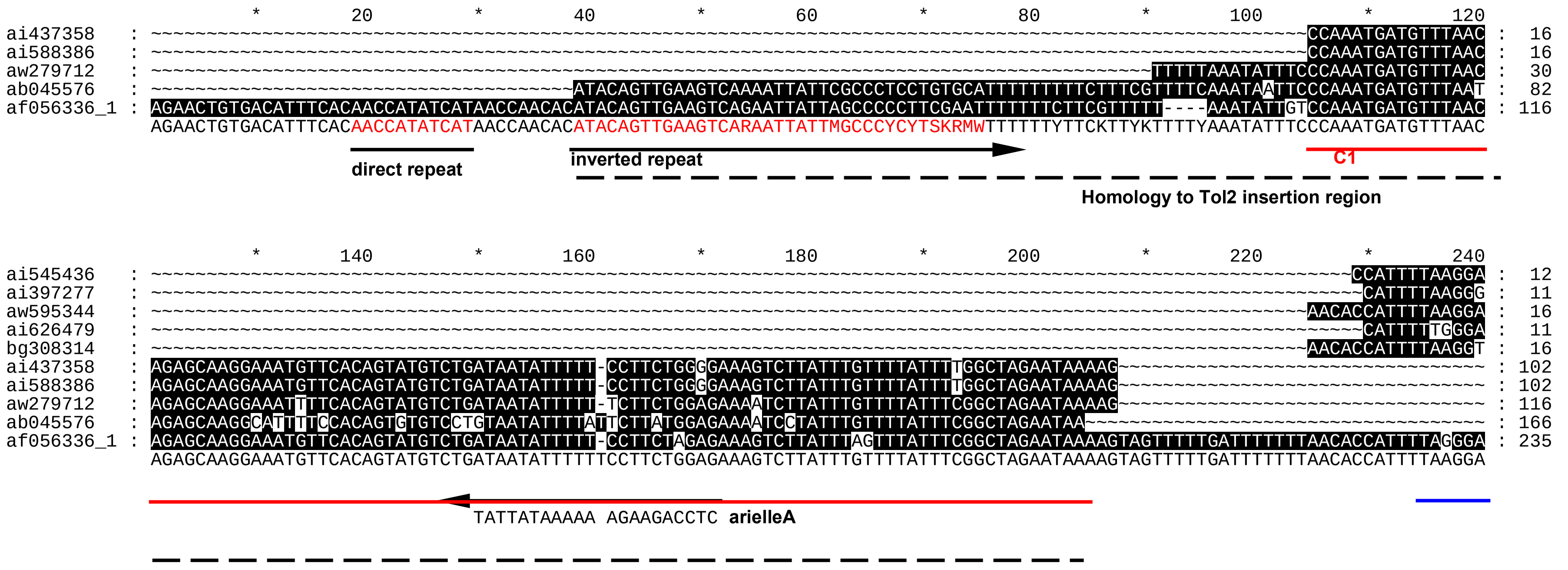


Figure 12

Organisation of the zebrafish *arielle* repeat copy found in the *bmp*4 gene. Conserved and variable regions, direct and inverted repeats are shown schematically. Sizes are not in scale. The repeat element consists of 4 conserved domains interrupted by variable regions. In two instances, a conserved sequence situated 3' of C4 is duplicated and found at the 5' end of the repetitive element in an inverted orientation. In the copy, which was first identified and used as query sequence, the repeat is flanked by a direct repeat. At the 3'-end, the direct repeat is separated from the inverted repeat by a nonrepetitive sequence.

Figure 13

(Next page)
Alignment of arielle repeats. The repeat copy in the zebrafish bmp4 gene was used as query (Genbank accession number AF056336). Primers used for repeat-anchored IRS PCR are shown at their binding site. Conserved nucleotides are shaded.



```
                      *        20         *        40         *        60         *        80         *       100         *       120
ai437358   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CCAAATGATGTTTAAC :  16
ai588386   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CCAAATGATGTTTAAC :  16
aw279712   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TTTTTAAATATTTCCCAAATGATGTTTAAC :  30
ab045576   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ATACAGTTGAAGTCAAAATTATTCGCCCTCCTGTGCATTTTTTTTTCTTTCGTTTTCAAATAATTCCCAAATGATGTTTAAT :  82
af056336_1 : AGAACTGTGACATTTCACAACCATATCATAACCAACACATACAGTTGAAGTCAGAATTATTAGCCCCCTTCGAATTTTTTTCTTCGTTTTT----AAATATTGTCCAAATGATGTTTAAC : 116
             AGAACTGTGACATTTCACAACCATATCATAACCAACACATACAGTTGAAGTCARAATTATTMGCCCYCYTSKRMWTTTTTTYTTCKTTYKTTTTYAAATATTTCCCAAATGATGTTTAAC

                      *       140         *       160         *       180         *       200         *       220         *       240
ai545436   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CCATTTTAAGGA :  12
ai397277   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CATTTTAAGGG :  11
aw595344   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AACACCATTTTAAGGA :  16
ai626479   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~CATTTTTGGGA :  11
bg308314   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AACACCATTTTAAGGT :  16
ai437358   : AGAGCAAGGAAATGTTCACAGTATGTCTGATAATATTTTT-CCTTCTGGGGAAAGTCTTATTTGTTTTATTTTGGCTAGAATAAAAG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 102
ai588386   : AGAGCAAGGAAATGTTCACAGTATGTCTGATAATATTTTT-CCTTCTGGGGAAAGTCTTATTTGTTTTATTTTGGCTAGAATAAAAG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 102
aw279712   : AGAGCAAGGAAATTTTCACAGTATGTCTGATAATATTTTT-TCTTCTGGAGAAAATCTTATTTGTTTTATTTCGGCTAGAATAAAAG~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 116
ab045576   : AGAGCAAGGCATTTTCCACAGTGTGTCCTGTAATATTTTATTCTTATGGAGAAAATCCTATTTGTTTTATTTCGGCTAGAATAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 166
af056336_1 : AGAGCAAGGAAATGTTCACAGTATGTCTGATAATATTTTT-CCTTCTAGAGAAAGTCTTATTTAGTTTATTTCGGCTAGAATAAAAGTAGTTTTTGATTTTTTTAACACCATTTTAGGGA : 235
             AGAGCAAGGAAATGTTCACAGTATGTCTGATAATATTTTTTCCTTCTGGAGAAAGTCTTATTTGTTTTATTTCGGCTAGAATAAAAGTAGTTTTTGATTTTTTTAACACCATTTTAAGGA
```

```
                    *       260         *       280         *       300         *       320         *       340         *       360
aw128711   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TAACTTGCCTAATTACCCTAACCTGCCTAGTCAACGTAATTAACCTAGTT :  50
ai601701   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TACAGAACAAACCATTGTTATACAATAACTTGCCTAATTACCCTAACC--TTTAGTTAACCTAATTAACCTAGTT :  73
be606085   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AGTCTTCAGAACAAACCATTGTTATAAAATAACTTGCCTAATTACCTTAACTTAACTAGTTAACTTGATTAACCTAGTT :  79
be606105   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AGTCTTCAGAACAAACCATTGTTATAAAATAACTTGCCTAATTACCTTAACTTAACTAGTTAACTTGATTAACCTAGTT :  79
bg307783   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ATAGTCTACAGAATAACCCATCATTTTACAATAACTTGCCTAATTACCCTAACCTGCCTAGTTAACCTAATTAACCTAGTT :  81
ai601354   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~TTTTTTTCGTTAGTCTACAGAACAAACCATCGTTATACAATAACTTGCCTAATTACCCTAAACTGCCTAGTTAACCTAATTAACCTAGTT :  90
ai641569   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~TATTTTTTCCGATAGTCTACAGAACAAACCATCGTTATAAAATAACTGGCCTAATTGCCCTATCCTGCCTAGTTAACCTAATTAACCTAGTT :  92
dre251640  : CAAAATGTTTAGCTTCTTTAAGCTATGTTTTTTTTT-CGATAGTCTGTCGAACAAACCATCCTTATACAATAACTTGCCTAATTACCCTAACCTGCCTAGTTACCCTAATTAACCTAGTT : 119
ai545436   : CAAAATTATTAGTCTCTTTAAGCAATATTTTTTT---CAATAGTATAATAAACAAACCATTATTATGTAATAACTTGCCTAAATACCCTAACCAGCCTAGTTAACCTAATTACCCTAGTT : 129
ai397277   : CAAAATTATTAGCCCCTTTAAGCGTTATATTTTTTT-CGATAGTCTACAGAACAAACCATCGTTATACAATAACTTGCCTAATTACCCTAACCAGCCAAGTTAACCTAATTAACCTAGTT : 130
aw595344   : CAAAATTATTAGCCCCTTTAGGCTAATTTTATTT--TCGATAGTCTACAGAACAAATCATCATTATACATTAACTTGCCTAATTACCCTAACCTGCCTAGTTAACCTAATTAACCTAATT : 134
ai626479   : CAAAATTATTAGCCCCTTTAAGCTATTTTTTTTT--TCCATATTCTTCAGAACAAACCATCGTTATACAATAACTTGCCTAATTACCCTAACCTGCCTAGTCCACCTTTTTAACTTAGTG : 129
bg308314   : CAATATTATTAGCCCCTTTAAGCTTAATTTTT----CGATTAGTCTCCAGAACAAACCATCATTATGCAATAACTTGCCTAATTACCCTAAC-------TTTACCCTAATTACCCTAGTG : 125
ai437358   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ai588386   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
aw279712   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ab045576   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
af056336_1 : CAAAATGGTTAGCCCCTTTAAGCTATATTTTT---CTCGATAGTCTACAGAACAAACCATCATTATACAATAACTTGTCTAATTACCCTAACCTGTCTAGTTAACCTAATTAACCTAGTT : 352
             CAAAATTATTAGCCCCTTTAAGCTATATTTTTTTTTTCGATAGTCTACAGAACAAACCATCGTTATACAATAACTTGCCTAATTACCCTAACCTGCCTAGTTAACCTAATTAACCTAGTT
```

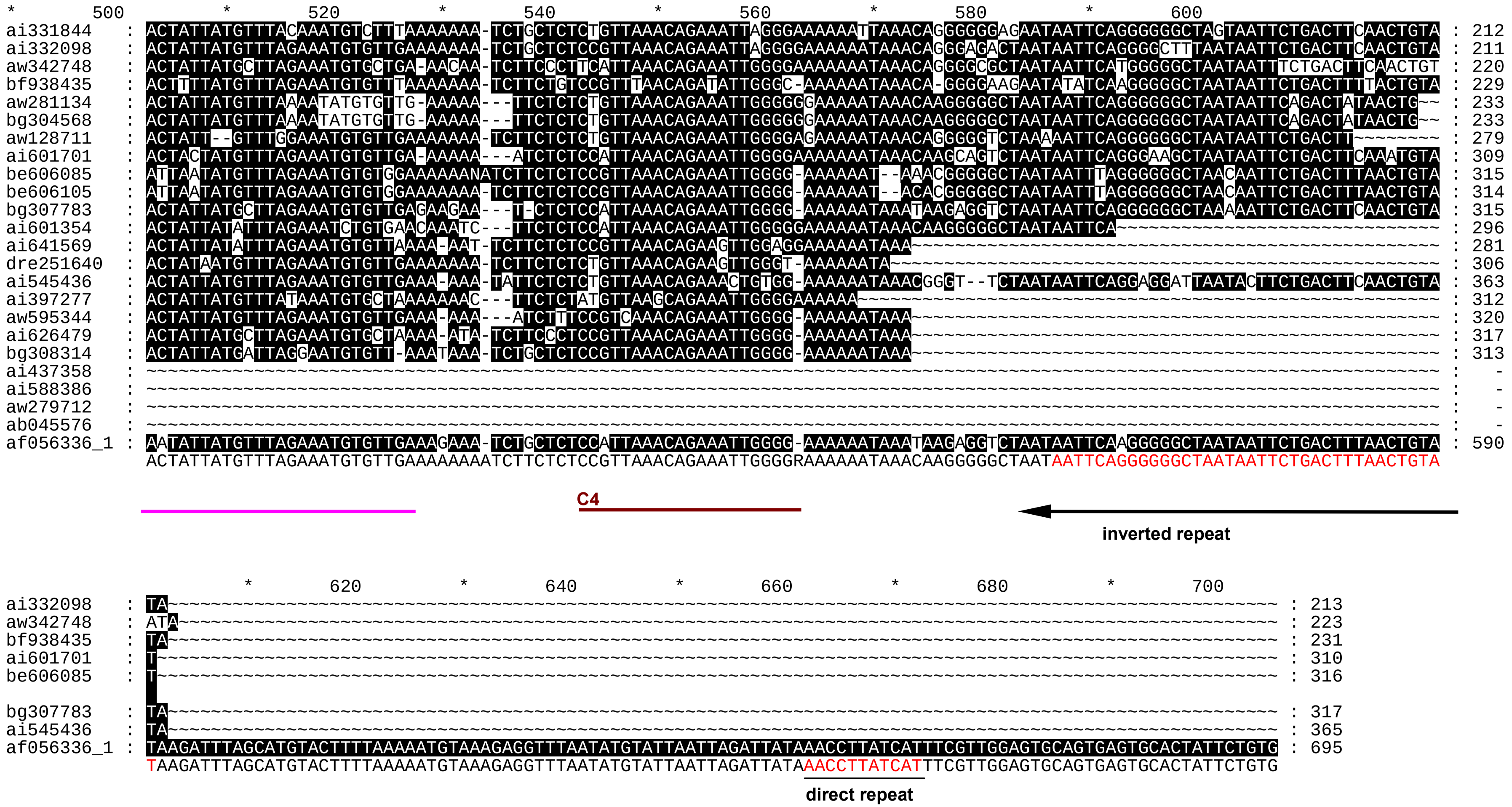
C2

C3

tandem repeats

```
                    *       380         *       400         *       420         *       440         *       460         *       480
ai331844   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~ATAGAAGTGCCTTGAAAAATATCAAGTCAGATATTATTTACTGTCATCATGGCAAAGATAAAAGAAATCAGCTATTAGAAATGAGTTATTAAA :  93
ai332098   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~TAGAAGTGTCTTGAAACATATCTAGTCTAATATTATTTACTGTCATCATGACAAAGATAAAATAAATGAGGTATTAGAAATGAGTTATTAAA :  92
aw342748   : ~~~~~~~~~~~~~~~~~~TTAAGCTGTATAGAATTGTCTGGAAAAATATCAAGTAAAATACTATATACTGTCATCATGTCAAAGATAAAATAAATCAGTTATTAGAAATGAGTTATTAAA : 102
bf938435   : ~~~~~~~~AAATATCACTTTAAGCTGTATAGAAGTGTCTCGAAAAATATCCAGTCAACTATTATTTACTGTGATCATGGCAAAGATTAAATAAATCAGTTATTAGAAATGAGTTATTAAA : 112
aw281134   : AAGCCTTTAAATGTCACTTGAAGCTGTATAGAAGTGACTT-AAAAATATCTAGTCAAATATTATTTACTGTCATCATGGCAAAAATAAAATATATCAGTTATTAGAAATGAGTTACTAAA : 119
bg304568   : AAGCCTTTAAATGTCACTTGAAGCTGTATAGAAGTGACTT-AAAAATATCTAGTCAAATATTATTTACTGTCATCATGGCAAAAATAAAATATATCAGTTATTAGAAATGAGTTACTAAA : 119
aw128711   : AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTATCTTGAAAAATATCTATTAAAATATTATTTATTGTCATCATGGCAAAGATAAAATAAATCAGTTATTGGAAATGAGTTATTAAA : 170
ai601701   : AAGCCTCTAAATGTCACTTTAAGATGTGTAGAAGTGTCTTGAAAAAAATCTAGTCAAATATTATATACTGTCATCATGGCAAAGATAAAATAAATCAGTTATTAGAAATTAGTTATTAAA : 193
be606085   : AGGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTT-AAAAATATCTAGTCAAATATTATTTACTGTCATCATGGCAAAGATAAAATAAATCAGTTTTTAGAAATGAGTTACTAAA : 198
be606105   : AGGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTT-AAAAATATCTAGTCAAATATTATTTACTGTCATCATGGCAAAGATAAAATAAATCAGTTTTTAGAAATGAGTTACTACC : 198
bg307783   : AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTT-AAAAATATCTAGTCAAATATTATTTACTGTCACCATGGTAAAGATAAAATAAATCCATTATTCAAAATGAGTTATTACA : 200
ai601354   : AAGCCTTTGAATGTCACTTTAAGCTGTATAGAAGTGTCTT-AAAAATATCTAGTCAAATATTATTTTCTGTAATTATGGCAAAGATAAAATAAATCAGTTATTAGAAATGAGTTATTAAA : 209
ai641569   : AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTTTAAAAATATCTAGTCAAATATTATGCACTGTGTTCATGGCAAAGATAAAATAAATCAGTTATTTTAAATGAGTTATTAAA : 212
dre251640  : AACTCTTTACATGTCACTTTAAGCTGTATAGAAGTGTCTTGAAAAATATCTAGTAAAATATTATTTACTCTCATCATGGCAAAGATAAATTAAATCAGTTATTAGAAATTAATTATTAAA : 239
ai545436   : AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTTGAAAAATATCTAGTC-AATATTATGTAGTGTCATAATGGCAAAAATAAAATAAATCAGTTCTTAGAGATGAGTTATTAAA : 248
ai397277   : AAGCTTTTAAATGTCCCTTTAAGCTGTATAGAAGTGTCTT-AAAAATACCTAGTCAAATATTATTTACTGTCATCATGGCAAATATAAAATAAATCAGTTATTAGAAATAAGTTATTAAA : 249
aw595344   : AAGCCCTTAAATGTCACTTTAAGATCTTTAGAAGTGTCTTGAAAAATATCTAGTCAAATATGATTTACTGTCATCATGGCAAAGATTAAATAAATCAGTTATTAGAAATTAGTTATTAAA : 254
ai626479   : AAGCCTTTAAATGTCACTGTAAGCTGTATAGAAGTGTCTTGAAAAATATCTAGTAAAATATTATTTACTGTCATCATGGCAAAGATAAAATAAATCAGTTATTAGAAATAAGTTTTTAAA : 249
bg308314   : AAGCATTTAAATGTCTCTTTAAGCTGTATAGAAGTGTCTTAAAGAATATCTAGTCTAATATTATGTGCTGTCATCATGACAAAGAGAAAATAAATCAGTTATTAGAGATGAGTTATTGAC : 245
ai437358   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ai588386   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
aw279712   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ab045576   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
af056336_1 : AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTTGAAAAATATAAAGTAAAATATTGTTTACAGTCATCGTGGTAAAGATAAAATTAATCCGTTATTAGAAATGAGTTATTAAA : 472
             AAGCCTTTAAATGTCACTTTAAGCTGTATAGAAGTGTCTTGAAAAATATCTAGTCAAATATTATTTACTGTCATCATGGCAAAGATAAAATAAATCAGTTATTAGAAATGAGTTATTAAA
```

ATATKATKTRCTGTCATCATG **arielleB**



```
*        500          *       520          *       540          *       560          *       580          *       600
ai331844   : ACTATTATGTTTACAAATGTCTTTAAAAAAA-TCTGCTCTCTGTTAAACAGAAATTAGGGAAAAAATTAAACAGGGGGGAGAATAATTCAGGGGGGCTAGTAATTCTGACTTCAACTGTA : 212
ai332098   : ACTATTATGTTTAGAAATGTGTTGAAAAAAA-TCTGCTCTCCGTTAAACAGAAATTAGGGGAAAAAATAAACAGGGAGACTAATAATTCAGGGGCTTTAATAATTCTGACTTCAACTGTA : 211
aw342748   : ACTATTATGCTTAGAAATGTGCTGA-AACAA-TCTTCCCTTCATTAAACAGAAATTGGGGAAAAAAATAAACAGGGGCGCTAATAATTCATGGGGGCTAATAATTTCTGACTTCAACTGT : 220
bf938435   : ACTTTTATGTTTAGAAATGTGTTTAAAAAAA-TCTTCTGTCCGTTTAACAGATATTGGGC-AAAAAATAAACA-GGGGAAGAATATATCAAGGGGGCTAATAATTCTGACTTTTACTGTA : 229
aw281134   : ACTATTATGTTTAAAATATGTGTTG-AAAAA---TTCTCTCTGTTAAACAGAAATTGGGGGAAAAAATAAACAAGGGGGCTAATAATTCAGGGGGGCTAATAATTCAGACTATAACTG~~ : 233
bg304568   : ACTATTATGTTTAAAATATGTGTTG-AAAAA---TTCTCTCTGTTAAACAGAAATTGGGGGAAAAAATAAACAAGGGGGCTAATAATTCAGGGGGGCTAATAATTCAGACTATAACTG~~ : 233
aw128711   : ACTATT--GTTTGGAAATGTGTTGAAAAAAA-TCTTCTCTCTGTTAAACAGAAATTGGGGAGAAAAATAAACAGGGGGTCTAAAAATTCAGGGGGGCTAATAATTCTGACTT~~~~~~~~ : 279
ai601701   : ACTACTATGTTTAGAAATGTGTTGA-AAAAA---ATCTCTCCATTAAACAGAAATTGGGGAAAAAAATAAACAAGCAGTCTAATAATTCAGGGAAGCTAATAATTCTGACTTCAAATGTA : 309
be606085   : ATTAATATGTTTAGAAATGTGTGGAAAAAANATCTTCTCTCCGTTAAACAGAAATTGGGG-AAAAAAT--AAACGGGGGCTAATAATTTAGGGGGGCTAACAATTCTGACTTTAACTGTA : 315
be606105   : ATTAATATGTTTAGAAATGTGTGGAAAAAAA-TCTTCTCTCCGTTAAACAGAAATTGGGG-AAAAAAT--ACACGGGGGCTAATAATTTAGGGGGGCTAACAATTCTGACTTTAACTGTA : 314
bg307783   : ACTATTATGCTTAGAAATGTGTTGAGAAGAA---T-CTCTCCATTAAACAGAAATTGGGG-AAAAAATAAATAAGAGGTCTAATAATTCAGGGGGGCTAAAAATTCTGACTTCAACTGTA : 315
ai601354   : ACTATTATATTTAGAAATCTGTGAACAAATC---TTCTCTCCATTAAACAGAAATTGGGGGAAAAAATAAACAAGGGGGCTAATAATTCA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 296
ai641569   : ACTATTATATTTAGAAATGTGTTAAAA-AAT-TCTTCTCTCCGTTAAACAGAAGTTGGAGGAAAAAATAAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 281
dre251640  : ACTATAATGTTTAGAAATGTGTTGAAAAAAA-TCTTCTCTCTGTTAAACAGAAGTTGGGT-AAAAAATA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 306
ai545436   : ACTATTATGTTTAGAAATGTGTTGAAA-AAA-TATTCTCTCTGTTAAACAGAAACTGTGG-AAAAAATAAACGGGT--TCTAATAATTCAGGAGGATTAATACTTCTGACTTCAACTGTA : 363
ai397277   : ACTATTATGTTTATAAATGTGCTAAAAAAAC---TTCTCTATGTTAAGCAGAAATTGGGGAAAAAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 312
aw595344   : ACTATTATGTTTAGAAATGTGTTGAAA-AAA---ATCTTTCCGTCAAACAGAAATTGGGG-AAAAAATAAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 320
ai626479   : ACTATTATGCTTAGAAATGTGCTAAAA-ATA-TCTTCCCTCCGTTAAACAGAAATTGGGG-AAAAAATAAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 317
bg308314   : ACTATTATGATTAGGAATGTGTT-AAATAAA-TCTGCTCTCCGTTAAACAGAAATTGGGG-AAAAAATAAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 313
ai437358   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ai588386   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
aw279712   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
ab045576   : ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ :   -
af056336_1 : AATATTATGTTTAGAAATGTGTTGAAAGAAA-TCTGCTCTCCATTAAACAGAAATTGGGG-AAAAAATAAATAAGAGGTCTAATAATTCAAGGGGGCTAATAATTCTGACTTTAACTGTA : 590
             ACTATTATGTTTAGAAATGTGTTGAAAAAAAATCTTCTCTCCGTTAAACAGAAATTGGGGRAAAAAATAAACAAGGGGGCTAATAATTCAGGGGGGCTAATAATTCTGACTTTAACTGTA
```

```
                      *       620          *       640          *       660          *       680          *       700
ai332098   : TA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 213
aw342748   : ATA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 223
bf938435   : TA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 231
ai601701   : T~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 310
be606085   : T~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 316

bg307783   : TA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 317
ai545436   : TA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ : 365
af056336_1 : TAAGATTTAGCATGTACTTTTAAAAATGTAAAGAGGTTTAATATGTATTAATTAGATTATAAACCTTATCATTTCGTTGGAGTGCAGTGAGTGCACTATTCTGTG : 695
             TAAGATTTAGCATGTACTTTTAAAAATGTAAAGAGGTTTAATATGTATTAATTAGATTATAAACCTTATCATTTCGTTGGAGTGCAGTGAGTGCACTATTCTGTG
```

4.1.3 Tests of IRS-PCR Primers

An optimal primer for IRS PCR is one that yields the largest amount of products with a wide size range. There should be no dominating single product. The desired PCR amplicon appears as a uniform smear on a gel. Different primer binding sites on different repeats were tested for their use in an IRS-PCR. Besides different mermaid primers, a primer binding to the angel repeat (Izsvak et al., 1999) and two primers binding to the arielle repeat were tested (Figure 14).

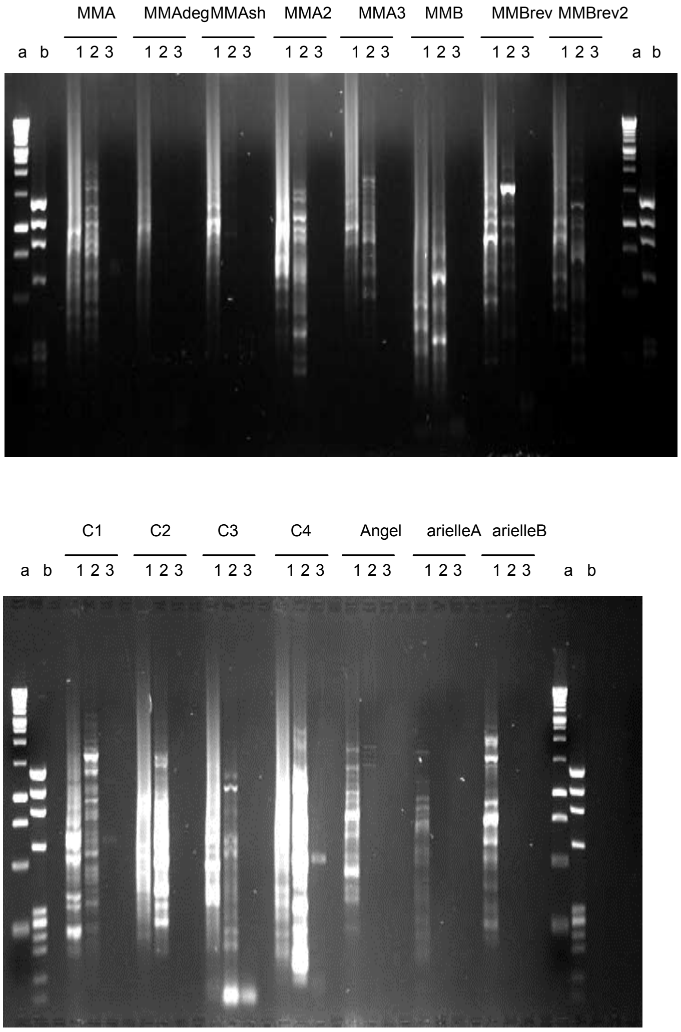

MMA
MMAdeg
MMAsh
MMA2
MMA3
MMB
MMBrev
MMBrev2
a b 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 a b
C1
C2
C3
C4
Angel
arielleA
arielleB
a b 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 1 2 3 a b

Figure 14

IRS PCR using different primers binding to the DANA/mermaid element (MMA-C4), the Angel repeat and the arielle repeat (arielleA and arielleB). As template was used: (1) 25 ng zebrafish genomic DNA, (2) 25 ng mouse genomic DNA, (3) no template. Mermaid specific primers generate PCR products in zebrafish and also in mouse DNA, which reflects the abundance of this repeat in vertebrates. In contrast primers binding the Angel and arielle repeat are specific for the zebrafish.

An additional test on the different PCR primers was performed using zebrafish PAC DNA as template. By this experiment, the number of potential PCR products for the whole genome can be estimated. Additionally, the tendency of a primer to amplify products from the vector is tested (Figure 15).

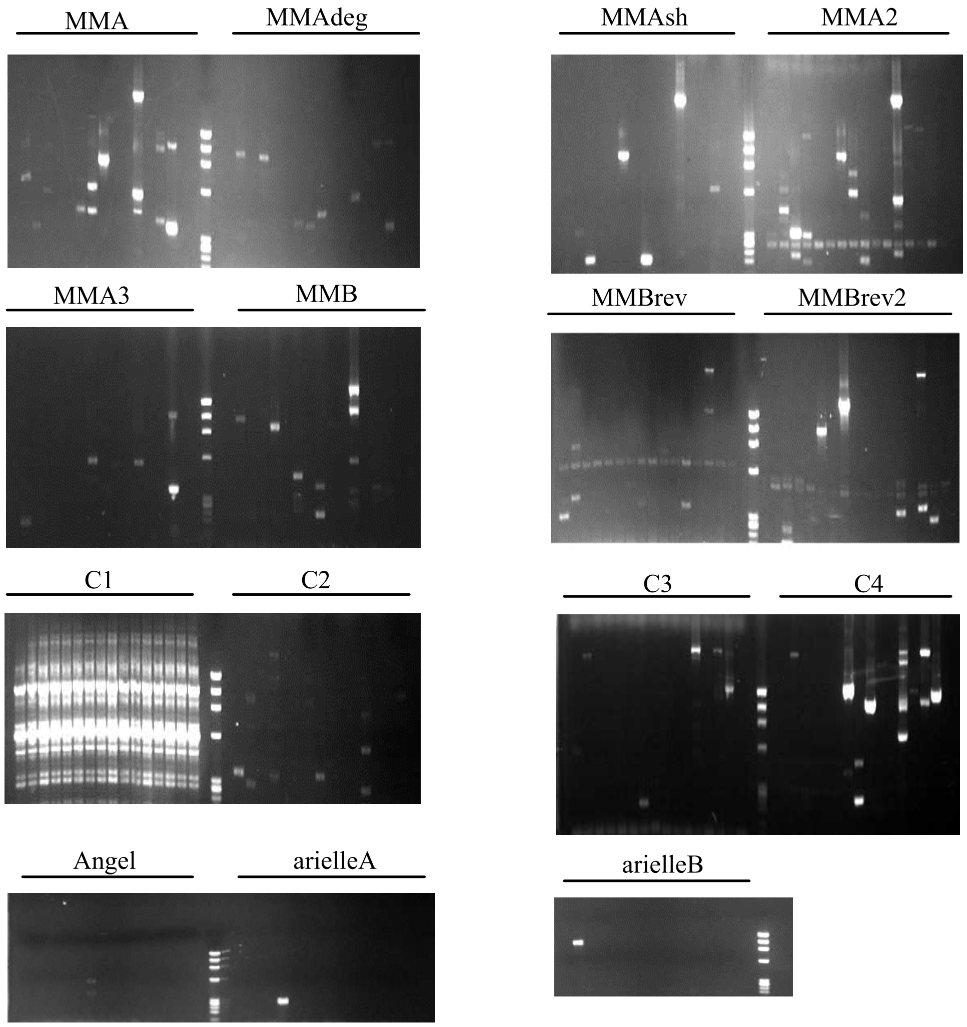


Figure 15

IRS-PCR on zebrafish PACs using DANA/mermaid (MMA-C4), Angel and arielle repeat primers. 15 different PACs and the PAC vector without insert were used as template. DNA size standard (middle lane): X174/BsuRI digest.

Figure 15 illustrates that IRS-PCR products from different primers differ in length and abundance, even when they have overlapping annealing sites. For example, the primer MMA amplifies 15 products from 15 PACs. Given an average insert length of 120 kb per PAC and a zebrafish genome size of 1.7x10^9 bp, this is equivalent to 14,000 potential IRS products per haploid genome using this primer. The real number of products using genomic DNA as template is expected to be considerably lower, because the amplification of IRS sequences that are long or do not have the exact primer binding site is suppressed in a multiplex PCR due to its competitive nature.

Other mermaid primers, for example MMAdeg have a significantly lower product yield. A uniform pattern of unspecific bands can be seen with the primer binding to the C1 region of DANA/mermaid. Angel and arielle repeat primers yield only very few products. From these results, it was concluded, that primers binding to the C2 and the C4 region of DANA/mermaid are best suited for IRS-PCR. Additionally, the use of combinations of different primers might increase the number of PCR products. During the further course of this work, the previously published primers MMA and MMB (Shimoda et al., 1996) were used. IRS-PCR on PACs and YACs was also used to determine the optimal PCR conditions for a given primer. The resulting protocol is described in 3.2.23.

4.1.4 Construction of an IRS marker library

IRS-PCR was performed on total genomic DNA of 5 zebrafish strains (AB, India, Tue, SJD, WIK) with the primers MMA and MMB having a $(CUA)_4$ extension at its 5'-end. A library of the complex amplicons was constructed by cloning the products using the pAMP10 cloning system (see 3.2.22). The ligation reaction was used to transform *E. coli* DH10B cells. Colonies were plated and picked robotically, and transferred in 384-well microtiter plates. From each strain 3072 Clones (8 plates) were picked from the MMA library, and 2304 Clones (6 plates) from the MMB library. This results in 27,648 clones, the number of spots which fit on a 22x22 cm membrane if spotted in duplicate in a 5x5 scheme.

Plasmid inserts of the library were PCR amplified in 384 well plates in a water bath using the vector specific primers M13 forward and 3/86. A subset of the PCR reactions was checked on a gel for testing the success of the cloning procedure, the amount of amplified DNA, and the average insert size Figure 16.

In addition to the insert, the primers also amplify ca. 230 base pairs of vector DNA. Because almost all PCR products exceed this length, the cloning efficiency can be considered high. Insert length is in the range of 200 and 2000 bp. The average insert length is estimated to be around 700 bp. For the further characterisation of the marker library by hybridisation, all of the 27,648 PCR products were spotted on 22x22 cm nylon membranes in duplicate. In the centre of each 5x5 block, concentrated genomic salmon sperm DNA is spotted as a guide spot.

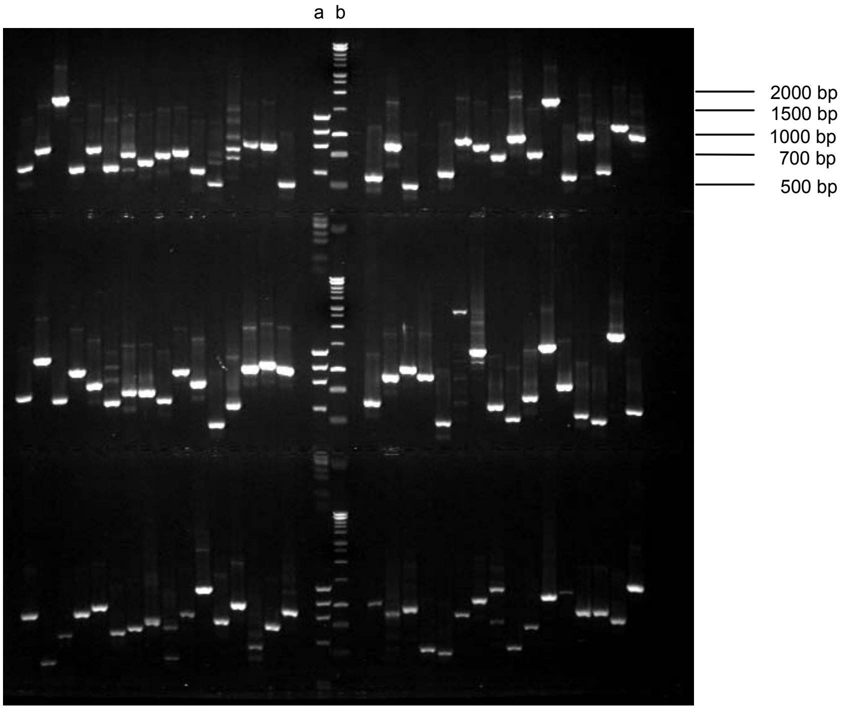


Figure 16

PCR of 96 clones of the mermaid marker library. Molecular weight standards are (a) X174/BsuRI digest, (b) 1 kb DNA ladder (Promega).

4.1.5 Characterisation and normalisation of the IRS marker library by oligonucleotide fingerprinting

By random cloning of a complex PCR amplicon, one can expect that products are cloned several times, depending, how abundant they are in the PCR amplicon. Identification of redundant clones can therefore increase the use of the library considerably. In addition to that, identification of identical products permits the determination of products, which are amplified in all five strains, and those products, which occur only in a subset of the strains used (polymorphic products). The latter can be used as genetic markers.

Such a normalisation can be carried out by sequential hybridisation of single clones on the library filters (backhybridisation). All clones hybridising with a given probe are listed and probes for the next round of hybridisation are chosen among the clones that have not been hit in previous hybridisations (sampling without replacement). This method becomes more and more inefficient, as the complexity and the size of the library increases, because thousands of hybridisations have to be done, and each hybridisation has to be followed by an analysis step.

A more parallelised approach is oligonucleotide hybridisation fingerprinting (ONF, Maier et al., 1994, Drmanac et al., 1996, Meier-Ewert et al., 1998, Radelof et al., 1998). This is based on the sequence-specific hybridisation of short oligonucleotide probes to high-density arrays of DNA. A short oligonucleotide is hybridised to a library filter under stringent conditions and those clones that are hit are identified by image analysis. Hybridisations are sequentially performed with 100-300 oligonucleotides. To each clone, a fingerprint of hybridisation results is assigned. The fingerprints (or vectors) of each clone are compared and clones having identical or similar fingerprints are grouped together (clustered), assuming that they have identical sequence (Herwig et al., 1999). Figure 17 shows a schematic outline of ONF analysis.

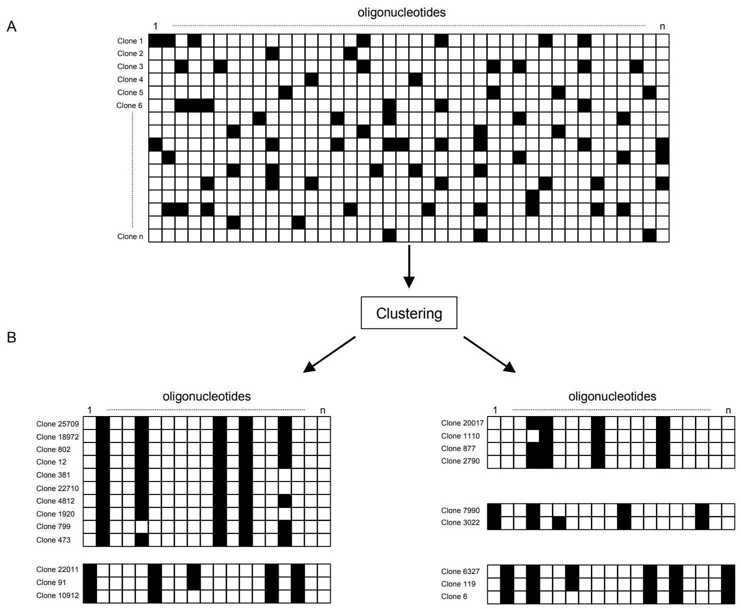


Figure 17

Schematic representation of the oligonucleotide fingerprinting process. A) Each horizontal line represents the nucleotide sequence of a clone. The individual clones are listed vertically (clone 1-n). Each vertical column shows which oligonucleotide sequence (oligonucleotides 1-n) matches the given clone (black = positive hybridisation result) The combination of positive and negative hybridisation results for a clone represents sequence-dependant fingerprint or vector.

B) The fingerprints of all clones (clones 1-n) are clustered by pairwise comparisons. Identical or similar fingerprints are grouped in clusters while unique fingerprints remain as singletons.

From an information theoretical point of view, oligonucleotides which match 50% of the clones would be the most efficient to use, because each probe would partition the library on two equally sized subsets, thus requiring a minimal number of hybridisations. A set of 20 such probes could partition the library into 2^{20} (1,048,576) unique fingerprints (Herwig et al., 2000). It has been shown experimentally using a cDNA library, that a hexamer oligonucleotide hybridises to 10-20% of the clones in a reliable and sequence-specific manner. To obtain an optimal stability and specificity of the duplex formation, together with a high hybridisation rate, pools of 16 decamers with a common octamer core were used (i.e. NXXXXXXXXN). The oligonucleotide probes were labelled with ^{33}P at the 5' end using T4 polynucleotide kinase. Figure 18 shows an oligonucleotide hybridisation of the zebrafish IRS marker library.

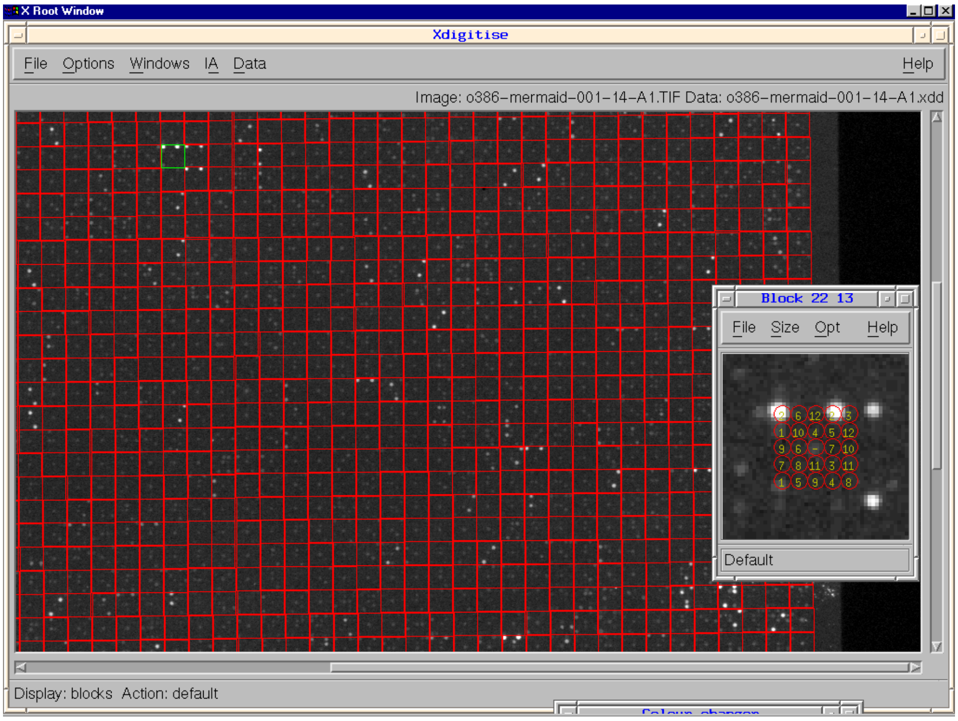


Image: o386–mermaid–001–14–A1.TIF Data: o386–mermaid–001–14–A1.xdd

Display: blocks Action: default

Figure 18

High-density grid filter hybridisation of the zebrafish IRS marker library using the oligonucleotide o386 (NCCATCTTCN). The image file is displayed by the image analysis package Xdigitise. Each 5x5 block is marked by a square. The small window shows a magnified single block. A duplicate clone with a positive hybridisation signal is visible. In the centre of each block, there is a weak signal of the salmon sperm guide dot.

Following hybridisation, the membranes are exposed to phosphor storage screens for 3-16 hours. The screens are scanned at 176 µm resolution on a Molecular Dynamics PhosphorImager. The resulting 16 bit TIF format image files are transferred to a local DEC Alpha Unix workstation, where they are image analysed using the HFA image analysis software. Block and clone positions are identified by means of the salmon sperm guide dots, and the intensity value of each clone is saved. Subsequently, the correlation of the duplicated spots is determined as a measure of hybridisation quality, and spot intensities are normalised for experimental variables such as position effects on the filter, different amounts of DNA of each clone etc. The normalisation is carried out using a double ranking method: In step one, the signals of all clones with the same oligonucleotide in a single hybridisation experiment are compared and ranked. The strongest signal is assigned a value of 1, the weakest 0, and all other clones in regular values of 1/N (where N equals the number of clones on a filter). In step two, the ranks of a clone in different hybridisations is ranked itself. The highest rank is assigned a value of 1, the weakest 0. The final output is a hybridisation matrix containing normalised intensities for all clones and probes. Hybridisation vectors of all clones were compared to each other, and clones with similar vectors were grouped together. These clusters represent clones having the same sequence. Clones with unique fingerprints are designated singletons.

A series of (ca. 120) oligonucleotide hybridisations was performed on the high-density IRS library filters, and 90 hybridisations, which were of sufficient quality for image analysis, were selected for cluster analysis. If large cDNA libraries are fingerprinted, the number of hybridisations is usually higher than that (around 200), but one can assume, that the analysis of IRS marker libraries requires less effort for several reasons: (1) The library is considerably less complex than a cDNA library. The upper limit for the number of products has been determined to be 14,000 by PCR on PACs (see 4.1.3). (2) The PCR products have a defined length, delimited by the mermaid binding sites, while cDNA clones from the same mRNA vary in length due to incomplete reverse transcription, making cluster analysis more difficult. (3) IRS PCR products are supposed to be random representations of the genome, while cDNAs belong to families of genes having sequence similarities or being splice variants. This also challenges the analysis and requires a large number of oligonucleotide hybridisations to be solved.

Clustering of data was performed repeatedly with different stringency parameters. As a control, hybridisations of IRS clones to the whole library (backhybridisations) were performed to identify homologous clones. Consistency of clustering and backhybridisation results was tested, determining the optimal clustering parameters.

The clustering analysis thought to be optimal partitions the mermaid marker library in 1604 cluster with two or more members and 9345 singletons. The clusters are ordered and numbered according to their size. Figure 19 illustrates how the hybridisation vectors look like for the clusters 12, 148, 712.

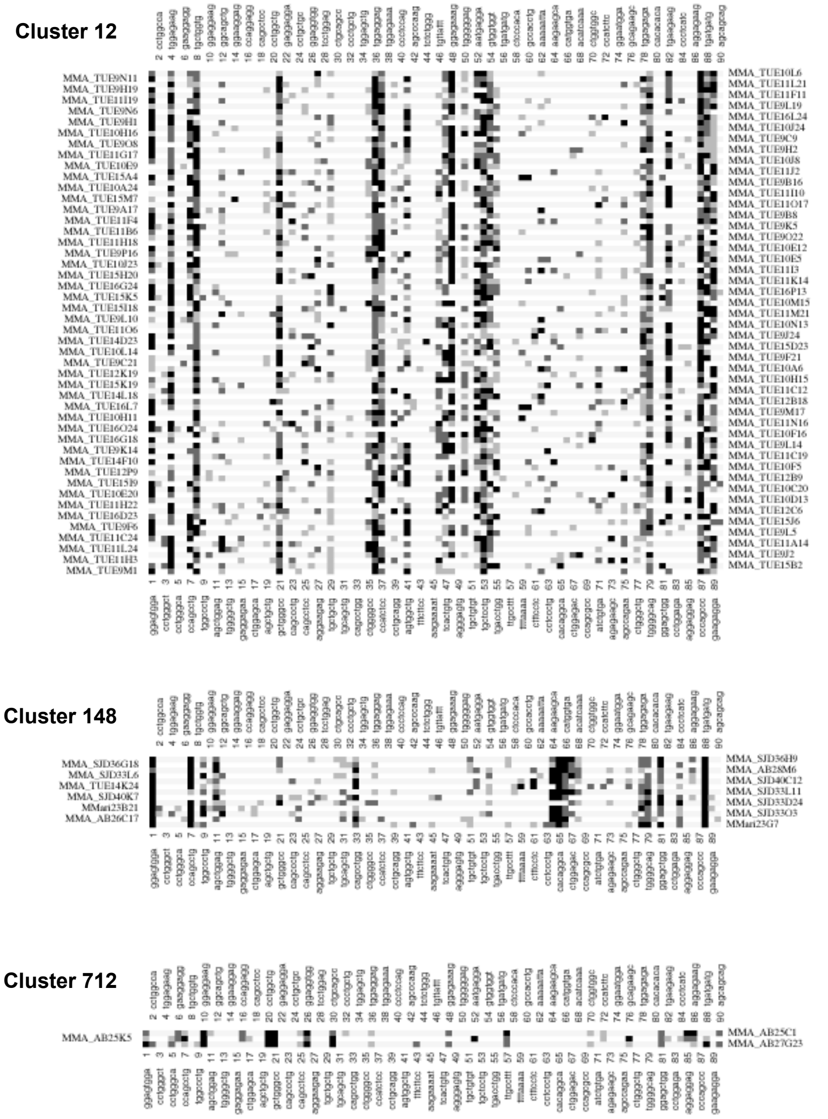


Figure 19

Visualisation of the hybridisation matrices of the clusters 12, 148 and 712. The vertical axis represents the clones in the cluster, the horizontal axis represents the oligos used. Grey levels represent the hybridisation intensities.

Cluster 1 consists of 316 clones, cluster 2 of 313 clones, cluster 3 of 273 clones etc. At the other end of the seize scale there are 130 clusters having 4 members, 239 clusters with 3 members and 789 clusters with 2 clones. The numbers of clusters is plotted against the cluster size in Figure 20

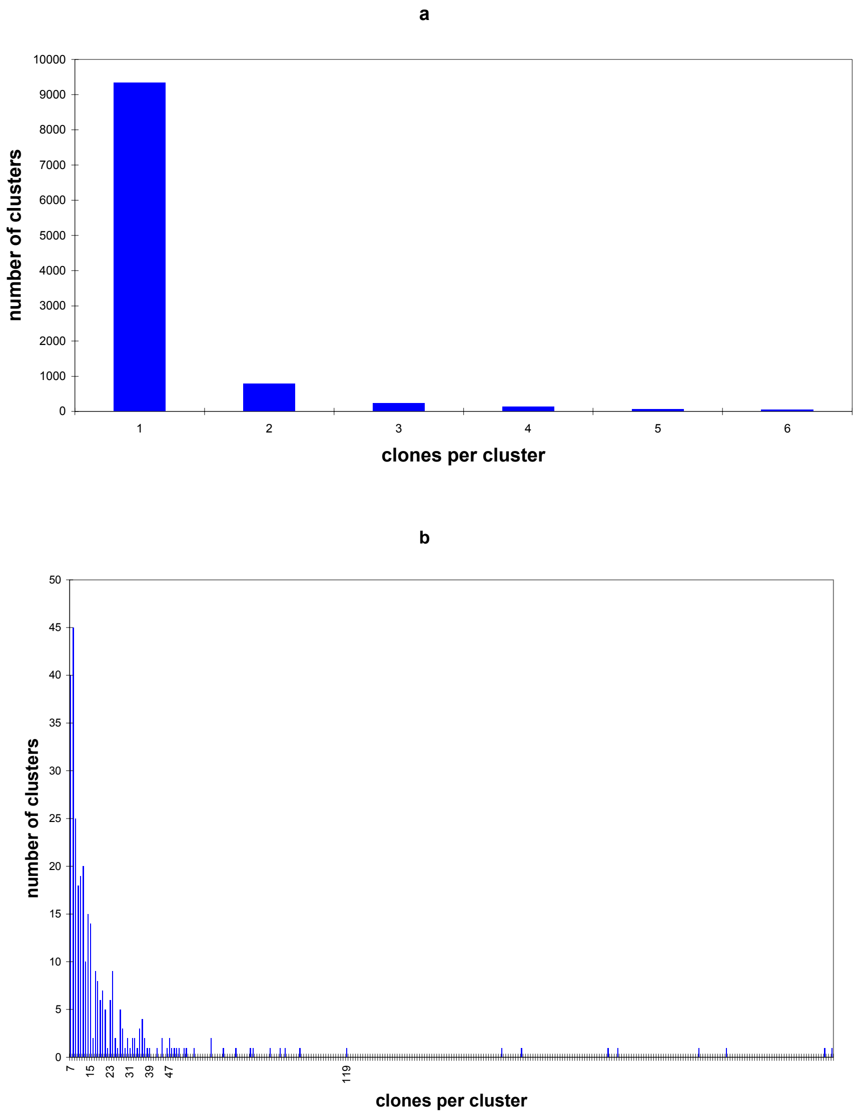


Figure 20

Numbers of clusters [y-axis] plotted against cluster size [x-axis]. (a) Singletons and the five smallest clusters (1-6 members) (b) Clusters with more than 6 members, scale differs from (a)

22 IRS clones from different clusters of different size were picked and used for control hybridisations. The image files were scored by hand using the Visual Grid software, developed at the Max-Planck-Institute for molecular genetics. A grid was laid on the image to be able to identify plate positions of the clones. Clones showing a hybridisation signal were marked. A Perl script (Meier-Ewert, pers. communication) was used to compare control hybridisation results with clustering results, to assess the reliability of the clustering. Figure 21 shows the output for two clones, one in cluster 148 and the other in cluster 712.

```
MMA_SJD33L11

mapping clones of level 1 or stronger!

MMA_SJD40K7:  148;  MMA_SJD33L11:  148;  MMA_SJD40C12:  148;  MMA_AB26C17:  148;
MMA_SJD36G18:  148;
MMA_AB25D4:  116;  MMA_SJD33D24:  148;  MMA_SJD33L6:  148;  MMA_SJD36H9:  148;  MMA_AB28M6:
148;


                                    Sets   1   2   3   4

1   clones in cluster 116   size 15     0   0   0   0
9   clones in cluster 148   size 13     0   0   0   0

                                  Total    0   0   0   0

10/10 clones are split into 2 clusters


MMA_AB25C1


mapping clones of level 1 or stronger!

MM2ariA1:    -;  MMA_IND20L15: 1605;  MMA_AB25K5:  712;  MMA_AB25C1:  712;  MMA_AB27G23:
712;  MMA_AB25O13: 1605;

                                    Sets   1   2   3   4

3   clones in cluster 712   size 3      0   0   0   0
2   clones in cluster 1605  size 9345   0   0   0   0

                                  Total    0   0   0   0

5/6 clones are split into 2 clusters
```

Figure 21

Comparison of clustering results and control hybridisations using the clone MMA_SJD33L11 (cluster 148) and MMA_AB25C1 (cluster 712). The program lists the clones hit by the probe, and which clusters they belong to. It shows that the clustering with the given data results in some error. In the first example, clone MMA_SJD33L11 was used as a probe. This clone is a member of cluster 148. It hybridises to 10 clones, including itself. Of these 10 clones, 9 belong to cluster 148 and one is from cluster 116. Cluster 148 on the other hand has 13 members, of which four are not hit by MMA_SJD33L11 in the hybridisation. The hybridisation confirms that the IRS products are strain specific. The IRS product MMA_SJD33L11 is only amplified in the strains AB and SJD, while the sequence of MMA_AB25C1 is specific for the strains AB and IND.

As a further control, the quality of ONF based clustering was tested by sequencing of clone inserts from the IRS marker library and aligning of homologous sequences (see next chapter).

4.1.6 Characterisation of IRS-PCR products by sequence analysis

Clone inserts of the IRS marker library were sequenced for a threefold purpose: (1) To asses the quality of ONF clustering by comparing it to sequencing results. (2) To characterise the structure of inter-mermaid products. (3) To identify single nucleotide polymorphisms (SNPs) in orthologous IRS-PCR products of different strains (discussed in chapter 4.2.3).

In total 180 clones from 23 clusters were picked. PCR of plasmid inserts was performed to analyse the insert length and to generate templates for sequencing.

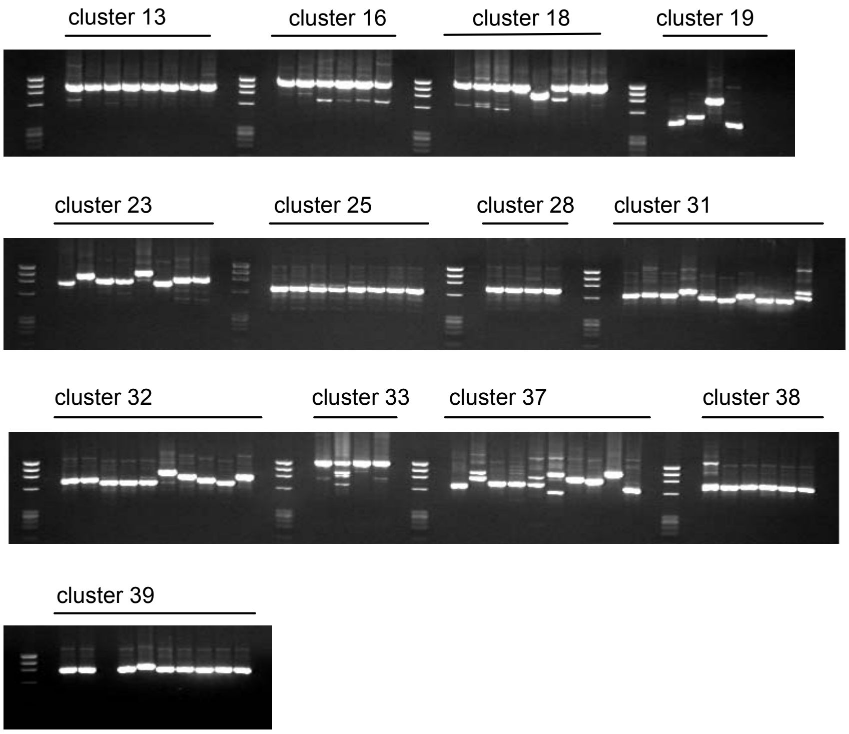


Figure 22

PCR of IRS clones for testing of uniformity of clusters. Uniformity of insert length corresponds well to sequence data (see Table 7).

Single-read sequencing of PCR products or purified plasmids was performed in both directions. Low quality parts of the sequences, i. e. base calls with a Phred value of less then 20 (i.e. error probability of > 0.01, Ewing and Green, 1998), were trimmed, and identical sequences were grouped together and aligned using the Staden software package (Staden et al., 2000).

Sequence analysis using *RepeatMasker* and *BLAST* reveals that in general, only one of the fragment ends is homologous to the mermaid sequence. This coincides with sequencing results from mouse IRS- products (L. Schalkwyk pers. communication). Obviously, the repetitive element serves as anchor for one primer, while the second site is misprimed. This is made possible by the rather non-stringent PCR priming conditions chosen. For this reason it is appropriate to speak of mermaid-anchored PCR rather than inter-mermaid PCR. Homology searches using BLAST showed that the highest sequence homologies exist to zebrafish ESTs and STSs (not shown). Some zebrafish genes were also found, but these genes are “the usual suspects” which are always found, when databases are queried with mermaid sequences. This shows that masking with RepeatMasker is not efficient in the case of zebrafish repeats.

ONF cluster	Clones sequenced	Clones in sequence clusters	Remarks
Cluster_1	10		CA repeat
Cluster_10	10	9;1	
Cluster_13	8	8	
Cluster_16	6	6	
Cluster_18	8	8	
Cluster_19	4	1;1;1	
Cluster_2	10	9;1	
Cluster_20	10	10	
Cluster_21	4	3;1	
Cluster_23	8		CA repeat
Cluster_25	8	8	
Cluster_28	4	4	
Cluster_3	4	1;1;1	
Cluster_31	10	1;1;4;3	
Cluster_32	10	1;2;7	
Cluster_33	4	3	
Cluster_37	12		CA repeat
Cluster_38	6	5	
Cluster_39	10	7;2	
Cluster_4	10	10	
Cluster_5	6	1;1;1;1;1;1	
Cluster_58	8	8	
Cluster_6	10	10	

Table 7

Clones from different ONF clusters were sequenced, and the sequences were clustered to confirm the ONF clusters. For example, 10 clones were sequenced from cluster 10, sequence clustering divided them into 2 groups, one cluster of 9 sequences, and one cluster of 1 sequence. From cluster 13, 8 clones were sequenced, and sequence clustering showed that all clones had the same sequence. In cases, where numbers of clustered sequences do not add up to the numbers of clones sequenced, sequence reactions failed. Clones of clusters 1, 23 and 37 turned out to consist mainly of a long CA repeat stretch without specific nonrepetitive sequence.

Both sequencing and control hybridisations suggest that oligonucleotide fingerprinting based clustering in general groups the right clones together. Nevertheless, clones grouped together in hybridisations are often split in different ONF clusters (underclustering). The reason for this might be erroneous clustering or cross-hybridisation, which occurs if clones contain stretches of similar or identical sequence. In this respect, ONF has a better discrimination than hybridisations using entire clones as probes. The second measure for the ONF process is the purity of clusters. The majority of clusters tested contain clones that are not correctly grouped in that cluster, as revealed by a negative hybridisation result (overclustering).

Sequencing confirms that ONF clusters are not pure. In 9 of 20 cases clusters contain unrelated sequences (Table 7). Because only a subset of clones in a cluster was sequenced, the real fraction of impure clusters is probably higher.

Impure clusters are due to erroneous ONF clustering. The resolution and discriminative power of ONF fingerprinting could be increased by hybridisation of more oligonucleotides. The quality of the clustering was however considered being sufficient for the purpose of this work. The IRS marker library is partially normalised and can be used as a source of non-redundant IRS markers. It is also possible to identify IRS products that are polymorphic between different strains (see Figure 21). These can be used for genetic mapping by hybridisation. An application for products occurring in more than one strain is the identification of SNPs for genotyping (see chapter 4.1.9).

4.1.7 Radiation hybrid mapping of IRS-PCR products

Two zebrafish radiation hybrid (RH) panels are available: LN54 (Hukriede et al. 1999) and T51 (Kwok et al. 1998, Geisler et al. 1999). IRS-PCR was performed on both panels, the IRS fragments were size-separated by gel electrophoresis (Figure 23A) and transferred on nylon membranes by the Southern blotting technique.

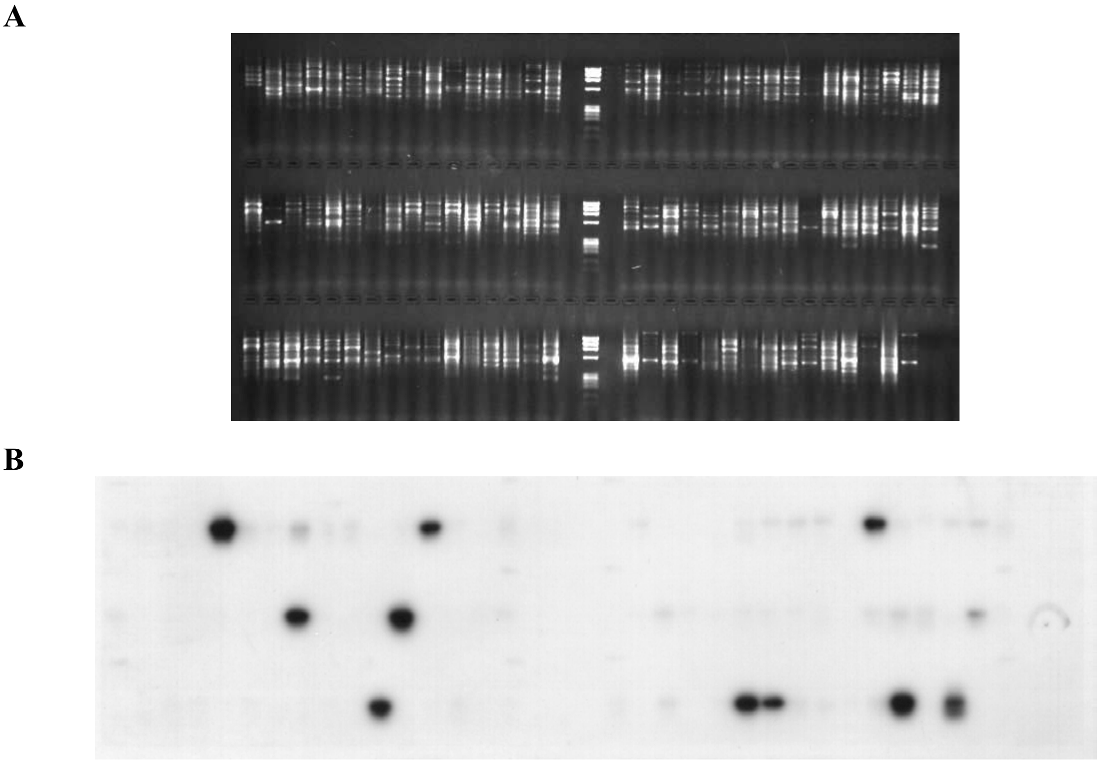


Figure 23

A: IRS PCR of the T51 Radiation hybrid panel.
B: Hybridisation of an IRS Fragment to IRS products of the radiation hybrid mapping panel immobilised on a nylon membrane.

For radiation hybrid mapping, both cloned IRS-products from the marker library, as well as IRS-products from large-insert clones were used. IRS-probes were radioactively labelled using random hexamer priming and hybridised on RH filters (Figure 23B).

Hybridisation results were scored and entered into a database using a web based scoring interface (M. Kramer, pers. communication). Radiation hybrid vectors of the T51 panel were submitted to the laboratory of Robert Geisler http://wwwmap.tuebingen.mpg.de:8082/rh/, data from the LN54 panel were sent to the radiation hybrid mapping server of the Dawid lab at the NIH http://mgchd1.nichd.nih.gov:8000/zfrh/beta.cgi, where map positions were calculated.

Marker	Linkage group	Linked to	LOD	RH panel
BUSMP706H052	24	z20051	10.57	T51
BUSMP706H1057	20	Z7933	9.8	LN54
BUSMP706N2419	20	20kx18	11.4	LN54
BUSMP706P1790	20	z20046	12.09	T51
BUSMP706P2222	7	z20576	17.76	T51
MMA_AB25A7	20	Z7803	9.6	LN54
MMA_AB25C1	17	Z9831	11.5	LN54
MMA_AB25C1	17	z22279	2.66	T51
MMA_AB25D1	6	Z265	10.3	LN54
MMA_AB25G1	8	Z20113	9.7	LN54
MMA_AB25L1	11	z8214	14.00	T51
MMA_AB25N1	11	Z9239	10.5	LN54
MMA_AB25O1	5	Z10663	16.8	LN54
MMA_AB25O1	5	z11496	12.23	T51
MMA_AB27P23	20	Z6804	10.8	LN54
MMA_AB28P11	20	Z6804	6.6	LN54
MMA_IND18E7	20	Z6804	8.1	LN54
MMA_SJD36E10	20	Z6804	12.7	LN54
MMA_SJD36E10	20	z11841	7.19	T51
MMA_SJD36H9	18	Z10008	7.3	LN54

Table 8

Radiation hybrid positions of IRS markers. Hybridisation probes were generated from the IRS marker library (MMA) or by IRS PCR of PACs (BUSMP). The table shows the marker names, the linkage group they map to, the framework marker which is next, the LOD score and the mapping panel used.

Radiation hybrid mapping of IRS products turned out to be rather inefficient. Only 20 % of the probes gave a hybridisation pattern, which could be scored with sufficient unambiguousness. Some probes had to be repeated more than two times. IRS clones from the marker library worked better than IRS products from PACs, probably because they represent IRS products, which are more abundant in a complex amplicon, and are thus easier to detect in the radiation hybrid IRS products. The biggest problem complicating the mapping is background hybridisation. This is probably due to the fact that IRS products often contain repetitive sequences (see 4.1.6). Hybridisation based mapping was therefore only done for a small number of clones, where it was of special interest (Table 8).

4.1.8 High-density arrays of IRS products of pooled large-insert clones on nylon membranes.

Physical mapping of a genome implies that large-insert genomic clones from a library are ordered according to their position in the genome (see 1.2.3). In clone based maps, this is accomplished by the identification of clone overlaps and construction of overlapping clone paths (contigs). The probability of detecting overlapping clones with a given number of tests (e.g. hybridisation experiments) increases with clone number and insert size. Therefore it is desirable to have a large number of clones as hybridisation targets. On the other hand, even when spotted in a high-density grid, it would take several 22x22 cm membranes to carry more than 100,000 colonies, which is hardly manageable for the sequential hybridisation of a large number of probes in a high-throughput manner. For this reason, clones are pooled to considerably reduce the number of spots on a filter. A disadvantage of pooling lies in the fact, that the complexity of the hybridisation target is increased and therefore the occurrence of background hybridisation is higher. This can be overcome by spotting IRS-PCR products instead of genomic DNA. Correspondingly, IRS-PCR products are used as hybridisation probes.

Pooling of genomic clones is usually performed in a three dimensional format (Hunter, 1997). Thus a plate, a row and a column address determine each clone. However, in the IRS-based physical map of the mouse genome a "six dimensional" pooling scheme was used, because it causes redundancy of the clone address, and thus makes it more tolerant against false negative scores. In the analysis step of the mouse project 40% of the complete clone addresses were determined using scores from dimensions 4-6, so the additional effort during the pooling step pays off during the later analysis (L. Schalkwyk and H. Himmelbauer, pers. communication).

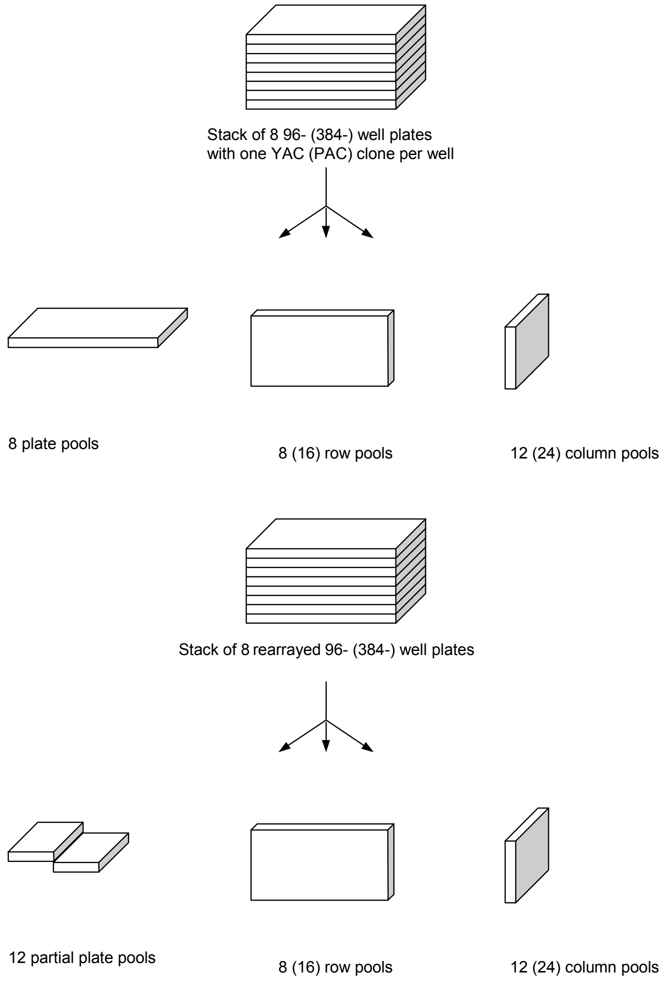

Stack of 8 96- (384-) well plates
with one YAC (PAC) clone per well
8 plate pools
8 (16) row pools
12 (24) column pools
Stack of 8 rearrayed 96- (384-) well plates
12 partial plate pools
8 (16) row pools
12 (24) column pools

Figure 24

6-fold pooling scheme. Pools are collected out of stacks of 8 microtitre plates. For the YAC library MGH_y932 96-well plates (i.e. 8 rows, 12 columns) were used, for the YAC-library HACHy914 and the PAC-library BUSMP706 394 well plates (16 rows, 24 columns) were used.

During the time of this work, three large-insert genomic libraries have been available (Table 3). 6-dimensional pooling of these libraries was realised according to the scheme outlined in Figure 24: 6 replicas of the libraries were produced, one for each dimension. Plates were grouped in stacks of eight, with stack I containing plates 1-8, stack II containing plates 9-16 etc. Subsequently, the clones of each plate were collected to form plate pools; respective rows of each stack were collected to row pools; and respective columns of each stack were collected to column pools. Subsequently, the stacks of the remaining replicas were rearrayed. Now stack I consisted of plates 1,9,25,….,57, stack II consisted of plates 2,10,26,….,58 etc. From these rearrayed stacks, plates were divided in thirds. Two non-overlapping thirds at a time from different plates were collected to form partial plate pools. From the rearrayed stacks, also row and column pools were collected. As a result, each clone is present in two plate pools, two row pools and two column pools. A missing address due to a false negative hybridisation result can be compensated by the second address.

Library	Pools 1-3	Pools 4-6
MGH_y932	698	786
HACHy914	530	784
BUSMP706	1591	1688

Table 9

Number of pools from the different libraries used. Pools 1-3 are pools from the initial blocks, while pools 4-6 were made from the rearrayed blocks.

The libraries combined consist of 157,624 clones. These were collected to 6077 pools (Table 9), which were stored in 65 96 well plates. IRS-PCR was performed on pools (Figure 25), and PCR products were transferred to 384-well plates for spotting.

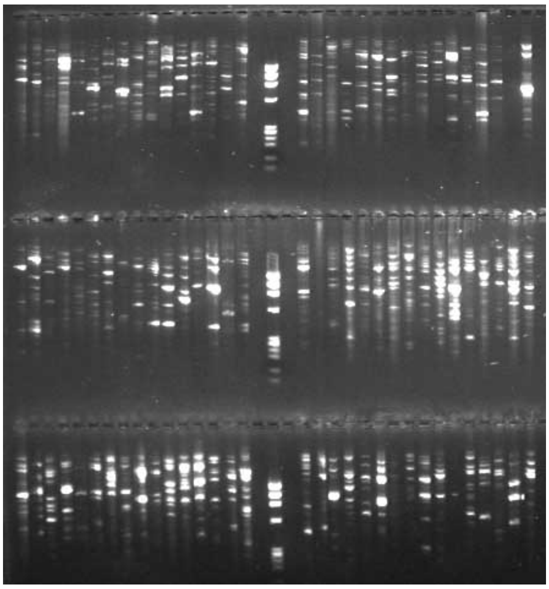
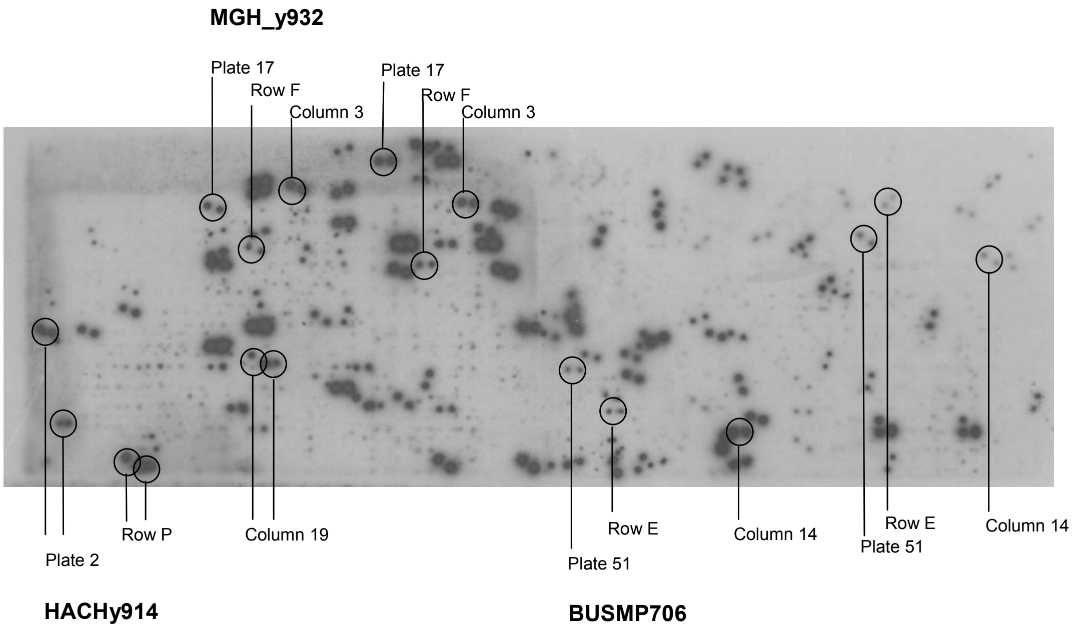


Figure 25

IRS-PCR of 96 PAC-pools. Efficiency of amplification varies strongly between the different IRS-PCR products. The weak bands are obscured by stronger bands and not visible on the gel.

If arranged in a 5x5 pattern, 4608 spots fit on an 11 x 7 cm field. For this reason, the pools were divided, with the two YAC-libraries spotted on one field, while the PAC-library was spotted on a second field. Thus one filter has the size of 22 x 7 cm which means, that hybridisation can be done in 15 ml Polypropylene tubes (Greiner), making it manageable for a high-throughput protocol. IRS-probes for hybridisation were either generated from the fingerprinted IRS-markerlibrary or from genomic clones. Isotopic labelling was performed by the random hexamer priming method. Figure 26 A shows the hybridisation of an IRS-clone on a pool filter. Hybridisation results were scored by aligning the autoradiographs to a grid and entering positive pools in a database using a Java based scoring tool (Figure 26 B).

A



B

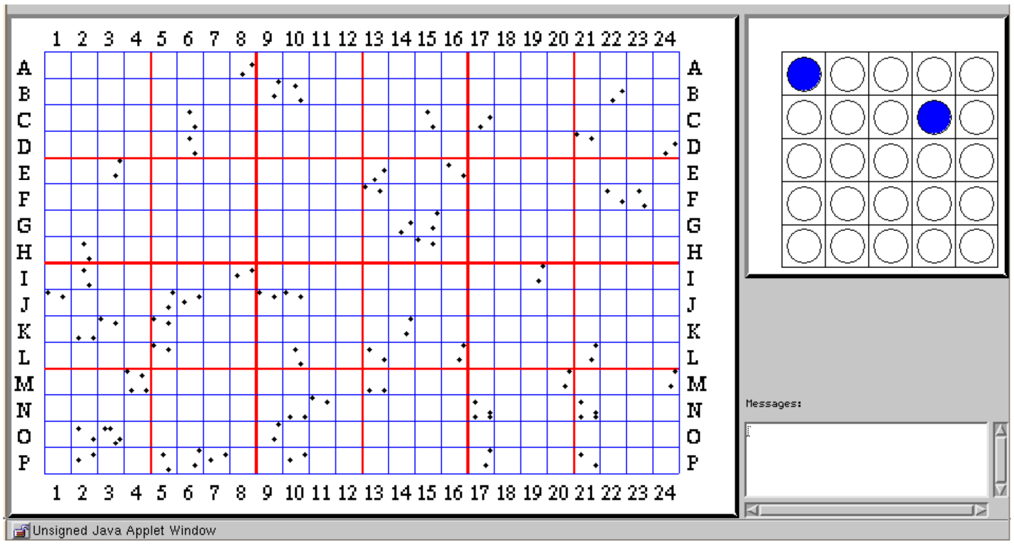
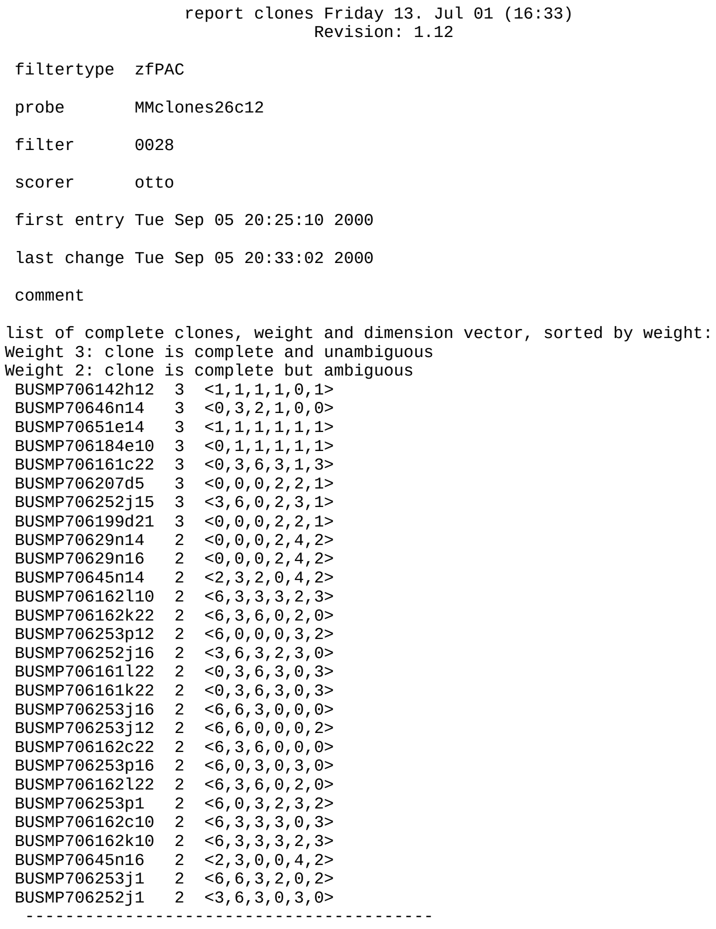


Figure 26

A: Hybridisation of the IRS clone MMA_AB26C12 on YAC and PAC pool filter. For each library, one complete and unambiguous clone address is outlined, using all pooling dimensions.
B: Java applet for entering hybridisation results, showing the data from the hybridisation on PAC pools (right half of the autoradiograph)

Positive clone addresses were determined by deconvolution of the pool data using software written in the *awk* scripting language (Kramer and Schalkwyk, personal communication, Figure 27). The output differentiates three classes of clone addresses: (1) Incomplete clone addresses; (2) complete, but ambiguous clone addresses (designated class 2 addresses); (3) complete and unambiguous clone addresses (class 3 addresses). Only the latter can be considered as reliable data and utilised for mapping purposes.



```
                  report clones Friday 13. Jul 01 (16:33)
                              Revision: 1.12

 filtertype  zfPAC

 probe       MMclones26c12

 filter      0028

 scorer      otto

 first entry Tue Sep 05 20:25:10 2000

 last change Tue Sep 05 20:33:02 2000

 comment

list of complete clones, weight and dimension vector, sorted by weight:
Weight 3: clone is complete and unambiguous
Weight 2: clone is complete but ambiguous
 BUSMP706142h12  3  <1,1,1,1,0,1>
 BUSMP70646n14   3  <0,3,2,1,0,0>
 BUSMP70651e14   3  <1,1,1,1,1,1>
 BUSMP706184e10  3  <0,1,1,1,1,1>
 BUSMP706161c22  3  <0,3,6,3,1,3>
 BUSMP706207d5   3  <0,0,0,2,2,1>
 BUSMP706252j15  3  <3,6,0,2,3,1>
 BUSMP706199d21  3  <0,0,0,2,2,1>
 BUSMP70629n14   2  <0,0,0,2,4,2>
 BUSMP70629n16   2  <0,0,0,2,4,2>
 BUSMP70645n14   2  <2,3,2,0,4,2>
 BUSMP706162l10  2  <6,3,3,3,2,3>
 BUSMP706162k22  2  <6,3,6,0,2,0>
 BUSMP706253p12  2  <6,0,0,0,3,2>
 BUSMP706252j16  2  <3,6,3,2,3,0>
 BUSMP706161l22  2  <0,3,6,3,0,3>
 BUSMP706161k22  2  <0,3,6,3,0,3>
 BUSMP706253j16  2  <6,6,3,0,0,0>
 BUSMP706253j12  2  <6,6,0,0,0,2>
 BUSMP706162c22  2  <6,3,6,0,0,0>
 BUSMP706253p16  2  <6,0,3,0,3,0>
 BUSMP706162l22  2  <6,3,6,0,2,0>
 BUSMP706253p1   2  <6,0,3,2,3,2>
 BUSMP706162c10  2  <6,3,3,3,0,3>
 BUSMP706162k10  2  <6,3,3,3,2,3>
 BUSMP70645n16   2  <2,3,0,0,4,2>
 BUSMP706253j1   2  <6,6,3,2,0,2>
 BUSMP706252j1   2  <3,6,3,0,3,0>
  ----------------------------------------
```

Figure 27

Deconvolution of the pool hybridisation data shown in Figure 26. The table shows addresses of clones positive in hybridisation. The digit following the clone name specifies if the clone address is ambiguous (2) or unambiguous (3). The digits in brackets describe the pools of each clone (plate, row, column, partial plate, row, column) values bigger than 0 indicate positive hybridisation results and the number of clone addresses which were deconvoluted using the respective pool.

4.1.9 Detection of Single Nucleotide polymorphisms in orthologous IRS-PCR products of different zebrafish strains

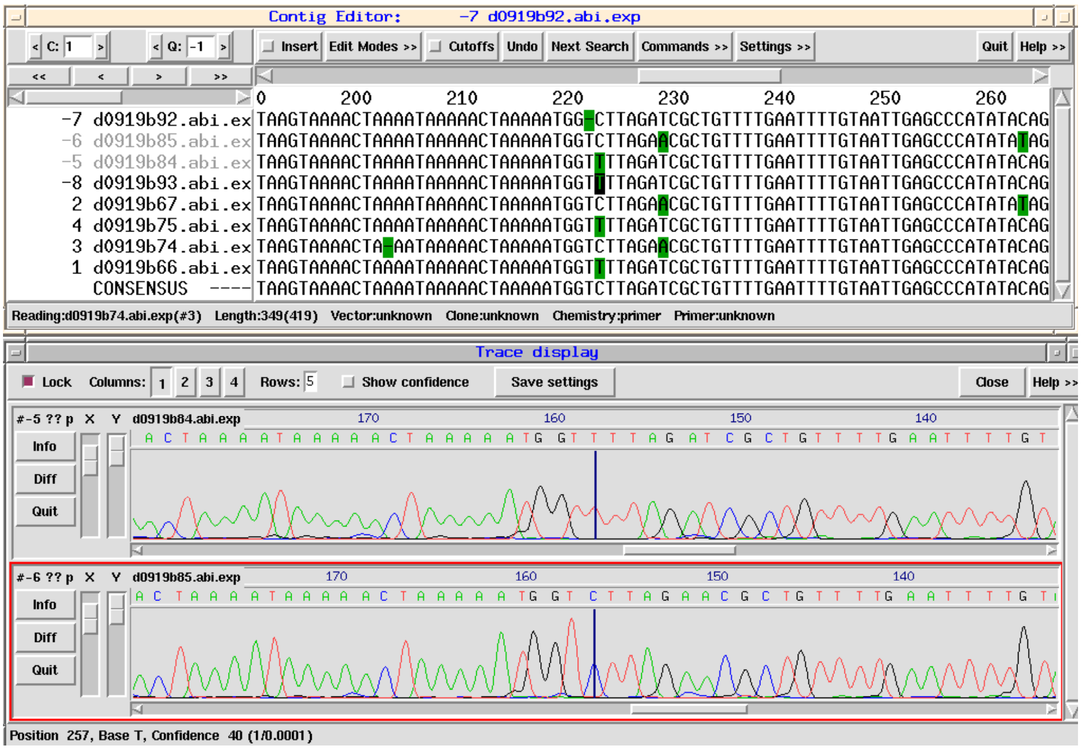
Base changes (transitions and transversions) in an otherwise homologous sequence are the most abundant DNA polymorphisms in a population (Stoneking, 2001). By definition, base changes with an allele frequency of more than 1% are termed single nucleotide polymorphisms (SNPs). SNPs can be divided in those, which are in coding sequences and cause amino acid substitutions and might therefore have an effect on the phenotype and those, which are presumed to be neutral, because they do not have an effect on amino acid sequence and no observable phenotype. A different classification distinguishes between cSNPs (SNPs in cDNAs) and other SNPs. While functional SNPs are the primary cause for phenotypic variation in a population, neutral SNPs can serve as a powerful tool in high-resolution genotyping, which is needed for mapping of quantitative trait loci and identification of linkage disequilibrium. Comparison of two human sequences is estimated to result in a SNP every 1000-2000 nucleotides, adding up to 1.6 million – 3.2 million SNPs. Obviously, the number of SNPs in the total human population is much higher. Recently, the International SNP Map Working group has published a map of 1.42 million single nucleotide polymorphisms in the human genome (Sachidanandam et al., 2001).

In order to detect SNPs it is necessary to isolate homologous DNA fragments from different individuals, and analyse the sequence. This ca be done by amplification and sequencing of specific loci. However, this requires the synthesis of oligonucleotide primers for each locus, limiting it to regions of known sequence and making it expensive for large-scale approaches.

It is therefore desirable to generate an orthologous reduced representative subsets of the genome of different individuals in an inexpensive way and sequence it (Altshuler et al., 2000). A simple way to do this is to isolate restriction fragments of a defined size range and clone them. This has the advantage, that the complexity of the subset can be adjusted by the selection of the restriction enzyme and by the size range of the fragments.

In this work, IRS-PCR was tested as an alternative method to generate a reduced representation of the genome for SNP detection. This takes advantage of the fact that an IRS library exists, which is normalised by oligonucleotide fingerprinting (see chapter 4.1.5). Orthologous sequences from different strains are grouped together in clusters and can therefore be selected for sequencing. The prevalence of sequence variation in IRS products is expected to be high, because usually IRS products consist of non-coding, evolutionary neutral sequence.

To test this, sequences originating from different strains, but grouped together in the same ONF cluster (see 4.1.5) were aligned using the gap4 program from the Staden package (Staden et al., 2000, Figure 28). Prior to analysis, low quality parts of the sequences were removed by eliminating all bases with a Phred value > 20, which means that the remaining bases have an error probability of < 0.01 (Ewing and Green, 1998). Different clones from the same strain were used to calculate strain specific consensus sequences. This step eliminates heterozygous alleles, which occur quite frequently, because the zebrafish strains are not completely inbred (Burgtorf, 1999). SNPs between different strains were detected by aligning the consensus sequences and searching for single base variations using the program trace_diff (Staden Package). All 5 strains were compared to each other and the SNP frequency is shown in Figure 29



```
Contig Editor:      -7 d0919b92.abi.exp

                   0       200       210       220       230       240       250       260
   -7 d0919b92.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGG-CTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
   -6 d0919b85.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTCTTAGAACGCTGTTTTGAATTTTGTAATTGAGCCCATATATAG
   -5 d0919b84.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTTTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
   -8 d0919b93.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTTTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
    2 d0919b67.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTCTTAGAACGCTGTTTTGAATTTTGTAATTGAGCCCATATATAG
    4 d0919b75.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTTTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
    3 d0919b74.abi.ex TAAGTAAAACTA-AATAAAAACTAAAAATGGTCTTAGAACGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
    1 d0919b66.abi.ex TAAGTAAAACTAAAATAAAAACTAAAAATGGTTTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
      CONSENSUS  ---- TAAGTAAAACTAAAATAAAAACTAAAAATGGTCTTAGATCGCTGTTTTGAATTTTGTAATTGAGCCCATATACAG
Reading:d0919b74.abi.exp(#3)  Length:349(419)  Vector:unknown  Clone:unknown  Chemistry:primer  Primer:unknown

Trace display
#-5 ?? p  X  Y  d0919b84.abi.exp
#-6 ?? p  X  Y  d0919b85.abi.exp
Position  257, Base T, Confidence  40 (1/0.0001)
```

Figure 28

Alignment of orthologous IRS sequences from different zebrafish strains. Base variations in different sequence runs are highlighted by green background. Two sequence trace files are shown, having a T-C polymorphism highlighted by the bar.

Detailed analysis and comparisons of all 5 strains against each other was only done for 5 clusters (2 kb of sequence in each strain): The average SNP frequency (0.013) however corresponds with the number seen in the other groups of orthologous sequences, which were not analysed in detail. The SNP frequency ranges from 0.007 (AB/WIK) to 0.026 (SJD/TUE).

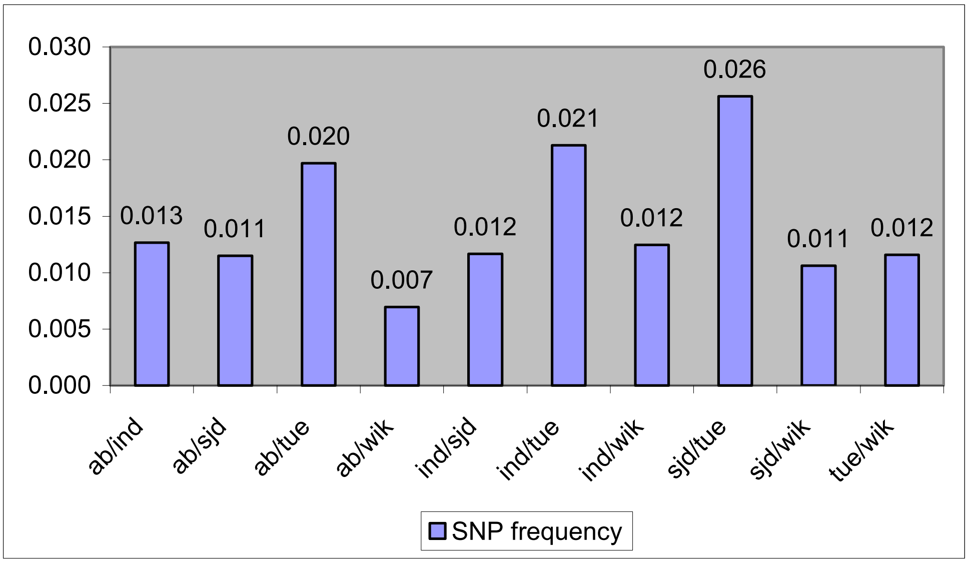


Figure 29

SNP frequencies in different strains. 5 sequence clusters were used for detailed analysis, adding up to 2 kb of DNA for each strain. Strains were compared pairwise, and SNP frequencies (Number of SNPs per bp) were calculated.

Sequences used for SNP detection were derived from cloned PCR products. Consequently, it is possible that a significant proportion of sequence variants is due to erroneous base incorporation by the *taq* polymerase. This was excluded by amplifying a fraction of the sequences from genomic DNA of different zebrafish strains using specific primers and sequencing these PCR products directly. The resulting sequence contained the same single nucleotide polymorphisms, as the IRS clones, suggesting that there is not a considerable amount of variation due to *taq* polymerase errors. Moreover, aligning reads from different clones and calculating a consensus suppresses artifactual SNP detection. Radiation hybrid mapping using specific primers confirms that the clusters analysed map to unique genomic locations and do not represent repetitive sequence paralogues. The results show that a library of IRS products is a suitable reduced representation of the genome for SNP detection by sequencing. It is particularly well suited, because the individual SNPs can be easily anchored to the physical clone map by hybridisation on IRS pool filters and due to the ONF analysis, orthologous fragments are already listed and grouped together.

4.2 Physical mapping of zebrafish linkage group 20

The goal of this project is to construct an STS based framework map of a Zebrafish chromosome 20, anchored to the genetic and radiation hybrid maps. This is the first map of its kind of a zebrafish chromosome and is therefore supposed to serve as a model for other physical mapping projects. Clone contigs will be templates for the zebrafish genomic sequencing project initiated this year (http://www.sanger.ac.uk/Projects/D_rerio/). They will also facilitate positional cloning of genes on this chromosome.

Beyond this immediate use, the map is also valuable as a control to empirically test parameters for assemblies based on restriction fragment fingerprinting and sequencing data. Once the clone map is assembled and the chromosome is sequenced, it can serve as a model for chromosome structure, repeat distribution etc. in the zebrafish. Thus, this project is complementary to the physical mapping project carried out at the Max-Planck-Institute for Developmental Biology in Tübingen, where a clone based map of the whole genome is constructed by restriction fingerprinting, (J. Rauch, pers. communication, http://www.eb.tuebingen.mpg.de/home/news/zf_genome.html see paragraph 5.2).

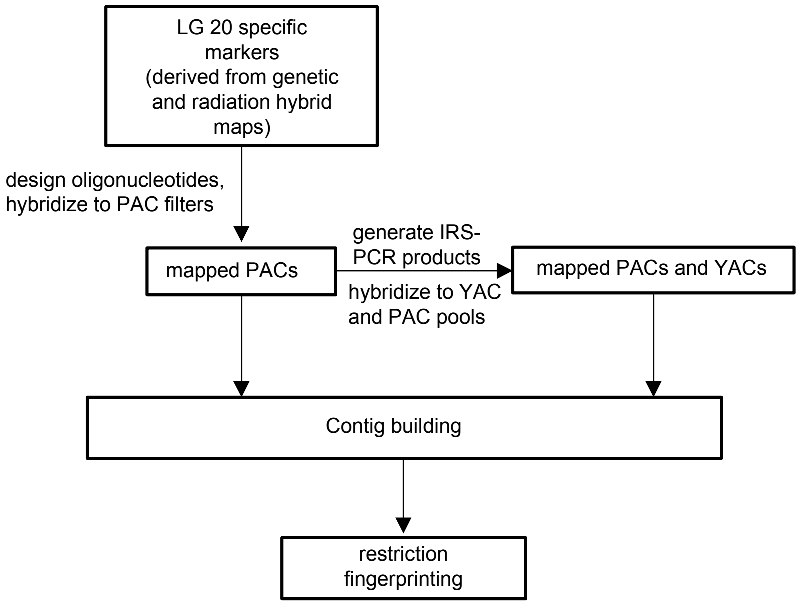

LG 20 specific markers (derived from genetic and radiation hybrid maps)
design oligonucleotides, hybridize to PAC filters
mapped PACs
generate IRS-PCR products
hybridize to YAC and PAC pools
mapped PACs and YACs
Contig building
restriction fingerprinting

Figure 30

Flow diagram of the physical mapping process.

Figure 30 illustrates the mapping procedure employed for this work. Initially, published markers from the LG20 maps were selected and oligonucleotide probes were generated for hybridisation on PAC filters. PACs mapped by these probes were picked and used for the generation of IRS probes, which were hybridised on IRS-pool filters of PAC and YAC libraries (4.1.8). Hybridisation data are used to assemble ordered contigs. To verify the assembly and to generate minimal tiling paths, the mapped clones are further analysed by restriction fragment fingerprinting.

The task of this Ph.D. work was to establish and supervise the mapping process. This implies, that, from a certain stage of the project, the majority of hybridisations was done by a technician and a diploma student. Restriction fingerprinting of mapped clones is currently carried out in a close collaboration with the MPI for Developmental Biology in Tübingen.

4.2.1 Mapping of PACs by hybridisation of oligonucleotides.

Data from genetic and radiation hybrid maps were used to select linkage group 20 specific markers (Gates et al., 1999; Geisler et al., 1999; Kelly et al., 2000; Shimoda et al., 1999). These markers consist of sequence tagged sites (STSs), expressed sequence tags (ESTs) and cloned genes. Sequences were retrieved from the GenEMBL database using *fetch* (GCG package). Repeats were masked by the RepeatMasker program. DNA sequences were clustered by the gap4 software (Staden et al., 2000) to identify markers, which have different names on different maps (for example accession numbers and EST names), but represent the same sequence. Paragraph 7.1 shows accession numbers and Genbank definitions of Genes, ESTs and STSs used for probe design until June 2001. In paragraph 7.2, oligonucleotide sequences are used.

Two different kinds of oligonucleotide probes were generated: 35mer oligonucleotides and “overgo” probes. 35mer oligonucleotides were designed using the gap4 software. For hybridisation, they were radioactively labelled using T4 polynucleotide kinase. Overgo probes consist of two oligonucleotides possessing an 8-bp complementary overlap at their 3’-termini (Figure 31). They were designed using the Overgo Maker program (http://genome.wustl.edu/gsc/overgo/overgo.html). The probes were labelled using a Klenow fill-in reaction with two different radioactive nucleotides.

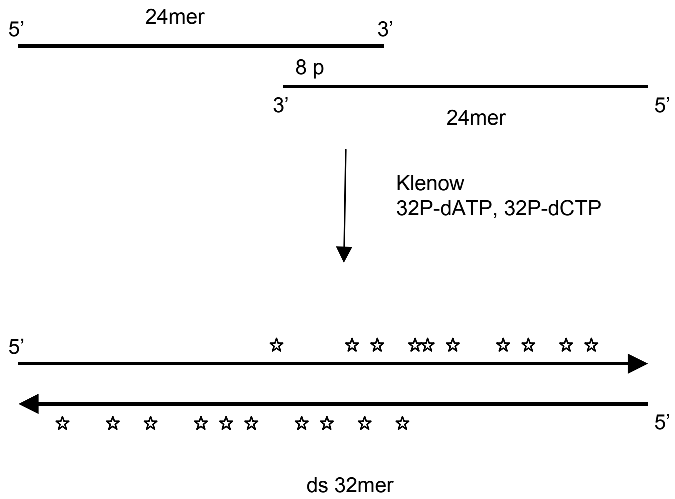


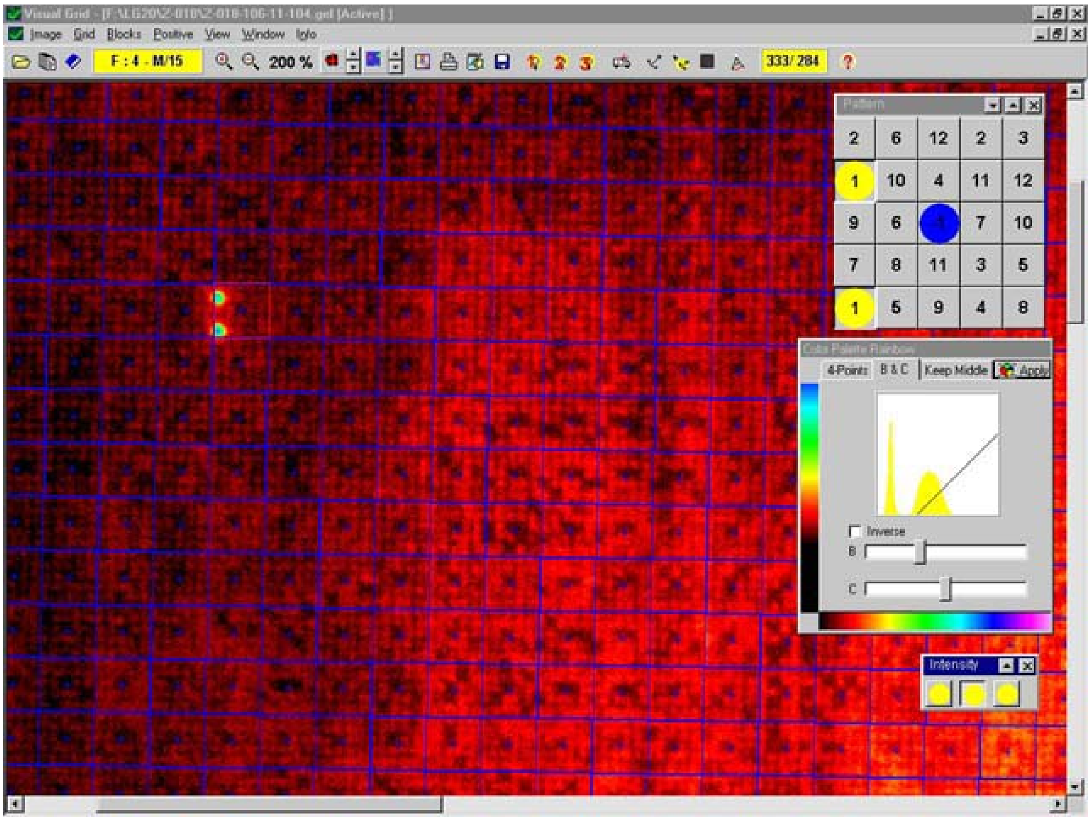
Figure 31

Principle of overgo labelling. Two oligonucleotides with an 8 bp complementary overlap are annealed. A Klenow fill in reaction is carried out using two radioactively labelled nucleotides. This way, a high specific activity of the probe is achieved.

Although overgo probes are supposed to have a particularly high specific activity compared to probes labelled by other methods (e. g. kinased oligonucleotides) initial testing of both protocols in hybridisation experiments showed no significant advantage in terms of specificity and signal strength using overgos. Hence it was decided to use mainly 35-mer oligonucleotides, which are cheaper and more convenient to handle.

Labelled probes were hybridised to three high-density PAC filters in a tube. Each filter carried 27,648 clones, adding up to 82,944 clones or 5.8x genome coverage. Details of the labelling and hybridisation process are described in 3.2.17. In order to evaluate hybridisation conditions and to identify non-recombinant clones, a control hybridisation was performed using an oligonucleotide probe which binds to the pUC insertion of the undigested PAC vector (Ioannou et al., 1994). Out of 82,944 clones on three filters used 102 (0.1%) gave a positive signal with the pUC probe.

Hybridised filters are subsequently exposed to PhosphorImager (Molecular Dynamics, Sunnyvale, CA) storage screens for 3-16 h. Screens are scanned at 176-µm resolution and results are stored as 16-bit TIF files. The image processing software Visual Grid (Kietzmann, personal communication, Clark et al., 1999) is used to visually inspect and score hybridisations. It semiautomatically aligns a grid onto the membrane image, and calculates the plate well coordinates for each clone highlighted by a click with the mouse. Each positive hybridisation is given a signal strength value between 1 (weak signal) and 3 (strong signal). Figure 32 shows a screenshot of a TIF image visualised using Visual Grid.



2	6	12	2	3
1	10	4	11	12
9	6	[illegible]	7	10
7	8	11	3	5
1	5	9	4	8

Figure 32

Hybridisation of an oligonucleotide derived from the STS marker Z9209 on a high-density colony filter, carrying 27000 PAC-bearing clones. The program Visual Grid is used to visualise the TIF format filter image, align a grid onto the image and assign clone addresses for positive hybridisation results. In this screenshot, a magnified sector of the entire filter is shown.

After saving coordinates as text files, all hybridisation data, including probe names, filter names and positive clones are stored and processed in an Oracle database (Nagel, pers. communication).

Hybridisation data were examined for consistency. Non-recombinant clones, as identified by hybridisation with a pUC specific probe, were excluded. Accordingly, clones which were hit by more than 12 probes were labelled as "unspecific clones" and not considered in the further analysis.

Up to June 2001, 344 STS and EST markers have been hybridised (7.1). 100 of these have been hybridised as pools of oligos, which have been generated using a 3-fold pooling scheme (P. Nierle, diploma thesis, in preparation) 333 hybridisations have been done in total, including repetitions and pools. The total set of oligonucleotide probes bound to 3196 clones (signal strength ≥1); 2147 clones (signal strength ≥2) and 720 clones (signal strength =3), respectively. With 344 probes and a genome coverage of 5.8, 1995 positive clones are expected. This suggests, that a hybridisation strength of ≥2 is necessary to exclude false positive clones. For that reason, clones with signal strength =1 were not considered for map assembly.

4.2.2 Mapping of PACs and YACs by hybridisation of IRS-PCR products

PACs identified by hybridisation of Chromosome 20 specific STS probes were picked and used as template for IRS-PCR. Strong IRS bands were cut out of a low melting point agarose gel and used for probe generation by random hexamer primed labelling. These probes were hybridised against high-density filters of IRS amplicons from PAC and YAC pools (described in 4.1.8)

Up to June 2001, 284 IRS PCR probes have been hybridised to pool filters (221 PAC derived probes, 58 YAC derived probes). 172 probes (60% of all probes; 59% of PAC derived probes 70% of YAC-derived probes) resulted in hybridisations, which gave complete and unambiguous clone addresses. In these hybridisations 561 clones were hit (344 PAC clones, 217 YAC clones), an average of 3.3 clones hit per probe. This is far less then expected from the 17 fold genome coverage represented on the filters. The high rate of false negatives is probably due to the following reasons: (1) IRS products of clone pools do not amplify evenly well in different pool backgrounds. This might result in missing positives, although some of them might be compensated by the redundant pooling scheme used. (2) Zebrafish inter-mermaid PCR products often contain repeat sequences (see 4.1.6), which can cause a high level of background noise. As a result, spots falsely scored as positive can introduce ambiguity in otherwise correctly addressed positive clones.

Hybridisation probes derived from large clusters of the IRS marker library 4.1.4 usually give rise to a better signal to noise ratio than PAC derived products, consistent with the idea, that those IRS products amplify well even from a complex background.

During the course of this work, different methods were tested to reduce background (raising of hybridisation temperature, competitive reannealing of probes in the presence of an excess amount of unlabelled DNA, etc., Baxendale et al., 1991). This has resulted in a certain improvement, but there is probably some potential left for optimisation.

4.2.3 Assembly of the physical framework map of zebrafish linkage group 20

Data from hybridisation on PAC filters and pools were retrieved from their respective databases and integrated using scripts designed for that purpose (D. Buzcek, A. Nagel and

M. Kramer, pers. communication). The result is a matrix of hybridisation results, where rows represent clones and columns represent probes. Hybridisation results are represented as integers at the probe- clone intersections. These integers can take a value between 0 and 3. In the PAC colony hybridisations, 0 stands for no signal and 3 means a strong signal. In the IRS pool hybridisations, 2 stands for a complete but ambiguous, 3 for a complete and unambiguous clone address. If a given clone was not present in a hybridisation with a given probe (e.g. YACs were not present on the PAC colony filters), this is given a value of 9. To construct a physical map, hybridisation data are ordered to the most likely path of overlapping clones by the software wprobeorder using a simulated annealing algorithm (Mott et al., 1993). The contigs produced can be sorted according to the chromosomal locations of mapped probes if a file containing the probes in order of their map position is provided. Results of map construction are visualised by the program "show" (Figure 33).

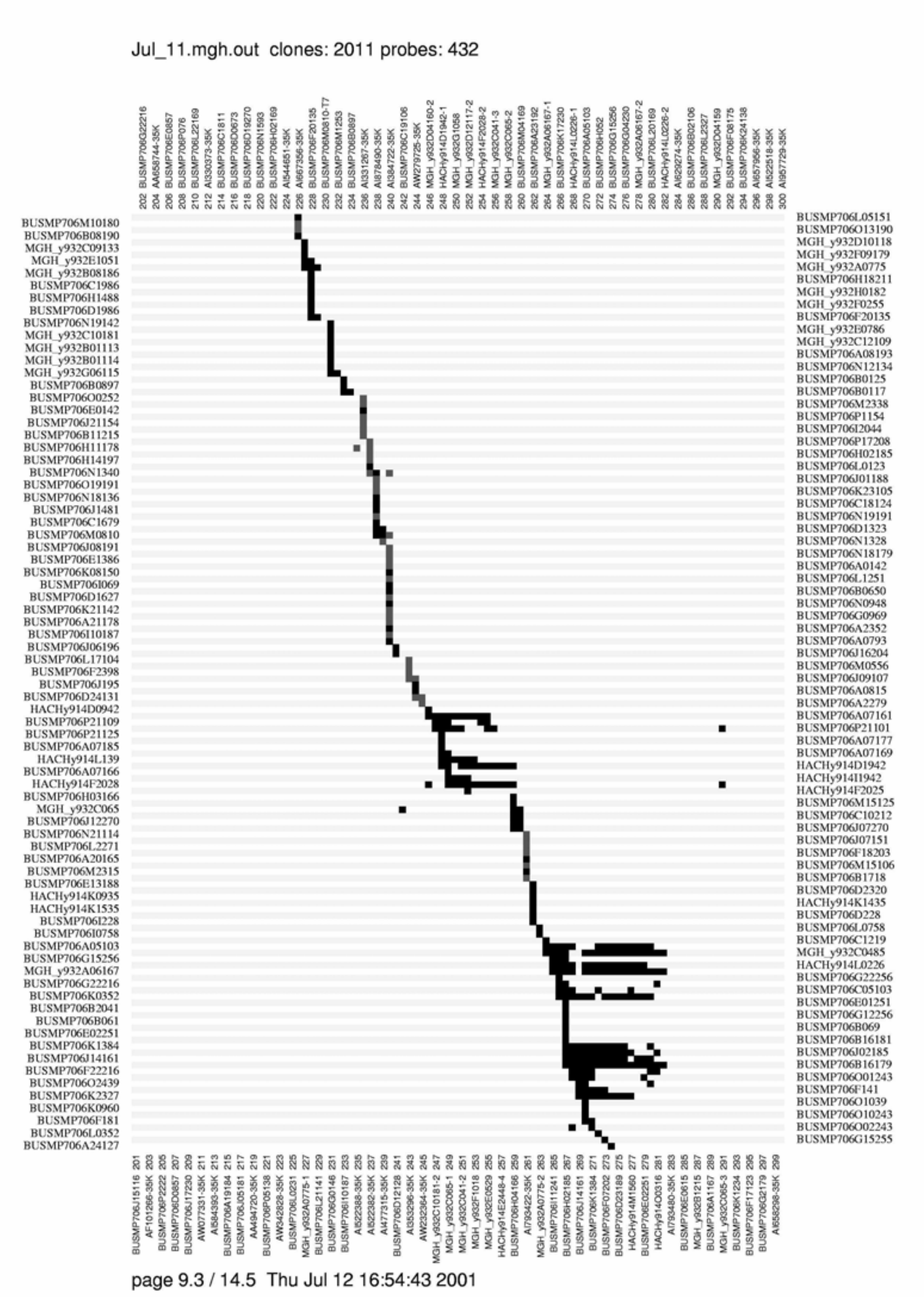


Figure 33

Partial view of the map of ordered hybridisation data as generated by wprobeorder and show, with probes as columns and clones as rows. Different hybridisation strengths (values between 0 and 3) are represented by different shades of grey. For this assembly only hybridisation results with strengths 2 and 3 were considered. Clone overlaps occur, where two different probes hit one clone.

Hybridisations with signal strengths ≥2 in the PAC colony hybridisations, as well as pool hybridisations resulting in complete and unambiguous clone addresses were considered for map construction only. Hybridisations with STS probe pools were treated separately and are not included here (P. Nierle, diploma thesis, in preparation). 388 probes fulfilled the selection criteria (216 STS probes; 172 IRS probes), hitting 2007 clones (1790 PACs, 217 YACs). This results in a marker density of 1 per 205 kb.

249 contigs were found. 201 (81%) of those are singletons, i.e., they share only one probe anchor. 22 (9%) share two anchors, 3% have 3 anchors, and 4% have four anchors, and 4% have more than 4 anchors. Figure 34 shows a contig linked by 11 anchors, two of which are mapped on the MGH genetic map. Z10756 is located at 68.7 cM, Z21123 is located at 69.8 cM from the top of the chromosome. Based on an average value of 740 kb per cM (Shimoda et al., 1999), this distance is equivalent to a physical contig size of 814 kb. As a control, the results were compared with a computer simulation of the experiment (discussed in paragraph 5.2).

```
contig  249 size:   11 location: 47 unknown
probes:
·Z10756-OV               mapped at position   42 on chrom 20
·AI415835-35K
·AI437240-35K
·AI476962-35K
·AI444373-35K
·AI416203-35K
·AI477017-35K
·AA606026-35K
·U57965-35K
·AI793363-35K
·Z21123-35K              mapped at position   47 on chrom 20

fitted clones:
BUSMP706J09156   3..........#
BUSMP706K147     3..........#
BUSMP706A1647    2..........#
BUSMP706N1348    3..........#
BUSMP706G2412    3..........
BUSMP706G1035    322223.....
BUSMP706B1450    22232332...
BUSMP706I0125    .3.........
BUSMP706I0127    ...22......
BUSMP706D051     ....2......
BUSMP706B071     ....2......
BUSMP706E1157    ....2222...
BUSMP706D0857    .2..3222...
BUSMP706E212     .....2.....
BUSMP706I0139    ......2....
BUSMP706B0248    .......3...
BUSMP706N0957    .......2...
BUSMP706B0250    .......3...
BUSMP706N0523    .......3..3
BUSMP706D0850    .......3...
BUSMP706N0857    ..2....2...
BUSMP706N0557    .......2...
BUSMP706L0419    .......33..
BUSMP706L0319    .......3333
BUSMP706P0426    ........223
BUSMP706D1427    .........2.
BUSMP706C1674    .........3.
BUSMP706K09140   .........23
BUSMP706H12140   .........23
BUSMP706F13185   .........22
BUSMP706F17123   ..........2#
BUSMP706O0426    ..........2#
BUSMP706P12139   ..........2#
BUSMP706F13173   ..........2

connections to other contigs:
BUSMP706G2412   (3) -> contig  108 probe BUSMP706J09156  not mapped
BUSMP706G1035   (2) -> contig   52 probe AA497290-35K  not mapped
BUSMP706B1450   (2) -> contig   52 probe AA497290-35K  not mapped
BUSMP706I0125   (3) -> contig  176 probe AI522382-35K  not mapped
```

```
BUSMP706F13185  (2) -> contig   66 probe AW128231-35K  not mapped
BUSMP706F13173  (2) -> contig  281 probe Z6425-35K  mapped at position   31 on chrom 20
```

Figure 34

Contig of 11 probes containing 34 clones. Probes Z10756-OV and Z10756-OV are anchored to the MGH microsatellite map. The table of fitted clones contains the hybridisation matrix of probes and clones, including hybridisation strengths. The third table shows connections to other contigs, i.e. clones which also hybridise with probes in other contigs. In these cases, the links have been not strong enough for the program to fuse these contigs.

To verify contigs assembled by wprobeorder and to determine overlap regions and physical contig sizes, restriction fingerprinting of all PACs identified by chromosome 20 specific probes is done during the time of this writing (data not shown here). All mapping data, including maps, marker sequences and mapped clones will be available to the public at our website (http://www.molgen.mpg.de/~ag_zebrafish/physical/physical.html, in preparation).

4.3 Radiation hybrid mapping of expressed sequence tags

In our group, gene expression in zebrafish is examined in a whole mount in situ hybridisation screen. This involves systematic screening of cDNA clones from normalised libraries as well as screening of clones which have been selected due to their sequence homology to human disease genes (see 1.2.4). During this Ph.D. work, the chromosomal locations of transcripts used in the screen were determined by radiation hybrid mapping. The rationale for this is the following: Clones showing specific (non-ubiquitous) expression patterns during embryonic development probably have tissue-specific functions and therefore represent candidate genes for mutations affecting these tissues. The link between a transcript and a mutant phenotype is the chromosomal location of the gene. Mutations are currently mapped genetically by several groups, among them the group of Robert Geisler, MPI for Developmental Biology, Tübingen. By contributing mapping information of transcripts with known expression domains, candidate genes for mutations are provided.

As for the human disease gene homologs, an additional, more direct benefit arises from resolving syntenic relationships through radiation hybrid mapping. If a zebrafish gene maps to a chromosomal region that corresponds to the location of the human homologue, they are likely to be true orthologues (Fitch, 2000) which is – together with data about functional analogy obtained from in situ hybridisation data – an important criterion for the use of the gene in the study of a disease.

These considerations led to different criteria in the selection of clones for mapping: The human disease gene homologues were, in principle, all considered for mapping. From the systematic in situ screen, only clones showing restricted expression patterns were chosen.
For PCR primer design, the 3'-untranslated region (UTR) of the cDNA was used, because this is the least conserved segment of a transcript, making crossmatching of PCR primers with the rodent background unlikely. 3'-sequencing of cDNA clones was carried out by using the anchored poly-T primer $(T)_{23}N$. Previous attempts to prime the sequencing reaction using a universal plasmid specific primer failed, because of polymerase slippage along the long poly-A tails of the transcripts. Sequencing using poly-T primers was also not very efficient, with only ca. 50% of the reactions resulting in usable reads (Phred value $\geq$ 30, longer than 100 bp). One reason for this might be mispriming at internal poly-A sequences. This effect is clone specific, i.e. most clones, which could not be sequenced in the first run, failed also in repetitions. Low quality segments of sequences were removed using the program trim_by_qual (S. Hennig, pers. communication). A BLAST homology search was done against the GenEMBL, Genbank-EST and SwissProt databases. Results were automatically compared to radiation hybrid data accessible through the web, and sequences already mapped by other groups were excluded. From unmapped, high-quality sequences PCR primers were designed using Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Radiation hybrid PCR reactions were performed in duplicate using the T51 radiation hybrid panel (Geisler et al., 1999, Figure 35) and repeated if ambiguous results were obtained.

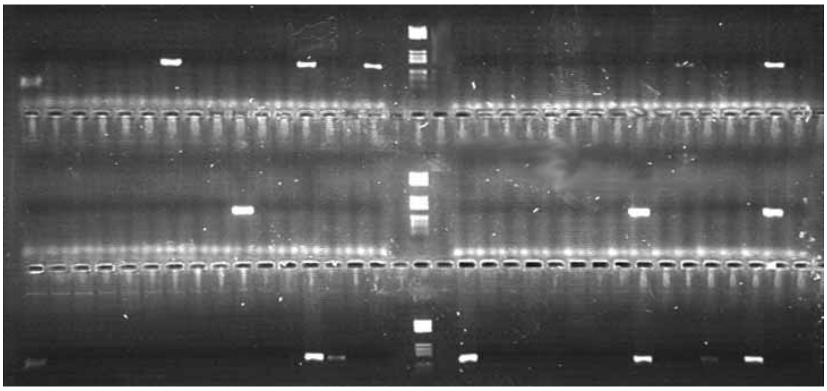


Figure 35
Radiation hybrid PCR of the marker ICRFp524F0916

PCR results were entered in a database using an interactive, web-based scoring tool (M. Kramer, pers. communication). Radiation hybrid vectors were submitted to Robert Geisler

(Max-Planck-Institute for Developmental Biology) for integrating the data in the existing map.

Table 10 shows mapping data of the human disease gene homologues (77 clones) sorted according their human chromosomal position. Only a subset of the 288 clones in the set of human disease gene homologues was mapped here. The reason for this is mainly the unavailability of 3' sequences due to a high failure rate in sequencing. Additionally, a large portion of the ESTs was already mapped on the T51 radiation hybrid panel. These clones were usually not considered for mapping. Mapping data in Table 10 agree with some established conserved syntenies between zebrafish and human (Barbazuk et al., 2000; Postlethwait et al., 2000), e.g. between Hsa 1 and Dre 2; Hsa 3 and Dre 11 and 23; Hsa 4 and Dre 1; Hsa7 and Dre 19; Hsa 11 and Dre 7 and 25; Hsa 16 and Dre 3.

This suggests that genes in these conserved synteny groups are orthologues. In cases were transcripts do not fall in known syntenic segments, it is difficult to decide, if they indicate new synteny groups, or if they are paralogues. This has to be determined by a careful phylogenetic analysis, using all sequence data available for zebrafish and human.

Clone-ID	Human homologue					Zebrafish homologue		
	Gene	Acc.-Nr.	Definition	Disease	location	LG	Closest marker	Marker position
MPMGp532D037	EPHX1	P07099	Epoxide hydrolase 1	Diphenylhydantoin toxicity	1q42.1	13	z9868	40.7
MPMGp609J1133	AGL	P35573	Glycogen debranching enzyme	Glycogen storage disease III	1p21	2	z9944	35.8
MPMGp609L1013	ACADM	P11310	acyl-coa dehydrogenase	Acadm deficiency	1p31	2	z6569	38.3
MPMGp609K1732	UROD	P06132	Uroporphyrinogen carboxylase	Porphyria cutanea tarda, erythropoietic porphyria	1p34	2	z6569	38.3
MPMGp609A0329	SLC2A1	P11166	Glucose transporter type1	Glucose transport defect, blood-brain barrier	1p35-p31.3	23	z15422	32.2
MPMGp609F2344	MTR	Q99707	Methionine synthase	Methylcobalamin deficiency	1q43			
MPMGp609L1832	MSH2	P43246	dna mismatch repair protein msh2	Colorectal cancer, hereditary, nonpolyposis	2p22-p21	12	z4847	42.1
ICRFp524J1472	NDUFS1	P28331	NADH-ubiquinone oxidoreductase	Lactic acidosis	2q33-q34	15	z13310	57.6
MPMGp609D0317	ITGA6	P23229	Integrin alpha6	Epidermolysis bullosa, junctional, with pyloric stenosis	2	9	z1777	3.6
MPMGp609E1015	DYSF	NM_003494	Dysferlin	Limb girdle muscular dystrophy 2B	2p13.3-p13.1	Unlinked *		
MPMGp609O0934	HADHB	P55084	Trifunctonal enzyme beta subunit	Trifunctional protein deficiency, type II	2p23	20*	Z13626	78.6
ICRFp524H089	GNAI2	P04899	G protein alpha inhibiting 2	Pituitary adenoma, Ventricular tachycardia	3p21	11	z868	25.3
MPMGp609C1513	GPX1	P07203	Glutathione peroxidase	Hemolytic anemia	3p21.3	11	z4353	47.6
MPMGp609P0516	MLH1	P40692	mutl protein homolog 1	Colon cancer, nonpolyposis type 2	3p21.3	13	z13250	38.4
ICRFp524C226	CTNNB1	P35222	Catenin beta 1 (88kD)	Colorectal cancer, Hepatoblastoma	3p21	16	z21155	15.5
ICRFp524H1811	GNAT1	P11488	G protein alpha transducing 1	Night blindness, congenital stationary	3p21	18	z13426	34.4
MPMGp609I0456	SLC2A2	P11168	Glucose transporter type 2	Diabetes mellitus, Fanconi-Bickel	3q26.1-q26.3	2	z6569	38.3
MPMGp609L0341	CACT	O43772	Carnitine/acylcarnitine translocase	Carnitine-acylcarnitine translocase deficiency	3p21.31	21	z3332	120.8
MPMGp567B2366	TKT	NM_001064	Transketolase	Wernicke-Korsakoff	3p14.3	23	z4003	16.6

				syndrome				
MPMGp532F1244	MYL3	P08590	Myosin, light polypeptide 3	Cardiomopathy, hypertrophic	3p21.3-p21.2	23	Z13363	34
MPMGp532L112	HGD	Q93099	Homogentisate 1,2-dioxygenase	Alkaptonuria	3q21-q23	24	z22927	45.3
MPMGp609B0817	UMPS	P11172	Uridine 5'-monophosphate synthase	Orotic aciduria	3q13	9	z3124	35.8
ICRFp524C2416	UMPS	P11172	Uridine monophosphate synthetase	Oroticaciduria	3q13	9	z5080	44
MPMGp609P0240	BCL6	P41182	b-cell lymphoma 6 protein	B-cell lymphoma	3q27	6	Z12094	49
MPMGp609H0631	COLQ	NM_005677	Acetylcholinesterase-associated collagen	Endplate acetylcholinesterase deficiency	3p24.2	19*	Z6079	39.4
MPMGp609L1635	UCHL1	P09936	Ubiquitin carboxyl-terminal esterase L1	Parkinson disease, familial	4p14	1	z7476	33.5
MPMGp609M1247	QDPR	P09417	Quinoid dihydropteridine reductase	Phenylketonuria	4p15.31	1	z6911	34.8
MPMGp609C1328	MSH3	P20585	dna mismatch repair protein msh3	Endometrial carcinoma	5q11-q12	5	z1167	55.9
MPMGp609A0636	WISP3	NM_003880	WNT1 inducible signaling pathway protein 3	Arthropathy, progressive pseudorheumatoid	6q22-q23	6	z11919	53.7
MPMGp609J0147	TWIST	Q15672	Twist	Saethre-Chotzen syndrome	7p21	19	z7	13
ICRFp524N1523	COL1A2	P08123	Collagen, type I, alpha 2	Osteogenesis imperfecta	7q22.1	19	z11403	53.5
MPMGp609E1935	BPGM	P07738	Phosphoglycerate mutase	Hemolytic anemia due to bisphosphoglycerate mutase deficiency	7q31-q34	21	z13467	32.4
MPMGp532D023	GCK	Q05810	Glucokinase	Hyperinsulinism	7p15-p13	8	z21115	57.4
MPMGp609N0536	PRKDC	NM_006904	Protein kinase, DNA-activated	Severe combined immunodeficiency	8q11	7	z11625	51.1
MPMGp609C1059	DPYS	Q14117	Dihydropyrimidinase	Dihydropyrimidinuria	8q22	Unlinked		
MPMGp609F1635	FTZF1	Q13285	fushi tarazu homologue	XY sex reversal	9q33	22	z4284	28
MPMGp609P2014	EGR2	P11161	krox-20	Neuropathy, congenital hypomyelinating	10q21.1-q22.1	12	z1473	29.1
MPMGp609H1115	LIPA	P38571	Lysosomal acid lipase	Wolman disease	10q24-q25	3*	Z21679	40.1
MPMGp609B1228	G6PT1	O43826	Glucose-6-phosphatase, transport protein 1	Glycogen storage disease type 1b	11q23	15	z6312	3.5
MPMGp609P0340	NDUFV1	P49821	nadh-ubiquinone	Leigh syndrome, Alexander	11q13	19	z3782	26

			oxidoreductase	disease				
MPMGp609L1129	CPT1A	P50416	Carnitine o-palmitoyltransferase	CPT deficiency, hepatic, type I	11q13	25	z3490	40.3
MPMGp609H2335	TSG101	Q99816	tumor susceptibility gene 101	Breast cancer	11p15.2-p15.1	25	z3490	40.3
MPMGp532N0518	LDHA	P00338	Lactate dehydrogenase A	Exertional myoglobinuria	11p15.4	25	z3632	40.3
MPMGp532L0213	MYO7A	Q13402	Myosin VIIA	Usher syndrome	11q13.5	6	Z9738	43.2
MPMGp609N0532	MEN1	O00255	Menin.	Multiple endocrine neoplasia I	11q13	7	z5649	34.3
ICRFp524O1915	PC	P11498	Pyruvate carboxylase	Pyruvate carboxylase deficiency	11q13.4-q13.5	7	z8604	57.8
ICRFp524M1095	CCND1	P24385	cyclin D1	Parathyroid adenomatosis, Centrocytic lymphoma	11q13	7	z8156	60.7
MPMGp609I1542	LMO2	P25791	Rhombotin-2	Acute T-cell Leukemia	11p13	18*	Z3853	45.9
MPMGp609A0832	SMPD1	P17405	Sphingomyelin phosphodiesterase	Nieman Pick disease	11p15.4-p15.1			
MPMGp609K0213	A2M	P01023	alpha-2-macroglobulin	Emphysema, Alzheimer	12p13.3-p12.3	15	z9214	43.3
ICRFp524D1482	MYL2	P10916	Myosin, regulatory light chain	Cardiomyopathy, hypertrophic	12q23-q24.3	3	z10934	53.8
MPMGp532F2310	PAH	P00439	Phenylalanine hydroxylase	Phenylketonuria	12q24.1	4	Z17278	53.1
MPMGp609F1211	PAH	P00439	Phenylalanine hydroxylase	Phenylketonuria	12q24.1	4	Z17278	53.1
ICRFp524N0565	PCCA	P05165	Propionyl Coenzyme A carboxylase	Propionic acidemia	13q32	1	Z1463	33.4
MPMGp609F0623	PYGL	P06737	Glycogen phosphorylase	Hers disease, glycogen storage disease type VI	14q21-q22	13	z15438	35.4
MPMGp532M184	FAH	P16930	Fumarylacetoacetatese	Tyrosinemia, type I	15q23-q25	7	z7836	0
MPMGp609I1256	TAT	P17735	Tyrosine aminotransferase	Tyrosinemia type II (richner-hanhart syndrome)	16q22.1-q22.3	18	z5442	28.4
MPMGp609C1349	HAGH	Q16775	Hydroxyacylglutathione hydrolase	Glyoxalase II deficiency	16p13	3	z5623	42
MPMGp609L0653	CLN3	Q13286	cln3	Batten, Spielmeyer-Vogt disease	16p12.1	3	z11227	57.4
MPMGp609A2343	PAFAH1B1	P43034	Platelet-activating factor acetylhydrolase	Lissencephaly-1	17p13.3	21	z3561	37.1
MPMGp609G1824	SSXT	Q15532	ssxt protein	Synovial sarcoma	18q11.2	2	z4875	0
MPMGp609N2130	AKT2	P31751	v-akt murine thymoma viral	Ovarian carcinoma	19q13.1-	10	z1450	4.7

			oncogene		q13.2			
MPMGp609E2135	LIG1	P18858	dna ligase 1	Immunodeficiencies and hypersensitivity to dna-damaging agents	19q13.2-q13.3	2	z1406	48.8
ICRFp524I095	BCAT2	O15382	Branched chain aminotransferase	Hypervalinemia or hyperleucine-isoleucinemia	19q13	3	z963	49.2
MPMGp609C2159	GCDH	Q92947	Glutaryl-coa dehydrogenase	Glutaricaciduria, type I	19p13.2	6	z17212	32.5
MPMGp609P1523	GAMT	Q14353	Guanidinoacetate	GAMT deficiency	19p13.3			
MPMGp609M2216	NOTCH3	NM_000435	zf notch 3	Arteriopathy with leukoencephalopathy	19p13.2-p13.1	Unlinked *		
ICRFp524A1223	PEPD	P12955	Peptidase D	Prolidase deficiency	19q12-q13.2	Unlinked *		
MPMGp609O0256	HNF4A	P41235	Hepatocyte nuclear factor	Diabetes mellitus, noninsulin-dependent	20q12-q13.1			
MPMGp609H2136	PFKL	P17858	6-phosphofructokinase	Hemolytic anemia	21q22.3	6	Z11919	53.7
MPMGp609D0540	CBS	P35520	Cystathionine beta-synthase	Homocystinuria, B6-responsive and nonresponsive types	21q22.3	9	Z3124	35.8
MPMGp609H0823	CRYAA	P02489	alpha crystallin a chain	Cataract, congenital, autosomal dominant	21q22.3	1*	Z5024	34.8
MPMGp609A1733	EWSR1	Q01844	rna-binding protein ews	Ewing sarcoma	22q12	10	z8146	27.9
MPMGp609N0326	ABCB7	O75027	abc transporter 7 protein	Anemia, sideroblastic, with ataxia	Xq13.1-q13.3	14	z7687	40.5
MPMGp609F0840	RPS4X	P12750	40s ribosomal protein s4	X-linked, Turner syndrome	Xq13.1	7	z8604	59.1
MPMGp609D1715	ATRX	P46100	Transcriptional regulator atrx	Alpha-thalassemia/mental retardation syndrome	Xq13			
MPMGp609P2134	PGK1	P00558	Phosphoglycerate kinase 1	Hemolytic anemia	Xq13	Unlinked *		

Table 10

(Previous page)

Comparison of map position of human disease genes and their zebrafish homologues: The table lists cDNA clones having sequences homologous to human genes involved in diseases. The orthologous human gene is shown (abbreviation, Accession number - either SWISS-PROT or RefSeq -, the gene definition, the disease it is involved in, the chromosomal location according to the GeneCards database), along with mapping data about the zebrafish gene (linkage group; closest framework marker, position of the framework marker on the MGH genetic map in cM from the top). Zebrafish sequences mapped by other groups are marked with an asterisk (*). In some cases, the zebrafish clones are in the mapping process, but map positions are not available yet.

Table 11 shows 34 zebrafish cDNA clones which showed restricted expression patterns in the systematic WMISH screen, and which were successfully mapped to the zebrafish genome. As in the human disease gene screen, only a subset of clones originally selected was finally mapped, due to unavailability of 3' sequences or already existing mapping data. We are now in the process of applying the mapping procedure to ESTs, which have been shown to be transcriptionally regulated in different developmental stages, mutants and treatments using cDNA array analysis, and which showed restricted expression patterns in whole mount in situ hybridisations (P. Aanstad, pers. communication). Together with those, the number of mapped clones adds up to > 120. All expression and mapping data are stored in the RZPD database (http://www.rzpd.de/) and will be made available to the community upon publication of the results (Musa et al., in preparation, Aanstad et al., in preparation).

Clonename	**Sequence homologue**	
	Accession-NR.	**Annotation**
ICRFp524B0117		
ICRFp524E0716	Q01703	Brachydanio rerio (zebrafish) (zebra danio). homeobox protein msh-c
ICRFp524G0916	P18669	homo sapiens (human). Phosphoglycerate mutase, brain form (ec 5.4.2.1) (pgam-b)
ICRFp524M0116	Q63429	rattus norvegicus (rat). Polyubiquitin
ICRFp524N0215	O76387	caenorhabditis elegans. c24g6.8 protein
MPMGp567B2216	P33731	canis familiaris (dog). signal recognition particle 72 kd protein (srp72)
MPMGp567C1616		
MPMGp567C232		
MPMGp567D0911	Q16658	homo sapiens (human). fascin (actin bundling protein)
MPMGp567G072	O00566	homo sapiens (human). m phase phosphoprotein 10
MPMGp567G222	Q91367	brachydanio rerio (zebrafish) (zebra danio). snail1
MPMGp567H012	Q16658	homo sapiens (human). fascin (actin bundling protein)
MPMGp567H0216	P91262	caenorhabditis elegans. cosmid f18f11
MPMGp567H0516		
MPMGp567I232	Q14520	homo sapiens (human). hgf activator like protein
MPMGp567J1810		
MPMGp567J182	O57592	fugu rubripes (japanese pufferfish) (takifugu rubripes). Ribosomal protein l7a
MPMGp567K042		
MPMGp567M0415		
MPMGp567M072	P32322	homo sapiens (human). pyrroline-5-carboxylate reductase
MPMGp567N072	Q13148	homo sapiens (human). tar dna-binding protein-43
MPMGp567N1115		
MPMGp567P1115		
MPMGp637A025	P18729	xenopus laevis (african clawed frog). gastrula zinc finger protein xlcgf57.1
MPMGp637A205	P45973	homo sapiens (human). Heterochromatin protein 1 homolog alpha (hp1 alpha) (antigen p25)
MPMGp637G075	Q18403	caenorhabditis elegans. similar to g beta repeats
MPMGp637G095		
MPMGp637I015		
MPMGp637I065		
MPMGp637K075	Q13151	homo sapiens (human). Heterogeneous nuclear ribonucleoprotein a0 (hnrnp a0)

MPMGp637K095	Q12906	homo sapiens (human). nf90 protein
MPMGp637K245	P05217	homo sapiens (human). tubulin beta-2 chain
MPMGp637M245	P41161	homo sapiens (human). ets-related protein erm (ets translocation variant 5)
MPMGp637O075		

Table 11

Zebrafish clones showing restricted expression patterns in the systematic WMISH screen successfully mapped on the T51 radiation hybrid panel. Sequence homologies in the SwissProt database and their accession numbers are shown.

4.4 Genetic mapping using amplified fragment length polymorphism (AFLP)

A modified amplified fragment length polymorphism (AFLP) protocol (Vos et al., 1995) was adapted for the zebrafish. It is particularly suited to genetically map large insert clones, as it was shown in mice, for which the method was developed (Himmelbauer et al., 1998). Genomic DNA of different strains is digested with a restriction enzyme (preferably a 6-base cutter to limit the number of fragments). Double stranded adapters are ligated to the restriction fragments and PCR is performed using primers that bind to the adapters and contain 2-3 additional nucleotides extending into the restricted DNA fragment. By this, only fragments with ends complementary to the extension and in the size range for PCR get amplified. Nucleotide exchanges at the enzyme cutting sites cause presence/absence polymorphisms of the PCR products.

In the zebrafish, hybridisation of AFLP amplicons on PAC filters showed only a very small fraction of +/- polymorphisms (not shown). In repeated experiments, most of them showed only a bad reproducibility. It was therefore concluded, that AFLP-derived complex probes are not suited for genetic mapping of the zebrafish. The reasons are probably the same as for the low efficiency of IRS-PCR based genetic mapping: Use of dominant markers, mapping strains that are not completely isogenic, and cross hybridisation due to repeat content of probes.

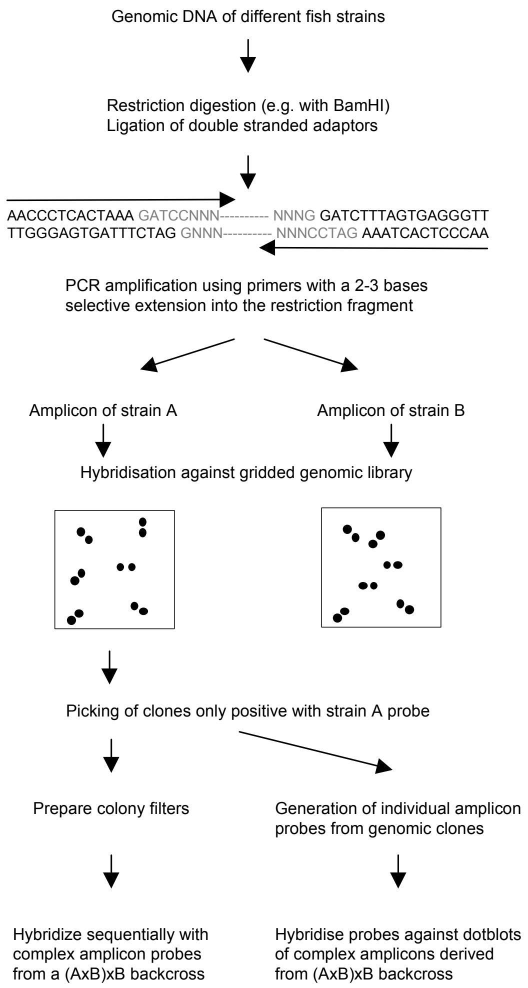

Genomic DNA of different fish strains
Restriction digestion (e.g. with BamHI)
Ligation of double stranded adaptors
AACCCTCACTAAA GATCCNNN---------- NNNG GATCTTTAGTGAGGGTT
TTGGGAGTGATTTCTAG GNNN---------- NNNCCTAG AAATCACTCCCAA
PCR amplification using primers with a 2-3 bases
selective extension into the restriction fragment
Amplicon of strain A
Amplicon of strain B
Hybridisation against gridded genomic library
Picking of clones only positive with strain A probe
Prepare colony filters
Generation of individual amplicon
probes from genomic clones
Hybridize sequentially with
complex amplicon probes
from a (AxB)xB backcross
Hybridise probes against dotblots
of complex amplicons derived
from (AxB)xB backcross

Figure 36

Outline of the AFLP based mapping procedure