Identification and analysis of genome-wide SNPs provide insight into signatures of selection and domestication in channel catfish (*Ictalurus punctatus*)

by

Luyang Sun

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 13, 2014

Key words: catfish; SNP; marker; genome selection; selective sweeps; domestication

Approved by

Zhanjiang Liu, Chair, Alumni Professor of School of Fisheries, Aquaculture, and Aquatic Sciences
Rex Dunham, Alumni Professor of School of Fisheries, Aquaculture, and Aquatic Sciences
Eric Peatman, Associate Professor, Professor of School of Fisheries, Aquaculture, and Aquatic Sciences
Joanna Diller, Associate Professor of Biological Sciences

Abstract

Domestication and selection for important performance traits can impact the genome, which is most often reflected by reduced heterozygosity in and surrounding genes related to traits affected by selection. In this study, analysis of the genomic impact caused by domestication and artificial selection was conducted by investigating the signatures of selection using single nucleotide polymorphisms (SNPs) in channel catfish (*Ictalurus punctatus*). A total of 8.4 million candidate SNPs were identified by using next generation sequencing. On average, the channel catfish genome harbors one SNP per 116 bp. Approximately 6.6 million, 5.3 million, 4.9 million, 7.1 million and 6.7 million SNPs were detected in the Marion, Thompson, USDA103, Hatchery strain, and wild population, respectively. The allele frequencies of 407,861 SNPs differed significantly between the domestic and wild populations. With these SNPs, 23 genomic regions with putative selective sweeps were identified that included 11 genes. Although the function for the majority of the genes remains unknown in catfish, several genes with known function related to aquaculture performance traits were included in the regions with selective sweeps. These included hypoxia-inducible factor 1β (*HIF1β*) and the transporter gene ATP-binding cassette sub-family B member 5 (*ABCB5*). HIF1β is important for response to hypoxia and tolerance to low oxygen levels that is a critical aquaculture trait. The large numbers of SNPs identified from this study are valuable for the development of high-density SNP arrays for genetic and genomic studies of performance traits in catfish.

Acknowledgments

First, I would like to extend my special thanks to my advisor Dr. Zhanjiang Liu for his valuable guidance, encouragement, support and patience during my Ph.D. study. I would also like to thank all my other committee members: Dr. Eric Peatman, Dr. Rex Dunham and Dr. Diller for making me learn more and be confidence in myself and my knowledge. My thanks also extend to the outside reader of my dissertation, Dr. Bernhard Kaltenboeck for his support and guidance. I am very grateful for the technical assistance by Dr. Huseyin Kucuktas and Ludmilla Kaltenboeck. And I would like to thank Dr. Shikai Liu, Dr. Ruijia Wang, Dr. Chao Li, Dr. Yu Zhang, Dr. Jiaren Zhang, Jun Yao, Lisui Bao, Xin Geng, Yun Li, Chen Jiang, Ailu Chen and all the other colleagues in the laboratory for their help, collaboration, and friendship. I would like to especially thank the Chinese Scholarship Council for the financial support.

Finally, I am grateful to my father Song Sun, my mother Guohua Liu and my beloved fiancée Li Zhou for their endless love and tremendous support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

CNV	Copy Number Variations
GWAS	Genome-wide association study
INDELs	Insertions and Deletions
MAF	Minor Allele Frequency
MAS	Marker Associated Selection
MSI	Molecular Selection Indices
NGS	Next Generation Sequencing Technologies
QTL	Quantitative Trait Loci
SNP	Single-Nucleotide Polymorphism
SSR	Simple Sequence Repeats

CHAPTER I

INTRODUCTION

Overview

Channel catfish (*Ictalurus punctatus*) is one of the most important aquaculture species in the U.S. Many commercial strains have been developed by the selection of traits with high economic values such as growth rate and disease resistance. However, the traditional selective breeding approaches are time-consuming and relatively inaccurate, especially in the selection of traits which exhibit low heritability or sex-related, because fish are selected only based on their phenotypes other than genotypes. Currently, genome-wide selection is the state of the art for genetic improvements of livestock species and poultry species, which is more precise in selection of genomic regions for the favorable alleles, and has the ability to shorten the time frame required for selection. However, genome-wide selection is not yet widely adopted with aquaculture species. With catfish, genome selection has not been conducted because tightly linked molecular markers with performance traits have not been identified. In order to identify markers that are closely linked with quantitative trait loci (QTL), large numbers of SNPs covering the whole genome are required.

In addition to the analysis of QTLs, SNPs are also useful for other genetic analysis such as starin identification, analysis of genetic variations within aquaculture populations, assessment of inbreeding within aquaculture populations, and analysis of selective sweeps after domestication and artificial selection.

SNP identification

DNA sequence variations are one of the key factors for understanding biological diversity, genome evolution and function (Kidd, Pakstis et al. 2004). Within a given species, genome sequences are highly similar among individuals, but there are sequence polymorphisms across the genomes including insertions and deletions (INDELs), inversions, translocations, copy number variations (CNVs), and of course, single nucleotide polymorphisms (SNPs). It is such genomic variations that form the basis of phenotypic differences. Of these genomic variations, SNPs have become the molecular markers of choice because of their high abundance, even genomic distribution, and suitability for automation. Of these characteristics, SNPs are best suited for automated genotyping using a number of platforms such as the Sequenom MassArray technology, the Illumina BeadArray technology and the Affymetrix Axiom Array technology (Oliphant, Barker et al. 2002, Gabriel, Ziaugra et al. 2009, Hoffmann, Kvale et al. 2011), all of which are capable of genotyping a very large number of SNPs, as well as a large number of samples at a relatively low cost on a per genotype basis. With the availability of a large number of SNPs, automated genotyping platforms can be developed to fit the situations of the species of interest (Matukumalli, Lawley et al. 2009, Ramos, Crooijmans et al. 2009, Groenen, Megens et al. 2011, Kranis, Gheyas et al. 2013).

The next generation sequencing technologies (NGS) provided great advantages for the identification of genome-wide SNP variations (Mardis 2008). To date, a large number of SNPs have been identified from a wide range of organisms. Over 187 million and 50 million SNPs have been identified in human and mouse, respectively (Keane, Goodstadt et al. 2011). Also, large numbers of SNPs were identified from agricultural species such as cattle (Gibbs, Taylor et

al. 2009, Matukumalli, Lawley et al. 2009, Zhan, Fadista et al. 2011), sheep (Kijas, Townley et al. 2009), chicken (Wong, Liu et al. 2004, Marklund and Carlborg 2010), pig (Wiedmann, Smith et al. 2008, Ramos, Crooijmans et al. 2009), turkey (Kerstens, Crooijmans et al. 2009, Aslam, Bastiaansen et al. 2012) as well as from aquatic animals such as zebrafish (Guryev, Koudijs et al. 2006, Bradley, Elmore et al. 2007), Pacific salmon (Smith, Elfstrom et al. 2005), common carp (Xu, Ji et al. 2012), Atlantic herring (Helyar, Limborg et al. 2012), and Atlantic cod (Hubert, Higgins et al. 2010).

Great efforts have been devoted to discovery of SNPs in catfish. Back in 2004, He et al. used an approach of comparative EST analysis to identify interspecific SNPs between channel catfish and blue catfish for applications in mapping using the interspecific hybrid system. Liu et al. (2011) conducted RNA-Seq analysis using pooled RNA samples from multiple individuals and identified several hundreds of thousands of gene-associated SNPs. In spite of such progress, genome-scale SNPs have not been available for catfish.

SNP application in aquaculture and ecology

The most important application of SNPs in aquaculture is marker-assisted selection (MAS) and whole genome selection. In a sense, whole genome selection is a type of MAS using markers distributed across the entire genome. MAS was first developed in 1990s. It is an improtant tool to supplement the traditional selection with trait-linked DNA markers. One basic form of MAS is to select the progeny of specific progenitors on the basis of molecular markers linked to the traits of interest (Dekkers and Dentine 1991, Arus and Moreno-González 1993). Another form of MAS is to establish molecular selection indices (MSI) using both information of molecular marker linked

to the traits of interest and the phenotypeic values of the traits of interest (Lande and Thompson 1990). MSI is more often used for the seelction of multiiple traits simultaneously. In MSI, a specific “value” is given to a specific trait such that selection of one trait will not negatively affect the selection of another trait. With MAS, selection is very effective, but selection based on partial knowledge on the whole genome can have negative outcome, simply because many other genomic regions could also affect the trait under selection. Therefore, genome wide selection is considered the most comprehensive approach using molecular marker and phenotypic information (Hayes and Goddard 2001).

In order to implement MAS or whole genome selection, SNPs associated with the traits of interest need to be identified first. A number of approaches have been developed for searching the associated SNPs or genome regions such as genome-wide association study (Hirschhorn and Daly 2005), QTL mapping (Goddard and Hayes 2009) and Bulk-segregant RNA sequencing (Wang, Sun et al. 2013). In livestock species, GWAS were conducted for the identification of SNPs related to body conformation (Wu, Fang et al. 2013), disease resistance (Purdie, Plain et al. 2011), feed efficiency (Abo-Ismail, Vander Voort et al. 2014), and milk production of cow (Raven, Cocks et al. 2014). With aquaculture species, few GWAS analysis have been done while there is a number of QTL mapping analysis using SSRs, SNPs or both. For instance, in rainbow trout, the QTLs related to the esmoregulation capacities and crowding responses were identified using SSR and SNP markers (Le Bras, Dechamp et al. 2011, Rexroad, Vallejo et al. 2012). In Atlantic salmon, fine-mapping was conducted to identify QTLs involved in resistance to infectious pancreatic necrosis (Houston, Haley et al. 2008, Moen, Baranski et al. 2009).

Strain differentiation, species differentiation and parentage analysis are also important applications of SNPs, and such analysis are useful for both aquaculture programs and ecological conservation programs. For instance, there are needs to differentiate individuals, families, strains and species. The use of phenotypes alone may not provide sufficient power of differentiation, especially at sub-species levels. However, there are always some differences in the genome, especially SNPs, even within the individuals of the same spawn, which provide the basis for molecular differentiation. For instance, channel catfish strains are difficult to distinguish as they are almost identical in their appearances. SNPs can be used to readily distinguish them.

SNPs can also be used in conservation and ecological studies, especially for population genetics (Etter, Bassham et al. 2011). By analyzing their allele frequencies, polymorphism level and linkage disequilibrium, population genetic structure, levels of in-breeding, selection pressure and evolutionary relations with other populations can be determined. When dealing with endangered species, such information can be used to protection the species effectively. (Li, Fan et al. 2009). In addition, SNP information of some aquatic microbes such as aquatic hyphomycete, which is a bio-indicator, can be used for monitoring and assessing anthropogenic stress and environmental ecosystems health (Krauss, Solé et al. 2011).

Another application of SNP markers is sex identification for aquaculture species. With aquaculture species, sex determination at early stages of life can be economically important. For instance, female half-smooth tongue sole grow several times faster than their male counterparts. Therefore, all female populations are desired for aquaculture. However, females

and males are not morphologically differentiable early in their life history. SNPs can be applied to differentiate females from males to allow culture of only females (Chen, Tian et al. 2009).

Advantages of SNPs as compared to microsatellites

Microsatellites are short tandem repeats. They are also highly abundant although not as abundant as SNPs in genomes. Microsatellites have been very popular for population and genetic studies because of their abundance, wide genome distribution, and small locus sizes allowing genotyping by PCR. Microsatellites are particularly well suited for population analysis because of their very high levels of polymorphisms. In population, a given microsatellite can have many alleles, and as many as 17 have been reported in catfish (Waldbieser and Bosworth 1997). However, automation of genotyping of microsatellites is difficult and therefore, genotyping of microsatellites can be laborious and expensive. SNPs, on the other hand, are even more abundant than microsatellites in the genome. Therefore, SNPs provide a better genome coverage than microsatellites. However, most SNPs are bi-allelic although as many as four alleles are theoretically possible in the population. Nonetheless, such shortcomings are compensated by their high adaptability for automation. Because of automated genotyping, simultaneous genotyping of millions of SNPs is possible by using high-density SNP arrays (Oliphant, Barker et al. 2002, Gabriel, Ziaugra et al. 2009, Hoffmann, Kvale et al. 2011).

Microsatellites have been successfully used for traceability for years and SNP are also increasingly used for this purpose. Both of them can be used as genetic markers for traceability. Due to the fact that many alleles of microsatellites exist in the population, microsatellites are more powerful in their differentiating powers for traceability type of applications. Herraez et al. (2005) compared the performances of microsatellites and SNPs in a Galloway cattle population

by analyzing exclusion power of both kinds of markers for individual identification and parental analysis (Herraez, Schafer et al. 2005). In general, the performance of 3-4 SNPs is equivalent to the one SSR and the paternity exclusion is over 99% for SSRs and about 98% for SNPs. Similar results were obtained by Fernández et al. (2013); they compared the effectiveness of microsatellites and SNP panels for traceability and genetic identification in an inbred angus herd (Fernández, Goszczynski et al. 2013). Two parameters were used to evaluate the performance of these two kinds of markers: cumulative SNP exclusion power values (Q) and sample matching probability (MP). Generally, the performance of 2-3 SNPs was equivalent to one SSR on Q value and MP. Both studies illustrated that both SNPs and SSRs are well-suited for traceability, but for each marker unit, the performance of SSRs is better than SNPs. However, when heterzygosity of populations is low, SNPs may function better than SSRs. A study focusing on the performance of SSRs and SNPs in the Lowland European bison (*Bison honasus*) have been conducted in 2009. The Lowland European bison are descended from just seven founders and two of them contributed more than 80% of gene pool. Under this situation, 17 SSRs and 960 SNPs were used for paternity and identity analysis and the results showed that SSR cannot successfully determine the paternity and identity in the European bison while SNPs can (Tokarska, Marshall, et al. 2009). These results indicated that SNP genotyping is more powerful than SSR for genetic analysis in related species and bottlenecked species, although a much larger number of SNPs are required to provide the differentiation power.

Genomic evolution and artificial selection

When a gene is under selection, the genetic diversity in the locus tends to decrease. Such reductions in genetic diversity have been observed not only within the gene under selection, but

also along the surrounding genomic regions because of genetic linkage. This phenomenon is described as hitch-hiking effect and genomic regions with low genetic diversity caused by hitch-hiking effect is referred to as selective sweep (Smith and Haigh 1974). A number of studies focused on selective sweep have been conducted in species like human (Diller, Gilbert et al. 2002, Hernandez, Kelley et al. 2011) mouse (Ihle, Ravaoarimanana et al. 2006, Teschke, Mukabayire et al. 2008), wheat (Raquin, Brabant et al. 2008), and maize (Palaisa, Morgante et al. 2004). Selective sweep analysis has been conducted with agricultural species to assess the impact of domestication and selection on genome composition. For instance, selective sweeps have been identified in chicken (Johansson, Pettersson et al. 2010, Rubin, Zody et al. 2010), pig (Rubin, Megens et al. 2012) and cattle (Boitard and Rocha 2013, Ramey, Decker et al. 2013). Unlike the livestock species, where selection has been taking place for a relatively a long period of time, the domestication and selection of aquaculture species including that of catfish has a relatively short history of less than 50 years. With fish species, selective sweeps have been reported in three-spined stickleback (*Gasterosteus aculeatus*) (Cano, Matsuba et al. 2006, Mäkinen, Shikano et al. 2008, Hohenlohe, Bassham et al. 2010), and Atlantic salmon (*Salmo salar*) (Vasemägi, Nilsson et al. 2012).

Catfish is an important aquaculture species in the United States. Its domestication and selection has a short history. In the last 50 some years, domestic populations have been established with channel catfish and blue catfish, and their selective breeding programs have focused on a number of performance traits including growth rates, feed conversion efficiency, low oxygen tolerance, and disease resistance among many other traits (Dunham and Smitherman 1983, Dunham, Brady et al. 1994, Geng, Feng et al. 2014). In spite of the progress, detailed genetic analyses of

domestication and selection have not been conducted. Selective sweeps in catfish are unknown at present.

Reference

1. Abo-Ismail, M. K., et al. (2014). "Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle." BMC Genetics 15(1): 14.
2. Arus, P. and J. Moreno-González (1993). Marker-assisted selection. Plant Breeding, Springer: 314-331.
3. Aslam, M. L., et al. (2012). "Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo)." Bmc Genomics 13: 391.
4. Boitard, S. and D. Rocha (2013). "Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed." Animal genetics.
5. Bradley, K. M., et al. (2007). "A major zebrafish polymorphism resource for genetic mapping." Genome Biol 8(4): R55.
6. Cano, J., et al. (2006). "The utility of QTL-Linked markers to detect selective sweeps in natural populations—a case study of the EDA gene and a linked marker in threespine stickleback." Mol Ecol 15(14): 4613-4621.
7. Chen, S.-L., et al. (2009). "Artificial gynogenesis and sex determination in half-smooth tongue sole (Cynoglossus semilaevis)." Marine biotechnology 11(2): 243-251.
8. Dekkers, J. and M. Dentine (1991). "Quantitative genetic variance associated with chromosomal markers in segregating populations." Theoretical and applied genetics 81(2): 212-220.
9. Diller, K. C., et al. (2002). "Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees." Molecular Biology and Evolution 19(12): 2342-2345.
10. Dunham, R. A., et al. (1994). "Response to challenge with Edwardsiella ictaluri by channel catfish, Ictalurus punctatus, selected for resistance to E. ictaluri." J Appl Aquaculture 3: 211-222.
11. Dunham, R. and R. Smitherman (1983). "Crossbreeding channel catfish for improvement of body weight in earthen ponds." Growth 47: 97-103.
12. Etter, P. D., et al. (2011). SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Molecular methods for evolutionary genetics, Springer: 157-178.
13. Fernández, M. E., et al. (2013). "Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd." Genetics and molecular biology 36(2): 185-191.
14. Gabriel, S., et al. (2009). "SNP genotyping using the Sequenom MassARRAY iPLEX platform." Curr Protoc Hum Genet Chapter 2: Unit 2 12.
15. Geng, X., et al. (2014). "Transcriptional regulation of hypoxia inducible factors alpha (HIF-α) and their inhibiting factor (FIH-1) of channel catfish (Ictalurus punctatus) under hypoxia." Comp Biochem Physiol B Biochem Mol Biol 228: 91-105.

16. Gibbs, R. A., et al. (2009). "Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds." Science 324(5926): 528-532.
17. Goddard, M. E. and B. J. Hayes (2009). "Mapping genes for complex traits in domestic animals and their use in breeding programmes." Nature Reviews Genetics 10(6): 381-391.
18. Groenen, M. A., et al. (2011). "The development and characterization of a 60K SNP chip for chicken." Bmc Genomics 12(1): 274.
19. Guryev, V., et al. (2006). "Genetic variation in the zebrafish." Genome Res 16(4): 491-497.
20. Hayes, B. and M. Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics 157(4): 1819-1829.
21. Helyar, S. J., et al. (2012). "SNP Discovery Using Next Generation Transcriptomic Sequencing in Atlantic Herring (Clupea harengus)." PLoS One 7(8): e42089.
22. Hernandez, R. D., et al. (2011). "Classic selective sweeps were rare in recent human evolution." Science 331(6019): 920-924.
23. Herraez, D. L., et al. (2005). "Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population." Zeitschrift fur Naturforschung C-Journal of Biosciences 60(7-8): 637-643.
24. Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nature Reviews Genetics 6(2): 95-108.
25. Hoffmann, T. J., et al. (2011). "Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array." Genomics 98(2): 79-89.
26. Hohenlohe, P. A., et al. (2010). "Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags." PLoS Genet 6(2): e1000862.
27. Houston, R. D., et al. (2008). "Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar)." Genetics 178(2): 1109-1115.
28. Hubert, S., et al. (2010). "Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua)." Bmc Genomics 11.
29. Ihle, S., et al. (2006). "An analysis of signatures of selective sweeps in natural populations of the house mouse." Molecular Biology and Evolution 23(4): 790-797.
30. Johansson, A. M., et al. (2010). "Genome-wide effects of long-term divergent selection." PLoS genetics 6(11): e1001188.
31. Keane, T. M., et al. (2011). "Mouse genomic variation and its effect on phenotypes and gene regulation." Nature 477(7364): 289-294.
32. Kerstens, H. H., et al. (2009). "Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey." Bmc Genomics 10: 479.
33. Kidd, K. K., et al. (2004). "Understanding human DNA sequence variation." J Hered 95(5): 406-420.
34. Kijas, J. W., et al. (2009). "A genome wide survey of SNP variation reveals the genetic structure of sheep breeds." PLoS One 4(3): e4668.

35. Kranis, A., et al. (2013). "Development of a high density 600K SNP genotyping array for chicken." Bmc Genomics 14(1): 59.

36. Krauss, G. J., et al. (2011). "Fungi in freshwaters: ecology, physiology and biochemical potential." FEMS microbiology reviews 35(4): 620-651.

37. Lande, R. and R. Thompson (1990). "Efficiency of marker-assisted selection in the improvement of quantitative traits." Genetics 124(3): 743-756.

38. Le Bras, Y., et al. (2011). "Detection of QTL with effects on osmoregulation capacities in the rainbow trout (Oncorhynchus mykiss)." BMC Genetics 12(1): 46.

39. Li, R., et al. (2009). "The sequence and de novo assembly of the giant panda genome." Nature 463(7279): 311-317.

40. Mäkinen, H. S., et al. (2008). "Hitchhiking mapping reveals a candidate genomic region for natural selection in three-spined stickleback chromosome VIII." Genetics 178(1): 453-465.

41. Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet 24(3): 133-141.

42. Marklund, S. and O. Carlborg (2010). "SNP detection and prediction of variability between chicken lines using genome resequencing of DNA pools." Bmc Genomics 11(1): 665.

43. Matukumalli, L. K., et al. (2009). "Development and characterization of a high density SNP genotyping assay for cattle." PLoS One 4(4): e5350.

44. Moen, T., et al. (2009). "Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait." Bmc Genomics 10(1): 368.

45. Oliphant, A., et al. (2002). "BeadArray technology: Enabling an accurate, cost-effective approach to high throughput genotyping." Biotechniques 32(6): 56-58.

46. Palaisa, K., et al. (2004). "Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep." Proceedings of the National Academy of Sciences of the United States of America 101(26): 9885-9890.

47. Purdie, A. C., et al. (2011). "Candidate gene and genome-wide association studies of< i> Mycobacterium avium</i> subsp.< i> paratuberculosis</i> infection in cattle and sheep: A review." Comparative immunology, microbiology and infectious diseases 34(3): 197-208.

48. Ramey, H. R., et al. (2013). "Detection of selective sweeps in cattle using genome-wide SNP data." BMC Genomics 14(1): 382.

49. Ramos, A. M., et al. (2009). "Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology." PLoS One 4(8): e6524.

50. RAQUIN, A. L., et al. (2008). "Soft selective sweep near a gene that increases plant height in wheat." Molecular ecology 17(3): 741-756.

51. Raven, L.-A., et al. (2014). "Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle." Bmc Genomics 15(1): 62.

52. Rexroad, C. E., et al. (2012). "QTL affecting stress response to crowding in a rainbow trout broodstock population." BMC Genetics 13(1): 97.

53. Rubin, C. J., et al. (2010). "Whole-genome resequencing reveals loci under selection during chicken domestication." Nature 464(7288): 587-591.

54. Rubin, C. J., et al. (2012). "Strong signatures of selection in the domestic pig genome." Proc Natl Acad Sci U S A 109(48): 19529-19536.

55. Smith, C. T., et al. (2005). "Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon." Mol Ecol 14(13): 4193-4203.

56. Smith, J. M. and J. Haigh (1974). "The hitch-hiking effect of a favourable gene." Genet Res 23(1): 23-35.

57. Teschke, M., et al. (2008). "Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans." Genetics 180(3): 1537-1545.

58. Tokarska, M., Marshall, T., Kowalczyk, R., Wójcik, J. M., Pertoldi, C., Kristensen, T. N., ... & Bendixen, C. (2009). Effectiveness of microsatellite and SNP markers for parentage and identity analysis in species with low genetic diversity: the case of European bison. Heredity, 103(4), 326-332.

59. Vasemägi, A., et al. (2012). "Screen for footprints of selection during domestication/captive breeding of Atlantic salmon." Comp Funct Genomics 2012: 628204.

60. Waldbieser, G. and B. Bosworth (1997). "Cloning and characterization of microsatellite loci in channel catfish, Ictalurus punctatus." Anim Genet 28(4): 295-298.

61. Wang, R., et al. (2013). "Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish." BMC Genomics 14(1): 929.

62. Wiedmann, R. T., et al. (2008). "SNP discovery in swine by reduced representation and high throughput pyrosequencing." BMC Genetics 9(1): 81.

63. Wong, G. K., et al. (2004). "A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms." Nature 432(7018): 717-722.

64. Wu, X., et al. (2013). "Genome wide association studies for body conformation traits in the Chinese Holstein cattle population." Bmc Genomics 14(1): 897.

65. Xu, J., et al. (2012). "Genome-wide SNP discovery from transcriptome of four common carp strains." PLoS One 7(10): e48140.

66. Zhan, B. J., et al. (2011). "Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping." Bmc Genomics 12(1): 557.

CHAPTER II

LITERATURE REVIEW

Molecular markers

Molecular marker is a genomic landmark that can be used for the tracing of a certain region of DNA (Vignal, Milan et al. 2002). It has a revolutionary impact on agriculture genetics and has usually been regarded as one of the most important sources of information in genetic enhancement. Genetic markers are useful in many area such as DNA fingerprinting, linkage mapping, parentage identification and measurement of genetic diversity. With DNA markers, a whole genome genetic selection can be detected theoretically (Schaeffer 2006). By using the genetic markers in breeding programs, phenotype based selection can be switched to genotype-based selection. The benefits of this have been obvious for decades. Also, DNA markers can be used instead of physical markers on breeding individuals, which in some circumstances is more reliable and convenient. Furthermore, DNA markers are strong genetic tools to identify species, strain, line and family. In 1970s, the DNA-based genetic markers were discovered. Many kinds of DNA markers have been used in genetic research including Restriction Fragment Length Polymorphism (RFLP), Random Amplified Polymorphic DNA (RAPD), Amplified Fragment Length Polymorphism (AFLP), microsatellite and single-nucleotide polymorphism (SNP). Among them, RFLP, Microsatellite and SNPs are co-dominance, which means heterozygosity and genotype allele frequency can be determined by using them.

During the development of molecular genetics, RFLP, RAPD and AFLP, which are relatively old faction genetic markers, are rarely used today. Microsatellites, also known as Simple Sequence Repeats (SSRs), are repeating sequences of 2-6 base pairs of DNA. They have been used increasingly in aquaculture species for the last 10 years, due to their elevated polymorphic information content (PIC), co-dominant mode of expression, Mendelian inheritance, abundance and broad distribution throughout the genome (Liu and Cordes 2004). Now, SNP has become the most widely used genetic marker in the genetics world. Theoretically, a SNP can have as many as four alleles, each containing one of four bases at the SNP site: A, T, C, and G. Practically, however, most SNPs are usually bi-allelic, which is restricted to one of two alleles. Obviously, their PIC is not as high as multi-allele microsatellites, but this shortcoming is balanced by their great abundance. SNP markers are inherited as co-dominant markers (Liu and Cordes, 2004).

SNP marker identification in catfish

SNP identification and analysis on catfish have been conducted from years ago. In 2008, Wang et. al, identified more than 33,000 putative SNPs from catfish EST (Wang, Sha et al. 2008). The catfish EST sequences used in the project were obtained from NCBI dbEST database, including both channel catfish and blue catfish ESTs, as the genome and appearance of these two species were highly conserved. CAP3 software were used for the contig assembly with the criteria set: 1) 95% overlap similarity; 2) a number of minmatch equal to 50. After the contigs were assembled, BLASTX was performed to identify the genes located in the contigs, with the E-value cutoff equal to e^{-10}. The autoSNP software were used for SNP isolation. It was found that the miner allele frequencies of identified SNPs were associated with the length of the assembled contig. The effect of assembled contig size and minor allele frequency on SNP filter were characterized

using a small part of putative SNPs. From the results part, more than 5,500 contigs were assembled with 4,387 contigs contained SNP. A total of 73% of the SNPs were isolated from the contigs that assembled by 2-3 EST sequences. The rest 27% SNPs were isolated from the contigs with 4 or more sequences. In total, over 33,000 SNPs were identified, with the average SNP rate of 0.79 SNP / 100 bp. Illumina Bead Arrays was used to verify the identified SNPs using 192 catfish. Among them, 63 fish were selected from three domestic strans and 63 fish were selected from three wild populations, and the other 66 fish came from the inter-specific mapping panel. A total of 384 putative SNPs were selected for SNP validation. Among them, 266 putative SNPs were successful genotyped and 156 SNPs were polymorphic. It was also reported that there was no significant association between Illumina's quality scores and the quality of SNP. Also, it is demonstrated that minor sequence allele frequency and contig size were the major two parameters that can affect the quality of SNPs. Another important parameter was the quality of SNP flanking region. Another important discovery was that the SNP genotyping successful rate was high associated with the presence of introns. This was one of the shortcoming of Bead Array technology. From the 118 putative SNPs that was not successful genotyped, 50 SNPs and their localized contigs can be aligned with zebrafish genome. And 64% of these SNPs were located at the exon-intron border, indicating that the presence of intron was one of the major reason that can result in the failures of SNP genotyping.

Another 48,000 high-quality SNPs were identified from over 300,000 putative SNPs. Contigs that used for SNP isolation were also assembled form EST sequences (Wang, Peatman et al. 2010). Instead of downloading the EST sequences from bioinformatic website, a total of 438,321 ESTs were sequenced using 4 blue catfish and 8 channel catfish libraries. After assembly, more than 45,000 contigs were generated, and over 14,000 unique genes were annotated in catfish.

This is the first genome level sequencing project on catfish, approximately 50% of the total catfish genes were identified. Evolutionary conservation analysis on identified genes were conducted by comparing the gene sequences to other teleost species including zebrafish, medaka, Tetraodon, as well as some high level vertebrate such as chicken, mouse and human. The results showed that 98% of catfish genes had at least one homolog in the fish species, suggesting the high conservation level on gene content among fish species. Among the 300,000 identified putative SNPs, 48,702 of them were isolated from blue catfish and 102,252 of them were isolated from channel catfish. The different SNP number in blue and channel catfish should be caused by the unequal number of sequence libraries used in the project. After quality filter, only 7.8% of blue catfish SNPs and 15.7% of channel catfish SNPs were left. The quality of the SNPs passed quality control were further assessed. The filtered SNP frequency was 0.25 SNP per kilobase in blue catfish, 0.64 SNP perl kilobase in channel catfish. And 90% of the high quality SNPs were came from the contigs with five or more ESTs.

Genotyping-by-sequencing technology has been used for SNP discovery in blue catfish (Li, Waldbieser et al. 2014). A total of 190 individuals from five domesticated and wild populations were used in the study. After SNP filtering, 4275 common SNPs were identified and used for population genetics and structure analysis. Sequenom MassARRAY were used for SNP validation. A number of 64 putative SNPs were successfully genotyped in all individuals from the populations used for SNP discovery and two new populations. The Genotyping-by-sequencing technology can provide individual genotype, which is important for population genetic structure analysis. However, compared with Illumina sequening, the data output is relatively small and some of the genomic information were lost during the sequencing.

In general, their study provide a new rapid, reliable and low cost approach for SNP identification in catfish as well as in other aquatic species with limited genetic background information.

With the advantages of next generation sequencing, gene associated SNPs has been identified from both channel catfish and blue catfish using RNA-Seq (Liu, Zhou et al. 2011). A total of 47 channel catfish and 19 blue catfish were used in the study. After transcriptome *de nove* assembly and mapping, SNPs and microsatellite markers were identified using CLC Genomics Workbench. The parameters of SNP identification were set as following: 1) The quality score of central base should larger than 25 and the quality scores of the flanking regions should large than 20; 2) the minimum read depth should large than 4; 3) minor allele count should larger than 2. In general, 24,440 unique protein coding genes were annotated from the assembly. A total of two million and 2.5 million gene associated SNPs were identified from the channel catfish and blue catfish, respectively. Among them, more than 340,000 channel catfish intra-specific SNPs, 366,269 blue catfish intra-specific SNPs, and over 420,000 common SNPs were identified. The SNPs were distributed all over the genome.

SNP applications in catfish

With a relatively large number of SNP markers identified, a number of genetic analysis and tools has been conducted/constructed using SNPs, including analysis of catfish disease resistance, analysis of hypoxia tolerance, high density linkage map development, catfish SNP array construction and GWAS analysis. A method of bulk segregant RNA-seq has been developed and applied for searching the genomic regions responsible for ESC disease (Wang, Sun et al. 2013). Bulk segregant RNA-seq was a combination of Bulk segregant analysis and RNA-seq using SNP markers. Because genes were differentially expressed during the disease challenge, transcriptome

were unevenly sequenced during RNA-seq analysis. Some genes were up-regulated and therefore more transcripts were sequenced. The uneven sequenced transcriptome could significantly affect the accuracy of significant SNP identification, as the gene depth may be not comparable. Bulk frequency ratio were introduced in the study to remove the bias caused by the unevenly sequencing. A number of SNPs with high bulk frequency ratio (>4) were identified and located in 359 genes. Among them, 337 genes had a SNP with bulk frequency ratio larger than 4 but smaller than 16, 23 genes had a SNP with bulk frequency ratio larger than 16 and 4 genes had at least one SNP with bulk frequency ratio larger than 32. The distribution of these genes with high bulk frequency ratio were analyzed. It was reported that eight linkage group harbored QTLs involved in ESC disease resistance including LG1, 3, 6, 9, 15, 17, 18 and 25. Among the, LG6, 15, and 17 contained the most genes with significant SNPs. Differentially expressed genes were also identified from the study using a normal RNA-Seq method. A total of 17 genes were listed as differentially expressed genes and genes with significant SNPs at the time, which should be more important for ESC disease resistance.

Analysis of catfish hypoxia resistance has been conducted using SNP markers (unpublished). Oxygen is required for life, and without oxygen, the human brain can survive for just six minutes. However, most organisms including humans can experience various levels of stresses due to low levels of oxygen, referred to as hypoxia. In humans, hypoxia occurs during acute and chronic vascular disease, pulmonary disease and cancer (Kondo, Hamada et al. 2005, Phillips, Mestas et al. 2005, Taylor and Sivakumar 2005). Although most hypoxia studies were conducted with mammals, hypoxia is a much more common phenomenon for fish, the most diverse group of extant vertebrates with over 25,000 species. Oxygen availability in water varies significantly over time and space. The oxygen content of water can change dramatically

depending on salinity, photosynthetic activity, pollution, wind, temperature, hour of the day and season. The survival of fish as a group depends on their ability to adapt rapidly to changing levels of environmental oxygen. Indeed, much of the diversity of fishes can be attributed to the adoption of specialized anatomic, behavioral, and physiological strategies to compensate for particular aquatic oxygen conditions (Powell and Hahn 2002, Nikinmaa and Rees 2005). In the study, SNPs with high bulk frequency ratios were generated from the RNA-seq data between the bulks of tolerant fish and sensitive fish. The genes with significant SNPs were identified and their genomic location were also assessed in linkage group level. The linkage groups contained more than 10 genes with significant SNPs and at least one gene harboring significant SNPs with BFR $\geq$ 4 were identified as potential genomic regions that harboring candidate genes for hypoxia resistance. Original analysis of the significant SNP alleles were also performed using the inter-species SNPs database of blue and channel catfish.

SNP array has been designed using the markers identified in this study (Liu, Sun et al. 2014), which is essential for genome wide association study and individual genotype screen. A catfish 250K SNP array has been development using Affymetrix Axiom genotyping technology. A total of 640,000 SNPs were selected based on their genomic location in order to have a good coverage of the genome. At last, a set of 250,000 SNPs was finalized for SNP array. The performance of the SNP array was then evaluated using wild channel catfish and hybrid catfish families. The SNPs conversion rates from different batches were from 79.4% to 87.3%, with the average SNP call rates greater than 99%. However, the polymorphic rate of the SNPs on the array was around 55%, which could be caused by the hybrid samples used for SNP array evaluation. The 250K SNP array has been successfully used for high density linkage map development and should be valuable for genome-wide association studies and whole genome selection.

A high density linkage map was developed using more than 50,000 SNPs, with their genotype screened by the catfish 250K SNP array (In press). The average inter-marker spacing was 0.4 cM across the whole genome and the female map were larger then male map, indicative of the higher recombination rate in the female. With the genetic information provided by the linkage map, 86% of the whole genome scaffolds can be allocated onto the 29 linkage groups, greatly facilitated the channel catfish whole genome assembly. In addition, the high density linkage map was extremely helpful when searching for genomic regions related to disease and stress responses and provided the basis for genomic comparative studies between catfish and other species.

Domestic analysis in aquatic species

Domestication is one kind of selection involving the removal of some selection pressure typical of natural environments but intensification of others relevant to farming conditions (Price 1999). For example, the anti-predator behavior of fish such as shoaling and schooling are essential for predator defense for wild fish (Magurran, Seghers et al. 1995, Pavlov and Kasumyan 2000). Under farm environments, there are either no or limited number of predators, and therefore the anti-predator behavior is no longer essential. Therefore, anti-predator behavior traits were reduced or totally lost in domesticated aquatic species such as rainbow trout (*Oncorhynchus mykiss*) (Berejikian, Mathews et al. 1996) and laboratory strains of zebrafish (*Danio rerio*) (Wright, Nakamichi et al. 2006), pumpkinseed sunfish (*Lepomis gibbosus*) (Coleman and Wilson 1998), and brown trout (*Salmo trutta*) (Johnsson, Petersson et al. 1996).

In rainbow trout, comparisons between individuals recently derived from wild stocks and domestic populations suggest significant genetic effects on mean swim level, hiding, foraging, startle response, and aggression level from domestication (Lucas, Drew et al. 2004). The results

of multiple comparisons demonstrated that the length of domestication history has a significant effect on the fish behavior patterns. The individuals of two populations which have been domesticated for more than 100 years showed an obvious reductions in predator avoidance behavior patterns. In contrast, the fish from two recently domesticated populations showed different behavior patterns and was less aggression when comparing with the long history domestic fish. These results provided insight into the genetic effect on domestication and the interactions between environment and genetic. Another experiment of rainbow trout showed that the fry from wild population displayed a higher level of agonistic behavior than did fry from domestic population (Berejikian, Mathews et al. 1996). In addition, the performance of wild fry was better than domestic fry in size-matched dyadic dominance challenges. It was reported that domestic fry cultured in a natural environment was more aggressive than those cultured in tanks. It was also more aggressive than the wild fry cultured in natural environment and tanks. The study demonstrated that within four to seven generations of domestication, behavioral changes could happen between domestic populations and its wild donor populations.

In zebrafish, behavioral and morphological differentiation were found between wild and laboratory zebrafish (Wright, Nakamichi et al. 2006). A total of 184 zebrafish were tested for shoaling tendency and boldness. QTLs responsible for growth rate, anti-predator behaviors, and boldness were identified on zebrafish chromosomes 23, 31 and 9, respectively. The results confirmed that domestication can affect the fish genome and resulted in behavioral and morphological changes. Another study measured the specificity of boldness and shyness in juvenile pumpkinseed sunfish. They found consistent individual differences on shyness and boldness within different context, while individual differences was not associated across

contests. Their finding indicated that boldness and shyness were context-specific and even within a single context, more than one factor could be exist in the regulation of behavioral phenotypes.

Similarly, Fine *et al.* (Fine, Lahiri et al. 2014) found that both spine and girdle exhibit negative allometric growth, and the pectoral spines and girdles are lighter in domesticated than in wild channel catfish (*Ictalurus punctatus*). It was reported there could be two different explanations for the negative growth of fish girdle and spine: 1) epigenetic effect caused by the pressure of predators; 2) artificial selection of spine growth during domestication. An experiment has been done to test if the negative growth of spine was caused by epigenetic effect. Domestic channel catfish fingerlings were culture with their predator, largemouth bass, for 13 weeks, and the results showed that there was no difference in pectoral spine growth between the case group and control group. Therefore, they concluded that the negative growth of channel catfish pectoral spine was likely caused by the selection pressure during domestication.

Genomic impact of domestication has not been well studied in fish species. Previous studies have shown morphological, behavioral and growth changes in channel catfish during domestication (Dunham 2011, Fine, Lahiri et al. 2014), but the molecular basis of such changes has not been elucidated, due, at least in part, to the lack of molecular markers capable of providing whole genome coverage. In regards to domesticated channel catfish selected for body weight, significant changes in allozyme and microsatellite allele frequencies were found (Hallerman, Dunham et al. 1986, Lamkom, Kucuktas et al. 2008).

Reference

1. Berejikian, B. A., et al. (1996). "Effects of hatchery and wild ancestry and rearing environments on the development of agonistic behavior in steelhead trout (Oncorhynchus mykiss) fry." Can J Fish Aquat Sci 53(9): 2004-2014.
2. Coleman, K. and D. S. Wilson (1998). "Shyness and boldness in pumpkinseed sunfish: individual differences are context-specific." Anim Behav 56(4): 927-936.
3. Dunham, R. A. (2011). Aquaculture and fisheries biotechnology: Genetic approaches. Auburn, CABI.
4. Fine, M. L., et al. (2014). "Reduction of the pectoral spine and girdle in domesticated channel catfish is likely caused by changes in selection pressure." Evolution 68: 2102-2107.
5. Hallerman, E., et al. (1986). "Selection or drift–isozyme allele frequency changes among channel catfish selected for rapid growth." Trans Am Fish Soc 115(1): 60-68.
6. Johnsson, J. I., et al. (1996). "Domestication and growth hormone alter antipredator behaviour and growth patterns in juvenile brown trout, Salmo trutta." Can J Fish Aquat Sci 53(7): 1546-1554.
7. Kondo, Y., et al. (2005). "Over expression of hypoxia-inducible factor-1alpha in renal and bladder cancer cells increases tumorigenic potency." J Urol 173(5): 1762-1766.
8. Lamkom, T., et al. (2008). "Microsatellite variation among domesticated populations of channel catfish (Ictalurus punctatus) and blue catfish (I. furcatus)." KU. Fish. Res. Bull 32: 37-47.
9. Li, C., et al. (2014). "SNP discovery in wild and domesticated populations of blue catfish, Ictalurus furcatus, using genotyping‐by‐sequencing and subsequent SNP validation." Molecular ecology resources.
10. Liu, S., et al. (2014). "Development of the catfish 250K SNP array for genome-wide association studies." BMC Res Notes 7: 135.
11. Liu, S. K., et al. (2011). "Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array." BMC Genomics 12(1): 53.
12. Liu, Z. and J. Cordes (2004). "DNA marker technologies and their applications in aquaculture genetics." Aquaculture 238(1): 1-37.
13. Lucas, M. D., et al. (2004). "Behavioral differences among rainbow trout clonal lines." Behav Genet 34(3): 355-365.
14. Magurran, A., et al. (1995). "The behavioral diversity and evolution of guppy, Poecilia reticulata, populations in Trinidad." Adv Study Behav 24: 155-202.
15. Nikinmaa, M. and B. B. Rees (2005). "Oxygen-dependent gene expression in fishes." Am J Physiol Regul Integr Comp Physiol 288(5): R1079-1090.
16. Pavlov, D. S. and A. O. Kasumyan (2000). "Patterns and mechanisms of schooling behavior in fish: A review." J Ichthyol 40: S163.

17. Phillips, R. J., et al. (2005). "Epidermal growth factor and hypoxia-induced expression of CXC chemokine receptor 4 on non-small cell lung cancer cells is regulated by the phosphatidylinositol 3-kinase/PTEN/AKT/mammalian target of rapamycin signaling pathway and activation of hypoxia inducible factor-1alpha." J Biol Chem 280(23): 22473-22481.

18. Powell, W. H. and M. E. Hahn (2002). "Identification and functional characterization of hypoxia-inducible factor 2alpha from the estuarine teleost, Fundulus heteroclitus: interaction of HIF-2alpha with two ARNT2 splice variants." J Exp Zool 294(1): 17-29.

19. Price, E. O. (1999). "Behavioral development in animals undergoing domestication." Appl Anim Behav Sci 65(3): 245-271.

20. Schaeffer, L. (2006). "Strategy for applying genome‐wide selection in dairy cattle." Journal of Animal Breeding and Genetics 123(4): 218-223.

21. Taylor, P. C. and B. Sivakumar (2005). "Hypoxia and angiogenesis in rheumatoid arthritis." Curr Opin Rheumatol 17(3): 293-298.

22. Vignal, A., et al. (2002). "A review on SNP and other types of molecular markers and their use in animal genetics." Genetics Selection Evolution 34(3): 275-306.

23. Wang, R., et al. (2013). "Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish." BMC Genomics 14(1): 929.

24. Wang, S., et al. (2010). "Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies." Genome biology 11(1): R8.

25. Wang, S., et al. (2008). "Quality assessment parameters for EST-derived SNPs from catfish." BMC Genomics 9: 450.

26. Wright, D., et al. (2006). "QTL analysis of behavioral and morphological differentiation between wild and laboratory zebrafish (Danio rerio)." Behav Genet 36(2): 271-284.

CHAPTER III

IDENTIFICATION AND ANALYSIS OF GENOME-WIDE SNPS PROVIDE INSIGHT INTO SIGNATURES OF SELECTION AND DOMESTICATION IN CHANNEL CATFISH

Materials and methods

Fish sources and sampling

All procedures involving the handling and treatment of fish used during this study were approved by the Auburn University Institutional Animal Care and Use Committee (AU-IACUC) prior to initiation of the project. A total of 150 channel catfish, with 30 individuals from each of Marion, Thompson, USDA103, one outbred commercial strain (hereafter referred to as Hatchery), and one wild population were used for this study. The four aquaculture strains were from different geographic locations within the United States, which possess different production traits such as growth rate, disease resistance and feed conversion efficiency (Dunham and Smitherman, 1984). The Marion strain was originally from the Marion National Fish Hatchery, which provided stock for many of the catfish farms in Alabama (Dunham and Smitherman, 1984). The original fish for this strain were collected from the Red River, Arkansas, and other strains. The Thompson strain was originally from Thompson-Anderson fingerling farms, which was one of the major fingerling farms in Mississippi. The origin of this strain can be traced primarily to the Yazoo River and to a lesser degree Red River and Kansas (Dunham and Smitherman, 1984). USDA103 was originally from US Department of Fish and Wildlife Hatchery in Uvalde, TX (Waldbieser

and Wolters, 2007). The Hatchery strain was originally from catfish farms in Mississippi, and was widely used in the catfish industry. The wild channel catfish used in this project were obtained from Coosa River, Alabama (Mickett et al., 2003; Simmons et al., 2006).

DNA extraction, library preparation and sequencing

The fish were euthanized with tricaine methanesulfonate (MS 222) at 300 mg/l before blood collection. For each individual, 500 µl blood was collected for DNA isolation, placed into 5 ml lysis buffer immediately, and then into a water bath at 55ºC for 12 h. Total DNA was isolated using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocol. Equal amounts of DNA (100 µg) from each individual were pooled for sequencing, one pool for each strain.

Sequencing was conducted commercially at HudsonAlpha Genomic Services Lab (Huntsville, AL, USA). Genomic libraries were prepared with the Paired-end Sequencing Sample Preparation Kit (Illumina, San Diego, CA) with 5 µg of genomic DNA for all strains, according to the manufacturer's instructions. For each strain, the prepared DNA library was sequenced on one lane of the Illumina HiSeq 2000 platform for 100-bp paired-end reads. The short reads were deposited in the NCBI Sequence Read Archive (SRA) under Accession number SRA075234 (http://www.ncbi.nlm.nih.gov/sra).

***De novo* assembly**

To fully utilize the next-generation sequencing data and provide insight into the completeness of our current whole genome assembly, *De novo* assembly was conducted using the 100 bp short

read dataset of Marion strain, Thompson strain, and USDA103 strain, separately. Preliminary assemblies were conducted to evaluate the performance of each popular assembler including ABySS, Velvet, Trinity and SOAP. According to the assembled contig length, run time, computer resource request, ABySS v 1.3.4 (http://www.bcgsc.ca/platform/bioinfo/software/abyss) was used for the final assembly. Assemblies were performed using multi-kmer strategy (from 30-90). Then, assembled contigs from the three strains were combined together. Homemade script was used to remove the contigs no longer than 200 bp. After that, CD-HIT software (http://weizhong-lab.ucsd.edu/cd-hit/) were used to remove repetitive contigs using the option –c 0.95 and –n 5, which meant the clustering threshold was equal to 95% identity and the size of word was set equal to 5. Finally, to evaluate the assembly and check the completeness of our previous whole genome assembly, mummer software (http://mummer.sourceforge.net/) was used to compare the newly assembled contigs and the whole genome scaffold. At least, the outputs were further processed by invoking 'show-coords -clor prefix.delta' for result table generation. Each field in the output table was defined in Mummer instructions as followed (http://mummer.sourceforge.net/manual/#coords): [S1] start position of the alignment area in the reference scaffold; [E1] end position of the alignment area in the reference scaffold; [S2] start position of the alignment area in the assembled contig sequence; [E2] end position of the alignment area in the assembled contig sequence; [LEN 1] length of the alignment area in the reference scaffold; [LEN 2] length of the alignment region in the assembled contig sequence; [% IDY] identity of the alignment shown in percentage; [% SIM] similarity of the alignment shown in percentage; [% STP] percent of stop codons in the alignment; [LEN R] length of the reference scaffold; [LEN Q] length of the assembled contig; [COV R] alignment coverage in the reference scaffold; [COV Q] alignment coverage in the assembled contig; [FRM] reading frame for the

scaffold and assembled contig alignments, respectively; [TAGS] the scaffold ID and assembled contig IDs respectively.

Reference mapping

Sequence mapping was performed using CLC Genomics Workbench (version 4.0.2; CLC bio, Aarhus, Denmark). Before mapping, raw sequence reads were trimmed to remove adaptor sequences, ambiguous nucleotides (N's), extreme short reads (< 30 bp) and low quality sequences (Quality score<20) using CLC Genomics Workbench. The quality of each sequence was assessed as follows: First, convert Q (base quality) was converted to an error probability (P): $P = 10^{\frac{Q}{-10}}$. Then, for every base a new value was calculated for every base: N =P(A)-P(Q), where A is the criterion of the minimal quality score. In this project, A=20 (Phred score); Q is the Phred quality score of each base. This value would be negative for bases with quality scores below 20. For every base, the software calculated the running sum of this value. The part of the sequence not trimmed was the region between the first positive value of the running sum and the highest value of the running sum. Everything before and after this region was trimmed.

The clean reads from each strain were then aligned with the catfish genome assembly. The mapping parameters were set as mismatch cost of 2, deletion cost of 3 and insertion cost of 3. The highest scoring matches that shared ≥ 95% similarity with the reference sequence across ≥ 90% of their length were included in the alignment. The mapping output was converted into BAM format for further analysis (Li et al., 2009a).

SNP identification and filtering

SNPs were identified from the pooled data from all the strains using the SAMtools (version 0.1.18) (Li et al., 2009a) and PoPoolation2 (Kofler et al., 2011) with the lowest criteria setting to obtain all potential SNPs. First, ambiguously mapped reads were removed using SAMtools with the command "`samtools view -q 20 -b`", the option "`-p 20`" means skip alignments with MAPQ score smaller then 20; the option "`-b`" means that the output will be written in the BAM format. Then, the output were sorted using the command "`samtools sort`", which can sort the alignments by leftmost coordinates. After that, all of the five mapping result files (one file for each strain) were piled up together using the command "`samtools mpileup –B pop1.bam pop2.bam pop3.bam pop4.ban pop5.bam > out.mpileup`". The option "-B" means to disable probabilistic realignment for the computation of base alignment quality. Using this option could greatly improve the results of SNP discovery by reducing false SNPs caused by misalignments. "`pop1.bam pop2.bam pop3.bam pop4.ban pop5.bam"` are the input files, which were generated form the last step. "`out.mpileup`" was the output file generated in this step. Synchronized file were then generated by a perl script provided in the PoPoolation2 toolkit. The synchronized file was the input file for PoPoolation2, which contained the allele frequencies for every base in the reference. A total of eight columns were generated in the synchronized file: the first field was the reference contig ID; the second field was the position within the reference contig; the third field was the reference genotype; the fourth field to the eighth filed were allele frequencies of each population respectively. Raw SNPs were identified using a perl provide by PoPoolation2 with command "`perl snp-frequency-diff.pl –input file.sync –output-prefix result`". A sample of the results were shown in

Table 1. A total of 19 columns were presented. Column 1 was the contig ID; Column 2 was the position in bp; Column 3 was the reference genotype. Here we did not insert the reference genotype information in the previous step, so all of them were "N". Column 4 was the number of alleles shown in the SNP; Column 5 was the genotype information of the SNP; Column 6 was deletion sum; Column 7 showed SNP type; Column 8 showed the major allele of each strain. Here we have five strains in total, so "AAAAA" meant the major allele in the five strain were all "A". In the same way, Column 9 was the minor allele of each strain. Then allele frequency was present in the next 10 columns. Column 10 to Column 14 showed the major allele frequency in the five strains, each column presented one strain. In the same way, Column 15 to Column 19 showed the miner allele frequency in the five strains, each column presented one strain.

Three factors that are important for excluding false SNPs caused by sequencing errors were set: 1) minimum read depth, 2) maximum read depth, and 3) minor allele read count. An optimal combination of these three factors was determined and used for screening quality SNPs. SNPs with the presence of both alleles in all five strains were defined as common SNPs. SNPs were defined as strain-specific SNPs if the SNP polymorphisms were found in only one strain. The information of identified SNPs were deposited in the National Animal Genome Research Program Aquaculture Genomics Data Repository (www.animalgenome.org/repository/pub/auburn2014.0530/).

Significant SNP analysis

SNPs with significantly different allele frequency ratios were identified between domestic catfish strains and the wild population (hereafter referred to as significant SNPs). Two-tailed Fisher's exact test was performed with the statistical significance level of false discovery rate corrected P

value ≤ 0.01. Significant SNPs were categorized into three groups based on their location: 1) in the coding regions, 2) near the coding regions and 3) in non-coding regions. Near the coding regions means the SNP is located in non-coding regions but within 100 bp from the coding region.

Selective sweep analysis

With the availability of significant SNPs, genomic regions with selective sweeps were identified from the four domestic strains by detecting the genome regions with extremely low heterozygosity. The pooled heterozygosity (H_p) score was calculated using the formula $H_p = 2\Sigma n_{MAJ}\Sigma n_{MIN}/(\Sigma n_{MAJ} + \Sigma n_{MIN})^2$ (Rubin et al., 2010; Rubin et al., 2012a). Σn_{MAJ} was the sum of the major allele reads, and Σn_{MIN} was the sum of the minor allele reads for all significant SNPs in one window. The H_p score was calculated based on 20 kb sliding window across the genome. Windows with less than five significant SNPs were not used for calculation. Putative selective sweeps were identified from windows with $-\log_2(H_p)$ score ≥ 4.

Table 1 Output format of the software Popoolation2

contig46	103	N	2	A/G	0	pop	AAAAA	GNGGG	18/21	40/40	56/57	48/49	34/36	3/21	0/40	1/57	1/49	2/36
contig46	170	N	2	G/C	0	pop	GCGGG	CGNNN	22/24	29/47	42/42	56/56	27/27	2/24	18/47	0/42	0/56	0/27
contig46	243	N	2	G/T	0	pop	GGGGG	TNTTT	31/32	38/38	22/27	36/41	37/41	1/32	0/38	5/27	5/41	4/41
contig46	299	N	2	G/A	0	pop	GGGGG	ANAAA	30/33	36/36	39/41	35/40	37/42	3/33	0/36	2/41	5/40	5/42
contig46	341	N	2	T/C	0	pop	TTTTC	CNCCT	17/33	33/33	30/42	30/47	28/41	16/33	0/33	12/42	17/47	13/41
contig46	398	N	2	A/C	0	pop	AAAAA	CCNCC	32/34	31/32	34/34	45/47	42/45	2/34	1/32	0/34	2/47	3/45
contig46	399	N	2	A/C	0	pop	CACAA	ACACC	18/34	28/33	22/33	26/49	26/45	16/34	5/33	11/33	23/49	19/45
contig46	402	N	2	A/C	0	pop	AAAAA	CNNCC	29/34	33/33	37/37	41/50	43/46	5/34	0/33	0/37	9/50	3/46
contig46	404	N	2	A/T	0	pop	TATAA	ATATT	17/28	27/32	25/36	23/46	22/43	11/28	5/32	11/36	23/46	21/43
contig46	798	N	2	C/T	0	pop	CCCCC	TTTTT	29/38	23/40	29/39	20/32	23/32	9/38	17/40	10/39	12/32	9/32
contig46	801	N	2	C/T	0	pop	CCCCC	TTNTN	39/40	38/39	37/37	28/29	31/31	1/40	1/39	0/37	1/29	0/31
contig46	811	N	2	C/A	0	pop	CCCCC	AANNN	33/36	36/38	42/42	27/27	35/35	3/36	2/38	0/42	0/27	0/35
contig46	834	N	2	T/A	0	pop	TTTTT	ANAAA	35/37	32/32	38/42	30/35	30/32	2/37	0/32	4/42	5/35	2/32
contig46	836	N	2	G/T	0	pop	GGGGG	TNNNN	34/37	30/30	44/44	36/36	34/34	3/37	0/30	0/44	0/36	0/34
contig46	868	N	2	C/T	0	pop	TCCCC	CTTNT	16/29	18/28	30/40	26/26	17/27	13/29	10/28	10/40	0/26	10/27
contig46	886	N	2	T/A	0	pop	TTAAT	AATTA	22/31	17/23	28/47	21/27	20/33	9/31	6/23	19/47	6/27	13/33
contig46	1004	N	2	T/A	0	pop	TTTTT	AANAA	50/57	26/36	44/44	32/44	48/49	7/57	10/36	0/44	12/44	1/49
contig46	1009	N	2	C/T	0	pop	CCCCC	NNNTN	63/63	38/38	48/48	42/44	50/50	0/63	0/38	0/48	2/44	0/50
contig46	1049	N	2	G/C	0	pop	CCGGG	GGCCC	28/48	27/42	23/42	24/39	35/47	20/48	15/42	19/42	15/39	12/47
contig113	102	N	2	G/T	0	pop	GGGGG	TNNTT	11/16	26/26	23/23	12/23	20/21	5/16	0/26	0/23	11/23	1/21
contig113	141	N	2	G/A	0	pop	GGGGG	ANNNN	34/38	45/45	41/41	42/42	43/43	4/38	0/45	0/41	0/42	0/43
contig113	144	N	2	A/G	0	pop	AAAAA	GNNGG	27/37	47/47	43/43	25/46	42/44	10/37	0/47	0/43	21/46	2/44
contig113	171	N	2	G/A	0	pop	GGGGG	AANAA	34/39	26/48	44/44	50/51	35/49	5/39	22/48	0/44	1/51	14/49
contig113	191	N	2	G/T	0	pop	GGGGG	NNNNT	33/33	46/46	51/51	50/50	50/52	0/33	0/46	0/51	0/50	2/52

Results

Illumina sequencing and reference mapping

A total of 40.6-44.7 Gb of sequences were generated from each strain (Table 2). Approximately 96% reads were clean after trimming. The average lengths of the clean reads varied from 94 to 95 nucleotides. Reference mapping was conducted by aligning sequence reads from each strain with the preliminary catfish genome assembly (unpublished data). A total of 30.7-34.6 Gb were aligned to the reference sequences (Table 2). On average, around 31X-35X genome coverage (read depth) were obtained for each of the five populations. When all the sequences were combined, the total read depth was 167X genome coverage (Table 2).

Table 2. Summary of genomic data generation of channel catfish using Illumina HiSeq 2000

Strains	Raw data	Trimmed reads	Average length	Reads mapped	Genome coverage
Hatchery	43.8 Gb	42.0 Gb	95.2 bp	32.6 Gb	33.3 X
USDA103	42.9 Gb	41.6 Gb	94.5 bp	33.7 Gb	34.4 X
Thompson	44.7 Gb	43.1 Gb	93.8 bp	34.6 Gb	35.3 X
Marion	42.3 Gb	40.8 Gb	94.2 bp	31.8 Gb	32.4 X
Wild population (Coosa River, AL)	40.6 Gb	39.3 Gb	94.8 bp	30.7 Gb	31.3 X
Total	214.3 Gb	206.8 Gb	94.5 bp	163.4 Gb	166.7 X

***De novo* assembly and comparative analysis**

Due to the large size of each raw read dataset, assembly were conducted using each of the pure domestic strain separately, including Marion, Thompson, and USDA103. *De novo* assembly was also conducted using the datasets of wild population and hatchery, while the assembled contigs were extremely short (data not shown). Preprimary analysis shown that k-mer = 51 and k-mer = 59 were the best k-mers for catfish genome assembly. Table 3 showed the statictial results of assembly using reads from Marion strain, k-mer = 51. A total of 4.2 million contigs were generated. Among them, 459k contigs (10.7%) were large than 200 bp; 76k contigs (1.8%) have the length larger than the length of N50 (2,583 bp); the maximum contig length was 35,874 bp. In total, approximately 691 million base pairs were assembled into contigs.

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	5,169,373	680,395	116,890	200	678	1,654	3,379	35,874	677.6e6
Contigs	4,296,586	459,176	76,480	200	1,102	2,583	5,270	35,874	691.5e6
Scaffolds	4,104,682	267,272	38,832	200	2,134	5,106	10,304	75,899	690.6e6

Table 3 Summary of *de nove* assembly using reads from Marion strain (k=51)

Table 4 showed the statictial results of assembly using reads from Marion strain, k-mer = 59. In general, the assembly results of k = 59 was better than such of k = 51. A total of 3.2 million contigs were generated. Among them, 432k contigs (13.4%) were larger than 200 bp; 65k contigs (2%) have the length larger than the length of N50 (3,086 bp); the maximum contig

length was 43,764 bp. In total, approximately 719 million base pairs were assembled into contigs.

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	3,936,175	657,953	106,207	200	755	1899	3908	27,138	710.2e6
Contigs	3,231,710	432,816	65,729	200	1,273	3,086	6,437	43,764	719.4e6
Scaffolds	3,093,329	294,435	38,779	200	2,089	5,210	10,850	61,907	718.5e6

Table 4 Summary of *de nove* assembly using reads from Marion strain (k = 59)

Table 5 showed the statictial results of assembly using reads from Thompson strain, k-mer = 51. A total of 5.1 million contigs were generated. Among them, 523k contigs (10%) were larger than 200 bp; 91k contigs (1.8%) have the length larger than the length of N50 (2,143 bp); the maximum contig length was 35,835 bp. In total, approximately 681 million base pairs were assembled into contigs.

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	6,001,408	732,677	132,022	200	605	1,447	2,927	28,682	667e6
Contigs	5,143,474	523,684	91,068	200	925	2,143	4,346	35,835	681.3e6
Scaffolds	4,947,412	327,622	50,430	200	1,652	3,866	7,839	41,611	680.4e6

Table 5 Summary of *de nove* assembly using reads from Thompson strain (k = 51)

Table 6 showed the statictial results of assembly using reads from Thompson strain, k-mer = 59. A total of 3.9 million contigs were generated. Among them, 508k contigs (12.9%) were larger than 200 bp; 81k contigs (2.1%) have the length larger than the length of N50 (2,463 bp); the maximum contig length was 35,834 bp. In total, approximately 711 million base pairs were assembled into contigs. Comparison between Thompson strain k =51 and k=59 showed that the assembly results using k = 59 was better, where the number of contigs larger than 200 bp was larger and the N50 and N80 were also larger than such of k = 51.

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	4,654,651	722,806	123,130	200	657	1,623	3,320	28,139	701.4e6
Contigs	3,950,311	508,235	81,734	200	1,032	2,463	5,086	35,834	711.5e6
Scaffolds	3,801,507	359,431	50,786	200	1,619	3,925	8,225	47,037	710.5e6

Table 6 Summary of *de nove* assembly using reads from Thompson strain (k = 59)

Table 7 showed the statictial results of assembly using reads from USDA103 strain, k-mer = 51. A total of 5.1 million contigs were generated. Among them, 512k contigs (10%) were larger than 200 bp; 89k contigs (1.7%) have the length larger than the length of N50 (2,190 bp); the maximum contig length was 35,910 bp. In total, approximately 684 million base pairs were assembled into contigs.

Table 7 Summary of *de nove* assembly using reads from USDA103 strain (k = 51)

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	6,024,425	726,489	129,721	200	611	1,472	2,991	35,910	667.7e6
Contigs	5,125,108	512,498	89,417	200	953	2,190	4,458	35,910	684.7e6
Scaffolds	4,903,028	290,418	43,310	200	1,906	4,514	9,142	48,987	683.7e6

Table 8 showed the statictial results of assembly using reads from Thompson strain, k-mer = 59. A total of 4 million contigs were generated. Among them, 495k contigs (12.2%) were larger than 200 bp; 79k contigs (2%) have the length larger than the length of N50 (2,554 bp); the maximum contig length was 36,048 bp. In total, approximately 717 million base pairs were assembled into contigs. Comparison between Thompson strain k =51 and k=59 showed that the assembly results using k = 59 was better, where the number of contigs larger than 200 bp was larger and the N50 and N80 were also larger than such of k = 51.

Type	N	n:200	n:N50	Minimum length	N80	N50	N20	Maximum length	Sum
Unitygs	4,784,998	720,136	120,717	200	660	1,650	3,393	32,261	702.3e6
Contigs	4,026,328	495,004	79,292	200	1,072	2,554	5,295	36,048	717.2e6
Scaffolds	3,853,149	321,825	43,427	200	1,876	4,609	9,658	59,777	716e6

Table 8 Summary of *de nove* assembly using reads from USDA103 strain (k = 59)

Contigs from the three assemblies were then combined together and removed duplicates to obtain the final comprehensive assembly. As shown in Table 9, vast majority of contigs before process (95%) were removed: 87% of them were contigs equal or smaller to 200 bp and 8% of them were contigs with more than 95% of identities. Finally, a total of 515 thousand contigs

were left, with the N50 equal to 3,510 base pair. Approximately 79 thousand contigs (15%) have the length greater than N50 and the maximum contig length was equal to 43,764 bp.

Table 9 Summary of the combination of three assemblies and the final assembly

	Before process	Final assembly
Number of contigs	11,200,000	515,051
Number of contigs greater than 200 bp	1,436,055	515,051
Number of contigs greater than N50	225,103	79,213
Minimum contig length	200	200
N80	1,116	1,503
N50	2,683	3,510
N20	5,608	6,931
Maximum contig length	43,764	43,764
Total	2.15E+09	9.52E+08

Comparison between the *de nove* assembly generated in this project and our whole genome assembly were conducted to evaluate the completeness of the whole genome assembly as well as to evaluate the quality of the *de nove* assembly. As shown in Table 10, results can be divided into seven groups including end, contains, contained, begin, identity, partial and novel. A total of 642 *de novo* contigs (0.1%) were completely equal to the scaffold of whole genome assembly.

Contained group contained the *de novo* contigs which was longer than its corresponding whole genome scaffold in both sides, in this way, the whole genome scaffold was part of the *de novo* contigs. A total of 4,409 contigs (0.5%) were classified into the group. Contains group have the *de novo* contigs which were than whole genome scaffolds and were part of them. A total of 450,515 contigs (54.2%) were classified into the group. Begin group meant that the latter half of the *de novo* contigs aligned to the beginning of the whole genome scaffold. A total of 9,574 contigs (1.2%) were classified into the group. End group meant that the front half of the *de novo* contigs aligned to the latter half of the whole genome scaffold. A total of 9,447 contigs (1.1%) were classified into the group. Partial group contained the *de novo* contigs which were part of the other longer *de novo* contigs. A number of 330,554 contigs (39.8%) can be classified into the group. The last group was novel group, it contained the *de novo* contigs that completely cannot align to the whole genome scaffold. In general, most of the *de novo* assembly contigs can be aligned to whole genome scaffold and the contigs in novel group, begin group and end group provide a useful candidate pool for the improvement of catfish whole genome assembly

Table 10 Summary of Mummer results

Group	**N**	**%**
Identity	642	0.1%
Contained	4,409	0.5%
Contains	450,515	54.2%
Begin	9,574	1.2%
End	9,447	1.1%
Partial	330,554	39.8%
Novel	25,680	3.1%
Total	830,821	100%

Optimization of the *in-silico* identification of SNPs

To reduce false SNPs derived from sequencing errors, a set of criteria was first developed, including the minimum read depth, the maximum read depth and minor allele read count. As shown in Figure 1, the impact of minimum read depth on SNP identification was tested in the 10-200 intervals with the increasing step of 10. Minimum read depth had only a small effect on the number of identified SNPs within the interval of 10-30. However, beyond this interval, the number of total SNPs was reduced gradually with the increase of minimum read depth (Figure 1).

Apparently, the greater the minimum read depth, the more reliable the SNPs are. However, the higher the minimum read depth, the fewer the reads that are qualified to be included in the analysis. A reasonable choice is to select the largest minimum read depth without significantly reducing the number of identified SNPs. Therefore, we set the minimum read depth at 30 for further analysis (Figure 1). Maximum read depth can have an impact on the quality of SNPs because extremely high numbers of reads are likely generated from non-unique sequences such as repetitive elements or paralogous sequences. Therefore, we evaluated the impact of maximum read depth on SNP identification. As shown in Figure 2, the total numbers of SNPs did not increase significantly when setting the maximum read depth greater than 300. We then examined the contents of repetitive elements for the reads included in these read-depth intervals. As shown in Table 11, the contents of repetitive elements within each read-depth range were similar, up to the maximum reads of 300. However, the content of repetitive elements increased significantly when the maximum read depth were set greater than 300, indicating that a larger proportion of

reads from retroelements and DNA transposons were included. To avoid the false SNPs caused by misalignment of reads from repetitive regions, we set the maximum read depth at 300 for further analysis.

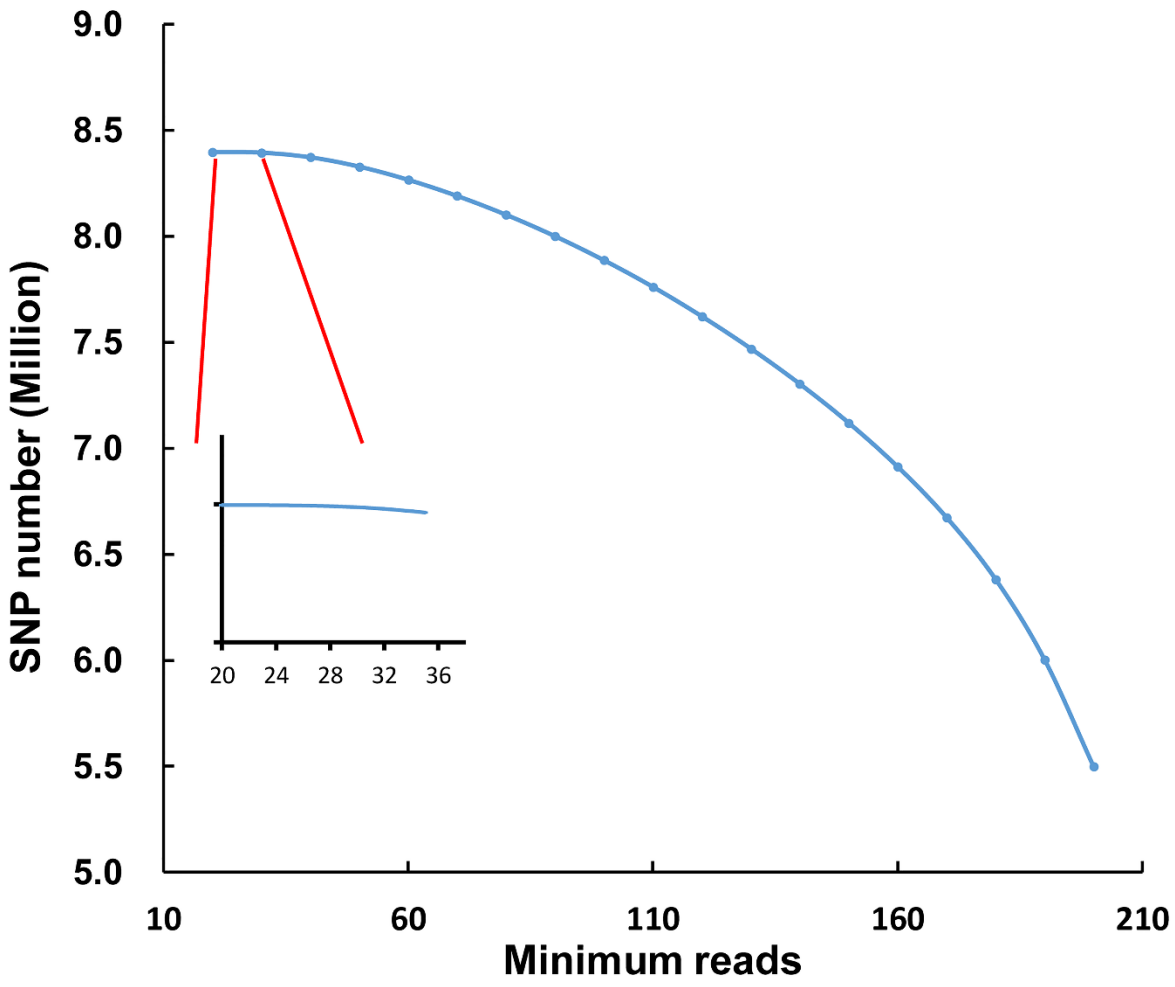

Figure 1. Influence of minimum reads on SNP identification. The x-axis represents the number of minimum reads used for SNP detection and the y-axis represents the number of SNP identified under a certain number of minimum reads.

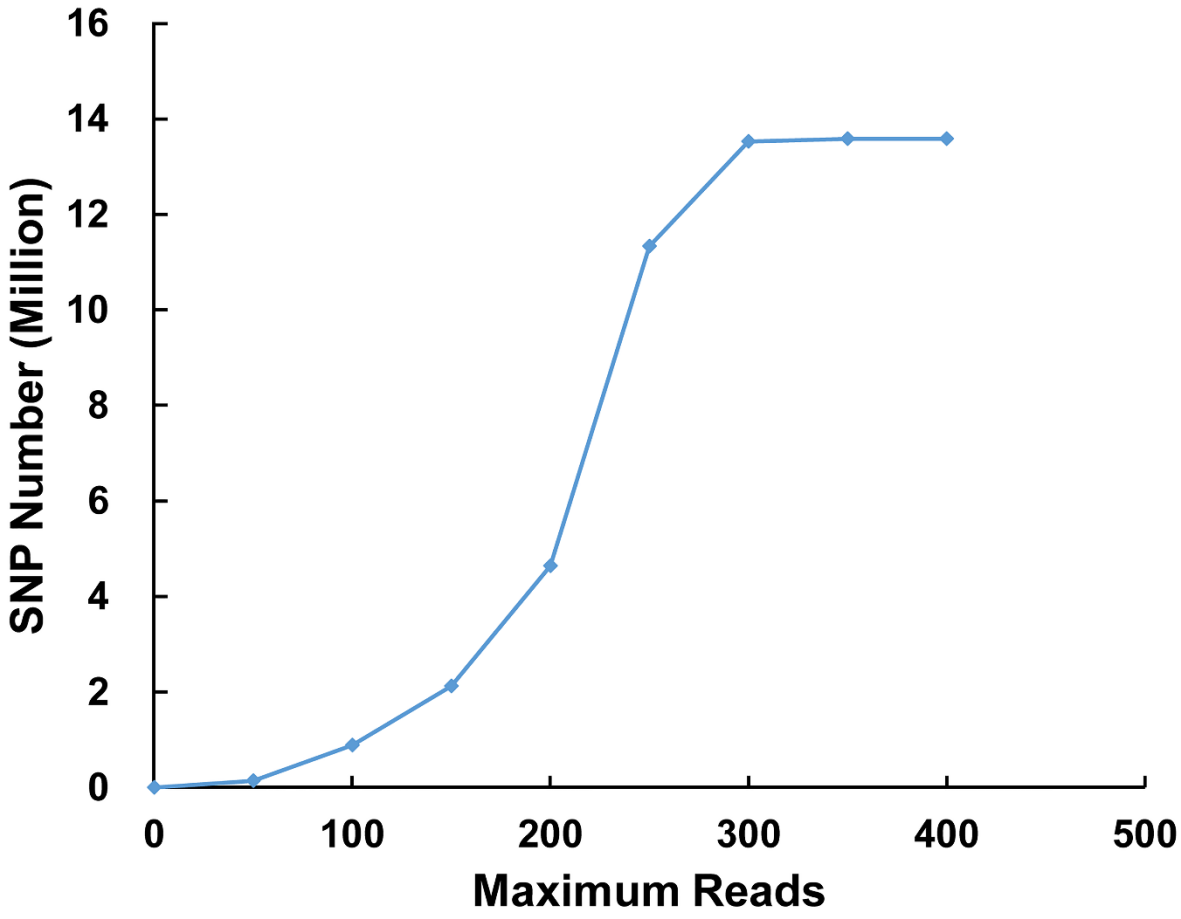

Figure 2. Influence of maximum reads on SNP identification. The x-axis represents the number of maximum reads used for SNP detection and the y-axis represents the number of SNP identified under a certain number of maximum reads.

Table 11 Summary of repetitive element analysis in the SNP flanking regions

Coverage range	Retroelements	DNA transposons	Unclassified
50-100	29	82	7
100-150	34	69	6
150-200	29	89	3
200-250	28	74	4
250-300	46	80	2
>300	101	195	13

Minor allele frequency (MAF) not only affects the SNP applicability for future genetic studies because it directly determines the polymorphism information content of the SNP markers, it also has an impact on the identification of quality SNP. In general, the relationship curve can be arbitrarily divided into two phases, in the first phase, when minor allele counts were set as 2-4, the total number of SNPs was reduced sharply, while in the second phase, when minor allele reads were set as greater than 4, the total number of SNPs was also reduced, but at a much reduced rate, suggesting that minor allele reads of 4-6 may be appropriate for data in the present work (Figure 3). Thus, the minor allele read counts were limited the minor allele read counts to be equal or greater than 5 for further analysis.

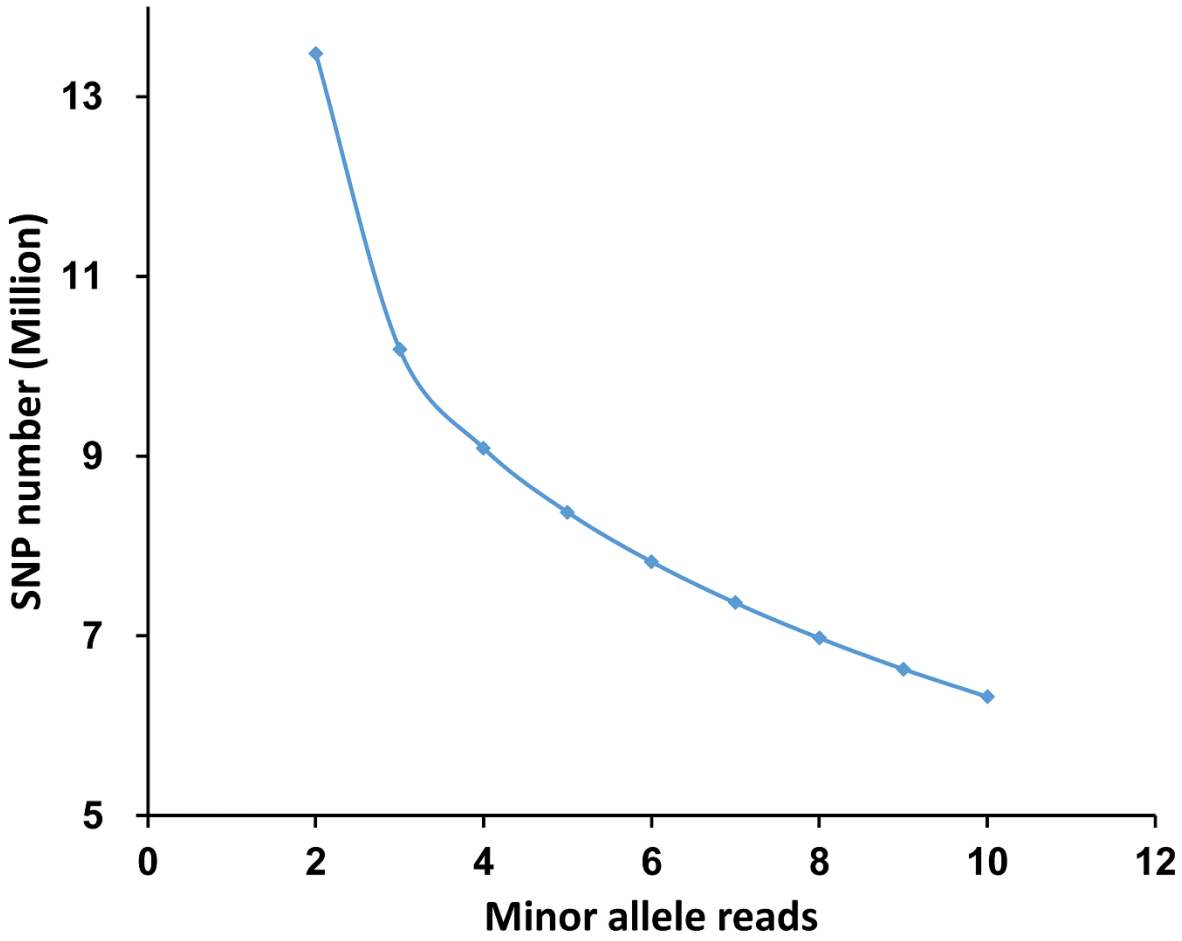

Figure 3. Influence of minor allele read counts on SNP identification. The x-axis represents the number of minor allele reads used for SNP detection and the y-axis represents the number of SNP identified under a certain number of minor allele reads.

In addition to the initial assessment of the factors control SNP quality, the percentage of sequences that were included for SNP identification were examined. As shown in Table 12, the setting of minimum read depth and the minor allele read count did not have a major impact on the percentage of sequences included in the analysis. In contrast, the maximum read depth can have a drastic impact on the percentage of sequences to be included for analysis. For instance, when the maximum read depth was limited to 150 (note that average read depth of this study is 166.7 X), only 4.4% of sequences were included (Table 12). When the parameters were set at 30 for minimum read depth, 300 for maximum read depth, and 5 for minor allele read counts, almost 58% of sequences were included (Table 12). This set of criteria was used for the identification of quality SNPs, the analysis of strain-specific SNPs and the analysis of selective sweeps.

Table 12 Optimization of criteria for SNP identification in channel catfish

Criteria set	Minimum reads	Maximum reads	Minor allele count	% Reads included	Total SNP number
1	20	Excluding top 2%	2	100%	13,582,677
2	30	Excluding top 2%	2	74.7%	13,576,132
3	30	300	3	74.2%	10,217,482
4	30	150	3	6.4%	1,703,297
5	30	300	5	57.6%	8,395,720
6	30	150	5	4.4%	1,295,156
7	50	300	5	57.5%	8,329,404
8	50	150	5	4.4%	1,228,840

SNP identification

A total of more than 13 million potential single nucleotide variations were observed at the most relaxed set of criteria, i.e., minimum read depth of 20, maximum read depth is set as excluding the top 2% of all reads, and minor allele read counts of 2. At our selected set of criteria, a total of 8,395,720 (~8.4 million) putative SNPs (hereafter referred to as SNPs) were identified (Table 12).

These 8.4 million SNPs were subsequently used for the assessment of the distribution of minor allele frequencies. The MAF of each identified SNP was estimated based on the reference number and variant allele reads observed in the reference mapping. Approximately 4 million SNPs have an estimated MAF ≤10% (Figure 4). Over 4.3 million SNPs have an estimated MAF > 10%, of which 2 million had a MAF of 10-20%; 992,502 had a MAF of 20-30%; 693,363 had a MAF of 30-40%; 606,046 had a MAF of 40-50%, and 9,305 SNPs had an equal minor and major frequencies at 0.5 (Figure 4).

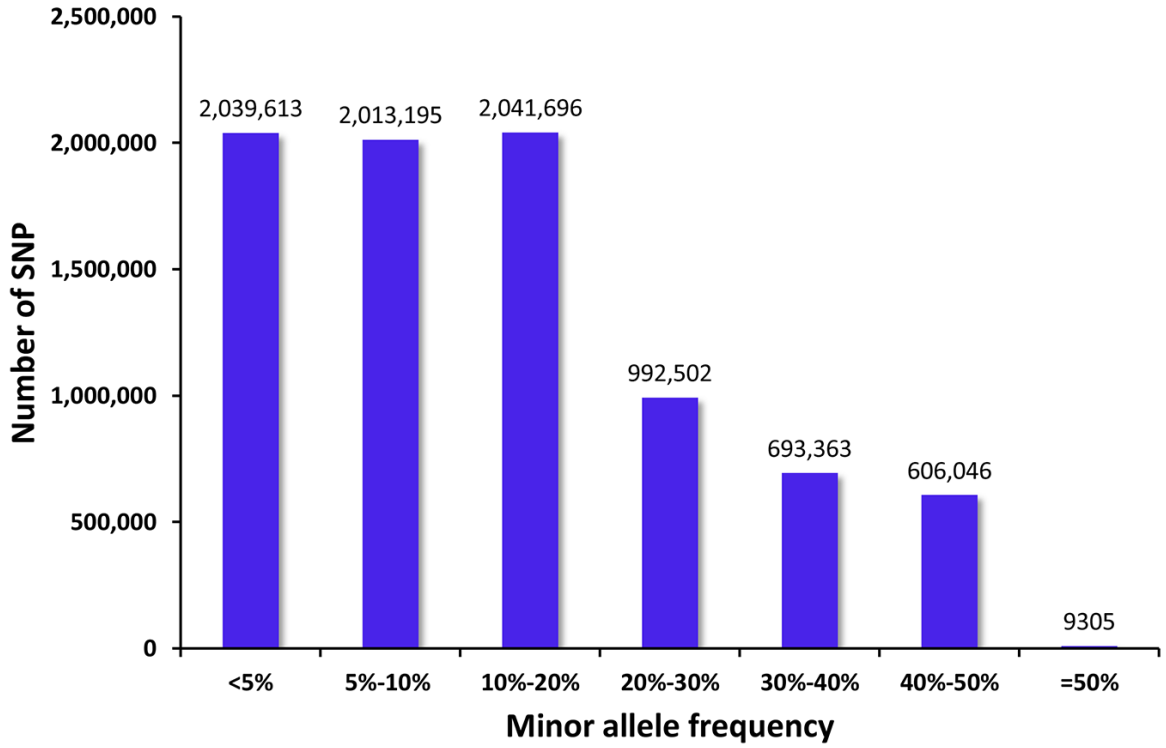

Figure 4. Distribution of SNP minor allele frequencies. SNPs were separated into six categories according to their MAF level. The first two categories contained the range of 5 percent and the other four categories contained the range of 10 percent.

Identification SNPs within and among strains

Putative SNPs identified from each of the five strains are shown in Table 13. Overall, 7.1 million, 4.9 million, 5.3 million, 6.6 million and 6.7 million SNPs, were identified from the Hatchery strain, USDA103, Thompson strain, Marion strain, and wild population, respectively (Table 13 and Figure 5). The largest number of SNPs was identified from the Hatchery strain,

followed by Wild population, Marion strain, and Thompson strain. USDA103 was the strain with the least number of SNPs identified (Table 13 and Figure 5).

SNPs that were observed from only one strain were considered as putative strain-specific SNPs. SNPs that were polymorphic in all strains were considered as common SNPs. Approximately, 2.7 million common SNPs were identified. The number of strain-specific SNPs identified from each of the five strains varied from 66,487 to 143,126, accounting for 0.9%, 2.9%, 2.2%, 1.3%, and 1.7% of SNPs that were identified from that strain, respectively (Table 13).

Table 13 Summary of strain-SNPs in channel catfish

Strain	Quality SNPs	Putative strain-specific SNPs	Percentage
Hatchery	7,100,489	66,487	0.9%
USDA103	4,898,477	143,126	2.9%
Thompson	5,263,008	116,793	2.2%
Marion	6,569,112	88,251	1.3%
Wild	6,654,504	109,998	1.7%

Inter-strain SNPs were also identified from each strain. Following was an example of inter-strain SNP: at a certain position, the genotype of population 1 was A/A; the genotypes of the other four populations were all T/T. Therefore, the genotype of inter-strain SNP in each strain was

homozygous, while the genotype among strains were heterozygous. As shown in Table 14, a number of inter-strain SNPs were identified in each strain, various from six SNPs to 87 SNPs. A total of 182 inter-strain SNPs were identified in all of the five strains. The inter-strain SNPs were very useful for strain differentiation. Because alleles were fixed in each population, inter-strain SNPs were more powerful and stable when being utilized in parentage analysis, population differentiation and original analysis.

Table 14 Summary of Fixed SNP in channel catfish

Strain	Inter strain SNP
Hatchery	6
USDA103	68
Thompson	87
Marion	9
Wild	11
Total	182

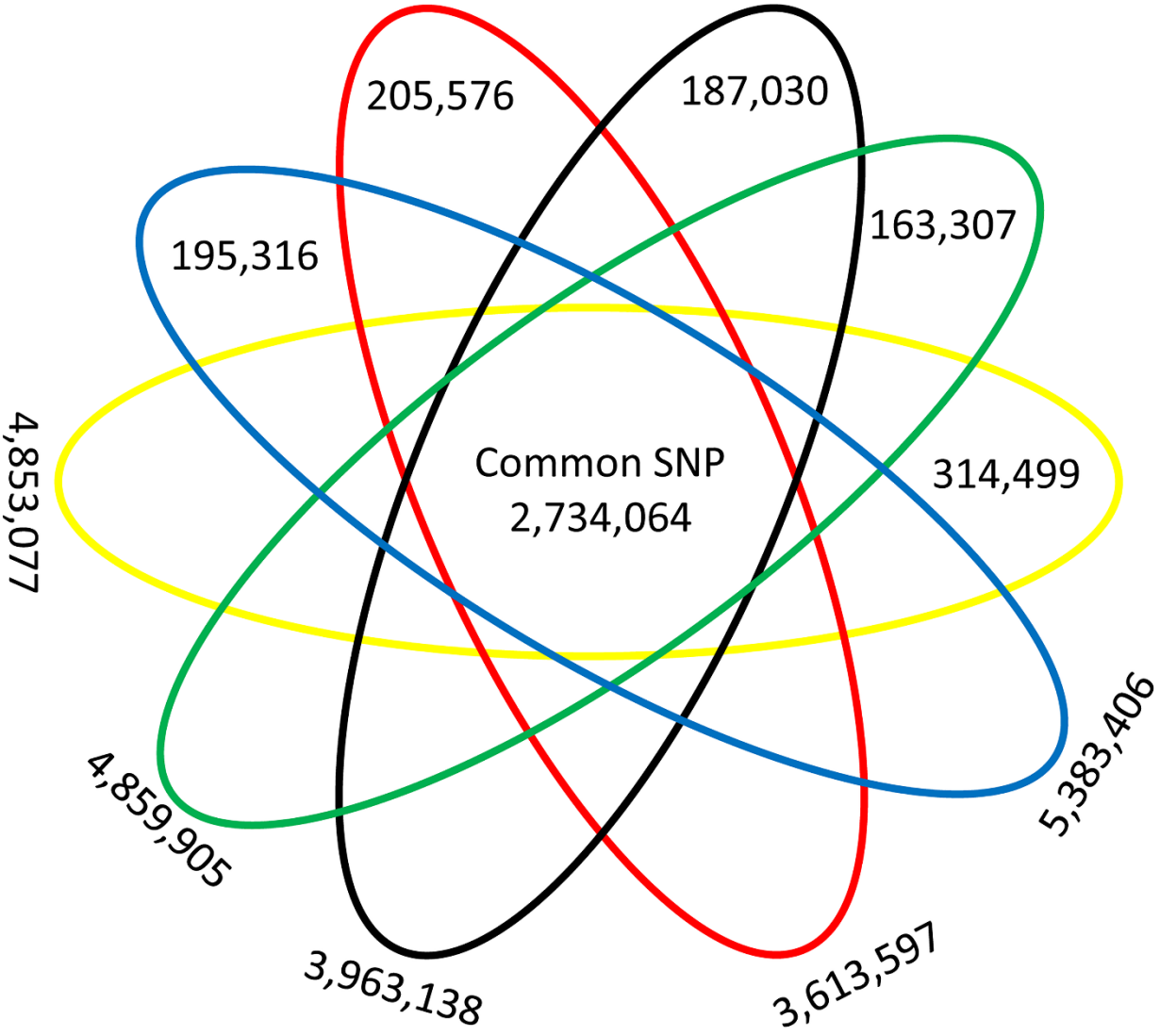

Figure 5. Presentation of common SNPs and strain-specific SNPs. Each color represents a strain. Blue, red, black, green and yellow represent Hatchery line, USDA103, Thompson, Marion and wild strain respectively. Numbers in the oval means number of strain-specific SNP; Numbers outside the oval means total SNPs identified from the strain.

Analysis of selective sweeps

As shown in Table 15, a total of 407,861 significant SNPs were identified, which had significant differences in allele frequencies between domestic catfish strains and the wild population (Fisher's exact test, FDR p-value ≤0.01). From them, 785 SNPs are only heterozygous in wild and homozygous in domestic strains (Appendix table 1); 164,306 SNP are only heterozygous in domestic strains and homozygous in wild strains. Of these 407,861 significant SNPs, 52,076 were located in coding regions, 21,232 were located within 100 bp of coding regions, and 334,553 were located in non-coding regions.

Table 15 Summary of SNPs with significant differences in allele frequencies

Category	**SNP number**
Significant SNPs	407,861
Significant SNPs in coding regions	52,076
Significant SNPs near coding regions	21,232
Significant SNPs in non-coding regions	334,553

A total of 237,655 (58.3%) significant SNPs were assigned to 29 tentative chromosomes based on the catfish linkage map (Ninwichian et al., 2012). The distribution of significant SNPs within chromosomes with the number of significant SNPs in 200 kb bins across each chromosome is illustrated in Figure 6. All of the 29 catfish chromosomes contained significant SNPs, with chromosome 3, chromosome 6 and chromosome 21 harboring the largest number of significant

SNPs (12,494, 12,417 and 12,340, respectively). Chromosome 29 contained the least number of significant SNPs (1,717). Regions with the largest number of significant SNPs were from chromosome 21.

Analysis for selective sweeps was performed as described by Rubin et al. (Rubin et al., 2012a; Rubin et al., 2010). The pooled heterozygosity (H_p) was calculated in 20-kb windows based on the major and minor alleles of significant SNPs, and were then log transformed. Most of the windows (73.5%) had the log-transformed H_p scores between 1 and 1.5, indicating high levels of heterozygosity (Figure 7). A total of 23 windows (0.1%) with log-transformed H_p score $\geqslant$ 4, indicating excessive levels of homozygosity in these regions, were identified as genomic regions with putative selective sweeps (Table 16).

The distribution of the 23 regions with selective sweeps in catfish genome was then analyzed. As shown in Figure 8, these regions were distributed among different chromosomes. Among them, chromosome 5, 12, 17 and 20 contained more than one region with selective sweeps. Chromosome 20 contained a region with the lowest level of heterozygosity. The H_p score of this region was 0 and therefore the log-transformed H_p score was infinite. Thus, a value of 7 was assigned, which was the highest log transformed H_p score (Figure 8) for the convenience of plotting.

A total of 11 genes were found from these genomic regions with selective sweeps (Table 17). These genes were located on eight chromosomes including chromosome 1, 3, 5, 7, 12, 17, 20 and 27. Among these genes, hypoxia-inducible factor 1-beta (*HIF-1β)* had the most significant H_p score, which was followed by ATP-binding cassette sub-family B member 5 (*ABCB5*).

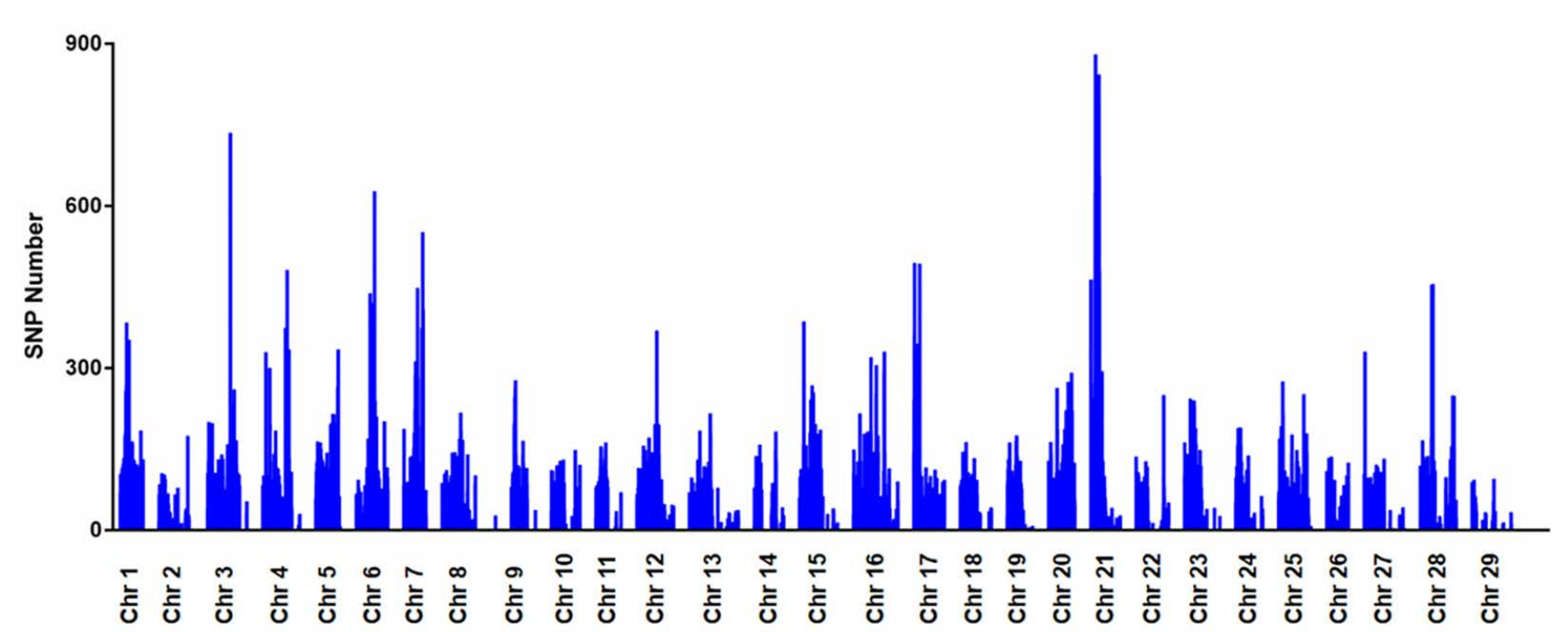

Figure 6. Genome-wide distribution of significant SNPs. Physical positions of all catfish 29 chromosomes are presented on the x-axis, and significant SNP numbers within a window size of 200 Kb is given on the y axis

Table 16 Summary of the 23 genomic regions with putative selective sweeps

Scffold ID	Window number	transformed Hp score	Start position	Protein ID	Gene Name
jcf7180003676417	361	-	20 kb	-	HIF-1-beta
jcf7180003676363	184	-6.38	20 kb	P35072	ABCB5
jcf7180003676363	196	-4.59	20 kb	-	-
jcf7180003676363	205	-4.62	20 kb	-	-
jcf7180003665128	1	-4.34	2.9 kb	-	-
jcf7180003676359	16	-4.31	20 kb	-	-
jcf7180003676305	60	-4.30	20 kb	Q5HZY0, Q6NUV0	Ubxn4, RAB3GAP1
jcf7180003676453	57	-4.30	20 kb	A4IFA3	GTF2IRD2
jcf7180003675342	1	-4.21	20 kb	-	-
jcf7180003676341	56	-4.21	20 kb	-	-
jcf7180003676323	16	-4.12	20 kb	-	-
jcf7180003669997	1	-4.11	10.9 kb	-	-
jcf7180003675277	9	-4.11	20 kb	Q9NQE7	PRSS16
jcf7180003676312	109	-4.07	20 kb	P20794,Q0P436	MAK,TMEM14C
jcf7180003675854	17	-4.06	20 kb	-	-
jcf7180003676350	31	-4.05	20 kb	-	-
jcf7180003662989	1	-4.05	5.1 kb	-	-
jcf7180003676337	177	-4.03	20 kb	P27546	MAP4
jcf7180003668664	1	-4.03	10.2 kb	Q8SPJ1	JUP
jcf7180003676121	31	-4.03	20 kb	-	-
jcf7180003668055	2	-4.02	20 kb	-	-
jcf7180003665961	1	-4.01	11.5 kb	Q5T3F8	TMEM63B
jcf7180003670939	2	-4.01	20 kb	-	-

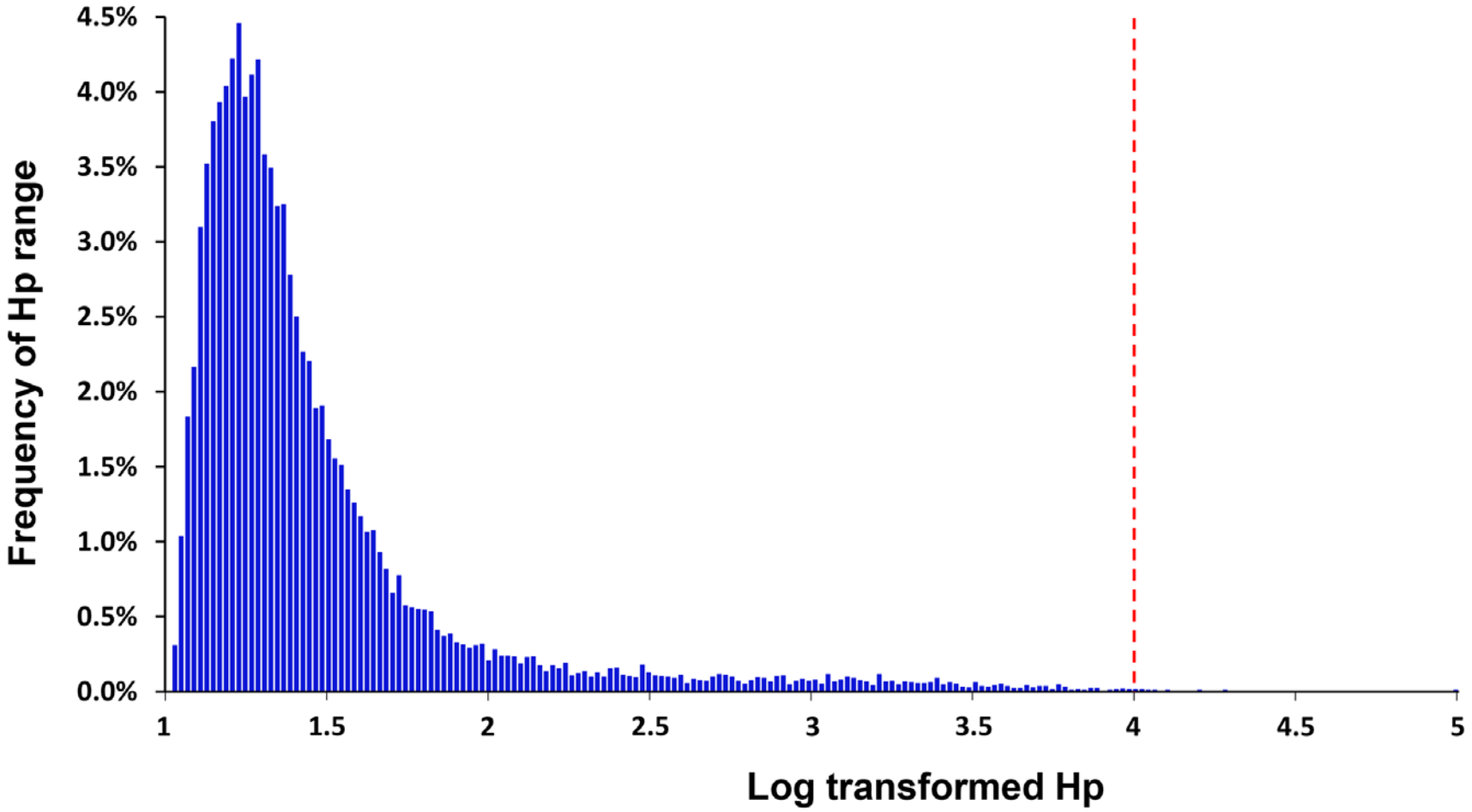

Figure 7. Histogram of log-transformed pooled heterozygosity (*Hp*) values. The x-axis is evenly divided into 200 bars from 1 to 5, and each bar represents a transformed *Hp* range of 0.02. The y-axis represents the percentage of each transformed *Hp* range in the total 200 transformed *Hp* ranges. All *Hp* values were transformed by $-\log_2$.

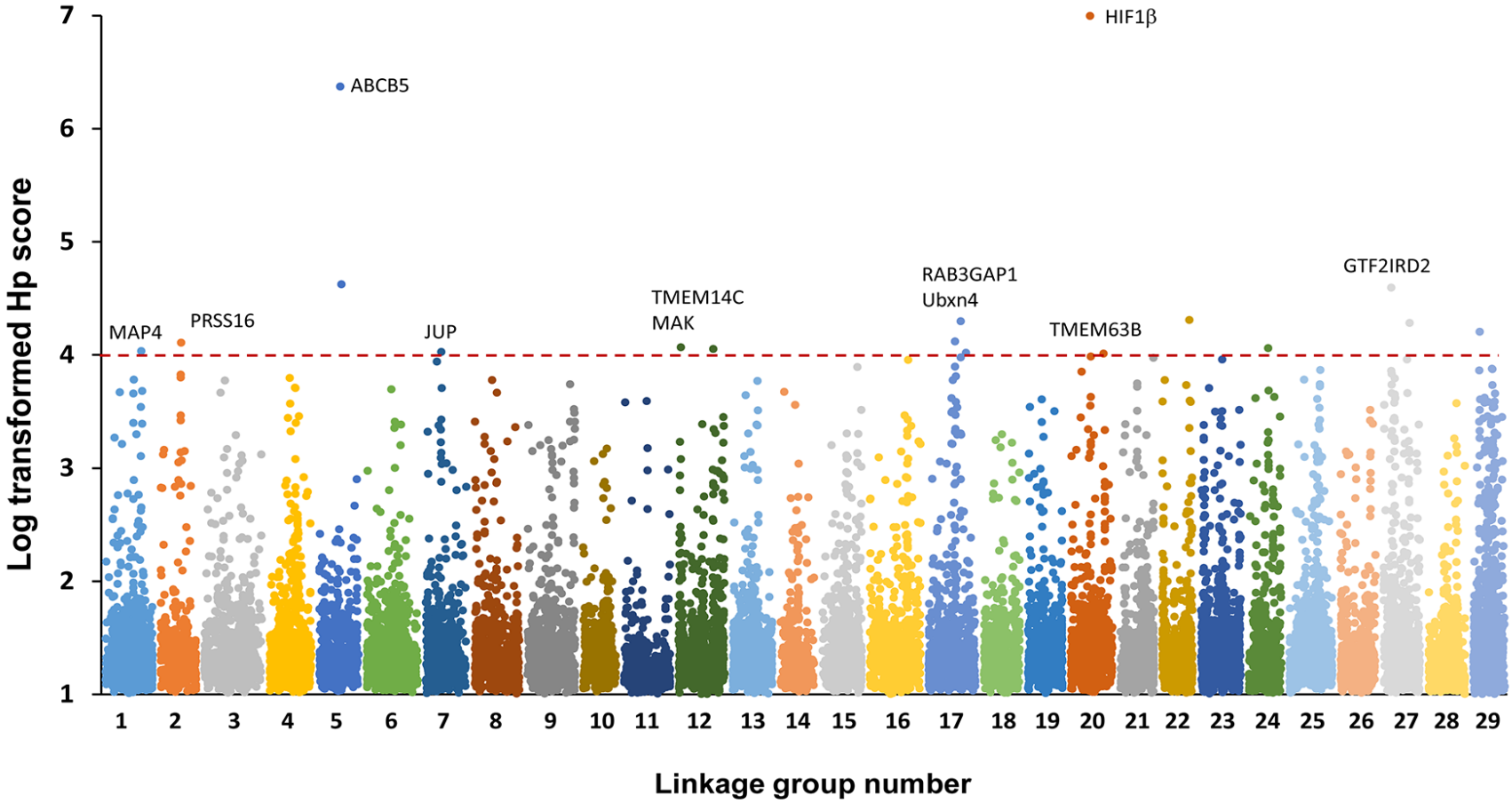

Figure 8. Genome-wide distribution of log-transformed pooled heterozygosity (*Hp*) values. The x-axis represents the positions of windows (20 Kb) along each chromosome, which is represented with different colors. The y-axis represents the *Hp* scores transformed by $-\log_2$. Windows of *HIF-1β* had the *Hp* score of 0, therefore, its transformed *Hp* score was defined as 7, the maximum score, for the convenience of plotting.

Table 17 List of genes identified from the regions with selective sweeps

Chromosome	Pooled heterozygosity	Log transformed *Hp*	Gene name	Putative function
Chr 20	0	-	*HIF-1β*	Stress response
Chr 5	0.012	6.38	*ABCB5*	Unknown
Chr 17	0.051	4.30	*RAB3GAP1*	Eye/brain development
Chr 17	0.051	4.30	*Ubxn4*	ERAD
Chr 27	0.051	4.30	*GTF2IRD2*	Transcription factor
Chr 3	0.058	4.11	*PRSS16*	T cell development
Chr 12	0.060	4.07	*TMEM14C,*	Heme biosynthesis
Chr 12	0.060	4.07	*MAK*	Spermatogenesis
Chr 1	0.061	4.03	*MAP4,*	Microtuble assembly
Chr 7	0.061	4.03	*JUP*	Junctional plaque protein
Chr 20	0.062	4.01	*TMEM63B*	Unknown

Discussion

In this study, next generation sequencing was conducted for multiple individuals from four aquaculture strains and one wild population to identify SNPs for determination of genomic impact of domestication. The large numbers of SNPs identified from this study will be useful for the development of high density SNP arrays for genetic and genomic analysis in catfish (Liu et al., 2014).

Pooled sequencing has been utilized as an efficient and reliable approach for detecting and genotyping SNPs from populations (Bansal et al., 2010). One of the challenges for this approach is to distinguish the real from false SNPs. Validation of millions of SNPs is not practical and extremely costly if not impossible. Strategies to increase SNP conversion rate need to be developed. To increase the likelihood for the identification of real SNPs, major factors affecting SNP identification need to be assessed, of which, the maximum reads, minimum reads and minor allele read count were the most important and common factors, incorporated into various SNP detection tools (Koboldt et al., 2009; Kofler et al., 2011; Li et al., 2009a; Wang et al., 2008).

Setting of maximum read depth mainly controls the mapping quality and filter SNPs located on repetitive elements, especially on interspersed repeats. On the genome scale, large numbers of sequences are repetitive elements. Nearly half of the human genome is made up of repeat sequences (Lander et al., 2001). Thus, genome-scale SNP identification usually results in a large number of false SNPs from misalignment of reads from repetitive elements. Therefore, the genome regions with extremely high read depth are more likely to represent repetitive regions. We studied the correlation between the read coverage and the proportion of interspersed repeats

(Table 11). The results demonstrated that when the read coverage >300, the number of interspersed repeats increased approximately three folds when compared with other read coverage intervals. The figure of maximum-total SNPs (Figure 2) indicated the main body of total SNPs is located on the middle area of maximum read intervals, which is approximately 1.5-fold of average read coverage. As a rule of thumb, setting the maximum reads at no more than twice the total average sequencing depth should reduce the chances of false SNPs. Therefore, we set 300 as the maximum read number to avoid the inclusion of interspersed repeats.

Setting of minimum read depth is used to remove low quality mapping positions caused by mapping error or insufficient coverage. It had a very limited effect on the SNP identification at first, demonstrated that almost all of the SNPs detected can pass this quality check (Figure 1). When minimum reads ≥30 (6 reads for each strain), the capacity of total SNPs begins to decrease proportionally with the increase of minimum reads. We set this point as the criterion of minimum reads to reduce the proportion of low quality SNPs and at the same time, to keep as many SNPs as possible.

Normally, minor allele read count is a quality control factor and can be used for neutralizing the effect of sequencing error rate. For SNP calling, of course there should be at least one non-reference allele count, but detection of large numbers of false SNPs will occur by using such a relaxed criterion. Obviously, increasing the standard of minor allele count would reduce the false SNP rate, but at the expense of eliminating some of the real SNPs. We found that minor allele read count had a very major effect on the number of total SNPs at first when it was under 5

(Figure 3), which demonstrated that numerous SNPs with only few reads of the variant allele were located in this interval, we classified these SNPs as low quality SNPs.

In that regards, reasonable criteria for SNP identification were set at a minimum read depth of 30, maximum read depth of 300, and minor allele count of 5, and 8.4 million putative SNPs were identified from five different catfish strains. On average, there are one SNP every 116 bp in catfish genome (Table 18). This level of genome sequence variation is similar to those in chicken, higher than in bovine, but lower than in medaka, mouse and human. Medaka has a very high SNP rate at 1 SNP/43 bp, and it is the most polymorphic vertebrate species reported to date (Kasahara et al., 2007). Chicken has 1 SNP every 133 bp (Rubin et al., 2010). Bovine has an estimated SNP rate of 1 SNP/378 bp (Stothard et al., 2011), three times less frequent as compared with the catfish genome. For mouse, 56.7 million SNP were identified from 17 inbred strains, i.e., approximately 1 SNP/61 bp (Keane et al., 2011). Human has an estimated SNP rate of 1 SNP/ 87 bp, identified from 1,092 individuals from 14 populations. Apparently, several factors would affect the SNP frequency including: 1) the number of populations involved in the analysis as well as the relatedness of these populations; 2) the number of chromosome sets tested; and 3) the sequencing depth for each project. Therefore, a direct comparison may prove to be difficult. However, this information can still provide us a rough assessment of the polymorphisms among species. In this regard, all the vertebrate animals are much less heterozygous than some of the invertebrate animals such as sea squirts, whose genome harbors 1 SNP every 20 bp (Small et al., 2007).

Table 18 Comparison of SNP frequencies in different species

Species	SNPs frequency in the genome	Populations /strains	Chromosome sets tested	References
Catfish	1 SNP per 116 bp	5	300	This study
Medaka	1 SNP per 43 bp	2	736	Kasahara et al. 2007
Chicken	1 SNP per 133 bp	8	174	Rubin et al., 2010
Bovine	1 SNP per 378 bp	2	4	Stothard et al. 2011
Mouse	1 SNP per 61 bp	17	34	Keane et al. 2011
Human	1 SNP per 87 bp	14	2048	Kidd et al., 2004

Approximately, 66,000-143,000 SNPs were identified as strain-specific for each strain (Table 13), which was approximately 6% of all SNPs. If more strains were evaluated than the 5 in this study, the proportion of strain-specific SNPs would likely be reduced. Catfish strains are almost impossible to distinguish based on phenotypes (Waldbieser and Wolters, 2007), therefore, these SNPs can be potentially used for strain identification, tracing the origin of commercial strains, and analyzing the genetic difference among strains and to mark fish for other genetic experiments. The 2.7 million common SNPs that are polymorphic in all five catfish populations will provide the main resources for SNP array design (Liu et al., 2014) and high-density linkage map development.

Liu *et al.* (Liu et al., 2011) sequenced 48 individuals of channel catfish from different strains (Marion, Pearson, Moyer, Holland and Noble) using pooled samples and detected more than two

million putative gene-associated SNPs with more than 0.5 million being high quality SNPs. Approximately, 66% (341,663) of the high quality SNPs were identified in our results, supporting the confidence of parameters used in this project. The remaining 34% of SNPs that were not shared by these two studies may be caused by the use of different strains, as well as the relatively stringent parameters used for SNPs calling in this study.

SNPs with significant differences in allele frequency between domestic and wild catfish populations were identified to provide insight into genomic impact of domestication and selection. Compared with all the SNPs identified from channel catfish, significant SNPs were approximately 5% of the total SNPs, indicating that the vast majority of genomic regions have not been affected by domestication or selection. Additional analysis was conducted to determine the position and genes associated with significant SNPs. The vast majority of significant SNPs (87.2%) were located in the non-coding DNA sequences, while 12.8% of the significant SNPs were found in coding regions of catfish genes. This proportion of SNPs associated with genes is greater than the proportion of gene sequences from the whole genome sequences, suggesting that domestication and selection may have had a greater impact on genes than on intergenic regions.

The significant SNPs were distributed on each of the catfish chromosomes (Figure 6). Chromosome 3, 6 and 21 contained a largest number of significant SNPs, but from which no putative selective sweeps were identified. Perhaps, the catfish genome harbors a large amount of genetic variation for further domestication and selective breeding given the relatively short domestication and history of selection. Also, recent studies indicate that soft sweeps are abundant in adaptation and may play a major role in the rapid adaptation in many species (Messer and Petrov, 2013). Because soft sweeps contain multiple adaptive alleles and they all

have relatively high frequencies, their genetic diversities should also be high. In this project, we only focused on hard selective sweeps from pooled sequencing data by searching the regions with low genetic diversity. Soft sweeps may be present in those chromosomes with abundant SNPs, but we only conducted our analysis with bi-allelic SNPs and our analysis does not provide any insight into soft sweeps.

A concern regarding the analysis of channel catfish was sampling since this species occupies a large geographical range, populations can be large and numerous domestic and wild populations exist. Assuming that all domesticated populations and a broad representation of wild populations can be achieved, significant SNPs between the domestic and wild populations could be used to reveal solid selective sweeps caused by domestication and selection. However, based on the nature of catfish industry, it is difficult to sequence large enough samples that can represent all genetic variations that exist in all domestic and wild strains. Therefore, we fully acknowledge the difficulties involved in the sampling of the domestic and wild populations for an aquatic species, however, analysis of putative selective sweeps should still provide insights into the potential impact of domestication on genome evolution. To identify hard type selective sweeps in domestic catfish caused by selective breeding, we analyzed the pooled heterozygosity (H_p scores) for the domestic populations using significant SNPs with the assumption that artificial selection by domestication tends to create runs of homozygosity (Kim et al., 2013).

When hard selective sweeps are analyzed using the method of Rubin et al (Rubin et al., 2012a; Rubin et al., 2010), two parameters could affect its accuracy and sensitivity. The first is the window size used for the calculation of H_p scores. Large window sizes could contain more SNPs and reduce the bias in the calculation of pooled heterozygosity, but it will also lose sensitivity

due to the uneven distribution of SNPs. In catfish, where the whole genome has not been fully assembled, the window size should be set smaller than those species with whole genome reference assemblies simply because very long contigs are not yet available. After reviewing variable window sizes, we used 20-kb siding windows. Another noteworthy parameter is the SNP number in each window. Obviously, windows with very small SNP number cannot provide the actual heterozygosity of the genome regions they represent. Therefore, we did not include the windows that contained less than five significant SNPs in the analyses.

Domestication and selection could change genetic variability, the genetic correlations among traits and the interactions among loci. Traits with high production values, such as growth rate, disease resistance and tolerance to low oxygen have been selected for generations in aquaculture species either intentionally or unintentionally. Resistance to low oxygen is an important aquaculture trait relevant not only for survival, but also growth and disease resistance. Hypoxia can cause high mortality for aquaculture species. Even if the fish survive under hypoxic conditions, exposures to low oxygen levels often trigger disease incidents that cause further major losses (Affonso et al., 2002; Guerriero et al., 2002). Variations in tolerance to low oxygen have been well studied with various aquaculture species (Anttila et al., 2013; Faust et al., 2004; Guan et al., 2011). However, genetic variation for low oxygen tolerance have not been systematically determined. In case of catfish, great efforts have been made on the genetic improvement of the important production traits, such as growth rates, disease resistance, tolerance to handling stress and hypoxia (Dunham and Smitherman, 1983; Dunham et al., 1994; Geng et al., 2014), but little is known of the genomic basis for such observed phenotypic improvements.

In the current study, a total of 23 genomic regions were identified that contained the signature of selective sweeps (log transformed H_p score >4, Table 16), which could be the strong candidates for further studies of domestication in channel catfish. These 23 regions were located in different chromosomes (Figure 8), suggesting that multiple traits or multiple loci controlling a few traits could have responded to domestication. A selective sweep caused by domestication was identified in channel catfish Chromosome 17 (Pooled heterozygosity = 0.051), which is highly homologous to zebrafish Chromosome 9 (Jiang et al., 2013). A QTL responsible for the anti-predator behavior on zebrafish Chromosome 9 was detected by three different measures (Wright et al., 2006). However, since those genomic regions are still large, it is not certain if the same genomic regions were under selection in zebrafish and in catfish. In three-spined stickleback, analysis for selective sweeps was conducted between ancestral oceanic populations and newly established freshwater populations (Hohenlohe et al., 2010). A total of nine regions were identified with adaptive significance, three of which were supported by the previous QTL analysis on fresh water adaption. Domesticated strains and wild populations of in Atlantic salmon were compared using 261 SNPs and 70 microsatellite markers (Vasemägi et al., 2012). A total of ten genomic regions were identified from different chromosomes with 14 genes identified from these regions. However, there was no overlap between these genes with our findings in channel catfish.

In the present study, we identified 11 genes from the 23 genomic regions with selective sweeps (Supplemental file 1). Two genes, hypoxia-inducible factor-1-beta (*HIF-1β*) and ATP-Binding Cassette, Sub-Family B, Member 5 (*ABCB5*), were located in the first two strongest hard sweeps (Figure 8). *HIF-1β* was located on the selective sweep region with H_p = 0, meaning that all the significant SNPs located in this region were homozygous in all domestic populations and were

heterozygous in the wild population. HIF-1β, also referred to as Aryl hydrocarbon receptor nuclear translocator (ARNT), mediates aryl hydrocarbon signaling and facilitates gene activation by dimerization with aryl hydrocarbon receptor (AHR) (Swanson, 2002). It is involved in the hypoxia response pathway where it forms heterodimers with HIF-1α, which in turn binds to P300 to activate a variety of hypoxia-responsive genes upon exposure to hypoxia (Semenza, 2003; Wilson and Hay, 2011). It is reasonable to conclude that selection for hypoxia tolerance under aquaculture conditions could have had a major genomic impact in this genomic region.

ABCB5 is a member of ATP-Binding Cassette transporter gene family that exists only in vertebrates (Annilo et al., 2006; Liu et al., 2013). The ABC transporters are membrane bound proteins and responsible for the transportation of substrates across biological membranes including sugars, amino acids, ions, polypeptides, and toxic metabolites. Proteins encoded by the ABC transporter gene family share a highly conserved domain structure. The uniqueness of domain structure among ABC transporters indicated their similarities in function. To transport the molecules, two transmembrane domains (TMDs) and two nucleotide-binding domains (NBDs, also called ATP-binding domains) were needed. The functions of ABC transporters can be classified into eight subfamilies according to their domain structure and primary sequence. Three functional groups, including importers, exporters and others, can be set up for functional classification of the ABC transporters. A total of 48 mammalian ABC transporters were first identified in Human (Dean, Hamon et al. 2001), with many of them discovered with hereditary diseases. In invertebrate species, such as worms and insects, ABC transporters have been associated with insecticide resistance and drug resistance (Leprohon, Légaré et al. 2006, Labbe, Caveney et al. 2011). In channel catfish, a total of 50 ABC transporter genes were identified, which can be divided into sever subfamilies (Liu, Li et al. 2013). The results of phylogenetic

analysis shown that the events of gene duplication and gene deletion were exist during the catfish genome evolution.

ABCB5 was highly expressed in melanocytes and may play an important role in melanomagenesis (Annilo et al., 2006; Lin et al., 2013). The expression of *ABCB5* was also significantly associated with tumor progression and recurrence, acting as an energy-dependent drug efflux transporter and function during the multidrug resistance process (Grimm et al., 2012; Szakács et al., 2006). Studies on childhood obesity reported a CNV region on *ABCB5* gene that was exclusively associated with childhood obesity (Glessner et al., 2010). For fish species, certain interspecific hybrids of *Xiphophorus* has been served as malignant melanoma models for years as they can induce melanoma spontaneously (Meierjohann, Schartl et al. 2004). Also, evidences of melanoma on wild coral trout (*Plectropomus leopardus*) have also been found. An interesting study of fish skin cancer reported that skin cancer can increase mating success in animals because female swordtail fish preferred males with black melanoma splotches (https://www.ohio.edu/research/communications/fish_cancer_gene.cfm). In our results, *ABCB5* was located in the second strongest selective sweep region (Table 16), suggesting extremely low genetic diversity block around the genomic region containing the ABCB5 gene.

For the genes with log transformed H_p scores around 4, which is not as significant as the other genes like HIF1β and ABCB5, their functions were also analyzed and listed in Table 17, which could provide some insights into the genetic reasons of domestic catfish traits such as the high production rate and the abilities of handling stress. However, we must stress that the sample size and radius were limited in this project, and even in the future, sampling and analysis of large

numbers of samples is cost prohibitive, and therefore, caution need to be exercised for the interpretation of such analysis with aquatic species with extremely large populations.

Rab3 GTPase activating protein subunit 1 encodes the catalytic subunit of a Rab GTPase activating protein. The heterodimer formed between RAB3GAP1 and a non0catalytic subunit could regulate the activity of small G proteins. The protein can also hydrolyze the GDP bound of Rab3. Mutations of Rab3 GTPase activating protein subunit 1 were reported that can result in Warburg Micro Syndrome (Warburg, Sjo et al. 1993). The features of the disease including significant visual impairment, postnatal microcephaly and intellectual disability. A total of 41% of Warburg Micro Syndrome were caused by homozygous mutations in Rab3 GTPase activating protein subunit 1, which was the most frequent mutation type in Warburg Micro Syndrome (Handley, Morris‐Rosendahl et al. 2013). In mice, individuals with rab3gap1 deletion showed to have abnormal release of synaptic vesicles and altered short-term synaptic plasticity in the hippocampus, indicating that basal synaptic transmission is suppressed in the mutant hippocampal synapses. (Sakane, Manabe et al. 2006). However, these mice were fertile, viable and no eye or brain abnormities. In our channel catfish data, the low heterozygous rate of domestic catfish rab3 GTPase activating protein subunit 1 gene indicated the existing of homozygous mutations in the gene and these mutations may contribute to the resistance of handing stress by suppression of synaptic transmission.

Ubiquitin regulator-X domain containing protein 4, also called erasin, is a membrane protein found in endoplasmic reticulum. The ubiquitin regulator X domain was first identified in 1996 in several eukaryotic proteins, as a protein domain similar to ubiquitin (Hofmann and Bucher

1996). In general, ubiquitin regulator X domain containing proteins are cofactors for Cdc48, which is also known as p97 (Decottignies, Evain et al. 2004, Hartmann-Petersen, Wallace et al. 2004). The ubiquitin regulator X domain contained about 80 amino acid residues and a number of proteins contained ubiquitin regulator X domain were identified and have been divided into several subfamilies including UBXD1, FAF1, SAKS1, TUG Rep-8 and UBXD3. Ubiquitin regulator-X domain containing protein 4 was belong to the subfamily of p47, which contained a central SEP domain and an ubiquitin regulator X domain. One of the common feature of the p47 subfamily was that it had two p97 binding site in the SEP domain and was important in the process of ER-associated protein degradation (Bruderer, Brasseur et al. 2004, Hitt and Wolf 2004). Ubiquitin regulator-X domain containing protein 4 is a highly conserved Erasin-like protein and play important roles in the ER-associated protein degradation process as a cofactor of Cdc48/p97(Liang, Yin et al. 2006, Schuberth and Buchberger 2008). In channel catfish, the extremely low genetic diversity in the genomic region contained Ubiquitin regulator-X domain containing protein 4 indicated that it may involve in channel catfish domestication and contributed in channel catfish cell autophagy process as well as stress responses such as handling stress, bacteria diseases and low oxygen stress.

GTF2IRD2 is a gene belong to I-repeat containing family of proteins (TFII-I family). It was identified in 2004 and was the latest member of TFII-I family (Tipney, Hinsley et al. 2004). In addition to its structural similarities to other I-repeat containing proteins, GTF2IRD2 is a fusion gene which contained a novel C-terminal transposon-like motif, which could be a result of transposable element random insertion. GTF2IRD2 was believed a gene related to Williams–Beuren syndrome, because it was located in the Williams–Beuren syndrome critical region on human chromosome 7, had the similar sequence structure with other genes located in Williams–

Beuren syndrome critical region, and its chromosomal location at the telomeric end of the Williams–Beuren syndrome breakpoint. The existence of the transposable element motif could allow the binding of other elements and lead to regional instability (McCarron, Duttaroy et al. 1994).

GTF2IRD2 had three copies in the critical region of Williams–Beuren syndrome, and was deleted in some Williams–Beuren syndrome patients, with classic clinical phenotypes, including cardiovascular system, memtal retardation, distinctive facial features, and tooth anomalies (Ohazama and Sharpe 2007). It has been reported that the TFII-I gene family was located on the genomic region that responsible for craniofacial anomalies. During the process of tooth development, GTF2IRD2 was expressed in the epithelial buds at the bud stage, it was also expressed in preameloblasts and preodontoblasts at the early bell stage (Ohazama and Sharpe 2007). Also, GTF2IRD2 can function as a regulator, which can inhibit the function of the other members in TFII-I gene family and GTF2IRD1 (Palmer, Taylor et al. 2012). Experiment results showed that transgenic expression of GTF2IRD1 and GTF2IRD2 in skeletal muscle leaded to significant shifts of fiber type in opposite direction. And the offspring of GTF2IRD1 and GTF2IRD2 mice showed a normal fiber type, suggesting interactions between them (Palmer, Taylor et al. 2012). Furthermore, it is reported that GTF2IRD2 was involved in higher-level abilities, for example, cognitive and behavioral functions. Analyses of these higher-level abilities showed that Williams–Beuren syndrome showed that patients with GTF2IRD2 deletion were significantly more cognitively impaired in executive functions including social reasoning, cognitive flexibility and spatial functioning (Porter, Dobson-Stone et al. 2012). In channel catfish, the genomic region with GTF2IRD2 gene of domestic populations showed significantly

less genetic diversity, suggesting that the mutations in this region may change the behavior of domestic fish and was selected during the breeding program for generations.

Thymus-specific serine protease (TSSP), which is encoded by PRESS16 gene, is one of the important proteins involved in intrathymic antigen presentation by MHC class II and involved in the positive selection of CD4+ thymocytes during the intrathymic T-cell discrete precess (Gommeaux et al., 2009). The CD4+ T cells, also known as T helper cells, can assist other immune cells in immunologic process for both type-1 and type-2 immunity. Previous studies on channel catfish immunity have shown that the catfish mucosal tissues such as skin and intestine are mainly responsible for the resistance of catfish disease such as enteric septicaemia of catfish (ESC) and columnaris (Li et al., 2012; Sun et al., 2012). Thus, the function of T helper cells in regulate type-2 immunity for the protection of mucosal sites from pathogens (Shinkai et al., 2002) indicates the potential roles of the PRSS16 gene in catfish disease immunity.

Heme, as a complex of protoporphyrin IX and iron, is extremely essential for most of the living organisms. In hemoproteins such as hemoglogin, myoglobin, and cytochromes, heme is a prosthetic group and function as a transporter for electrons and oxygen (Wijayanti et al., 2004). However, It could also be deleterious because free heme can generate reactive oxygen species that lead to oxidative stress. Therefore, the levels of cellular heme are tightly controlled by a well-organized balance between heme biosynthesis and heme catabolism (PONKA, 1999).

Transmembrane protein 14C is a gene coding for a transmembrane protein functioned as mitochondrial transporter. Studies on heme biosynthesis showed that the gene was essential for understanding inherited anemia and hemoglobin production (Nilsson et al., 2009; Yien et al., 2014). In 2009, a total of five genes, including TMEM14C, SLC25A39, SLC22A4, C1orf69 and

ISCA1 were identified as candidate genes that involved in heme biosynthesis (Nilsson et al., 2009). Gene knock-down experiments in zebrafish showed that individuals with all five genes knocked down showed profound anemia, without modifications in erythroid lineage specification (Nilsson et al., 2009). Another study on TMEM14C reported that it was enriched in vertebrate hematopoietic tissues and is important for erythropoiesis and heme synthesis in vivo and in vitro (Yien et al., 2014). Because TMEM14C was important mitochondrial transporter, TMEM14C deficiency mice showed prophyrin accumulation in the mice fetal liver due to profound anemia, accompany with the phenotypes of erythroid maturation arrest and embryonic lethality. In general, their research illustrated that TMEM14C involved in the terminal steps of the heme synthesis pathway and facilitates the transport of protoporphyrinogen IX for heme biosynthesis and hemoglobin production. In channel catfish, we identified significant difference in gene HIF-1β, which is the most significant gene between domestic channel catfish and wild channel catfish. The identification of TMEM14C, who primary functions in heme biosynthesis, provide another evidence that domestic channel catfish was more tolerant to wild channel catfish. It also guaranteed further work in the field of genetic mechanisms of low oxygen tolerance.

Androgen, working together with androgen receptor, control the development, maintenance and transformation of prostate. It was also related to the development of male sex organs and secondary sex characteristics. The main function of androgen including testes formation, androgen production, spermatogenesis and muscle mass regulation. Male germ cell-associated kinase is a serine/threonine protein kinase that play a role in cell cycle regulation. Human male germ cell-associated kinase was identified in 2002 as an androgen associated kinase protein (Xia, Robinson et al. 2002). The results of Real-time PCR showed that the expression of male germ cell –associated kinase was 9-fold induced by the androgenic hormone

5alpha-dihydrotestosterone 24h post-stimulation (Xia, Robinson et al. 2002). Also, male germ cell –associated kinase had a higher expression level in prostate cancer cell lines than in normal cell lines, indicating that male germ cell –associated kinase is a protein kinase that involved in androgen synthesis and should be participate in androgen-mediated signaling in cell lines of prostate cancer cell. Another study reported that male germ cell-associated kinase has physical contact with androgen receptor, a type of nuclear receptor and most closely related to the progesterone receptor. Also, male germ cell –associated kinase can improve the ability of androgen receptor transactivation in different prostate cancer cell lines and can interact with steroid receptor coactivator-3 co-activator. Individuals with male germ cell –associated kinase gene knock-down can result in the reduction of androgen receptor transactivation ability. Furthermore, cells with male germ cell –associated kinase deficiency showed a phenotype of growth reduction. The expression analysis of the cells illustrated that the androgen receptor pathway was significantly impeded, suggesting that male germ cell –associated kinase may be a general co-activator of androgen receptor and involved in androgen receptor function in prostate cancer cells (Ma, Xia et al. 2006). In addition to androgen receptor-dependent function, male germ cell-associated kinase also has androgen receptor-independent function in mitosis. The overexpression of male germ cell-associated kinase gene could result in mitotic defects, for example, centrosome amplification and lagging chromosomes, through the decreasing of anaphase promoting complex (Wang and Kung 2011). Overall, male germ cell-associated kinase was function in both androgen receptor-dependent and –independent and participate in the development of prostate cancer from the early stage to late stage (Wang and Kung 2011).

Microtubule-associated protein 4 is a gene that encode a major non-neuronal microtubule-associated protein. The protein is involved in microtubule assembly and the

phosphorylation of microtubule-associated protein 4 could affect microtubule properties and cell cycle progression. It has been reported that low free tubulin concentration could lead to down-regulation of microtubule-associated protein 4 (Holmfeldt et al., 2003).

This protein is also involved in hypoxia response through the regulation of mitochondrial membrane permeability, which plays a key role in apoptosis and necrosis induced by hypoxia. It is reported that microtubule-associated protein 4 phosphorylation increased after hypoxia and resulted in microtubules disruption, although its protein levels do not change (Hu et al., 2010). The subsequent study demonstrated the overexpression of microtubule-associated protein 4 can promote the stabilization of microtubule network through increased production and polymerization of tubulin under low oxygen condition (Fang et al., 2011). Also, the overexpression of microtubule-associated protein 4 can improve cell viability and ATP under low oxygen condition (Fang et al., 2011). However, the actual mechanisms related to microtubule-associated protein 4 has not been determined.

Microtubule-associated protein 4 was also identified as HIV-1 dependency factors. It was reported that knock-down of dynein, axonemal, light chain 1 and microtubule-associated protein 4 inhibited HIV-1 infection regardless of envelope (Gallo and Hope, 2012). It was also demonstrated that dynein, axonemal, light chain 1 and microtubule-associated protein 4 affected reverse transcription other than unclear translocation. These results indicated that dynein, axonemal, light chain 1 and microtubule-associated protein 4 may related to the HIV life cycle at reverse transcription (Gallo and Hope, 2012).

Junction plakoglobin, also called gamma-catenin, is a protein that encode by gene JUP. The protein can bind to classic cadherins as well as desmosomal cadherins. It was also a critical

protein involved in the morphogenesis of the skin and heart (Breuninger et al., 2010). It has been reported that junction plakoglobin is a tumor suppressor gene in a number of cancers including cervical, breast and bladder cancer (Denk et al., 1997; Giroldi et al., 1999; Sommers et al., 1994). The decreased expression of junction plakoglobin during prostate cancer progression may related to the invasion and metastasis of junction plakoglobin, while the detailed role of junction plakoglobin in prostate cancer is still unknown (Franzen et al., 2012). The down-regulation of junction plakoglobin suppressed the proliferation and colony formation of chromic myeloid leukemia cells (Niu et al., 2013). The down-regulation can also inhibited the phosphorylation of glycogen synthase kinase-3-beta. These results indicated that junction plakoglobin is an oncogene protein in chronic myeloid leukemia (Niu et al., 2013).

Transmembrane protein 63B is a protein-coding gene. There is not much studies about this gene. Several GO annotations of this gene were available including lysosomal membrane (GO:0005765), membrane (GO:0016020), intergral component of membrane (GO:0016021), and extracellular vesicular exosome (GO:0070062). A genome-wide associate study demonstrated that transmembrane protein 63B, together with transmembrane protein 217 and glutamate receptor, ionotropic, kainate 2, was associated with diabetic retinopathy. Both of these genes were located on the same loci of human chromosome 6 (Lin, Huang et al. 2013).

Considering the smaller effective population size of domestic strains at research institutions compared to wild populations, some random genetic changes may take place due to founder effect and genetic drift. However, commercial populations are much larger than wild populations, but still could be impacted by founder effects. These would be partially offset by crossbreeding as many commercial populations originated from multiple strains (Dunham and

Smitherman, 1984). Our findings of domestication related regions and genes could provide some insights into the genetic explanation of the differences between domestic and wild channel catfish in performance, morphology and behavior traits. For instance, the smallest numbers of SNPs were detected in USDA103. This may have been a result of historically small population sizes, founder effects from one or more brood stock transfers between hatcheries and research institutions, and intense selection for growth as this was one of the fastest growing domestic strains even before the recent directed selection (Dunham and Smitherman, 1983). Additionally, a large number of SNPs identified in this project using stringent criteria have been included in the construction of catfish SNP array (Liu et al., 2014) and will be further utilized in analysis of population diversity, development of high-density linkage maps and genome-wide selection.

Reference

1. Abo-Ismail, M.K., Vander Voort, G., Squires, J.J., Swanson, K.C., Mandell, I.B., Liao, X., Stothard, P., Moore, S., Plastow, G., Miller, S.P., 2014. Single nucleotide polymorphisms for feed efficiency and performance in crossbred beef cattle. Bmc Genet 15, 14.
2. Affonso, E., Polez, V., Corrêa, C., Mazon, A., Araujo, M., Moraes, G., Rantin, F., 2002. Blood parameters and metabolites in the teleost fish *Colossoma macropomum* exposed to sulfide or hypoxia. Comp Biochem Physiol C Toxicol Pharmacol 133, 375-382.
3. Annilo, T., Chen, Z.-Q., Shulenin, S., Costantino, J., Thomas, L., Lou, H., Stefanov, S., Dean, M., 2006. Evolution of the vertebrate ABC gene family: analysis of gene birth and death. Genomics 88, 1-11.
4. Anttila, K., Dhillon, R.S., Boulding, E.G., Farrell, A.P., Glebe, B.D., Elliott, J.A., Wolters, W.R., Schulte, P.M., 2013. Variation in temperature tolerance among families of Atlantic salmon (*Salmo salar*) is associated with hypoxia tolerance, ventricle size and myoglobin level. J Exp Biol 216, 1183-1190.
5. Arus, P., Moreno-González, J., 1993. Marker-assisted selection, Plant Breeding. Springer, pp. 314-331.
6. Aslam, M.L., Bastiaansen, J.W., Elferink, M.G., Megens, H.J., Crooijmans, R.P., Blomberg le, A., Fleischer, R.C., Van Tassell, C.P., Sonstegard, T.S., Schroeder, S.G., Groenen, M.A., Long, J.A., 2012. Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). BMC genomics 13, 391.
7. Bansal, V., Harismendy, O., Tewhey, R., Murray, S.S., Schork, N.J., Topol, E.J., Frazer, K.A., 2010. Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome research 20, 537-545.
8. Boitard, S., Rocha, D., 2013. Detection of signatures of selective sweeps in the Blonde d'Aquitaine cattle breed. Animal genetics.
9. Bradley, K.M., Elmore, J.B., Breyer, J.P., Yaspan, B.L., Jessen, J.R., Knapik, E.W., Smith, J.R., 2007. A major zebrafish polymorphism resource for genetic mapping. Genome biology 8, R55.
10. Bruderer, R. M., et al. (2004). "The AAA ATPase p97/VCP interacts with its alternative co-factors, Ufd1-Npl4 and p47, through a common bipartite binding mechanism." Journal of Biological Chemistry 279(48): 49609-49616.
11. Cano, J., Matsuba, C., Mäkinen, H., Merilä, J., 2006. The utility of QTL-Linked markers to detect selective sweeps in natural populations—a case study of the EDA gene and a linked marker in threespine stickleback. Mol Ecol 15, 4613-4621.
12. Chen, S.-L., Tian, Y.-S., Yang, J.-F., Shao, C.-W., Ji, X.-S., Zhai, J.-M., Liao, X.-L., Zhuang, Z.-M., Su, P.-Z., Xu, J.-Y., 2009. Artificial gynogenesis and sex determination in half-smooth tongue sole (Cynoglossus semilaevis). Marine biotechnology 11, 243-251.
13. Clabecq, A., Henry, J.-P., Darchen, F., 2000. Biochemical characterization of Rab3-GTPase-activating protein reveals a mechanism similar to that of Ras-GAP. Journal of Biological Chemistry 275, 31786-31791.

14. Dean, M., et al. (2001). "The human ATP-binding cassette (ABC) transporter superfamily." Journal of lipid research 42(7): 1007-1017.

15. Decottignies, A., et al. (2004). "Binding of Cdc48p to a ubiquitin‐related UBX domain from novel yeast proteins involved in intracellular proteolysis and sporulation." Yeast 21(2): 127-139.

16. Dekkers, J., Dentine, M., 1991. Quantitative genetic variance associated with chromosomal markers in segregating populations. Theoretical and applied genetics 81, 212-220.

17. Diller, K.C., Gilbert, W.A., Kocher, T.D., 2002. Selective sweeps in the human genome: a starting point for identifying genetic differences between modern humans and chimpanzees. Molecular Biology and Evolution 19, 2342-2345.

18. Dunham, R., Smitherman, R., 1983. Crossbreeding channel catfish for improvement of body weight in earthen ponds. Growth 47, 97-103.

19. Dunham, R.A., Brady, Y., Vinitnantharat, S., 1994. Response to challenge with *Edwardsiella ictaluri* by channel catfish, *Ictalurus punctatus*, selected for resistance to *E. ictaluri*. J Appl Aquaculture 3, 211-222.

20. Dunham, R.A., Smitherman, R.O., 1984. Ancestry and breeding of catfish in the United States. Alabama Agri. Exp. Station Circular, Auburn.

21. Etter, P.D., Bassham, S., Hohenlohe, P.A., Johnson, E.A., Cresko, W.A., 2011. SNP discovery and genotyping for evolutionary genetics using RAD sequencing, Molecular methods for evolutionary genetics. Springer, pp. 157-178.

22. Faust, H.A., Gamperl, A.K., Rodnick, K.J., 2004. All rainbow trout (*Oncorhynchus mykiss*) are not created equal: intra-specific variation in cardiac hypoxia tolerance. J Exp Biol 207, 1005-1015.

23. Fernández, M.E., Goszczynski, D.E., Lirón, J.P., Villegas-Castagnasso, E.E., Carino, M.H., Ripoli, M.V., Rogberg-Muñoz, A., Posik, D.M., Peral-García, P., Giovambattista, G., 2013. Comparison of the effectiveness of microsatellites and SNP panels for genetic identification, traceability and assessment of parentage in an inbred Angus herd. Genetics and molecular biology 36, 185-191.

24. Fukui, K., Sasaki, T., Imazumi, K., Matsuura, Y., Nakanishi, H., Takai, Y., 1997. Isolation and characterization of a GTPase activating protein specific for the Rab3 subfamily of small G proteins. Journal of Biological Chemistry 272, 4655-4658.

25. Gabriel, S., Ziaugra, L., Tabbaa, D., 2009. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.] Chapter 2, Unit 2 12.

26. Geng, X., Feng, J., Liu, S., Wang, Y., Arias, C., Liu, Z., 2014. Transcriptional regulation of hypoxia inducible factors alpha (HIF-α) and their inhibiting factor (FIH-1) of channel catfish (*Ictalurus punctatus*) under hypoxia. Comp Biochem Physiol B Biochem Mol Biol 228, 91-105.

27. Gibbs, R.A., Taylor, J.F., Van Tassell, C.P., Barendse, W., Eversoie, K.A., Gill, C.A., Green, R.D., Hamernik, D.L., Kappes, S.M., Lien, S., Matukumalli, L.K., McEwan, J.C., Nazareth, L.V., Schnabel, R.D., Taylor, J.F., Weinstock, G.M., Wheeler, D.A.,

Ajmone-Marsan, P., Barendse, W., Boettcher, P.J., Caetano, A.R., Garcia, J.F., Hanotte, O., Mariani, P., Skow, L.C., Williams, J.L., Caetano, A.R., Diallo, B., Green, R.D., Hailemariam, L., Hanotte, O., Martinez, M.L., Morris, C.A., Silva, L.O.C., Spelman, R.J., Taylor, J.F., Mulatu, W., Zhao, K.Y., Abbey, C.A., Agaba, M., Araujo, F.R., Bunch, R.J., Burton, J., Gill, C.A., Gorni, C., Olivier, H., Harrison, B.E., Luff, B., Machado, M.A., Mariani, P., Morris, C.A., Mwakaya, J., Plastow, G., Sim, W., Skow, L.C., Smith, T., Sonstegard, T.S., Spelman, R.J., Taylor, J.F., Thomas, M.B., Valentini, A., Williams, P., Womack, J., Wooliams, J.A., Liu, Y., Qin, X., Worley, K.C., Gao, C., Gill, C.A., Jiang, H.Y., Liu, Y., Moore, S.S., Nazareth, L.V., Ren, Y.R., Song, X.Z., Bustamante, C.D., Hernandez, R.D., Muzny, D.M., Nazareth, L.V., Patil, S., Lucas, A.S., Fu, Q., Kent, M.P., Moore, S.S., Vega, R., Abbey, C.A., Gao, C., Gill, C.A., Green, R.D., Matukumalli, L.K., Matukumalli, A., McWilliam, S., Schnabel, R.D., Sclep, G., Ajmone-Marsan, P., Bryc, K., Bustamante, C.D., Choi, J., Gao, H., Grefenstette, J.J., Murdoch, B., Stella, A., Villa-Angulo, R., Wright, M., Aerts, J., Jann, O., Negrini, R., Sonstegard, T.S., Williams, J.L., Taylor, J.F., Villa-Angulo, R., Goddard, M.E., Hayes, B.J., Barendse, W., Bradley, D.G., Boettcher, P.J., Bustamante, C.D., da Silva, M.B., Lau, L.P.L., Liu, G.E., Lynn, D.J., Panzitta, F., Sclep, G., Wright, M., Dodds, K.G., 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. Science 324, 528-532.

28. Glessner, J.T., Bradfield, J.P., Wang, K., Takahashi, N., Zhang, H., Sleiman, P.M., Mentch, F.D., Kim, C.E., Hou, C., Thomas, K.A., 2010. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. Am J Hum Genet 87, 661-666.

29. Goddard, M.E., Hayes, B.J., 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381-391.

30. Gommeaux, J., Grégoire, C., Nguessan, P., Richelme, M., Malissen, M., Guerder, S., Malissen, B., Carrier, A., 2009. Thymus‐specific serine protease regulates positive selection of a subset of CD4+ thymocytes. European journal of immunology 39, 956-964.

31. Grimm, M., Krimmel, M., Polligkeit, J., Alexander, D., Munz, A., Kluba, S., Keutel, C., Hoffmann, J., Reinert, S., Hoefert, S., 2012. ABCB5 expression and cancer stem cell hypothesis in oral squamous cell carcinoma. Eur J Cancer 48, 3186-3197.

32. Groenen, M.A., Megens, H.J., Zare, Y., Warren, W.C., Hillier, L.W., Crooijmans, R.P., Vereijken, A., Okimoto, R., Muir, W.M., Cheng, H.H., 2011. The development and characterization of a 60K SNP chip for chicken. BMC genomics 12, 274.

33. Guan, B., Ma, H., Wang, Y., Hu, Y., Lin, Z., Zhu, Z., Hu, W., 2011. Vitreoscilla hemoglobin (VHb) overexpression increases hypoxia tolerance in zebrafish (*Danio rerio*). Mar Biotechnol 13, 336-344.

34. Guerriero, G., Di Finizio, A., Ciarcia, G., 2002. Stress-induced changes of plasma antioxidants in aquacultured sea bass, *Dicentrarchus labrax*. Comp Biochem Physiol A Mol Integr Physiol 132, 205-211.

35. Gunbin, K., Ruvinsky, A., 2013. Evolution of general transcription factors. Journal of molecular evolution, 1-20.

36. Guryev, V., Koudijs, M.J., Berezikov, E., Johnson, S.L., Plasterk, R.H., van Eeden, F.J., Cuppen, E., 2006. Genetic variation in the zebrafish. Genome research 16, 491-497.

37. Handley, M. T., et al. (2013). "Mutation Spectrum in RAB3GAP1, RAB3GAP2, and RAB18 and Genotype–Phenotype Correlations in Warburg Micro Syndrome and Martsolf Syndrome." Human mutation 34(5): 686-696.

38. Hartmann-Petersen, R., et al. (2004). "The Ubx2 and Ubx3 cofactors direct Cdc48 activity to proteolytic and nonproteolytic ubiquitin-dependent processes." Current biology 14(9): 824-828.

39. Hayes, B., Goddard, M., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.

40. Helyar, S.J., Limborg, M.T., Bekkevold, D., Babbucci, M., van Houdt, J., Maes, G.E., Bargelloni, L., Nielsen, R.O., Taylor, M.I., Ogden, R., Cariani, A., Carvalho, G.R., Panitz, F., Consortium, F., 2012. SNP Discovery Using Next Generation Transcriptomic Sequencing in Atlantic Herring (Clupea harengus). PloS one 7, e42089.

41. Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., Sella, G., Przeworski, M., 2011. Classic selective sweeps were rare in recent human evolution. Science 331, 920-924.

42. Herraez, D.L., Schafer, H., Mosner, J., Fries, H.-R., Wink, M., 2005. Comparison of microsatellite and single nucleotide polymorphism markers for the genetic analysis of a Galloway cattle population. Zeitschrift fur Naturforschung C-Journal of Biosciences 60, 637-643.

43. Hirschhorn, J.N., Daly, M.J., 2005. Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics 6, 95-108.

44. Hitt, R. and D. H. Wolf (2004). "Der1p, a protein required for degradation of malfolded soluble proteins of the endoplasmic reticulum: topology and Der1 ‐ like proteins." FEMS yeast research 4(7): 721-729.

45. Hoffmann, T.J., Kvale, M.N., Hesselson, S.E., Zhan, Y.P., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J., Ewing, M., Gollub, J., Henderson, M., Hubbell, E., Iribarren, C., Kaufman, J., Lao, R.Z., Lu, Y.T., Ludwig, D., Mathauda, G.K., McGuire, W., Mei, G.W., Miles, S., Purdy, M.M., Quesenberry, C., Ranatunga, D., Rowell, S., Sadler, M., Shapero, M.H., Shen, L., Shenoy, T.R., Smethurst, D., Van den Eeden, S.K., Walter, L., Wan, E., Wearley, R., Webster, T., Wen, C.C., Weng, L., Whitmer, R.A., Williams, A., Wong, S.C., Zau, C., Finn, A., Schaefer, C., Kwok, P.Y., Risch, N., 2011. Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. Genomics 98, 79-89.

46. Hofmann, K. and P. Bucher (1996). "The UBA domain: a sequence motif present in multiple enzyme classes of the ubiquitination pathway." Trends in biochemical sciences 21(5): 172-173.

47. Hohenlohe, P.A., Bassham, S., Etter, P.D., Stiffler, N., Johnson, E.A., Cresko, W.A., 2010. Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. PLoS Genet 6, e1000862.

48. Houston, R.D., Haley, C.S., Hamilton, A., Guy, D.R., Tinch, A.E., Taggart, J.B., McAndrew, B.J., Bishop, S.C., 2008. Major quantitative trait loci affect resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar). Genetics 178, 1109-1115.

49. Hubert, S., Higgins, B., Borza, T., Bowman, S., 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (Gadus morhua). BMC genomics 11.

50. Ihle, S., Ravaoarimanana, I., Thomas, M., Tautz, D., 2006. An analysis of signatures of selective sweeps in natural populations of the house mouse. Molecular Biology and Evolution 23, 790-797.

51. Jiang, Y., Gao, X., Liu, S., Zhang, Y., Liu, H., Sun, F., Bao, L., Waldbieser, G., Liu, Z., 2013. Whole genome comparative analysis of channel catfish (*Ictalurus punctatus*) with four model fish species. BMC Genomics 14, 780.

52. Johansson, A.M., Pettersson, M.E., Siegel, P.B., Carlborg, Ö., 2010. Genome-wide effects of long-term divergent selection. PLoS genetics 6, e1001188.

53. Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., Jindo, T., Kobayashi, D., Shimada, A., Toyoda, A., Kuroki, Y., Fujiyama, A., Sasaki, T., Shimizu, A., Asakawa, S., Shimizu, N., Hashimoto, S.I., Yang, J., Lee, Y., Matsushima, K., Sugano, S., Sakaizumi, M., Narita, T., Ohishi, K., Haga, S., Ohta, F., Nomoto, H., Nogata, K., Morishita, T., Endo, T., Shin-I, T., Takeda, H., Morishita, S., Kohara, Y., 2007. The medaka draft genome and insights into vertebrate genome evolution. Nature 447, 714-719.

54. Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., Furlotte, N.A., Eskin, E., Nellaker, C., Whitley, H., Cleak, J., Janowitz, D., Hernandez-Pliego, P., Edwards, A., Belgard, T.G., Oliver, P.L., McIntyre, R.E., Bhomra, A., Nicod, J., Gan, X., Yuan, W., van der Weyden, L., Steward, C.A., Bala, S., Stalker, J., Mott, R., Durbin, R., Jackson, I.J., Czechanski, A., Guerra-Assuncao, J.A., Donahue, L.R., Reinholdt, L.G., Payseur, B.A., Ponting, C.P., Birney, E., Flint, J., Adams, D.J., 2011. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature 477, 289-294.

55. Kerstens, H.H., Crooijmans, R.P., Veenendaal, A., Dibbits, B.W., Chin, A.W.T.F., den Dunnen, J.T., Groenen, M.A., 2009. Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC genomics 10, 479.

56. Kidd, K.K., Pakstis, A.J., Speed, W.C., Kidd, J.R., 2004. Understanding human DNA sequence variation. The Journal of heredity 95, 406-420.

57. Kijas, J.W., Townley, D., Dalrymple, B.P., Heaton, M.P., Maddox, J.F., McGrath, A., Wilson, P., Ingersoll, R.G., McCulloch, R., McWilliam, S., Tang, D., McEwan, J., Cockett, N., Oddy, V.H., Nicholas, F.W., Raadsma, H., International Sheep Genomics, C., 2009. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PloS one 4, e4668.

58. Kim, E.S., Cole, J.B., Huson, H., Wiggans, G.R., Tassell, C.P.V., Crooker, B.A., Liu, G., Da, Y., Sonstegard, T., 2013. Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. PLoS One 8, e80813.

59. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L., 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25, 2283-2285.

60. Kofler, R., Pandey, R.V., Schlotterer, C., 2011. PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27, 3435-3436.

61. Kranis, A., Gheyas, A.A., Boschiero, C., Turner, F., Yu, L., Smith, S., Talbot, R., Pirani, A., Brew, F., Kaiser, P., Hocking, P.M., Fife, M., Salmon, N., Fulton, J., Strom, T.M., Haberer, G., Weigend, S., Preisinger, R., Gholami, M., Qanbari, S., Simianer, H., Watson, K.A., Woolliams, J.A., Burt, D.W., 2013. Development of a high density 600K SNP genotyping array for chicken. BMC genomics 14, 59.

62. Krauss, G.J., Solé, M., Krauss, G., Schlosser, D., Wesenberg, D., Bärlocher, F., 2011. Fungi in freshwaters: ecology, physiology and biochemical potential. FEMS microbiology reviews 35, 620-651.

63. Labbe, R., et al. (2011). "Genetic analysis of the xenobiotic resistance‐associated ABC gene subfamilies of the Lepidoptera." Insect molecular biology 20(2): 243-256.

64. Lande, R., Thompson, R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124, 743-756.

65. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R.,

Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., International Human Genome Sequencing, C., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

66. Le Bras, Y., Dechamp, N., Krieg, F., Filangi, O., Guyomard, R., Boussaha, M., Bovenhuis, H., Pottinger, T.G., Prunet, P., Le Roy, P., 2011. Detection of QTL with effects on osmoregulation capacities in the rainbow trout (Oncorhynchus mykiss). Bmc Genet 12, 46.

67. Leprohon, P., et al. (2006). "Modulation of Leishmania ABC protein gene expression through life stages and among drug-resistant parasites." Eukaryotic cell 5(10): 1713-1725.

68. Li, C., Zhang, Y., Wang, R., Lu, J., Nandi, S., Mohanty, S., Terhune, J., Liu, Z., Peatman, E., 2012. RNA-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following *Edwardsiella ictaluri* infection in channel catfish, *Ictalurus punctatus*. Fish & shellfish immunology 32, 816-827.

69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009a. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078-2079.

70. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., 2009b. The sequence and de novo assembly of the giant panda genome. Nature 463, 311-317.

71. Liang, J., et al. (2006). "Characterization of erasin (UBXD2): a new ER protein that promotes ER-associated protein degradation." Journal of cell science 119(19): 4011-4024.

72. Liang, J., Yin, C., Doong, H., Fang, S., Peterhoff, C., Nixon, R.A., Monteiro, M.J., 2006. Characterization of erasin (UBXD2): a new ER protein that promotes ER-associated protein degradation. Journal of cell science 119, 4011-4024.

73. Lin, H.-J., et al. (2013). "TMEM217, TMEM63B, GRIK2 genes on chromosome 6 are associated with diabetic retinopathy."

74. Lin, J.Y., Zhang, M.F., Schatton, T., Wilson, B.J., Alloo, A., Ma, J., Qureshi, A.A., Frank, N.Y., Han, J.L., Frank, M.H., 2013. Genetically determined ABCB5 functionality correlates with pigmentation phenotype and melanoma risk. Biochem Bioph Res Co 436, 536-542.

75. Liu, S., et al. (2013). "Genome-wide identification, characterization and phylogenetic analysis of 50 catfish ATP-binding cassette (ABC) transporter genes." PLoS One 8(5): e63895.

76. Liu, S., Li, Q., Liu, Z., 2013. Genome-Wide Identification, Characterization and Phylogenetic Analysis of 50 Catfish ATP-Binding Cassette (ABC) Transporter Genes. PloS one 8, e63895.

77. Liu, S., Sun, L., Li, Y., Sun, F., Jiang, Y., Zhang, Y., Zhang, J., Feng, J., Kaltenboeck, L., Kucuktas, H., Liu, Z., 2014. Development of the catfish 250K SNP array for genome-wide association studies. BMC Res Notes 7, 135.

78. Liu, S.K., Zhou, Z.C., Lu, J.G., Sun, F.Y., Wang, S.L., Liu, H., Jiang, Y.L., Kucuktas, H., Kaltenboeck, L., Peatman, E., Liu, Z.J., 2011. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics 12, 53.

79. Ma, A.-H., et al. (2006). "Male Germ Cell–Associated Kinase, a Male-Specific Kinase Regulated by Androgen, Is a Coactivator of Androgen Receptor in Prostate Cancer Cells." Cancer research 66(17): 8439-8447.

80. Makeyev, A.V., Erdenechimeg, L., Mungunsukh, O., Roth, J.J., Enkhmandakh, B., Ruddle, F.H., Bayarsaihan, D., 2004. GTF2IRD2 is located in the Williams–Beuren syndrome critical region 7q11. 23 and encodes a protein with two TFII-I-like helix–loop–helix repeats. Proceedings of the National Academy of Sciences of the United States of America 101, 11052-11057.

81. Mäkinen, H.S., Shikano, T., Cano, J.M., Merilä, J., 2008. Hitchhiking mapping reveals a candidate genomic region for natural selection in three-spined stickleback chromosome VIII. Genetics 178, 453-465.

82. Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. Trends in genetics : TIG 24, 133-141.

83. Marklund, S., Carlborg, O., 2010. SNP detection and prediction of variability between chicken lines using genome resequencing of DNA pools. BMC genomics 11, 665.

84. Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S., Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS one 4, e5350.

85. McCarron, M., et al. (1994). "Drosophila P element transposase induces male recombination additively and without a requirement for P element excision or insertion." Genetics 136(3): 1013-1023.

86. Meierjohann, S., et al. (2004). "Genetic, biochemical and evolutionary facets of Xmrk-induced melanoma formation in the fish< i> Xiphophorus</i>." Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 138(3): 281-289.

87. Messer, P.W., Petrov, D.A., 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol 28, 659-669.

88. Mickett, K., Morton, C., Feng, J., Li, P., Simmons, M., Cao, D., Dunham, R., Liu, Z., 2003. Assessing genetic diversity of domestic populations of channel catfish (*Ictalurus punctatus*) in Alabama using AFLP markers. Aquaculture 228, 91-105.

89. Moen, T., Baranski, M., Sonesson, A.K., Kjøglum, S., 2009. Confirmation and fine-mapping of a major QTL for resistance to infectious pancreatic necrosis in Atlantic salmon (Salmo salar): population-level associations between markers and trait. BMC genomics 10, 368.

90. Ninwichian, P., Peatman, E., Liu, H., Kucuktas, H., Somridhivej, B., Liu, S., Li, P., Jiang, Y., Sha, Z., Kaltenboeck, L., 2012. Second-generation genetic linkage map of catfish and its integration with the BAC-based physical map. G3 (Bethesda) 2, 1233-1241.

91. Ohazama, A. and P. T. Sharpe (2007). "TFII‐I gene family during tooth development: Candidate genes for tooth anomalies in Williams syndrome." Developmental Dynamics 236(10): 2884-2888.

92. Oliphant, A., Barker, D.L., Stuelpnagel, J.R., Chee, M.S., 2002. BeadArray technology: Enabling an accurate, cost-effective approach to high throughput genotyping. Biotechniques 32, 56-58.

93. Palaisa, K., Morgante, M., Tingey, S., Rafalski, A., 2004. Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proceedings of the National Academy of Sciences of the United States of America 101, 9885-9890.

94. Palmer, S. J., et al. (2012). "GTF2IRD2 from the Williams–Beuren critical region encodes a mobile-element-derived fusion protein that antagonizes the action of its related family members." Journal of cell science 125(21): 5040-5050.

95. Porter, M. A., et al. (2012). "A role for transcription factor GTF2IRD2 in executive function in Williams-Beuren syndrome." PLoS One 7(10): e47457.

96. Purdie, A.C., Plain, K.M., Begg, D.J., De Silva, K., Whittington, R.J., 2011. Candidate gene and genome-wide association studies of< i> Mycobacterium avium</i> subsp.< i> paratuberculosis</i> infection in cattle and sheep: A review. Comparative immunology, microbiology and infectious diseases 34, 197-208.

97. Ramey, H.R., Decker, J.E., McKay, S.D., Rolf, M.M., Schnabel, R.D., Taylor, J.F., 2013. Detection of selective sweeps in cattle using genome-wide SNP data. BMC genomics 14, 382.

98. Ramos, A.M., Crooijmans, R.P.M.A., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P., Hansen, M.S., Hedegaard, J., Hu, Z.L., Kerstens, H.H., Law, A.S., Megens, H.J., Milan, D., Nonneman, D.J., Rohrer, G.A., Rothschild, M.F., Smith, T.P.L., Schnabel, R.D., Van Tassell, C.P., Taylor, J.F., Wiedmann, R.T., Schook, L.B., Groenen, M.A.M., 2009. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PloS one 4, e6524.

99. RAQUIN, A.L., Brabant, P., Rhoné, B., Balfourier, F., Leroy, P., Goldringer, I., 2008. Soft selective sweep near a gene that increases plant height in wheat. Molecular ecology 17, 741-756.

100. Raven, L.-A., Cocks, B.G., Hayes, B.J., 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC genomics 15, 62.

101. Rexroad, C.E., Vallejo, R.L., Liu, S., Palti, Y., Weber, G.M., 2012. QTL affecting stress response to crowding in a rainbow trout broodstock population. Bmc Genet 13, 97.

102.Rubin, C.J., Megens, H.J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, O., Jern, P., Jorgensen, C.B., Archibald, A.L., Fredholm, M., Groenen, M.A., Andersson, L., 2012a. Strong signatures of selection in the domestic pig genome. Proceedings of the National Academy of Sciences of the United States of America 109, 19529-19536.

103.Rubin, C.J., Megens, H.J., Martinez Barrio, A., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, O., Jern, P., Jorgensen, C.B., Archibald, A.L., Fredholm, M., Groenen, M.A., Andersson, L., 2012b. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci U S A 109, 19529-19536.

104.Rubin, C.J., Zody, M.C., Eriksson, J., Meadows, J.R.S., Sherwood, E., Webster, M.T., Jiang, L., Ingman, M., Sharpe, T., Ka, S., Hallbook, F., Besnier, F., Carlborg, O., Bed'hom, B., Tixier-Boichard, M., Jensen, P., Siegel, P., Lindblad-Toh, K., Andersson, L., 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464, 587-591.

105.Sakane, A., et al. (2006). "Rab3 GTPase-activating protein regulates synaptic transmission and plasticity through the inactivation of Rab3." Proceedings of the National Academy of Sciences 103(26): 10029-10034.

106.Sakane, A., Manabe, S., Ishizaki, H., Tanaka-Okamoto, M., Kiyokage, E., Toida, K., Yoshida, T., Miyoshi, J., Kamiya, H., Takai, Y., 2006. Rab3 GTPase-activating protein regulates synaptic transmission and plasticity through the inactivation of Rab3. Proceedings of the National Academy of Sciences 103, 10029-10034.

107.Schuberth, C. and A. Buchberger (2008). "UBX domain proteins: major regulators of the AAA ATPase Cdc48/p97." Cellular and molecular life sciences 65(15): 2360-2371.

108.Schuberth, C., Buchberger, A., 2008. UBX domain proteins: major regulators of the AAA ATPase Cdc48/p97. Cellular and molecular life sciences 65, 2360-2371.

109.Semenza, G.L., 2003. Targeting HIF-1 for cancer therapy. Nat Rev Cancer 3, 721-732.

110.Shinkai, K., Mohrs, M., Locksley, R.M., 2002. Helper T cells regulate type-2 innate immunity in vivo. Nature 420, 825-829.

111.Simmons, M., Mickett, K., Kucuktas, H., Li, P., Dunham, R., Liu, Z., 2006. Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact. Aquaculture 252, 133-146.

112.Small, K.S., Brudno, M., Hill, M.M., Sidow, A., 2007. Extreme genomic variation in a natural population. Proceedings of the National Academy of Sciences 104, 5698-5703.

113.Smith, C.T., Elfstrom, C.M., Seeb, L.W., Seeb, J.E., 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol Ecol 14, 4193-4203.

114.Smith, J.M., Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genet Res 23, 23-35.

115.Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X.P., Moore, S.S., 2011. Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. BMC genomics 12, 559.

116.Sun, F., Peatman, E., Li, C., Liu, S., Jiang, Y., Zhou, Z., Liu, Z., 2012. Transcriptomic signatures of attachment, NF-κB suppression and IFN stimulation in the catfish gill following columnaris bacterial infection. Developmental & Comparative Immunology 38, 169-180.

117.Swanson, H.I., 2002. DNA binding and protein interactions of the AHR/ARNT heterodimer that facilitate gene activation. Chem Biol Interact 141, 63-76.

118.Szakács, G., Paterson, J.K., Ludwig, J.A., Booth-Genthe, C., Gottesman, M.M., 2006. Targeting multidrug resistance in cancer. Nat Rev Drug Discov 5, 219-234.

119.Teschke, M., Mukabayire, O., Wiehe, T., Tautz, D., 2008. Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. Genetics 180, 1537-1545.

120.Tipney, H. J., et al. (2004). "Isolation and characterisation of GTF2IRD2, a novel fusion gene and member of the TFII-I family of transcription factors, deleted in Williams–Beuren syndrome." European journal of human genetics 12(7): 551-560.

121.Vasemägi, A., Nilsson, J., McGinnity, P., Cross, T., O'Reilly, P., Glebe, B., Peng, B., Berg, P.R., Primmer, C.R., 2012. Screen for footprints of selection during domestication/captive breeding of Atlantic salmon. Comp Funct Genom 2012, 628204.

122.Waldbieser, G., Bosworth, B., 1997. Cloning and characterization of microsatellite loci in channel catfish, Ictalurus punctatus. Animal genetics 28, 295-298.

123.Waldbieser, G.C., Wolters, W.R., 2007. Definition of the USDA103 strain of channel catfish (Ictalurus punctatus). Anim Genet 38, 180-183.

124.Wang, L.-Y. and H.-J. Kung (2011). "Male germ cell-associated kinase is overexpressed in prostate cancer cells and causes mitotic defects via deregulation of APC/CCDH1." Oncogene 31(24): 2907-2918.

125.Wang, R., Sun, L., Bao, L., Zhang, J., Jiang, Y., Yao, J., Song, L., Feng, J., Liu, S., Liu, Z., 2013. Bulk segregant RNA-seq reveals expression and positional candidate genes and allele-specific expression for disease resistance against enteric septicemia of catfish. BMC genomics 14, 929.

126.Wang, S., Sha, Z., Sonstegard, T.S., Liu, H., Xu, P., Somridhivej, B., Peatman, E., Kucuktas, H., Liu, Z., 2008. Quality assessment parameters for EST-derived SNPs from catfish. BMC Genomics 9, 450.

127.Warburg, M., et al. (1993). "Autosomal Recesssive Microcephaly, Microcornea, Congenital Cataract, Mental Retardation, Optic Atrophy, and Hypogenitalism: Micro Syndrome." Archives of Pediatrics & Adolescent Medicine 147(12): 1309.

128.Wiedmann, R.T., Smith, T.P.L., Nonneman, D.J., 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. Bmc Genet 9, 81.

129.Wilson, W.R., Hay, M.P., 2011. Targeting hypoxia in cancer therapy. Nat Rev Cancer 11, 393-410.

130.Wong, G.K., Liu, B., Wang, J., Zhang, Y., Yang, X., Zhang, Z., Meng, Q., Zhou, J., Li, D., Zhang, J., Ni, P., Li, S., Ran, L., Li, H., Zhang, J., Li, R., Li, S., Zheng, H., Lin, W., Li, G.,

Wang, X., Zhao, W., Li, J., Ye, C., Dai, M., Ruan, J., Zhou, Y., Li, Y., He, X., Zhang, Y., Wang, J., Huang, X., Tong, W., Chen, J., Ye, J., Chen, C., Wei, N., Li, G., Dong, L., Lan, F., Sun, Y., Zhang, Z., Yang, Z., Yu, Y., Huang, Y., He, D., Xi, Y., Wei, D., Qi, Q., Li, W., Shi, J., Wang, M., Xie, F., Wang, J., Zhang, X., Wang, P., Zhao, Y., Li, N., Yang, N., Dong, W., Hu, S., Zeng, C., Zheng, W., Hao, B., Hillier, L.W., Yang, S.P., Warren, W.C., Wilson, R.K., Brandstrom, M., Ellegren, H., Crooijmans, R.P., van der Poel, J.J., Bovenhuis, H., Groenen, M.A., Ovcharenko, I., Gordon, L., Stubbs, L., Lucas, S., Glavina, T., Aerts, A., Kaiser, P., Rothwell, L., Young, J.R., Rogers, S., Walker, B.A., van Hateren, A., Kaufman, J., Bumstead, N., Lamont, S.J., Zhou, H., Hocking, P.M., Morrice, D., de Koning, D.J., Law, A., Bartley, N., Burt, D.W., Hunt, H., Cheng, H.H., Gunnarsson, U., Wahlberg, P., Andersson, L., Kindlund, E., Tammi, M.T., Andersson, B., Webber, C., Ponting, C.P., Overton, I.M., Boardman, P.E., Tang, H., Hubbard, S.J., Wilson, S.A., Yu, J., Wang, J., Yang, H., International Chicken Polymorphism Map, C., 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432, 717-722.

131. Wright, D., Nakamichi, R., Krause, J., Butlin, R.K., 2006. QTL analysis of behavioral and morphological differentiation between wild and laboratory zebrafish (*Danio rerio*). Behav Genet 36, 271-284.

132. Wu, X., Fang, M., Liu, L., Wang, S., Liu, J., Ding, X., Zhang, S., Zhang, Q., Zhang, Y., Qiao, L., 2013. Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. BMC genomics 14, 897.

133. Xia, L., et al. (2002). "Identification of human male germ cell-associated kinase, a kinase transcriptionally activated by androgen in prostate cancer cells." Journal of Biological Chemistry 277(38): 35422-35433.

134. Xu, J., Ji, P., Zhao, Z., Zhang, Y., Feng, J., Wang, J., Li, J., Zhang, X., Zhao, L., Liu, G., Xu, P., Sun, X., 2012. Genome-wide SNP discovery from transcriptome of four common carp strains. PloS one 7, e48140.

135. Zhan, B.J., Fadista, J., Thomsen, B., Hedegaard, J., Panitz, F., Bendixen, C., 2011. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. BMC genomics 12, 557.

Appendix Table 1 Genotype of SNPs that only heterozygous in wild populations and homozygous in domestic strains

Contig ID	Position	Contig Length	Wild Genotype	Domestic Genotype
contig1066305	626	1100	C/G	C
contig1195013	1408	1833	T/G	T
contig126172	2484	4428	C/T	C
contig1399105	179	2436	C/T	C
contig1399105	1927	2436	C/T	C
contig1399617	3039	6936	T/C	T
contig1399617	3819	6936	T/C	T
contig1399617	3820	6936	T/C	T
contig1433077	652	988	T/G	T
contig1439377	609	857	G/T	G
contig1504121	477	755	T/C	T
contig1504121	664	755	T/G	T
contig1504121	712	755	T/A	T
contig1513397	230	4905	G/C	G
contig1531335	705	801	A/G	A
contig1543499	517	743	G/A	G
contig1579874	14	2908	T/C	T
contig1618517	386	447	A/T	A
contig1636537	487	930	T/A	T
contig1703213	63	747	G/A	G
contig1721951	612	632	T/G	T
contig1731249	2493	4460	G/A	G
contig1765380	755	847	A/G	A
contig1886046	661	730	C/G	C
contig1951086	43	1541	G/A	G
contig2032394	814	928	T/C	T
contig2151939	8	809	T/A	T
contig2166825	1882	2088	A/T	A
contig220162	1607	4556	A/T	A
contig2205018	328	668	C/A	C
contig2228667	1195	1797	C/G	C
contig2301338	14	1237	T/A	T
contig2344297	633	970	G/T	G
contig2344297	830	970	G/C	G
contig2344297	905	970	T/C	T
contig2379502	431	1941	C/A	C
contig2379502	1094	1941	G/A	G
contig2437365	623	736	G/C	G
contig2531632	3342	3973	A/G	A
contig2637452	245	1140	A/T	A
contig2641966	586	620	T/C	T
contig2648672	124	977	T/C	T
contig2740858	24	408	C/G	C
contig279244	16	1385	G/C	G
contig2800813	21	1173	C/A	C

contig2875231	1520	4265	G/A	G
contig2875231	2299	4265	T/C	T
contig2947412	153	568	G/T	G
contig2987655	2610	2620	T/C	T
contig3049463	1126	1736	C/T	C
contig3049463	1255	1736	C/T	C
contig3056119	983	1158	A/G	A
contig3096814	2891	3094	T/C	T
contig3115236	1498	1770	C/A	C
contig3151415	639	2246	A/C	A
contig3537664	7455	7476	A/C	A
contig3538691	3917	4851	G/A	G
contig3539402	87	5096	G/T	G
contig3539402	492	5096	T/G	T
contig3539402	1078	5096	G/T	G
contig3539402	2994	5096	C/T	C
contig3541108	3146	3160	T/G	T
contig3541227	5739	7012	A/G	A
contig3541227	5858	7012	C/T	C
contig3541227	6309	7012	T/C	T
contig3542254	4253	6306	C/G	C
contig3542729	2078	4657	A/G	A
contig3542890	1736	2823	C/G	C
contig3542890	2260	2823	C/T	C
contig3543078	665	1310	A/T	A
contig3543405	1048	2465	G/A	G
contig3543824	49	1676	T/A	T
contig3543824	736	1676	T/C	T
contig3544371	99	914	C/A	C
contig3545465	8522	16648	C/T	C
contig3545465	9992	16648	C/T	C
contig3545465	13840	16648	T/A	T
contig3546159	2164	3502	T/A	T
contig3546196	2060	2396	C/T	C
contig3546329	1104	3342	C/T	C
contig3547572	3669	5173	G/A	G
contig3547613	64	12773	G/T	G
contig3547613	1197	12773	G/A	G
contig3547613	2415	12773	A/G	A
contig3547613	2447	12773	T/A	T
contig3547613	2449	12773	T/C	T
contig3547613	3064	12773	C/T	C
contig3547613	3340	12773	C/A	C
contig3547613	3356	12773	C/T	C
contig3547613	3627	12773	A/T	A
contig3547613	8328	12773	C/A	C
contig3547613	8401	12773	A/G	A
contig3547613	8602	12773	A/T	A
contig3547613	8603	12773	G/T	G
contig3547613	10090	12773	G/T	G

contig3547613	10772	12773	G/A	G
contig3547613	10844	12773	T/C	T
contig3547860	247	2954	G/A	G
contig3547860	269	2954	G/C	G
contig3547860	589	2954	A/C	A
contig3548265	4778	14391	G/T	G
contig3548905	304	1847	G/A	G
contig3549615	3735	3881	C/A	C
contig3550705	4005	9266	C/T	C
contig3550705	8639	9266	C/G	C
contig3550705	8813	9266	T/A	T
contig3551004	1589	4270	T/A	T
contig3551600	741	1407	G/A	G
contig3551761	338	4228	G/C	G
contig3551761	339	4228	G/T	G
contig3552204	1180	4312	A/G	A
contig3553100	2016	4181	A/T	A
contig3553100	2287	4181	G/T	G
contig3553100	3927	4181	G/A	G
contig3553692	30	497	G/A	G
contig3553978	576	2565	G/A	G
contig3554764	967	2764	G/A	G
contig3555327	5766	11210	A/C	A
contig3555746	3786	3800	T/A	T
contig3556346	87	516	G/A	G
contig3556346	92	516	C/T	C
contig3556373	6379	7271	G/A	G
contig3556665	1094	5695	C/T	C
contig3557245	846	11821	T/C	T
contig3557245	8970	11821	C/T	C
contig3557245	10063	11821	G/A	G
contig3557674	2942	5179	A/C	A
contig3557708	1018	4587	T/C	T
contig3557708	3753	4587	A/T	A
contig3560254	677	3634	C/T	C
contig3561560	537	543	C/T	C
contig3561623	151	2010	C/G	C
contig3562262	449	714	G/A	G
contig3563290	685	2564	A/G	A
contig3563767	296	520	C/A	C
contig3565612	2638	2730	C/A	C
contig3566125	9854	9888	G/A	G
contig3566806	12010	19552	A/G	A
contig3566806	15818	19552	T/C	T
contig3566806	16911	19552	G/A	G
contig3567167	23	1503	C/T	C
contig3567509	3602	3611	C/T	C
contig3568199	2537	3799	G/A	G
contig3568199	2552	3799	G/A	G
contig3568712	134	8547	G/T	G

contig3568712	4888	8547	A/C	A
contig3569712	4522	4909	G/A	G
contig3569726	217	428	G/A	G
contig3570485	52	1107	A/G	A
contig3570624	900	963	A/G	A
contig3571234	2325	3228	A/T	A
contig3572929	627	1892	T/G	T
contig3573475	5408	10573	A/T	A
contig3573597	5085	5912	C/T	C
contig3574432	3581	12255	T/C	T
contig3574432	3587	12255	G/T	G
contig3574432	5015	12255	A/C	A
contig3574432	5046	12255	T/A	T
contig3574432	5173	12255	G/A	G
contig3574586	3538	8269	A/T	A
contig3574586	6455	8269	C/G	C
contig3575640	1276	1603	C/G	C
contig3575837	1179	1697	C/T	C
contig3577336	56	863	C/T	C
contig3577336	88	863	C/T	C
contig3577926	4016	11135	G/T	G
contig3578464	5157	5211	C/A	C
contig3579265	2216	12126	T/G	T
contig3579555	804	2125	G/A	G
contig3579555	1951	2125	T/C	T
contig3579666	780	3747	A/T	A
contig3580135	1233	1284	T/A	T
contig3581105	183	720	C/A	C
contig3581408	512	835	A/C	A
contig3581577	2053	2059	T/A	T
contig3581931	3333	6274	C/A	C
contig3583709	5124	7662	C/A	C
contig3583709	5142	7662	T/C	T
contig3583709	5554	7662	T/A	T
contig3583709	5665	7662	T/A	T
contig3583709	5839	7662	C/A	C
contig3583709	6344	7662	G/A	G
contig3583709	6347	7662	G/T	G
contig3583709	6571	7662	G/A	G
contig3583709	6628	7662	G/C	G
contig3583709	6743	7662	C/T	C
contig3585736	1025	1136	G/A	G
contig3585739	73	700	C/G	C
contig3585739	116	700	A/T	A
contig3585739	148	700	A/T	A
contig3586409	1553	4591	A/T	A
contig3586844	319	1404	C/G	C
contig3587278	1525	2765	G/A	G
contig3587481	4192	5067	C/T	C
contig3587600	737	2179	C/A	C

contig3588717	436	740	A/G	A
contig3588824	3254	5516	A/G	A
contig3589033	4068	5727	C/A	C
contig3589106	2718	3354	A/G	A
contig3589253	1717	4478	A/T	A
contig3589618	4015	4030	T/A	T
contig3590525	321	806	C/T	C
contig3591085	982	1081	C/T	C
contig3591138	2089	5991	A/T	A
contig3594086	629	1831	G/C	G
contig3594109	7	3860	A/T	A
contig3594109	570	3860	A/G	A
contig3594109	741	3860	G/A	G
contig3594109	3530	3860	A/G	A
contig3595140	233	6329	A/T	A
contig3595477	2675	2890	G/A	G
contig3595823	1404	2452	A/T	A
contig3596770	435	1067	G/T	G
contig3597306	1379	1441	T/C	T
contig3597693	882	1450	T/C	T
contig3598079	582	3587	C/A	C
contig3598079	2954	3587	G/T	G
contig3598079	3006	3587	T/C	T
contig3598099	3051	3691	T/A	T
contig3598099	3053	3691	G/T	G
contig3598425	1541	2892	G/A	G
contig3600982	4639	5097	G/T	G
contig3601107	1754	1848	G/T	G
contig3601736	1354	5092	C/T	C
contig3601904	789	3482	A/G	A
contig3601997	396	766	G/A	G
contig3602486	4201	6026	A/G	A
contig3602946	1207	7606	T/G	T
contig3603372	822	4022	T/A	T
contig3603759	1932	2556	T/A	T
contig3604769	672	6648	T/C	T
contig3604985	784	1048	T/A	T
contig3605540	1090	2364	T/C	T
contig3605699	256	9897	G/T	G
contig3606644	1966	5138	A/G	A
contig3606644	5108	5138	G/A	G
contig3606926	4571	5649	G/A	G
contig3606926	5256	5649	A/G	A
contig3607764	1086	4341	G/A	G
contig3608431	543	921	C/A	C
contig3608703	349	3621	T/A	T
contig3608828	167	1379	G/A	G
contig3608828	1025	1379	T/G	T
contig3609990	1298	5970	A/T	A
contig3609990	5449	5970	C/T	C

contig3610134	964	11697	T/A	T
contig3610134	1189	11697	A/C	A
contig3610243	2143	7374	G/C	G
contig3611588	1055	1946	C/T	C
contig3611588	1058	1946	A/T	A
contig3611783	1789	2472	A/T	A
contig3612325	391	1486	C/T	C
contig3613854	535	1271	G/A	G
contig3615075	859	2662	C/T	C
contig3615075	2544	2662	C/T	C
contig3615380	432	3306	T/G	T
contig3615380	441	3306	T/C	T
contig3615380	3074	3306	A/T	A
contig3615380	3094	3306	G/A	G
contig3615525	346	1044	A/C	A
contig3616093	243	1714	G/A	G
contig3616437	8969	11001	C/T	C
contig3616437	9024	11001	G/A	G
contig3616614	1331	3187	C/T	C
contig3617085	186	2335	G/C	G
contig3617573	2169	3149	C/T	C
contig3618032	254	3237	G/T	G
contig3618551	44	9497	A/T	A
contig3619312	57	5364	T/A	T
contig3619752	967	2897	G/T	G
contig3620506	36	3671	A/T	A
contig3621720	105	834	G/A	G
contig3621913	3665	5999	G/T	G
contig3622079	383	1484	A/G	A
contig3622799	792	1926	G/A	G
contig3623205	642	3319	A/G	A
contig3623863	2710	3255	A/G	A
contig3624099	3326	5201	T/C	T
contig3624099	3904	5201	T/C	T
contig3624099	5149	5201	A/T	A
contig3624394	258	1229	C/T	C
contig3624394	391	1229	T/C	T
contig3625068	595	696	A/C	A
contig3626133	1607	1644	T/A	T
contig3628089	3841	5984	C/T	C
contig3628089	4077	5984	C/T	C
contig3628963	4496	9278	G/A	G
contig3629766	1822	2939	T/C	T
contig3629834	348	2199	C/A	C
contig3629834	405	2199	A/T	A
contig3629834	442	2199	T/A	T
contig3629834	546	2199	A/G	A
contig3629834	552	2199	C/A	C
contig3629834	575	2199	G/A	G
contig3629834	1513	2199	G/A	G

contig3629834	1595	2199	G/A	G
contig3629834	1625	2199	G/T	G
contig3630136	1694	1939	C/A	C
contig3630591	1769	7674	G/A	G
contig3630632	3727	4241	G/A	G
contig3631579	1478	2726	G/A	G
contig3632676	1438	2058	T/A	T
contig3632798	3263	4915	T/C	T
contig3632798	3297	4915	A/C	A
contig3632798	3662	4915	T/G	T
contig3632798	4305	4915	G/T	G
contig3632819	523	562	A/T	A
contig3632819	524	562	C/T	C
contig3632835	730	1234	A/T	A
contig3633489	30	1323	T/A	T
contig3633589	2220	2981	A/T	A
contig3634463	1918	1946	A/C	A
contig3634615	48	4786	T/C	T
contig3635446	6421	12253	C/T	C
contig3635774	772	6939	C/A	C
contig3635774	1056	6939	A/G	A
contig3635774	2398	6939	C/A	C
contig3635958	1189	8543	T/A	T
contig3636142	4373	6265	A/T	A
contig3636212	236	2968	C/T	C
contig3636701	293	4484	G/T	G
contig3636701	614	4484	T/C	T
contig3636701	1254	4484	C/T	C
contig3636701	2409	4484	G/T	G
contig3636701	3049	4484	C/T	C
contig3637141	923	1027	A/C	A
contig3637176	857	4226	A/C	A
contig3638260	1456	13062	T/A	T
contig3638260	5499	13062	G/T	G
contig3638468	235	1823	G/A	G
contig3638468	270	1823	C/A	C
contig3638468	306	1823	C/T	C
contig3638468	1600	1823	G/A	G
contig3638468	1659	1823	G/T	G
contig3639549	48	10549	A/G	A
contig3639626	1025	1759	T/C	T
contig3640180	1316	3047	A/T	A
contig3640188	30	3712	T/C	T
contig3640862	1556	4438	C/T	C
contig3640914	801	1194	A/T	A
contig3641358	1489	2650	C/T	C
contig3641358	1494	2650	G/A	G
contig3641579	1425	2283	C/T	C
contig3641933	666	1779	C/T	C
contig3641933	1142	1779	G/C	G

contig3641933	1215	1779	A/G	A
contig3641933	1292	1779	A/G	A
contig3642495	445	1915	G/T	G
contig3642798	98	4883	C/T	C
contig3643511	575	3021	G/C	G
contig3643511	1063	3021	A/G	A
contig3644413	1313	1949	C/T	C
contig3644413	1342	1949	C/T	C
contig3644563	468	3521	A/T	A
contig3644644	58	2363	G/C	G
contig3644703	585	606	A/T	A
contig3645833	186	512	G/T	G
contig3646271	1147	2430	T/C	T
contig3647955	7	8174	A/C	A
contig3648259	883	913	A/T	A
contig3648600	910	2436	C/T	C
contig3649096	770	8527	T/G	T
contig3649492	2001	2517	G/A	G
contig3649571	28	3326	A/C	A
contig3650094	2042	7232	A/G	A
contig3650094	2091	7232	C/T	C
contig3650094	3980	7232	A/C	A
contig3650094	4016	7232	A/T	A
contig3650094	5296	7232	A/C	A
contig3650274	5361	6215	T/G	T
contig3650555	446	2399	G/T	G
contig3650555	905	2399	A/G	A
contig3650598	862	2048	G/C	G
contig3650953	1808	7442	A/T	A
contig3651298	159	7141	C/A	C
contig3651445	473	5495	C/T	C
contig3651445	849	5495	C/T	C
contig3651445	3160	5495	A/C	A
contig3652150	1620	1667	G/T	G
contig3654109	1441	5938	A/C	A
contig3654855	1932	3751	G/T	G
contig3654855	3188	3751	A/G	A
contig3654855	3619	3751	C/T	C
contig3654855	3667	3751	C/T	C
contig3655013	66	1035	G/T	G
contig3655718	7546	8037	C/T	C
contig3655739	429	851	A/G	A
contig3656160	122	482	C/A	C
contig3656243	1103	6066	G/A	G
contig3656243	1283	6066	G/C	G
contig3656692	870	5393	C/T	C
contig3658370	1831	6282	A/T	A
contig3658790	55	1608	G/T	G
contig3659012	2053	2092	C/T	C
contig3659282	2241	5854	C/G	C

contig3659282	2245	5854	T/A	T
contig3659607	590	610	C/T	C
contig3660654	2137	7782	C/T	C
contig3660732	280	967	C/T	C
contig3660732	286	967	T/A	T
contig3661377	3626	4211	G/T	G
contig3661464	5409	7387	A/G	A
contig3661517	280	6487	A/T	A
contig3661517	906	6487	G/C	G
contig3661517	1179	6487	C/T	C
contig3661517	1227	6487	T/A	T
contig3661517	1780	6487	T/G	T
contig3661517	3197	6487	A/G	A
contig3661517	4465	6487	C/T	C
contig3661517	4929	6487	A/T	A
contig3662148	6497	9381	A/G	A
contig3662148	6544	9381	A/T	A
contig3662571	340	16649	G/T	G
contig3662571	457	16649	A/G	A
contig3662571	693	16649	T/G	T
contig3662571	1101	16649	C/A	C
contig3662571	1183	16649	T/C	T
contig3662571	1434	16649	A/G	A
contig3662571	1848	16649	A/C	A
contig3662571	1903	16649	T/A	T
contig3662571	1953	16649	T/C	T
contig3662571	2105	16649	C/T	C
contig3662571	2196	16649	C/G	C
contig3662571	5365	16649	C/T	C
contig3662571	6339	16649	G/A	G
contig3662571	6344	16649	C/G	C
contig3662571	7698	16649	G/A	G
contig3662571	12101	16649	A/G	A
contig3662571	13314	16649	C/T	C
contig3662571	15767	16649	T/C	T
contig3662571	16130	16649	C/T	C
contig3662571	16229	16649	C/T	C
contig3663977	3717	5541	T/A	T
contig3663977	4738	5541	T/A	T
contig3664194	5866	6104	G/A	G
contig3664734	290	2069	C/A	C
contig3664734	323	2069	G/A	G
contig3665212	856	1011	C/A	C
contig3665589	1797	2070	A/T	A
contig3665606	1643	6113	T/A	T
contig3666589	3116	3282	G/C	G
contig3666900	1157	5795	T/C	T
contig3666900	1991	5795	T/C	T
contig3666900	2401	5795	A/G	A
contig3666900	3911	5795	G/T	G

contig3667396	527	5819	C/T	C
contig3667783	3283	4814	T/A	T
contig3668023	933	2004	C/A	C
contig3668194	239	4087	T/C	T
contig3668194	346	4087	T/C	T
contig3668194	645	4087	T/C	T
contig3668194	1027	4087	T/A	T
contig3668194	1068	4087	G/C	G
contig3668339	262	2114	T/A	T
contig3668407	205	2104	C/T	C
contig3668547	4744	5713	T/A	T
contig3668588	1796	2075	G/A	G
contig3669118	11378	14547	C/A	C
contig3669224	120	2394	G/A	G
contig3669224	122	2394	C/T	C
contig3669224	124	2394	G/C	G
contig3669224	170	2394	G/A	G
contig3669224	182	2394	A/T	A
contig3669224	283	2394	C/A	C
contig3669224	306	2394	A/C	A
contig3669224	319	2394	T/G	T
contig3669224	320	2394	T/A	T
contig3669224	387	2394	G/A	G
contig3669224	817	2394	C/T	C
contig3669224	1220	2394	A/G	A
contig3669363	914	3107	T/C	T
contig3669363	2784	3107	T/C	T
contig3669463	864	2301	G/A	G
contig3669980	2302	3122	A/G	A
contig3670919	407	1376	G/C	G
contig3670919	480	1376	G/T	G
contig3671048	1533	6755	C/T	C
contig3671132	660	2674	T/A	T
contig3672109	401	687	G/T	G
contig3672353	3709	3773	C/T	C
contig3674615	873	1946	G/C	G
contig3674685	82	2461	G/A	G
contig3675884	3778	5182	G/A	G
contig3675884	4006	5182	G/A	G
contig3675884	4052	5182	T/G	T
contig3675884	4059	5182	G/A	G
contig3675884	4772	5182	T/C	T
contig3675884	4785	5182	A/G	A
contig3675884	4853	5182	T/A	T
contig3675884	4854	5182	C/T	C
contig3676191	4312	9598	G/A	G
contig3676262	374	2061	C/T	C
contig3676284	473	2190	G/A	G
contig3676284	921	2190	C/T	C
contig3677243	1320	4224	T/C	T

contig3677243	1358	4224	A/G	A
contig3677243	1368	4224	T/G	T
contig3677243	3551	4224	A/T	A
contig3677243	3555	4224	C/T	C
contig3677333	345	761	T/C	T
contig3677361	1906	2005	T/G	T
contig3677414	125	2377	C/T	C
contig3677520	23	1176	G/A	G
contig3677520	564	1176	T/C	T
contig3677520	630	1176	A/G	A
contig3677520	786	1176	C/T	C
contig3677520	841	1176	T/C	T
contig3677520	1062	1176	T/C	T
contig3677520	1075	1176	C/T	C
contig3678640	7541	9213	G/A	G
contig3678741	2066	6020	C/T	C
contig3678957	1079	4570	G/C	G
contig3680582	5032	7883	C/T	C
contig3680582	6876	7883	C/T	C
contig3681274	237	593	C/T	C
contig3681583	5750	6278	G/C	G
contig3682774	8616	8833	C/T	C
contig3683409	4307	4323	G/A	G
contig3684254	1728	5123	T/A	T
contig3684254	3583	5123	G/A	G
contig3684368	4826	4894	G/A	G
contig3684547	829	2352	G/A	G
contig3685782	1048	3020	T/C	T
contig3685878	715	3191	G/A	G
contig3685942	165	4576	C/T	C
contig3687718	360	2235	T/A	T
contig3687968	586	976	C/T	C
contig3688665	6245	11664	A/G	A
contig3689110	567	855	A/G	A
contig3689643	1908	3989	C/A	C
contig3690539	354	622	G/A	G
contig3691226	2931	3614	A/G	A
contig3692240	1601	3835	T/C	T
contig3692273	241	2003	G/A	G
contig3693026	3051	4944	C/A	C
contig3693069	154	5607	C/T	C
contig3693380	900	1395	T/G	T
contig3693380	901	1395	T/G	T
contig3693447	180	1220	T/A	T
contig3693447	221	1220	A/C	A
contig3693447	226	1220	T/G	T
contig3693447	666	1220	C/T	C
contig3693448	594	1058	C/A	C
contig3693475	820	1026	C/T	C
contig3693654	64	1172	C/T	C

contig3693669	33	605	C/T	C
contig3694733	1752	2200	T/C	T
contig3694896	1273	3171	G/A	G
contig3694896	1493	3171	G/A	G
contig3694896	1981	3171	A/G	A
contig3695257	135	406	T/A	T
contig3695301	919	1337	A/T	A
contig3695301	920	1337	C/T	C
contig3695973	4175	4198	C/T	C
contig3696013	84	2312	A/G	A
contig3696096	12199	12511	T/A	T
contig3697234	2400	7772	A/T	A
contig3697234	6743	7772	G/C	G
contig3697532	466	2351	A/T	A
contig3697635	1101	1108	G/T	G
contig3697793	1171	6248	A/T	A
contig3698576	802	2542	T/A	T
contig3698576	811	2542	C/T	C
contig3699751	1221	2861	C/T	C
contig3699777	900	6296	C/T	C
contig3699777	5446	6296	C/T	C
contig3699906	5212	10600	G/A	G
contig3700554	4501	12413	A/G	A
contig3700592	696	2051	T/G	T
contig3700996	3239	12544	G/A	G
contig3700996	8201	12544	C/T	C
contig3700996	11160	12544	T/C	T
contig3701014	563	1467	C/G	C
contig3701065	329	3266	G/C	G
contig3701065	434	3266	G/C	G
contig3701065	482	3266	C/A	C
contig3701065	709	3266	A/C	A
contig3701065	1318	3266	G/T	G
contig3701065	1970	3266	A/T	A
contig3701065	2404	3266	C/A	C
contig3701065	2418	3266	G/A	G
contig3701065	2687	3266	G/A	G
contig3701065	2896	3266	G/T	G
contig3701066	967	2644	A/T	A
contig3701066	1033	2644	G/C	G
contig3701066	1161	2644	A/G	A
contig3701066	1195	2644	G/A	G
contig3701066	1202	2644	A/C	A
contig3701066	1571	2644	T/C	T
contig3701066	1671	2644	G/C	G
contig3701066	2036	2644	A/T	A
contig3701066	2055	2644	G/C	G
contig3701066	2448	2644	G/C	G
contig3701066	2505	2644	T/A	T
contig3701067	591	5625	T/C	T

contig3701067	605	5625	T/A	T
contig3701067	681	5625	T/C	T
contig3701067	709	5625	A/T	A
contig3701067	876	5625	T/C	T
contig3701067	1756	5625	A/G	A
contig3701067	3183	5625	C/T	C
contig3701475	1371	2977	G/A	G
contig3701633	487	5367	T/C	T
contig3701633	3182	5367	G/T	G
contig3701633	3726	5367	C/T	C
contig3701633	4013	5367	A/G	A
contig3701633	4209	5367	G/A	G
contig3701633	4210	5367	G/A	G
contig3701633	4524	5367	C/G	C
contig3701633	5043	5367	A/G	A
contig3701633	5045	5367	T/A	T
contig3701894	442	981	C/T	C
contig3702906	5029	5990	G/T	G
contig3703072	80	2198	C/T	C
contig3703072	81	2198	A/G	A
contig3704041	1506	2613	A/G	A
contig3704799	68	14340	T/A	T
contig3704916	1229	4268	C/T	C
contig3704916	1295	4268	A/G	A
contig3705995	1556	1879	G/C	G
contig3707212	1109	1563	A/T	A
contig3707212	1115	1563	A/C	A
contig3707212	1163	1563	C/G	C
contig3707591	1289	2311	C/T	C
contig3707790	298	1278	C/A	C
contig3707862	284	1157	C/T	C
contig3707862	285	1157	T/C	T
contig3707941	2137	6416	C/T	C
contig3709434	1636	1790	T/C	T
contig3710084	68	2070	C/T	C
contig3710084	549	2070	C/T	C
contig3710084	728	2070	C/T	C
contig3710230	2275	4759	C/T	C
contig3710385	479	493	C/T	C
contig3711531	1056	1723	G/A	G
contig3712398	73	3260	C/G	C
contig3712690	6713	8241	G/A	G
contig3713146	2503	2560	C/T	C
contig3714004	3085	4297	C/T	C
contig3714482	4510	9897	T/A	T
contig3715161	340	461	G/A	G
contig3715374	764	2482	T/C	T
contig3715374	856	2482	C/A	C
contig3715374	857	2482	C/T	C
contig3715374	1726	2482	T/A	T

contig3715374	1775	2482	T/A	T
contig3715782	706	1116	G/C	G
contig3716283	2496	10347	T/C	T
contig3716601	3690	3717	C/G	C
contig3716673	15892	24478	G/A	G
contig3716787	332	7992	A/C	A
contig3716863	1161	6748	G/A	G
contig3717235	5734	6518	C/A	C
contig3718168	1712	2559	C/G	C
contig3718231	2347	4065	A/G	A
contig3718651	3428	6367	T/C	T
contig3718722	3187	3385	G/A	G
contig3718785	2324	4180	C/T	C
contig3720303	523	901	G/C	G
contig3720303	545	901	G/T	G
contig3720924	662	1494	C/A	C
contig3721726	5836	7950	T/C	T
contig3721726	5850	7950	T/A	T
contig3723289	1072	5025	G/A	G
contig3723307	2547	3311	A/T	A
contig3723674	2261	2845	C/T	C
contig3723674	2576	2845	G/A	G
contig3724317	2024	8607	T/A	T
contig3724612	13168	13770	C/T	C
contig3724769	420	1191	C/T	C
contig3725833	1585	3931	T/C	T
contig3725837	871	902	C/A	C
contig3726227	909	1260	C/A	C
contig3726227	913	1260	C/A	C
contig3726457	9587	12082	T/C	T
contig3727237	591	1765	G/A	G
contig3727275	268	837	C/T	C
contig3729978	71	1691	T/C	T
contig3730488	284	5110	G/A	G
contig3730719	382	3355	G/T	G
contig3730719	477	3355	T/A	T
contig3730719	938	3355	T/C	T
contig3730719	1965	3355	A/T	A
contig3730719	3076	3355	A/C	A
contig3730719	3090	3355	G/A	G
contig3730719	3091	3355	A/C	A
contig3731126	579	2278	C/T	C
contig3731476	1338	2863	A/G	A
contig3731604	2062	3316	A/T	A
contig3731716	2035	3424	T/A	T
contig3731922	798	3717	A/G	A
contig3731922	3067	3717	T/G	T
contig3731996	533	1418	A/T	A
contig3732034	254	9133	C/T	C
contig3732553	13	468	A/C	A

contig3733020	6891	6894	T/A	T
contig3733325	1878	5826	C/G	C
contig3733325	4706	5826	C/T	C
contig3733373	3096	3977	A/T	A
contig3733543	792	1598	C/A	C
contig3733938	1197	1860	A/G	A
contig3734203	5625	8227	C/A	C
contig3734535	1982	3954	A/G	A
contig3734535	3093	3954	C/T	C
contig3734535	3690	3954	A/G	A
contig3734777	6544	6723	G/T	G
contig3734979	11	975	G/A	G
contig3735208	764	3536	C/T	C
contig3735427	4003	7906	C/T	C
contig3735427	4443	7906	G/A	G
contig3736361	3138	6511	A/T	A
contig3736576	1679	4909	G/A	G
contig3736576	1693	4909	G/A	G
contig3737510	2450	6674	T/C	T
contig3737664	1330	3290	G/A	G
contig3737664	1370	3290	T/G	T
contig3737783	2452	4001	A/G	A
contig3737783	3014	4001	G/C	G
contig3737783	3015	4001	C/T	C
contig3737783	3082	4001	G/A	G
contig3738769	2145	6316	G/T	G
contig3739171	2056	2093	A/T	A
contig3739605	2633	2636	T/C	T
contig3739632	358	400	G/A	G
contig3740124	106	443	G/T	G
contig3740141	68	6063	T/A	T
contig3740141	1505	6063	G/A	G
contig3740141	1912	6063	A/G	A
contig3741067	64	1690	G/A	G
contig3741067	128	1690	G/A	G
contig3741305	182	2654	C/G	C
contig3741305	372	2654	C/T	C
contig3741312	1592	2882	C/A	C
contig3742585	516	3248	C/T	C
contig3742765	265	877	T/C	T
contig3742834	15996	19529	A/G	A
contig3742834	16090	19529	G/A	G
contig3742834	16342	19529	A/C	A
contig3745104	403	499	G/C	G
contig3746238	1714	6785	G/T	G
contig3746238	6630	6785	G/A	G
contig3746569	2522	4707	C/T	C
contig3746951	2647	2652	C/A	C
contig3747624	794	3026	T/A	T
contig3747624	1057	3026	C/T	C

contig3747624	1536	3026	T/C	T
contig3747715	9708	9715	G/A	G
contig3750229	3708	5851	T/C	T
contig3750918	515	1126	A/G	A
contig3751430	54	8307	C/T	C
contig3751430	105	8307	G/T	G
contig3751774	3388	6057	A/G	A
contig3751989	2174	11961	A/G	A
contig3751989	9405	11961	C/G	C
contig3752370	986	2059	G/T	G
contig3753006	941	1068	G/A	G
contig3753306	46	1411	C/T	C
contig3753391	1361	4736	T/A	T
contig3753406	2026	2053	A/G	A
contig3753406	2034	2053	G/A	G
contig3753647	309	1223	C/G	C
contig3753908	2693	5765	C/A	C
contig3754754	2192	2583	A/G	A
contig3755207	374	642	C/A	C
contig3755207	380	642	C/A	C
contig3755207	467	642	G/T	G
contig3755253	2367	11689	T/A	T
contig3755587	16755	18010	T/G	T
contig414579	391	591	C/A	C
contig416785	1333	2220	C/T	C
contig432575	17	456	G/T	G
contig463424	1049	1057	G/A	G
contig467439	1123	1576	C/T	C
contig673366	246	966	T/A	T
contig684391	12	553	C/A	C
contig772665	8	569	T/C	T
contig803878	3401	7358	T/C	T
contig803878	4049	7358	C/A	C
contig803878	5046	7358	T/A	T
contig803878	5083	7358	A/G	A
contig803878	5139	7358	C/T	C
contig803878	5360	7358	G/A	G
contig803878	5806	7358	C/A	C
contig807490	723	1094	T/A	T
contig811166	662	1699	A/G	A
contig828497	2079	3275	A/G	A
contig828497	2105	3275	A/T	A