Detailed report of the project NKIF- 105393: Examination of the molecular genetic background of sex determination in fish

1. The objective of the project:

The fish species exhibit enormous taxonomic, phenotypic and genetic diversity. Such genetic variation can be noticed in the sex determining genetic factors of fish, too. This means that chromosome systems and genes involved in the sex determination of fish species do not only differ from those of mammals, birds, and insects, but also from each other. However information of these is very scarce. The objective of the research was to find out more about the sex determination in economically and scientifically important fish species. The research programme was divided into three sub-tasks. First, searching and interspecific testing of sex specific genetic markers from economically important fish species in which sex determination is genetic. Second, preparation of a highly inbred line for the analyses of the zebrafish (Danio rerio) multichromosomal sex determination mechanism. Third, the *de novo* genome sequencing and analysis of the data from the African catfish (*Clarias gariepinus*) genome and identification the sex chromosome specific molecular genetic markers on contigs.

2. Searching and interspecific testing of sex specific genetic markers in fish species

The objective of this subtask was the identification of useable genetic markers for the sex determination of fish species in which sex determination proved to be genetic but the mechanism is not yet fully understood.

Tail fin clips were collected from the relevant fish species for marker search and testing. All the fish were sexed based on gametes or gonads. The following samples were used: burbot (*Lota lota*;17 spawner and 15 milter samples were collected from a Polish population -Olstyn), perch (*Perca fluviatilis*; 138 sexed sample from Biatorbágy, Dunaföldvár and Olstyn), pike perch (106 sample from Balaton, Dunaföldvár and Danube Delta) and carp (*Cyprinus carpio*-15 male and 15 female- from Ráckeve) individuals. Samples from brown trout (*Salmo trutta m. fario*), Siberian sturgeon (*Acipenser baerii*), African catfish (*Clarias gariepinus*), caras (*Carassius auratus gibelio*) and bullhead catfishes (*Ictalurus sp*p) were also available from previous sample collections.

To reveal the genetic differences between the two sexes, bulk segregant analysis was used. For the analysis, pooled DNA samples were generated from the DNA of 12 to 25 individuals of the same sexes. Two female and two male pools per species were used. The amount and concentration of DNA samples in each pool were identical, so individuals contributed equally to the DNA clusters. Pools were used for comparative analyses to reveal differences between sexes. Since pools were generated on the basis of common phenotype (on the basis of gender in this case), the individual differences „cancel each other" while differences that connected with the selected phenotype were „magnified". Thus, the observed differences between sexes are more likely to be associated with sex than the individual difference from a single sample.

To identifying differences between pools – a relatively new method, the PCR (Polymerase Chain Reaction) based FluoMEP (Fluorescent Motif Enhanced Polymorphism) was chosen, as a

high-throughput and improved method. The method can be successfully used for identifying genetic markers or for genotyping experiments, enabling the analysis of genomes where no previous information is available.

At first, 2 spawner pools and 2 milter pools from the Siberian sturgeon were tested by 47 primer combinations. All in all, comparative analysis of 4937 fragments was carried out. However, only two fragments seemed to be related to sex.

Approximately 25000 fragments were generated from perch pools by 140 primer combinations and 12 putative markers were identified.

In case of burbot approximately 11000 fragments were generated from these pools by 76 primer combinations. At first, 9 and later additional 7 putative sex linked fragments were identified.

More detailed linkage analyses of the putative markers from Siberian sturgeon, perch and burbot were carried out on individual DNA samples. Unfortunately only one of the perch putative markers showed sex specific pattern. This DNA fragment of the perch was isolated from the FluoMEP pattern and cloned to plasmid vector. Then, the amplified fragments were sequenced by capillary sequencing. The identified sequences did not contained real differences between the male and females sequences.

As an alternative strategy for sex specific marker search, in case of pike perch and bullhead catfish we tested different newly isolated genomic DNA markers (10 microsatellites and 2 SCARs) in different populations, but none of these showed sex specificity. The tested bullhead catfish markers were published at the “PeerJ”.

As a part of the task, we also tested the usability, the interspecific and inter population adaptability of previously described sex specific markers. Fourteen sex linked marker (5 microsattelite, 2 RAPD, 7 SCAR) from 9 species were tested in taxonomically related species. The tested combinations are summarised in table 1.

Table 1.: interspecific tests of different published sex-specific markers in related species.

Original species Marker ID and type	Tested species	Claria gariepinus	Cyprinus carpis	Perca fluviatilis	Ameiurus nebulosus	Carassius auratus gibelio	Lota lota	Salmo trutta
Pelteobagrus fulvidraco YelloCat_XY SCAR		x			x			
Ictalurus punctatus AUEST0678 Microsatellite		x			x			
Ictalurus punctatus IpCG02579 Microsatellite		x			x			
Oncorhynchus mykiss SDY SCAR								x
Oplegnathus fasciatus OpI286 Microsatellite				x		x		
Ciprinus carpio S2107 RAPD			x					
Ciprinus carpio Ccmf1 SCAR			x					
Ciprinus carpio Ccmf2 SCAR			x					
Ciprinus carpio Ccmf3 SCAR			x			x		
Carassius auratus gibelio MSM SCAR			x			x		
Scleropages formosus OPA-019 RAPD			x					
Brycon amazonicus SCAR			x					
Gadus morhua MS 89 Microsatellite							x	
Gadus morhua MS128 Microsatellite							x	

Although, the primers amplified the putative regions of the genomes from the closely related species, the fragments did not show the expected differences between the two sexes. Even the three published sex-linked markers from Asian subspecies of common carp (*Ciprinus carpio hemopterus*) did not show any sex-specific differences in the most related European subspecies (*Ciprinus carpio carpio*). The only exception was found in the *Salmonidae* family. Brown trout samples were checked with the described *Salmonidae* sex determining gene specific PCR and we found some controversial result compared with phenotypic sex. For this reason we have made more detailed testing. Altogether, 301 morphologically sexed adult brown trout individuals were tested with the sex specific PCR from the Lillafüred and Szilvásvárad populations. The phenotyping were made by experts, based on secondary sex characteristics. Only 42% of the spawners and 69% of the milters were phenotyped correctly compared to the molecular genetic test. While in a more detailed experiment the sex of 60 surgically phenotyped individuals were 100% identical with the PCR test. Presumably the previous differences were caused by imperfect phenotyping caused by hardly expressed sex specific traits. These results indicate strong unreliability of phenotypic sexing and the practical importance of the molecular genetic method. The results of the analyses of brown trout (*Salmo trutta*) broodstocks were published at the "Halászat". In addition, an MSC and a BSC thesis were written from the topic of the subtask..

3. Analyses of the zebrafish (*Danio rerio*) sex determination

Zebrafish is the most important model animal among fish, but despite of its widely analysed biological background and described genome sequences the exact mechanism of sex determination or the genes involved in this process, are not fully explored. For the easier identification of genomic regions responsible for sex determination, we started the development of a specific stock having a very narrow genetic background. Gynogenesis is one of the fastest ways to produce highly inbred strains and it has been successfully utilized for the generation of 'double haploids' or 'dihaploids' in a number of teleost species. To establish a stock, dihaploid individuals were generated by mitotic gynogenesis. In the gynogenesis, irradiated carp sperm was used (carp and zebrafish can not produce viable hybrids). Genetic material of the sperms was inactivated by gamma radiation, and the first cell division was blocked by heat shock, according to the protocol of Streisinger and collaborators (1981). More than 110 female zebrafish from the AB strain were striped gently (approx. 200 egg/individual) following anaesthethesia, without sacrificing the females. After the fertilization and the heat shock, survival rate was less than 1%, as expected. Haploid and hybrid embryos died and only diploid larvae could survive, in which the inhibition of the first cell division was successful. Only 184 larvae hatched from more than 18,000 eggs. However, most of them were dead in later developmental stages, due to different developmental disturbances (probably these were caused by genetic defects, which are lethal in the homozygous state, those have been revealed the presumable presence of dihaploid genome). The survival rate was lower than 1% as it was expected. Finally, 14 individuals reached adulthood. Five of them (4 males and one putative female) were not able to produce offspring, while 7 males and 2 females were fertile. These individuals were used to produce the F1 generation. The F1 generation consisted of 46 individuals and only one individual displayed the male phenotype. This male was used to produce the F2 generation. In the F2 generation 42 individuals (offspring from one spawning) reached the adulthood, and all of them were female

We observed a sex-biased nature of the highly homogenous lines both in the gynogenetic (male biased) and the F1 and F2 generations (female biased; ~98-100%). Presumably this is a result of the multi chromosomal sex determination of *Danio rerio*. Unfortunately, it was not possible to continue the inbreeding process. Our results show that irradiated and cryopreserved sperm of carp can be used for gynogenesis in zebrafish.

The gynogenetic and the F1 generation were genotyped by 20 microsatellites (those were originated from different chromosome arms) and every one of them proved that they were true di-haploids. 20 microsatellites were analysed. Every one of the gynogenotes showed homozygosity for all examined loci (they were 100% dihaplods), while the F1 generation had higher genetic diversity, due to the fact that the different dihaploid gynogenotes are not identical clones. The random segregation and the recombination of the maternal chromosomes resulted different dihaploid genomes in the G0 generation.

Our results show that irradiated and cryopreserved sperm from carp can be used for the production of di-haploid zebrafish. However, adequate control groups were not available to check the sperm and egg quality before the treatments, as the unfertilized, haploid and interspecific hybrid individuals (from carp x zebrafish crosses) have a similar phenotype and they are not viable. As expected, the survival rate was lower than 1%. We observed a sex-biased nature of the highly homogenous lines both in the gynogenetic (male biased) and the F1 generation (female biased; ~98%). Presumably this is a result of the polygenic sex determination system that was hypotetized in domesticated strains of zebrafish. To the best of our knowledge, this is the first experiment that produced gynogenetic zebrafish individuals with irradiated, cryopreserved sperm from related species. In addition our approach has solved the sperm quantity/quality, logistic and handling problems, but the sex biased nature of the highly inbred lines of zebrafish remains an issue to be tackled. A BSc thesis were written about the microsatellite analyses of the gynogenotes. A scientific paper from these results is under construction.

4. Genome wide analyses of the African catfish

The original plan for the sequence analysis of sex determining regions in the African catfish was to build a BAC (Bacterial artificial chromosome)-based genomic DNA library which would be screened for sex chromosome representing BAC clones harbouring previously identified sex-specific markers and determine their sequences. In the meantime, the cost of new-generation sequencing significantly fall, thus setting a switch to a full genome-wide sequencing to replace the lengthy process of BAC cloning. This new method provided substantially more information for our research. For the sequence analysis, individuals were obtained from different stocks and tissue samples were collected. For further DNA and RNA analyses, two tissue or organ samples from the tail fin, gonad, brain, kidney, head kidney, heart, skin, liver, gills and barbs were collected from all individuals. Samples were frozen in liquid nitrogen and stored at -80 °C until analysis. DNA was isolated from tail fin clips for further investigation. Then the applicability and integrity of DNA samples were checked by fluorimetry (Qubit) and spektrofotometry (IMPLEN). Concentration and purity of samples were determined and integrity was re-checked by agarose gel electrophoresis. Useability of previously developed sex specific markers were tested on the samples of collected individuals, and species identity was checked by the sequence analysis of cytochrome-oxidase I. and cytochrome-b-gens. From the available samples, one male DNA was selected for whole genome sequencing on Illumina HiSeq 2000 platform .A „paired-end" (PE1)

and a „mate-pair”(MP) genomic library were sequenced, using TruSeq chemistry. The average insert size was smaller than 350 bp et PE library while in case of the MP library it was 5000bp. The two library sequencing yield 71,1 billion basis information (281447634 basis from the PE and 430208098 basis from MP library). This amount of data can cover the African catfish genome approximately 60 times. We have started building up the draft *de novo* sequence of African catfish genome. NIIF (National Information Infrastructure Development Institute) made the necessary “supercomputer” available for us at University of Pécs and later at Miskolc. The quality of the sequence data were checked by Trimmomatic software and more than 89% of the data were usable for the later process. During the data processing we compared three assembler softwares. These are the Allpaths-lg and SOAPdenovo, and the minia program. The best result was generated by the Allpaths-lg. This draft genome contains 12477 scaffold. The total length of the sequences is 939385245 bp. This is approximately 80% of the predicted 1173 Mb long full genome sequence. The program predicted the presence of 96 820 genes, nevertheless this number is ~4 times higher than expected based on the known genomes. Both Y chromosome specific markers were identified in the draft genome, however the carrier contigs are too small (CgaY1 were identified on a 6kb long, the CgaY2 on a 44Kb long contig).

To improve the quality of the draft genome two additional genomic library were generated and sequenced. One of the later was a “long jump” (LJ) nuclear genomic library (with 8Kb insert size), while the other was a “Shotgun” library sequencing (PE2). HiSeq 2500 platform of Illumina was used for data collections. For the library preparation the DNA of the same individual was used as previously. The “paired –end” library yield 37,1 Mbp data from the male sample and 28,6 Mbp data were gained from the “long jump” library. The two library contain 65,7 Mbp data. The second genome alignment contained all four libraries with ~114x coverage, (136,8 billion bases). The second sequence assembly was performed with Allpaths-lg and Ray. The best result was generated by Allpaths-lg (8783 scaffolds, N50:848271). All estimated quality values were enhanced significantly by the raised coverage. The total length of the estimated genome size is 1070503 bp. This is aproximetly 90% of the C value based predicted genome size, which corresponds to the expected level.

For the assembly of the mitochondrial genome PE1 reads were processed with the MITObim (mitochondrial baiting and iterative mapping) pipeline. In the first step, mirabait program from the MIRA version 4.0.2 package was used to select mitochondrial reads. For baiting we used the *Clarias fuscus* mitochondrium genome sequence. In the second step we used the Mitobait.pl script, which carried out three iterative steps to build the mitochondrial genome. The final sequence was further trimmed at the ends manually. The MITOS software was used for the annotation of the genes and the result was corrected by hand. The resulted mithochondrial genome of African catfish is 16,505 base pair. The overall base composition of African catfish is 24,7% for T, 32,6 % for A, 27,8% for C, and 14,9% for G. It means somewhat lower GC (42,7%) content. The mitogenome comprised the 37 typical mitochondrial genes including the 12 S rRNA and 16 S rRNA genes, 22 tRNA genes and 13 protein-coding genes (PCGs) and the control region. The order and structure of the genes corresponded to the teleost’s mitochondrion. Except for ND6 and eight tRNA genes (tRNAGln, tRNAAla, tRNAAsn, tRNACys, tRNATyr, tRNASer, tRNAGlu and tRNAPro), encoded on the H-strand, all other genes were encoded on the L strand. In all 13 PCGs, normally ATG was used as the start codon except COX1, which began with GTG. Similarly, the end codons of the protein coding genes were varied with TAA, TAG, or incomplete stop codon T, which presumably become complete termination codons by polyadenylation. However the TAG -stop codon was not used by *C. gariepinus*. The major non-coding region (D-

loop) between the tRNAPro and tRNAPhe genes found to be 862 bp in length. A second, 31-nucleotide long, non-coding sequence was identified between the tRNAAsn and tRNACys genes. The annotation of the genomic DNA was started using the NCBI BLAST suite (v2.2.8).

For the annotation of the genome a BLAST database was created using the repeat masked version of the African catfish genome using the makeblastdb command (the RepeatMasker version4.0.5. was used for masking). All protein sequences originated from the *Clariidae* family and the *Siluridei* order were downloaded from the UniProt database, while all cDNA and CDS sequences of the *Danio rerio* and the three-spined stickleback (*Gasterosteus aculeatus*) were downloaded from the ENSEMBL database for gene identification in the draft genome. All groups of sequences were aligned separately. 72% (518) of the 722 *Clariidae* sequences and 75% (11867) of the 15808 *Siluridei* proteins were found in the draft genome with 80% or higher identity. On the other hand, zebrafish and three-spined stickleback CDS/cDNA showed much lower homology, only 23-31% (10281-8366) of the query (44487-27576) sequences were found.

To get more information from the differentially expressed genes, male and female transcriptomes were sequenced from mixed brain and gonad tissues, by Illumina HiSeq2500. The transcriptome sequencing yielded 20,8 Mbp data from the male and 22,5MBp data from the females samples. During the comparative analyses of the data (after the quality check, the selection and the assembly) approximately 160 million reads were accepted. Filtered reads were combined for the *de novo* transcriptome assembly. A total of 311,187 genes were assembled, with 367,161 transcripts. The assembled transcriptome was annotated based on the Trinotate pipeline. Coding regions were identified in the assembled transcripts using Transdecoder and PFAM. Protein and DNA homologies were identified using the standalone BLAST suite. Protein domains were identified using HMMER. Signal peptides and transmembrane regions were identified using SignalP and tmHMM. Functional annotation was performed using the KEGG, GO and egg NOG data. Transcripts were filtered based on annotation similarity, of which 35,000 had an identity of at least 30% to a known gene, and 15,129 transcripts had an annotation identity above 70%. Since the two samples were isolated from female and male specimen, we compared the gene expression values of the assembled and annotated genes. We first aligned the RNA-seq reads of the two samples to the reference transcriptome using the bowtie algorithm, then quantified expression values using the RSEM package. Since the expression value of many genes showed high fold changes between the samples, we filtered for genes where the FPKM value is over 100 in at least one of the samples and a fold change above 2, to reduce the high false positive count. By applying our hard filters, we identified the genes that showed strong expression change in the two samples. More than 570 genes had expression change between the two samples, and had an annotation closely related to the gender. Based on the expression differences and pathway analyses of these genes we have used 19 genes (amhr2, dmrt2, amh, msl1, fancd2, sox3, akap11 x4, cyp19, dmrt3, hsd11b, alpl, fsh, akap11 x1-3, dmrt1, pten, fst, ar, sox9, wnt4) to confirm the expression results by real time PCR. For the analyses, pooled RNA samples were used from the male and female gonads and brains in three repetitions. Six individual samples were used for each pool. The analyses showed good regiment with the transcriptome analyses, however one gene, the, akap11 x1-3 did not show significant differences. Most of the genes showed elevated gene expression in the male gonad (dmrt1, dmrt2, dmrt3, amh, fst, ar, wnt4,akap11x4) while only the amhr2 showed higher expression in the female gonad. Some gene had higher expression in the brain compared to the gonads (sox9, hsd11b, alpI, cyp19, sox3) and the fancd2 had higher expression in the gonads.

These results confirm the expression data of the transcriptomes and add useful information to the sex determination background of the African catfish. One publication from the *de novo*

genome sequencing and one from the transcriptome analyses are under construction. In addition, the data of the expression analyses are part of an MSc dissertation.

5. Summary

In the project we have searched sex specific genetic markers in 3 fish species wit FlupMEP analyses, and by interspecific adaptation of known markers in 7 species, but only the *Salmonidae* family specific SDY marker showed difference between the two sexes of brown trout. These results show that, the genetic backgrounds of the sex determination systems in different fish families differ much more than it was expected. In addition the complex sex determination mechanism of zebrafish was also analysed. Dihaploid individuals were produced, by an improved method of mitotic gynogenesis, and two offspring generations were created as well. Both, the gynogenetic (~80% male) and the F1 and F2 generations (~98-100% female) were sex-biased. Presumably, this is a result of the suspected multi chromosomal sex determination of *Danio rerio* laboratory lines. The deepest analysis was made on the African catfish for identification of the sex determining region and genes that are involved in the process. The first draft genome of the species was performed by using four genomic libraries of one heterogametic male individual. Using the available genetic sex markers of the species, two sex related scaffolds were identified. The male and the female *de novo* transcriptomes were determined and compared from mixed brain and gonad tissues. More than 570 differentially expressed genes were identified. These are revealed the disparity between the male and female sample. All these results are demonstrating the wide genetic diversity of the sex determination systems and provide important information to further analyses.