Statistical methods for identification of NF-κB related genes and their expression profiles in channel catfish after bacterial infection

by

Xiaozhu Wang

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 7, 2017

Keywords: Maximum likelihood, hypothesis test, Immune response, Gene expression, Fish

Approved by

Ash Abebe, Chair, Professor of Mathematics and Statistics
Nedret Billor, Professor of Mathematics and Statistics
Peng Zeng, Associate Professor of Mathematics and Statistics

Abstract

Interactions of NF-κB family, IκB family and IKK complex are key components of NF-κB pathway that is essential for many biological processes including innate and adaptive immunity, and inflammation and stress responses. Despite their importance, systematic analysis of these genes in fish has been lacking. In this project, statistical analysis of NF-κB related genes and their gene expression after bacterial infection was conducted with channel catfish.

A total of thirteen NF-κB related genes were identified in the channel catfish genome, including five NF-κB family genes, five IκB family genes and three IKK complex genes. To confirm the annotation of these thirteen NF-κB related genes, maximum likelihood methods were applied to construct phylogenetic tree of each gene with various species. The reliability of these constructed phylogenetic trees was confirmed by the bootstrap test. Syntenic analysis was used to further determine the annotation of those genes which failed to be confirmed by phylogenetic analysis. To determine the evolutionary patterns of these NF-κB related genes, likelihood ratio test was applied for detecting their selective pressures during evolution. The result indicates that negative selection may play a crucial role in the adaptive evolution of these NF-κB related genes.

Expression profiles of NF-κB related genes after *Flavobacterium columnare* (columnaris) infection were determined by conducting both real-time PCR analysis of different tissues and Kal's Z test with the existing RNA-Seq dataset. The majority of NF-κB related genes were significantly regulated in mucosal tissues of gill, skin and intestine after columnaris infection, indicating their potential involvement in host defense responses to bacterial infection. Distinct expression patterns

of NF-κB related genes were observed in the susceptible and resistant catfish in response to columnaris infection, suggesting that expression of these genes may contribute to the variations in disease resistance/susceptibility of catfish.

Acknowledgments

I would like to express my sincere appreciation to my committee chair, Dr. Ash Abebe, for his guidance and persistent help in training me for this interdisciplinary research. I really appreciate his support and encouragement, taking me through hard times. Without his support, this thesis would not have been possible.

I would like to thank my committee members, Dr. Nedret Billor and Dr. Peng Zeng, for their suggestion and assistance of my thesis. Their guidance not only improve my research but also deep my understanding of statistical application. I would also like to thank my PhD advisor Dr. Zhanjiang Liu, for his technique training and valuable assistance.

In addition, I am grateful for all the professors in Department of Mathematics and Statistics who enhanced my knowledge of statistical theory in various classes. My thank also goes to all my friends in Auburn for their support and friendship.

Finally, I would like to express my deepest appreciation to my beloved parents for their endless love and tremendous support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
CDS	Coding sequences
FDR	False discovery rate
IκB	Nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor
IKK	Inhibitor of nuclear factor kappa-B kinase
MEGA	Molecular Evolutionary Genetics Analysis
ML	Maximum likelihood
NF-κB	Nuclear factor kappa-light-chain-enhancer of activated B cells
ORF	Open reading frame
PDF	Probability density function
PMF	Probability mass function
RPKM	Reads per kilobase of exon model per million mapped reads
RNA-Seq	RNA-sequencing

Chapter 1: Introduction

As an applied thesis, this project aimed to apply statistical theory and method to solve problems in fish genetics and genomics. This introduction chapter focuses on: 1) problem statement; 2) the methodology of main statistical analysis used in this project; 3) literature reviews of immune response to bacterial infection in channel catfish; and 4) literature review of biological functions of NF-κB related genes and pathways.

1.1 Problem statement

As a dominant species in the aquaculture industry of the United States, catfish is one of the most economically important agricultural commodities. However, the catfish industry faces a persistent challenge from bacterial diseases, which cause huge economic losses for farmers each year. It is of great interest to understand catfish immune responses to bacterial infection and to improve catfish disease resistance through the traditional breeding program or genetic improvement. Many immune genes and pathways are involved in the host defense to bacterial infection. A better understanding of the functions of these genes or pathways will set a solid foundation for genetic improvement of disease resistance in catfish.

NF-κB pathway has long been believed to play an important role in many biological processes including innate and adaptive immunity, inflammation and stress responses. In particular, the interaction of the NF-κB family, IκB family and IKK complex is a key component of NF-κB pathway. For better understanding of the immune function of NF-κB related genes in catfish host

defense to bacterial infection, this project was conducted to address the following questions of interest: 1) How many gene members of NF-κB family, IκB family and IKK family can be identified in the channel catfish genome? 2) What are the evolutionary patterns of these NF-κB related genes? 3) After bacterial infection, will expression levels of these NF-κB related genes significantly change? 4) Is there significant variation on expression levels of these NF-κB related genes between the disease resistant catfish and the disease susceptible catfish? Various statistical approaches were applied to address these questions. In particular, maximum likelihood (ML) estimation was used to determine members of NF-κB family, IκB family and IKK family in channel catfish. The likelihood ratio test was used to identify the evolutionary patterns of these NF-κB related genes among various species. The Kal's Z test was applied to identify the expression profiles of NF-κB related genes after bacterial infection in catfish by using existing RNA-Seq dataset.

1.2 Construction of phylogenetic trees

In this section, the theory and methodology of ML methods used in construction of phylogenetic trees are reviewed. The notation and approach of ML estimation are closely following the books *Statistical Inference* written by (Casella and Berger, 2002) and *Molecular Evolution and Phylogenetics* written by (Nei and Kumar, 2000).

Phylogenetic analysis of DNA or protein sequences is a crucial tool to identify the evolutionary history of organisms. It is widely used to study the evolutionary relationships of all levels of classification of organisms by estimating the relationship of different genes or DNA segments. Moreover, it can be used to identify unknown genes in one organism by looking at the relationships of the sequences in this organism with the corresponding genes in other organisms.

In addition, clarification of the evolutionary pattern of multigene families and the process of adaptive evolution at the molecular level can also be achieved by using phylogenetic analysis. Phylogenetic relationships of genes or organisms are usually presented in a tree-like form, which is the phylogenetic tree, and the branch pattern of a tree is treated as its topology.

Phylogenetic tree is a graphical mean to reconstruct evolution based on similarity of individuals under study; the length of a horizontal branch of the tree reveals the amount of change between an individual and its nearest ancestor (Bush, 2001). There are two types of phylogenetic tree: 1) a rooted tree, which reflects the most basal ancestor of the tree in question (Figure 1A) or 2) an unrooted tree which implies no known ancestral root (Figure 1B).

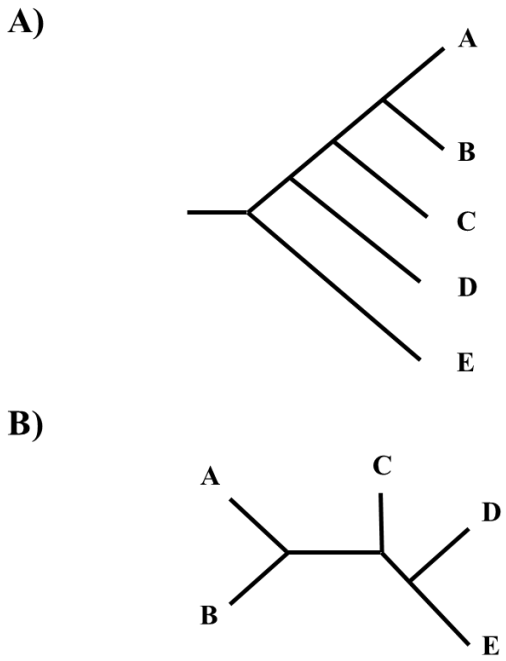

Figure 1. (A) a rooted tree and (B) an unrooted tree. A-E can represent species or genes.

The construction of a phylogenetic tree is considered to be a statistical inference of a true phylogenetic tree, which is unknown. In general, this statistical inference involves two steps: 1)

estimation of the true topology by using a certain optimization principles such as the maximum likelihood principle or the minimum evolution principle; then 2) estimation of branch lengths for the chosen topology by using several statistical methods such as least squares or maximum likelihood methods.

There are three major approaches for inferring phylogenies including distance methods, parsimony methods and likelihood methods. In distance methods, evolutionary distances are computed for all pairs of taxa, and a phylogenetic tree is built based on the relationships among these distance values by using various strategies. These strategies include least squares methods (Cavalli-Sforza and Edwards, 1967; Rzhetsky and Nei, 1992b), minimum evolution method (Edwards and Sforza, 1963), and neighbor joining method (Saitou and Nei, 1987). In maximum parsimony methods, unweighted maximum parsimony and weighted maximum parsimony can be conducted for construction of phylogenetic trees by choosing the topology with the smallest minimum number of substitutions and then estimating the branch length of the chosen topology (Czelusniak et al., 1990).

1.2.1 Phylogenetic analysis using ML methods

In this project, maximum likelihood methods were applied to construct the phylogenetic trees. The idea of using the maximum likelihood estimation for phylogenetic inference was first introduced by Cavalli-Sforza and Edwards (1967). Construction of a phylogenetic tree using DNA sequence data with the ML method was developed by Felsenstein (1981), then extended to protein sequence data by Kishino et al. (1990). Before detailed in the strategy of ML methods for construction of phylogenetic trees, the theoretical foundation of ML estimation and likelihood function is introduced.

ML estimation is a point estimation method to estimate the parameters of a statistical model given observed samples by finding the parameter values that maximize the likelihood of making the observations given the parameters. As one of the most popular techniques for deriving estimators, it is known to be optimal in the sense that it provides the smallest variance of a parameter estimate when sample size is large. ML estimation is used for a wide range of statistical models such as linear models, generalized linear models, structural equation analysis and discrete choice models. It is also applied in a wide range of disciplines including economics, engineering, social sciences, and of course, biology.

The likelihood function is defined as follows: assuming $X_1, \dots, X_n$ is an independent and identically distributed sample from a population with pdf or pmf $f(x|\theta_1, \dots, \theta_k)$, the likelihood function is

$$L(\theta|x) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \dots, \theta_k)$$

For a given sample point x, let $\hat{\theta}(x)$ be a parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x held fixed. $\hat{\theta}(X)$ is a ML estimator of the parameter θ based on a random sample X. If the likelihood function is differentiable (in θ_i), possible candidates for the ML estimator are the values of $(\theta_1, \dots, \theta_k)$ that solve

$$\frac{\partial}{\partial \theta_i} L(\theta|x) = 0, \qquad i = 1, \dots, k.$$

The solutions which make the first derivatives 0 are only potential candidates for the ML estimate since the first derivative being 0 is only a necessary condition for a maximum but not a sufficient condition. For instance, points make the first derivative be 0 may be local or global

extreme points (minima or maxima), or inflection points. The ML method is aimed to find a global maximum for a parameter.

The general idea of construction phylogenetic tree using ML estimation is that: 1) the likelihood of observing a given set of sequence data for a specific substitution model is maximized for each potential topology; then 2) the topology with the highest maximum likelihood is chosen as the final tree. Since the likelihood function in current ML methods of phylogenetic inference includes no parameters for topologies, a topology with the highest ML value is chosen under the assumption that the topology with good estimates of branch lengths is more likely to be the true tree (Nei, 1987).

For one topology of m taxa, the likelihood function L can be generally written as

$$L = f(x; \theta)$$

where x is a set of observed DNA or protein sequences, θ is a set of parameters such as branch length parameter, nucleotide frequencies, and substitution model parameters (e.g. substitution rate parameter, time parameter). All these parameters can be estimated using ML estimation by maximizing L for a given set of observed data.

There are two major search strategies for ML tree including: 1) construct a ML tree directly using ML estimation; 2) construct a ML tree by searching a correct topology based on an initial tree. When the first strategy is chosen, the likelihood values are needed to compute for all potential topology using a certain set of sequence data. The procedure to compute the likelihood values for each of potential topology is detailed in Nei and Kumar (2000), which can be very time-consuming as all possible topologies have to be considered and the likelihood values of them need to be compared for the maximum value. When the second strategy is chosen, an initial tree is built by

another simple method (e.g. star tree), then it is decomposed into a bifurcating tree step-by-step by computing the likelihood value at each step of taxon pairing and by choosing a pair of neighbors with the highest ML value.

1.2.2 Statistical test of phylogenetic trees

In a constructed phylogenetic tree, two types of error may exist: 1) topological errors which encompasses the differences in branching pattern between a constructed tree and the true tree; and 2) branch length errors which are the deviations of estimated branch lengths from the true branch lengths. Statistical tests are conducted to examine the reliability of the obtained phylogenetic trees.

To test the branch length errors, one statistical test for the reliability of an inferred tree is interior branch test, which is initially developed for trees constructed by distance methods (Sitnikova et al., 1995). In this method, each interior branch length for all topologies is tested to see if the value is 0 or positive. Only the length values for interior branches in the corrected topology should be 0 or positive, all interior branch length in all other topologies should have at least one negative value. Therefore, if a constructed phylogenetic tree has an interior branch whose length is significantly negative, then it is highly likely that the current topology of the tree is incorrect. To test the branch length, there are several statistical methods such as the normal deviate (*Z*) test (Rzhetsky and Nei, 1992a), the analytical method (Li, 1989; Nei et al., 1985) and the likelihood ratio test (Felsenstein, 1988).

Another statistical test, which is the most commonly used to test for the reliability of an inferred tree, is the bootstrap test introduced by Felsenstein (1985) and Efron (1982). The bootstrap resampling technique is conducted to evaluate the reliability of a constructed tree. If there are *m* sequences, each with *n* sites (nucleotides or codons or amino acids), a phylogenetic tree can be

reconstructed using some tree building method. In bootstrap resampling, *n* sites are chosen randomly with replacement, thus some sites may be chosen more than once while other sites may not be chosen throughout the whole procedure. These randomly chosen *n* sites along with *m* rows of *n* columns each constitute a new set of sequences. A new tree can be reconstructed with these new sequences using the same tree building method as before. The topology of this new tree is then compared to that of the original tree. Next a score of 0 is given if each interior branch of the original tree is different from that of the bootstrap tree the sequence it partitions; whereas a score of 1 (identity value) is given when the interior branch of the original tree is same as that of the bootstrap tree. This resampling and reconstruction procedure is repeated several hundred times, and the percentage of times each interior branch is given a value of 1 is recorded, which is the bootstrap value. As a general rule, if the bootstrap value for a given interior branch is 95% or higher, then the topology at that branch is considered "correct".

To test topological differences, which involves comparing an ML tree and a suboptimal tree, can be tricky. A standard likelihood ratio test with an χ^2 approximation cannot be applied in this comparison because of the zero degree of freedom in this test and the violation of the asymptotic properties of ML estimators (Felsenstein, 1988; Yang et al., 1995). Instead of the standard likelihood ratio test, the test of variation in single-site log-likelihood between an ML tree and a suboptimal tree can represent the difference in log-likelihood values between two trees (Kishino and Hasegawa, 1989).

1.2.3 Molecular evolutionary genetics analysis (MEGA)

In this project, ML estimation is used to construct sequence alignments and phylogenetic trees by using MEGA 6.0 software. MEGA 6.0 is a molecular evolutionary genetics analysis

software, which is powerful for inferring phylogenetic trees, selecting best-fit substitution models (nucleotide or amino acid), and estimating evolutionary rates site-by-site by using a collection of maximum likelihood analysis (Tamura et al., 2011). As detailed above, in MEGA 6.0, construction of phylogenetic trees using ML methods includes several steps: 1) an initial tree is first built using a fast but suboptimal method such as neighbor-joining method; 2) its branch lengths are adjusted to maximize the likelihood of the data set for that tree topology under the desired model of evolution; 3) variants of the topology are created using the NNI (nearest neighbor interchange) method to search for topologies that fit the data better; 4) maximum-likelihood branch lengths are computed for all potential tree topologies and the greatest likelihood retained as the best choice so far; 5) this search continues until no greater likelihood can be found. After an ML tree is obtained, bootstrap test is conducted to examine if the current tree is the best-fit for the data set.

1.3 Estimation of selective pressure

The speed of evolution of various genes depends on evolutionary pressure making some genes highly conserved while others more polymorphic. High conservation in genes might indicate the occurrence of purifying selection, which is a force that preserves the adapted condition, and typically observed in functionally important genes. By contrast, for genes with an adaptive function, extensive variation in genes indicates the occurrence of positive selection, which is a force that prefers the amino-acid replacements, and the encoded protein might benefit from these genetic variations.

1.3.1 Synonymous and nonsynonymous substitutions

There are two types of codon substitution: 1) synonymous substitution is a nucleotide substitution that does not change the encoded amino acid; while 2) nonsynonymous substitution is a nucleotide substitution that changes the encoded amino acid. Natural selection operates mainly on the protein level, including 1) purifying selection which is a natural selection against deleterious mutations with negative selective coefficients; and 2) positive selection which is a Darwinian selection fixing advantageous mutations with positive selective coefficients. Comparison of the relative rates of synonymous and nonsynonymous substitutions is widely used to detect whether purifying selection or positive selection is operating on a protein. The nucleotide substitution rate (d_S) of synonymous is neutral to selective pressure of a protein, whereas the nucleotide substitution rate (d_N) of non-synonymous is a function of selective pressure of a protein. The ratio of these two rates ($\omega = d_N/d_S$) can be used as a measure of selective pressure, because synonymous and nonsynonymous mutations are under very different selective pressures and are fixed at very different rates (Yang, 2007). For instance, in the absence of selection, the synonymous and non-synonymous substitution rates should be equal. When purifying selection is operating to preserve the current structure and function of a protein, the excess of synonymous substitution makes d_N/d_S less than 1; while when positive selection is operating for more amino-acid replacements, the excess of non-synonymous substitution makes d_N/d_S more than 1 (Bielawski and Yang, 2003; Bush, 2001).

There are multiple statistical methods for detecting selective pressure, which can be divided into two classes: (1) intuitive methods which count synonymous (S) and nonsynonymous (N) sites in each sequence, and their differences among sequences, then correct for multiple substitutions at the same site (Comeron, 1995; Miyata and Yasunaga, 1980; Nei and Gojobori, 1986; Yang and

Nielsen, 2000); (2) the ML method based on explicit models of codon substitution (Goldman and Yang, 1994; Muse, 1996). In this project, the selective pressures of NF-κB related proteins were detected by using PAML package, which conducts phylogenetic analysis by maximum likelihood. The following paragraphs will be focus on the methodology of ML estimation of selective pressure, which are closely following several papers published by Yang's group (e.g. Bielawski and Yang, 2003; Yang, 2007; Yang and Nielsen, 2000).

1.3.2 ML estimation of selective pressure

The first step for detecting molecular adaptation is to construct a model of codon substitution with Markov processes, as the codon is considered the unit of evolution (Goldman and Yang, 1994; Yang, 1998; Yang and Bielawski, 2000). The substitution model is formulated at the level of instantaneous rates, which does not allow multiple changes. Here, a Markov process, which is a simple stochastic process, describes substitutions between the 61 sense codons within a protein coding sequence. The substitution rate from codons i to j ($i \neq j$) is given as:

$$q_{ij} = \begin{cases} 0 \\ \pi_j \\ \kappa\pi_j \\ \omega\pi_j \\ \omega\kappa\pi_j \end{cases}$$

where $q_{ij} = 0$ if i and j differ at more than one position, $q_{ij} = \pi_j$ for synonymous transversion, $q_{ij} = \kappa\pi_j$ for synonymous transition, $q_{ij} = \omega\pi_j$ for nonsynonymous transversion, $q_{ij} = \omega\kappa\pi_j$ for nonsynonymous transition. Parameter κ represents the transition/transversion rate ratio, π_j represents the equilibrium frequency of codon j and ω ($= d_N/d_S$) measures the selective pressure on the protein.

The q_{ij} are relative rates because time and rate are confounded in such an analysis. Given the rate matrix $Q = \{q_{ij}\}$ the transition probability matrix over time t is calculated as:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

where $p_{ij}(t)$ is the probability that codon i becomes codon j after time t. Likelihood calculation on a phylogeny involves summing over all possible codons in extinct ancestors (internal nodes of the tree).

All the parameters such as transition/transversion rate ratio κ, d_N/d_S the ratio ω in the model of codon substitution are estimated from the data by ML and corrected for multiple hits. Pathways of change between codons are also weighted within the same step. These estimated parameters are then used to calculate d_N and d_S according to their definitions. For example, suppose there are *n* codon sites in a protein, and a certain site (*m*) has codons GCG and GUG. The data at site *m*, which are linked to an ancestor with codon k by branch lengths t_0 and t_1 (Figure 2), is denoted $x_m = \{GCG, GUG\}$. Due to the unknown ancestral codon, the summation is over all 61 possible codons for k. Thus, the probability of site *m* is

$$f(x_m) = \sum_k \pi_k p_{k,GCG}(t_0) p_{k,GUG}(t_1) = \pi_{GCG} p_{GCG,GUG}(t_0 + t_1).$$

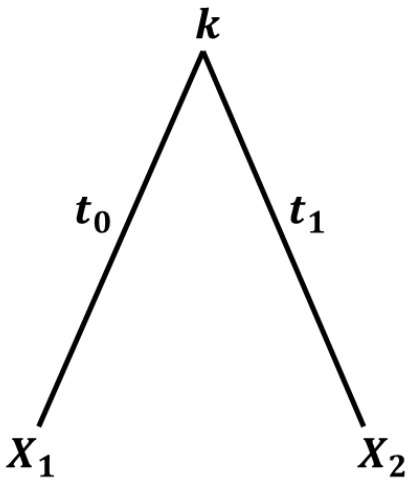

Figure 2. Rooted phylogenetic tree reveals the data at site m linked to an ancestor with codon k by branch lengths t_0 and t_1.

The log likelihood function is a summation of all codon sites in the sequence

$$l(t, \kappa, \omega) = \sum_{m=1}^{n} \log f(x_m)$$

As mentioned above, all these parameters, such as t, κ, ω, are estimated by maximizing the log likelihood function.

Likelihood-ratio test, which is used for comparing the goodness of fit of the null model and the alternative model, can be used to test whether d_N is significantly higher than d_S. This test is based on the likelihood ratio, which expresses how many times more likely the data are under one model than the other. In this case, the hypothesis test has completely specified models under both the null and alternative hypotheses, which for convenience are written in terms of fixed values of parameter ω:

H_0: $\omega = 1$ in null model

H_a: $\omega \neq 1$ in alternative model

Each of the two competing models, the null model (ω fixed at 1) and the alternative model (ω treated as a free parameter), is separately fitted to the data and the log-likelihood recorded. The test statistic (often denoted by D) is twice the log of the likelihoods ratio:

$$\mathrm{D} = -2\ln\left(\frac{likelihood\ for\ null\ model}{likelihood\ for\ alternative\ model}\right)$$

$$= 2\times[\ln(likelihood\ for\ alternative\ model) - \ln(likelihood\ for\ null\ model)]$$

Then twice the log-likelihood difference (D) between the null model and the alternative model is compared with a χ^2 distribution with $df_{alt} - df_{null} = 1$ degree of freedom to test whether ω in null model is different from that in the alternative model.

Noticeably, to use χ^2 approximation to the likelihood ratio statistic, certain conditions are required: (1) the hypotheses must be nested; (2) to maintain the success of χ^2 approximation, the sample should be sufficiently large; (3) the "boundary" problem should be avoided, in which H_a may not be related to H_0 by fixing one or more of its parameters at the boundary of parameter space. If the above conditions cannot be met, Monte Carlo simulation needs to be used for estimation of the exact distribution (Anisimova et al., 2001).

The power of the likelihood-ratio test is high when the sequences are not highly similar or divergent, however, it is modulated by the length of the sequence and the strength of positive selection (Bielawski and Yang, 2003). Because of the use of χ^2 approximation, the false positive rate predicted for the likelihood-ratio test is less than that predicted by the specified significance level of the test (Anisimova et al., 2001). The robustness of the likelihood-ratio test, which refers to the stability of results to changes in the model assumptions, can be improved by several methods:

(1) use of multiple models; (2) assumptions such as the ω distribution should be evaluated relative to their effects on ML parameter estimation; (3) care is needed when identifying results obtained from local optimal (Bielawski and Yang, 2003).

1.4 Hypothesis testing for a proportion

As a method of statistical inference, statistical hypothesis test is used to determine whether the null hypotheses (H_0) or the alternative hypotheses (H_a) is the outcome of a study for a pre-specified level of significance. After statement of the relevant null and alternative hypotheses, statistical assumptions along with the choice of test statistic should be made for the hypothesis testing based on the sample. There are several test statistics can be used including one-sample or two-sample z-test, one-sample or paired *t*-test, F test, test for proportion, Chi-squared test for variance and Chi-squared test for goodness of fit. The distribution of the chosen test statistic then can be derived under the null hypothesis from the assumptions, and the observed value t_{obs} is computed from the observations. Finally, the observed value t_{obs} can be determined to see if it in the critical region (the probability of the critical region is the significant level α). The decision rule is to reject the null hypotheses H_0 if the observed value t_{obs} is in the critical region, and to accept or "fail to reject" the null hypotheses otherwise.

Identification of transcript and the quantification of gene expression by using RNA-sequencing (RNA-Seq) are distinct core activities in molecular biology. One major interest of RNA-Seq technology is the differential expression analysis, which requires the comparison of gene expression values among samples. RNA-seq quantification is based on expression level that are absolutely or probabilistically assigned to transcripts. Reads per kilobase of exon model per million mapped reads (RPKM) (Brockman et al., 2008) or read counts (Trapnell et al., 2012) can

be calculated as the original expression values. The original expression values need to be scaling normalized across samples for comparison. There are several approaches to identify differential expression. The first set of approaches compute differential expression by using discrete probability distributions such as the Poisson or negative binomial distribution (known as the gamma-Poisson distribution). The second set compute differential expression by treating expression value (e.g. read count or RPKM) of a specific mRNA per tissue as a fraction or proportion of the total read count of mRNA molecules in that tissue. This approach was first introduced by Kal et al., (1999). The proportion (p) of a specific mRNA can be present as:

$$p = \frac{n_{specific\ mRNA/sample}}{N_{total\ mRNA/sample}} = \frac{n_{specific\ mRNA}}{N_{total\ mRNA}}$$

The distribution for this approach is a binomial distribution, which can be well approximated by a normal distribution with mean p and $SD_P = \sqrt{p(1-p)}$ for the high-throughput RNA-Seq dataset. Thus, the process to identify differential expression genes is a hypothesis test for a proportion to test if the difference $p_1 - p_2$ (which are expression values of a specific gene in different groups) is significant or not. The standard error (SE) of the difference $p_1 - p_2$, in which the proportions p_1 and p_2 are from samples with sizes N_1 and N_2, respectively, is present as:

$$SE_{P_1-P_2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$$

Then the test statistic Z can be calculated by using the difference $p_1 - p_2$ and SE:

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2}}$$

where p_0 is calculated as $p_0 = (n_1 + n_2)/(N_1 + N_2)$, which is the estimate of the proportions when the null hypothesis is true. Under the null hypothesis, this test statistic *Z* is followed a normal distribution and can serve as a statistical test for the difference between the proportions p_1 and p_2.

The relation between the difference $p_1 - p_2$ that can also be detected with by a two-sided probability of a type I error of less than α, as well as a type Ⅱ error of less than β:

$$p_1 - p_2 > \left(Z_{\alpha/2} SE_{H_0}\right) + \left(Z_{\beta/2} SE_{H_1}\right)$$
$$> \left(Z_{\alpha/2}\sqrt{p_0(1-p_0)/N_1 + p_0(1-p_0)/N_2}\right)$$
$$+ \left(Z_{\beta/2}\sqrt{p_1(1-p_1)/N_1 + p_1(1-p_1)/N_2}\right)$$

where p_1 is calculated as $p_1 = (n_1 + n_2)/(N_1 + N_2)$, which is the estimate of the proportions when the alternative hypothesis is true. Then the critical value can be calculated for either 0.01 or 0.05 levels of significance. A major advantage of a hypothesis test for proportions is that it can be used for samples with different sizes.

For multi-group experiments, pair comparison can be tested by choosing either to test all pairs of groups or to have each group compared to a specified reference group (control group). Since a large number of multiple comparisons are conducted, some tests will have P values less than 0.05 purely by chance, even if all the null hypotheses are true. Thus, a use of single-inference procedures can lead to the increase of false positive rate. Therefore, a multiple testing correction procedure is crucial to control this multiplicity effect.

The classic approach to solve the multiple comparison problem is to control the familywise error rate (the probability of making at least one false positive call) by using Bonferroni correction

(Cabin and Mitchell, 2000). Simply, the critical value α for an individual test from the multiple tests can be calculated by dividing the familywise error rate (usually 0.05) by the number of tests:

$$\alpha_{Bonferroni\ correction} = \frac{0.05}{N_{total\ number\ of\ tests}}$$

The Bonferroni correction is appropriate for a fairly small number of multiple comparisons, when a single false positive in a set of tests would be a problem. However, if a large number of multiple comparisons are conducted, the use of the Bonferroni correction may lead to a very high rate of false negatives.

An alternative approach is to control the false discovery rate (FDR), which is the proportion of "discoveries" (significant results) that are false positives (the rate of Type I errors in null hypothesis testing) (Benjamini and Hochberg, 1995; Reiner et al., 2003). FDR correction provides less stringent control of Type I errors compared to the Bonferroni correction. Thus, FDR corrected procedures have greater power with the cost of increasing rates of Type I errors. In the following paragraphs, the Benjamini-Hochberg FDR procedure will be introduced (Benjamini and Hochberg, 1995).

Suppose m independent null hypotheses are testing simultaneously, of which m_0 are true. The situation can be summarized as a traditional form in Table 1. The number of hypotheses rejected is R, which is an observable random variable; U, V, S and T are unobservable random variables. For a given significant level α, if each individual null hypothesis is tested separately at level α, then $R = R(\alpha)$ is increasing in α.

Table 1. Number of errors committed when testing m null hypotheses.

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

The proportion of errors (FDR), in which the null hypotheses are falsely rejected, can be present by the unobserved random variable Q. As the proportion of the rejected null hypotheses which are erroneously rejected, $Q = V/(V + S)$. When $V + S = 0$, $Q = 0$ indicating there is no false positive committed. Then the expectation of FDR Q is

$$E(Q) = E\{V/(V + S)\} = E(V/R)$$

The FDR controlling procedure is given as follows. Given testing $H_1, H_2, H_3, \dots, H_m$ based on the corresponding P values $P_1, P_2, P_3, \dots, P_m$, put the individual P values in the order from smallest to largest, $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)}$, and denote the P value for $H_{(i)}$ null hypothesis is $P_{(i)}$. Suppose it is desired to control the FDR at q^*. Let k be the largest i for which

$$P_{(i)} \leq \frac{i}{m} q^*$$

Then all $H_{(i)}$ $i = 1,2, \dots, k$ should be rejected, even the ones that are not less than their Benjamini-Hochberg critical value.

1.5 Immune response to bacterial infection in catfish

Channel catfish is a primary species in the aquaculture industry of the United States. the production of catfish has suffered massive economical losses due to pathogen infection. Two major catfish diseases are ESC disease by *E. ictaluri* and columnaris disease by *F. columnare.* Both bacterial pathogens cause acute to chronic infections, which result in massive loss in catfish

industry. *E. ictaluri* enters the host through intestine and other mucosal epithelial sites (Hawke et al., 1981) while *F. columnare* infects the external tissues such as gill and skin (Shoemaker et al., 2008).

Innate and adaptive immune responses play important roles on host defense to various diseases in channel catfish. A complete defense system is built by innate and adaptive immunity, increasing the resistance to pathogenic organisms, such as E. ictaluri and Flavobacterium columnare (LaFrentz et al., 2012). To understand the immune response to bacterial infections, numerous studies have been conducted. For instance, 212 unique, differentially expressed transcripts were identified in the liver of channel catfish at the acute phase response stage following ESC infection (Peatman et al., 2007). Transcriptome analysis of catfish intestine after ESC infection revealed 1,633 differentially expressed genes, which involved in different stages of immune responses such as muscle fiber dynamics, inflammation, pathogen recognition and growth disruption (Li et al., 2012). Moreover, by coupling of RNA-seq with bulked segregant analysis (BSA), 1,255 differentially expressed genes with allele-specific expression were found between resistant and susceptible fish, along with 56,419 significant SNPs residing on 4,304 unique genes were identified in response to ESC infection (Wang et al., 2013). Transcriptome analysis of catfish gill identified over 2600 differentially expressed genes at three timepoints after columnaris infection, including genes with putative roles in facilitating pathogen adhesion and invasion and the concomitant host immune response (Sun et al., 2012). In addition, transcriptome analysis in the gill of resistant and susceptible catfish following columnaris infection indicated the importance of a polarization in mucosal status prior to infection, which may impact early pathogenic adherence, entry, and host inflammatory processes (Peatman et al., 2013).

As many genes have been identified to associate with immune responses to bacterial infection in catfish, all immune-related genes along with other members in the corresponding gene families were further characterized, which included pathogen recognition receptors (Rajendran et al. 2012a; 2012b; Zhang et al., 2013; Sun et al., 2014), claudins (Sun et al., 2015), serpins (Li et al., 2015), chemokines (Bao et al., 2006; Peatman et al., 2006; Fu et al., 2017), complementary proteins (Jiang et al., 2015), tumor suppressor proteins (Mu et al., 2015), Rho GTPase genes (Tan et al., 2017), apolipoprotein genes (Yang et al., 2017), and a number of regulatory and response genes involved in disease responses (Yao et al., 2014; 2015; Yuan et al., 2016). However, immune function of NF-κB related genes still remains unknown in channel catfish. In this study, we identified NF-κB related genes in the channel catfish genome and determined their expression in response to bacterial infections of *E. ictaluri* and *F. columnare*.

1.6 Biological functions of NF-κB related genes

Nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) are transcription factors that regulate innate and adaptive immune responses (Pahl, 1999). They are among the key regulators of inflammation after bacterial infection (Gerondakis et al., 1999; Pahl, 1999). There are five NF-κB family proteins in mammals: RelA (p65), RelB, c-Rel, p50 (p105/NFκB1) and p52 (p100/NFκB2), encoded by *RELA*, *RELB*, *REL*, *NFKB1*, and *NFKB2* gene, respectively (Lawrence, 2009). NF-κB proteins are sequestered in cytoplasm by inhibitory molecules of IκB family, which includes five proteins: IκBα, IκBβ, IκBε, Bcl3 and IκBζ, encoded by *NFKBIA*, *NFKBIB*, *NFKBIE*, *BCL3* and *NFKBID* gene, respectively (Baldwin, 1996). However, p105 and p100, the precursor proteins of p50 and p52 respectively, can be assigned to both NF-κB and IκB families as they have the characteristics of both NF-κB and IκB proteins (Oeckinghaus and Ghosh, 2009). The activation of NF-κB is mainly triggered by the upstream IκB kinase (IKK) complex, which includes three

major members: IKKα, IKKβ and IKKγ, encoded by *CHUK*, *IKBKB*, and *IKBKG* gene, respectively (Israel, 2010). NF-κB family, along with its inhibitors (IκB family) and activators (IKK complex), work together to exert important roles in pro- and anti-inflammatory responses after bacterial infection.

NF-κB proteins are DNA sequence-specific transcription factors which are associated with each other to form functional complexes and regulate inflammatory and innate immune responses (Karin, 1999). They all share a conserved Rel homology domain (RHD), which is functionally important for dimerization, IκB binding, and nuclear translocation (Baldwin, 1996). Although NF-κB proteins were originally thought to be only produced in mature B cells and plasma cells, they are now known to be present in cytoplasm of most cells in an inactive form due to their binding with IκBs (Baeuerle and Baltimore, 1988). Activation of NF-κB has been demonstrated to be achieved by degradation of IκBs through their phosphorylation by IKK complex, which allows free NF-κB to enter the nucleus and activate transcription of target genes (Israel, 2010). Expression of the genes regulated by NF-κB is mostly responsive to environmental changes. Such genes include tumor necrosis factor alpha (TNFα), MHC pathway genes, granulocyte colony-stimulating factor (G-CSF), and inhibitor of apoptosis proteins (IAPs) (Oeckinghaus and Ghosh, 2009). In addition, post-translational modifications such as phosphorylation, ubiquitination, or acetylation can significantly affect the regulation of the NF-κB pathway (Perkins, 2006).

IκB proteins were originally known as inhibitors of NF-κB with the function of cytoplasmic retention, while subsequent discoveries suggested that IκB proteins act as complex regulators of NF-κB (Grilli et al., 1993). IκB proteins can be further divided into two groups: 'cytoplasmic IκBs' (IκBα, IκBβ and IκBε) and 'nuclear IκBs' (Bcl3 and IκBζ) (Yamamoto and Takeda, 2008). They all contain five to seven Ankyrin repeats (ANK) which are essential for interactions with the RHD

in NF-κB proteins. The inhibitory function of IκBs is delivered through their binding to NF-κB proteins to form dimers, thereby preventing NF-κB to be transported into the nucleus. Upon IKK activation, IκBα is rapidly degraded, leading to an oscillatory NF-κB activation, while the degradation of IκBβ and IκBε is much slower, providing a long-term oscillatory response of NF-κB (Hoffmann et al., 2002). After ubiquitination and proteasomal degradation of phosphorylated cytoplasmic IκBs, free NF-κB dimers translocate from cytoplasm to nucleus, where nuclear IκBs can either activate or inhibit NF-κB transcriptional activity by interacting with NF-κB subunits (Oeckinghaus et al., 2011; Yamamoto and Takeda, 2008).

IKK complex, as a core element of NF-κB activation, is composed of two catalytically active kinases (IKKα and IKKβ) and a regulatory subunit (IKKγ) (Chen et al., 1996). There are two separate NF-κB activation mechanisms depend on different IKK proteins: 1) the canonical pathway requires IKKβ associated with IKKγ to rapidly degrade IκB proteins in response to upstream TNFα and IL-1 signaling; and 2) the non-canonical pathway relies on activation of IKKα by upstream NF-κB-inducing kinase (NIK), and then IKKα specifically phosphorylates p52-RelB complexes (Oeckinghaus et al., 2011; Pomerantz and Baltimore, 2002; Scheidereit, 2006; Senftleben et al., 2001).

Functions of NF-κB related genes in immune system has been well studied in mammals (Baeuerle and Henkel, 1994; Yamamoto and Gaynor, 2001). Toll-like receptors (TLRs), as primary sensors, are the key pattern recognition receptors (PRRs) for pathogen associated molecular patterns (PAMPs) of bacterial pathogens (Philpott and Girardin, 2004). TLRs can active NF-κB pathway, leading to the subsequent innate immune responses (Kawai and Akira, 2007). The roles of NF-κB pathway in the production of inflammatory mediators against bacterial infection is well documented in mammalian species, but only a few studies in teleost fish were

reported. For instance, three NF-κB molecules (p65, c-rel, and p52) and two IκBα orthologues (IκBαa and IκBαb) were found to be involved in notochord development in zebrafish (Correa et al., 2004; Pradhan et al., 2012). NF-κB subunits and their activation by aryl hydrocarbon receptor agonists (AhR) were reported in marine fish (Schlezinger et al., 2000). In spite of these studies, systematic analysis of NF-κB related genes in teleost fish are still lacking, especially their expression patterns after bacterial infection.

Chapter 2: Experimental design and methods

In this section, the experimental design and statistical method used in this project were detailed. The steps of the experimental design include identification of NF-κB related genes in the channel catfish genome, bacterial infection challenge and RNA extraction, Real-time PCR and statistical analysis using RNA-Seq datasets from previous studies.

2.1 Identification of channel catfish NF-κB related genes

To identify the catfish sequences of NF-κB related genes, the first step is to search amino acid sequences of NF-κB related genes from other species such as human (*Homo sapiens*), mouse (*Mus musculus*), zebrafish (*Danio rerio*), fugu (*Takifugu rubripes*) and stickleback (*Gasterosteus aculeatus*) in online database such as Ensembl (http://useast.ensembl.org/index.html). After amino acid sequences from these species were obtained, they were used as queries to search the channel catfish RNA-Seq dataset (Liu et al., 2012) by using TBLASTN. For the most comprehensive BLAST, we included additional RNA-Seq datasets in our analysis (Li et al., 2012; Sun et al., 2012). All BLAST procedures were conducted with BLAST+ software on Linux terminal. Duplicates in the initial sequence pool were eliminated by using Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/), and the resulting unique set of sequences was subjected to further analysis.

The open reading frame (ORF) of each unique sequence was identified by ORF finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and then verified by BLASTP against NCBI non-

redundant protein sequence database. BLASTN was conducted against the catfish reference genome sequence (Liu et al., 2016) for validation and for the determination of copy numbers. The genes within the catfish reference genome were predicted by FGENESH program (Solovyev et al., 2006).

2.2 Sequence and evolutionary analysis of NF-κB related genes

The gene structures such as intron, exon or UTR of these NF-κB related genes were investigated based on each coding sequence (CDS) and corresponding genomic sequence by using the online Gene Structure Display Server (http://gsds.cbi.pku.edu.cn/index.php). The functional domain characteristics of NF-κB related genes were identified using NCBI Conserved Domain Search (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). To provide insights into the evolutionary patterns of NF-κB related genes, patterns of selection pressure were examined with maximum likelihood analysis by using the YN00 program in the PAML X package (Xu and Yang, 2013).

2.3 Phylogenetic and syntenic analysis

The full-length amino acid sequences of NF-κB related genes from several representative vertebrates, including human, mouse, cow, lizard, frog, and several fish species such as zebrafish, fugu, tilapia, stickleback, Amazon molly, and Atlantic cod were retrieved from ENSEMBL genome browser (http://www.ensembl.org/) or NCBI database (http://www.ncbi.nlm.nih.gov/) for phylogenetic analysis. Multiple protein sequence alignments were conducted using MUSCLE (MUltiple Sequence Comparison by Log-Expectation) with default parameters (Edgar, 2004). Phylogenetic analysis was conducted with maximum likelihood method using MEGA 6.0 software

(Tamura et al., 2013). Bootstrap tests with 1,000 replicates were performed to evaluate the phylogenetic trees.

In cases where phylogenetic analysis was not sufficient to establish the orthologies (e.g., *NFKB2* and *BCL3*), syntenic analysis was also performed to properly annotate the genes. Conserved syntenic blocks of human and zebrafish were first identified from the Genomicus database version 84.01 (Louis et al., 2015; Muffato et al., 2010), and then compared to those of catfish.

2.4 Experimental fish

All procedures involved in the handling and treatment of fish during this study were approved by the Auburn University Institutional Animal Care and Use Committee (AU-IACUC) prior to initiation of the study. One-year old experimental fish were moved to the Auburn University Hatchery Challenge Facility one week before hypoxia challenge. Fish were maintained in four 50-gallon tanks and acclimatized for 2 weeks at a temperature of 28°C with aerated flow-through water. Experimental fish were fed once daily and the DO concentration in water was monitored twice a day using the YSI dissolved oxygen meter.

2.5 Bacterial infection and sample collection

F. columnare bacteria were cultured from a single colony and re-isolated from a symptomatic fish. Challenge experiments were conducted by immersion exposure for 2 h at a final concentration 1×10^5 CFU/mL. Gill, skin, and intestine tissues were collected before challenge (0 h) and at 8 h after challenge. Equal amounts of tissue samples (50 mg) were collected from each

fish within the three replicate pools (3 pools of 5 fish each) at each sampling time-point. Samples were flash-frozen in liquid nitrogen and stored at -80°C until RNA extraction.

2.6 RNA extraction and quantitative real-time qPCR

Total RNA was extracted using RNeasy Mini kit (Qiagen, USA). RNA concentration and quality were checked using a NanoDrop ND-1000 UV-VIS spectrophotometer (NanoDrop Technologies, USA). For real-time PCR, first strand cDNA was synthesized using the iScript cDNA Synthesis Kit (Bio-Rad, USA). All the cDNA products were diluted to 250 ng/μL for real-time PCR reaction using the PerfeCTa® SYBR® Green FastMix® Reaction Mixes (Bio-Rad, USA) on a CFX96 real-time PCR Detection System (Bio-Rad, USA). The primers used in quantitative real-time PCR were listed in Table 1. The 18S rRNA and 28S rRNA were used as the internal reference controls. The thermal cycling profile consisted of an initial denaturation at 95°C (for 30 s), followed by 40 cycles of denaturation at 94°C (5 s), an annealing/extension temperature at 60°C (5 s), and 72°C for 5 s, followed by dissociation curve analysis to verify the specificity of amplified products. Relative fold changes were calculated based on the expression levels of 18S rRNA and 28S rRNA using a randomization test in Relative Expression Software Tool (REST) version 2009 (Pfaffl et al., 2002). Expression differences between groups were considered significant when p-value < 0.05. Three biological replicate RNA samples and three technical replicates were conducted to confirm expression patterns.

Table 2. Primers of catfish NF-κB related genes used in this study.

Gene name	Primer name	Sequences (5' - 3')	Position
rela	rela-F	GAGCAGGCGGCGAGAAAC	1582-1599
	rela-R	CAACATGGACATGAGGCGATC	1741-1761
relb	relb-F	ACGGCACGCCTGAAATAGC	29-47
	relb-R	GGCCACGAGGTGATGTAAGC	237-256
rel	rel-F	ACCTGGTGGGGAAAGACTGC	173-192

	rel-R	AGGTCGTATTCCTCCGTCTGC	354-374
nfkb1	nfkb1-F	ATCAATGCCTTCCGCCTCA	1237-1255
	nfkb1-R	GCCTACTGTGGTCAGATTTTCGT	1409-1431
nfkb2	nfkb2-F	AGCGAAGCAGCCTGGATACC	495-514
	nfkb2-R	GCTGTCGTAGATGGGGTTGG	662-681
nfkbiaa	nfkbiaa-F	CAGCGGCATGGACTCGTTA	96-114
	nfkbiaa-R	TGTCTGTTGAGGAATGGGTCG	306-326
nfkbiab	nfkbiab-F	AGGACGACACGGAGGATTACTG	122-143
	nfkbiab-R	GTTTGTTCAGGAACGGGTGC	315-334
nfkbib	nfkbib-F	GGTTTGGGCTCGGATAAGCT	103-122
	nfkbib-R	CTCGTCGTTGATCGAATCCC	299-318
nfkbie	nfkbie-F	GGAGGGAACGAGTGGGAAAA	768-787
	nfkbie-R	TGGTGATGGCGGTGTCTTG	913-931
bcl3	bcl3-F	CTACCCTTTCCCCTGCGATAA	158-178
	bcl3-R	CTGCTGCACTTGGGCACTTT	412-431
chuk	chuk-F	AGTGTAGTTGCGGTTTTCGTCC	641-662
	chuk-R	CAGGGTGTATTATTCCTCCCCTC	861-883
ikbkb	ikbkb-F	CCTCTGCCCTCACATACCTTCA	389-410
	ikbkb-R	CTCCTTAGCATAGCCCAAGTCAA	494-516
ikbkg	ikbkg-F	AGAACGGGCTGTCATTGTGG	1137-1156
	ikbkg-R	GGTCTCCTGTAAGGTCCTCTGCT	1355-1377

2.7 Expression analysis of NF-κB related genes using RNA-Seq dataset

Expression of the NF-κB related genes after disease infection were determined using existing RNA-Seq dataset. To analyze expression patterns of these genes after columnaris infection, the Illumina-based RNA-Seq dataset was retrieved from the SRA database (SRA Accession SRP017689). These RNA-Seq short reads were generated from gill samples after *F. columnare* infection. The details of bacterial challenge, RNA extraction, library construction and transcriptome sequencing were described (Peatman et al., 2013). Trimmed reads of the RNA-Seq dataset were analyzed using CLC Genomics Workbench (v5.5.2). The trimmed RNA-Seq reads were aligned with all genes identified in channel catfish genome along with all cDNA sequences of channel catfish NF-κB related genes. Mapping parameters were set to allow reads identity $\geq 95\%$ and mismatches ≤ 2. The number of total mapped reads for each NF-κB related gene was determined and RPKM (Reads Per Kilobase of the transcript per Million mapped reads) was

calculated. The expression fold-change of each NF-κB related gene was calculated based on the normalized RPKM using proportions-based Kal's Z test. NF-κB related genes with absolute expression fold change values ≥ 1.5, total read number ≥ 5 and p-value ≤ 0.05 were regarded as differentially expressed genes.

Chapter 3: Results and discussion

In this section, the major biological findings of this project were summarized including: 1) gene members of three NF-κB related gene families, 2) phylogenetic trees and syntenic analysis of these genes, and 3) expression patterns of these NF-κB related genes after bacterial infection. The results were fully discussed to reveal the importance of these NF-κB related genes in host defense to bacterial infection.

3.1 Results

3.1.1 Identification of NF-κB related genes in channel catfish

Thirteen NF-κB related genes were identified in channel catfish including: five genes in NF-κB family (*RELA*, *RELB*, *REL*, *NFKB1*, *NFKB2*), five genes in IκB family (*NFKBIAA*, *NFKBIAB*, *NFKBIB*, *NFKBIE*, *BCL3*), and three genes in IKK complex (*CHUK*, *IKBKB*, *IKBKG*). *NFKBID* in IκB family was not found in the channel catfish genome. The characteristics of these genes, including transcript length, coding sequences, conserved domains, chromosomal locations, and accession numbers, are listed in Table 3. Gene length of all NF-κB related genes ranged from 1,392 to 3,776 bp in channel catfish. Three genes (*REL*, *NFKB1* and *CHUK*) were located in linkage group (LG) 6, and two genes (*NFKBIAA* and *NFKBIE*) were located in LG28; the remaining genes were each in various linkage groups. Members of NF-κB family shared the IPT and RHD domains, while IκB proteins contained 4 to 6 ANK repeats. *CHUK* and *IKBKB* shared

the PKc domain, whereas *IKBKG* had two unique domains UBAN and NEMO. In all cases, except *NFKBIA* from IκB family, where gene duplications were observed in teleost fish, catfish had just one copy of all NF-kB related genes (Table 4). It is interesting that *NFKBIB* was only identified in zebrafish and catfish, while *NFKBID* was represented in all other analyzed teleost species but not in zebrafish and catfish.

Table 3. Characteristics of NF-κB related genes identified in channel catfish. LG: linkage group.

Family name	Genes	mRNA (bp)	CDS (aa)	Conserved Domain	Location	Accession NO.
NF-κB Family	RELA	2372	601	IPT, RHD	LG29	JT417958
	RELB	2861	557	IPT, RHD	LG22	JT407102
	REL	2237	580	IPT, RHD	LG6	JT317830
	NFKB1	3776	507	IPT, RHD, ANK, DD	LG6	JT316856
	NFKB2	3383	909	IPT, RHD, ANK, DD	LG3	JT319661
IkB Family	NFKBIAA	1392	308	ANK	LG28	JT419244
	NFKBIAB	1541	320	ANK	LG16	JT408574
	NFKBIB	2304	397	ANK	LG14	JT416051
	NFKBIE	1525	360	ANK	LG28	JT413485
	BCL3	2489	631	ANK	LG1	JT463618
IKK complex	CHUK	3706	756	PKc	LG6	JT406644
	IKBKB	3129	776	PKc, S_TKc	LG11	JT411949
	IKBKG	2816	568	UBAN, NEMO	LG15	JT417502

The abbreviations of the functional domains were immunoglobulin-like fold, plexins, transcription factors (IPT); Rel homology domain (RHD); Ankyrin repeats (ANK); death domain (DD); protein kinases, catalytic domain (PKc); Serine/Threonine protein kinases, catalytic domain (S_TKc); NF-Kappa-B essential modulator (NEMO); ubiquitin binding in ABIN and NEMO (UBAN).

Table 4. Copy numbers of NF-κB related genes in several representative vertebrate genomes.

Gene	Catfish	Zebrafish	Fugu	Stickleback	Tilapia	Frog	Lizard	Mouse	Cow	Human
RELA	1	1	1	1	1	1	1	1	1	1
RELB	1	1	1	1	1	1	1	1	1	1
REL	1	1	1	0	1	1	1	1	1	1
NFKB1	1	1	1	1	1	1	0	1	1	1
NFKB2	1	1	1	1	1	1	1	1	1	1
NFKBIA	2	2	2	2	2	1	1	1	1	1
NFKBIB	1	1	0	0	0	0	1	1	1	1
NFKBID	0	0	1	1	1	1	1	1	1	1
NFKBIE	1	1	1	1	1	1	0	1	1	1
BCL3	1	1	1	0	1	1	1	1	1	1
CHUK	1	1	1	1	1	1	1	1	1	1
IKBKB	1	1	1	1	1	1	1	1	1	1
IKBKG	1	1	1	1	1	1	1	1	1	1

To further investigate the structural diversity of NF-κB related genes in channel catfish, we searched exon/intron organization within each family. As shown in Figure 3, NF-κB family members had 7-22 introns, and *RELA* had a relatively long intron length. IκB family members had 5-9 introns, and *BCL3* had a much longer intron than other IκB members. IKK complex members had 12-20 introns which were considerably longer than that in any members of the other two families. To measure the selective pressure on NF-κB related genes, pairwise estimation of d_N/d_S values was obtained by the method of maximum likelihood with genes from human, mouse, cow, frog, zebrafish and channel catfish. All d_N/d_S values were much smaller than 1, indicating that negative selection is the major selection pressure in the evolution of the NF-κB related genes (Table 5-16).

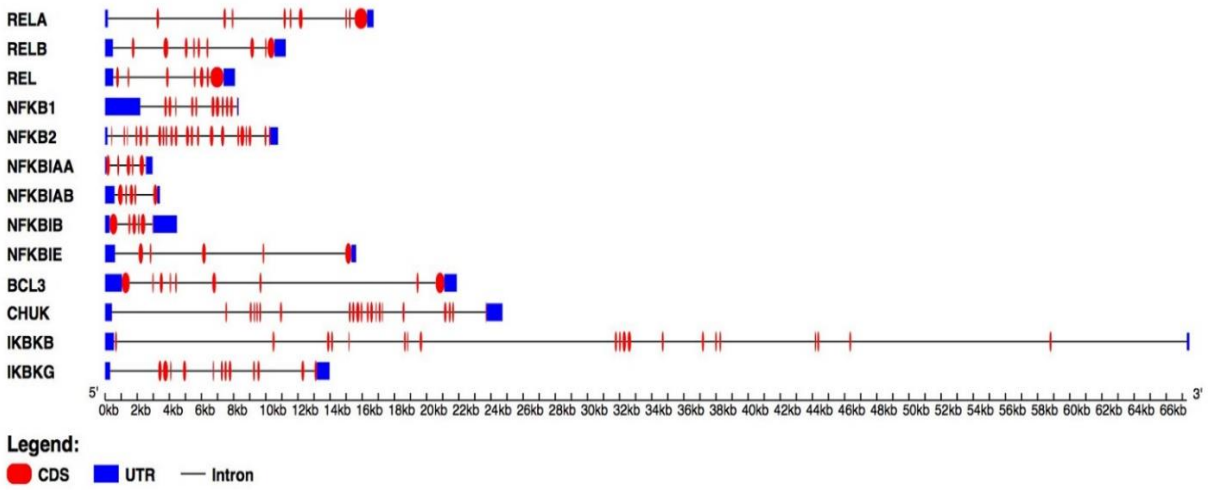

Figure 3. Structural analysis of NF-κB related genes. The exons are represented by red round-cornered rectangles. The untranslated regions (UTRs) are represented by blue rectangles. The black lines connecting two exons represent introns.

Table 5. Maximum likelihood (ML) estimates of Dn/Ds values of *RELA* gene.

RELA	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.08					
cow	0.07	0.06				
frog	0.09	0.09	0.09			
zebrafish	0.15	0.12	0.11	0.12		
catfish	0.12	0.12	0.12	0.13	0.13	

Table 6. Maximum likelihood (ML) estimates of Dn/Ds values of *RELB* gene.

RELB	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.02					
cow	0.02	0.03				
frog	0.12	0.12	0.12			
zebrafish	0.15	0.16	0.15	0.28		
catfish	0.15	0.16	0.15	0.18	0.18	

Table 7. Maximum likelihood (ML) estimates of Dn/Ds values of *REL* gene.

REL	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.17					
cow	0.24	0.15				
frog	0.11	0.11	0.11			
zebrafish	0.14	0.22	0.13	0.13		
catfish	0.14	0.16	0.13	0.14	0.18	

Table 8. Maximum likelihood (ML) estimates of Dn/Ds values of *NFKB1* gene.

NFKB1	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.08					
cow	0.07	0.06				
frog	0.09	0.09	0.09			
zebrafish	0.15	0.12	0.12	0.12		
catfish	0.12	0.12	0.12	0.13	0.13	

Table 9. Maximum likelihood (ML) estimates of Dn/Ds values of *NFKB2* gene.

NFKB2	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.07					
cow	0.11	0.09				
frog	0.19	0.22	0.22			
zebrafish	0.24	0.27	0.25	0.32		
catfish	0.25	0.26	0.28	0.23	0.1	

Table 10. Maximum likelihood (ML) estimates of Dn/Ds values of *NFKBIA* gene.

NFKBIA	human	mouse	cow	frog	zebrafish-*NFKBIAA*	zebrafish-*NFKBIAB*	catfish-*NFKBIAA*	catfish-*NFKBIAB*
human								
mouse	0.06							
cow	0.06	0.05						
frog	0.07	0.07	0.07					
zebrafish-*NFKBIAA*	0.11	0.11	0.11	0.12				
zebrafish-*NFKBIAB*	0.11	0.11	0.11	0.12	0.09			
catfish-*NFKBIAA*	0.12	0.11	0.12	0.13	0.07	0.09		
catfish-*NFKBIAB*	0.12	0.12	0.12	0.12	0.1	0.06	0.09	

Table 11. Maximum likelihood (ML) estimates of Dn/Ds values of *NFKBIB* gene.

NFKBIB	human	mouse	cow	zebrafish	catfish
human					
mouse	0.14				
cow	0.11	0.14			
zebrafish	0.56	0.79	0.7		
catfish	0.71	0.78	0.53	0.21	

Table 12. Maximum likelihood (ML) estimates of Dn/Ds values of *NFKBIE* gene.

NFKBIE	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.15					
cow	0.1	0.09				
frog	0.13	0.25	0.19			
zebrafish	0.25	0.29	0.19	0.39		
catfish	0.28	0.21	0.19	0.28	0.18	

Table 13. Maximum likelihood (ML) estimates of Dn/Ds values of *BCL3* gene.

BCL3	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.07					
cow	0.04	0.06				
frog	0.14	0.25	0.21			
zebrafish	0.1	0.11	0.11	0.2		
catfish	0.1	0.18	0.11	0.27	0.24	

Table 14. Maximum likelihood (ML) estimates of Dn/Ds values of *CHUK* gene.

CHUK	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.06					
cow	0.05	0.05				
frog	0.06	0.04	0.03			
zebrafish	0.14	0.14	0.13	0.19		
catfish	0.08	0.11	0.06	0.18	0.06	

Table 15. Maximum likelihood (ML) estimates of Dn/Ds values of *IKBKB* gene.

IKBKB	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.04					
cow	0.03	0.03				
frog	0.1	0.08	0.04			
zebrafish	0.07	0.14	0.07	0.18		
catfish	0.07	0.07	0.07	0.06	0.07	

Table 16. Maximum likelihood (ML) estimates of Dn/Ds values of *IKBKG* gene.

IKBKG	human	mouse	cow	frog	zebrafish	catfish
human						
mouse	0.04					
cow	0.09	0.03				
frog	0.21	0.23	0.15			
zebrafish	0.17	0.31	0.34	0.52		
catfish	0.3	0.41	0.34	0.43	0.08	

3.1.2 Phylogenetic and syntenic analysis of catfish NF-κB related genes

The channel catfish NF-κB related genes were annotated through phylogenetic and syntenic analysis. Phylogenetic trees were constructed for each of the three families. In cases where phylogenetic analysis alone failed to provide sufficient evidence for the gene annotation, syntenic analysis was also conducted with the sequences of channel catfish, human and zebrafish.

Phylogenetic analysis of NF-κB family genes is shown in Figure 4. Catfish *RELA*, *RELB*, *REL* and *NFKB1* genes were grouped with their respective counterparts in other species, and most closely related with the corresponding zebrafish genes. However, catfish *NFKB2* gene was not clustered with zebrafish *NFKB2* gene, requiring additional analysis for its annotation. Syntenic analysis revealed conservation of the syntenic blocks, suggesting proper annotation of the *NFKB2* gene (Figure 7A).

Phylogenetic analysis of IκB family genes was grouped into five clades as expected from the phylogenetic relationship with various organisms (Figure 5). All catfish IκB family genes except *BCL3* were most closely grouped with their respective counterparts in zebrafish. Further, syntenic analysis was conducted to provide additional evidence for the annotation of *BCL3*. As shown in Figure 7B, the catfish *BCL3* was located within the conserved syntenic block as that of the zebrafish *BCL3* gene, suggesting its proper annotation.

As shown in Figure 6, phylogenetic analysis placed every IKK complex gene into its expected phylogenetic cluster with other organisms. Catfish *CHUK*, *IKBKB* and *IKBKG* were most closely associated with their respective zebrafish genes.

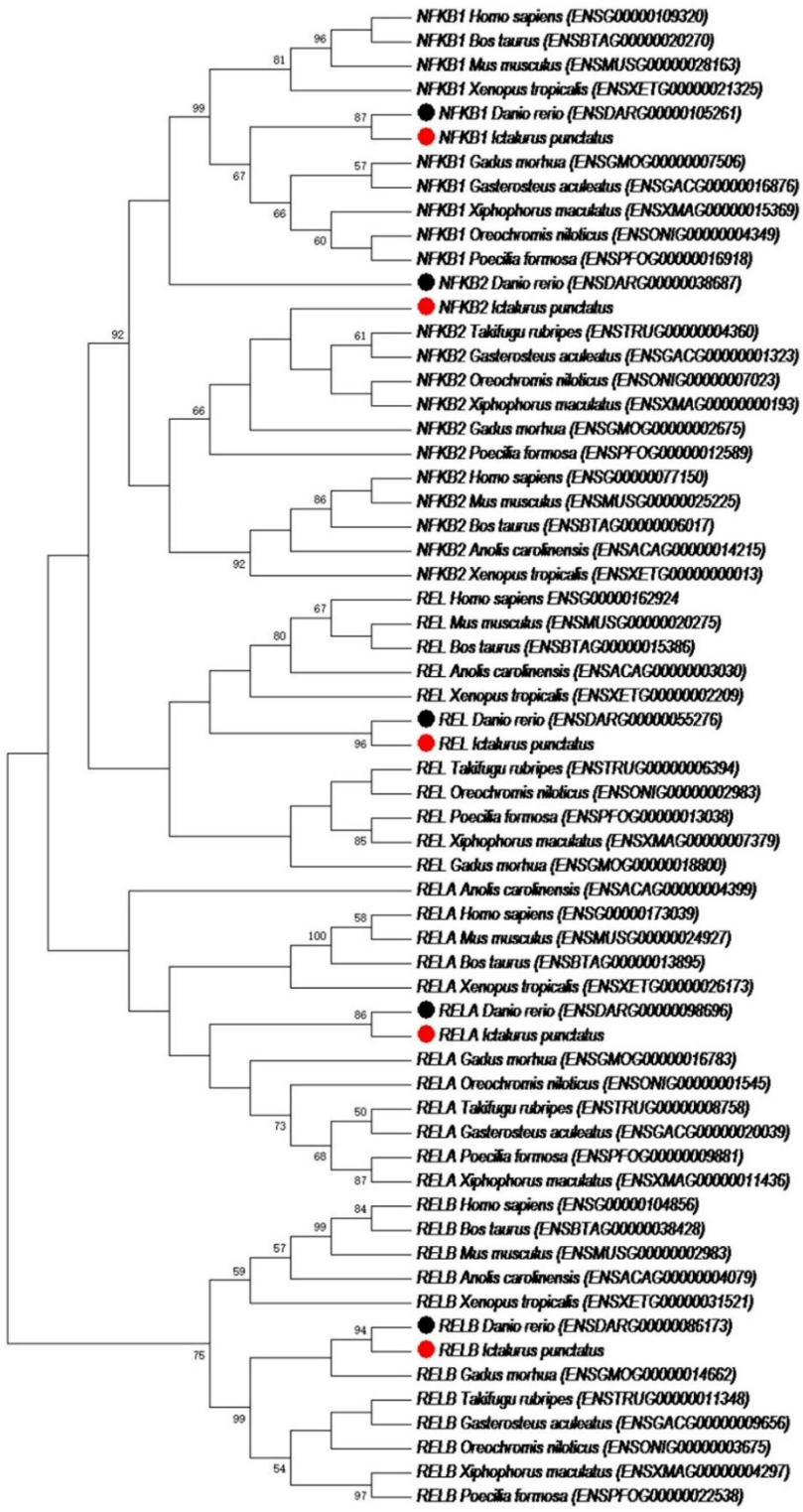

Figure 4. Phylogenetic analysis of NF-κB family genes. The phylogenetic tree was constructed with amino acid sequences from selected fish and tetrapod species using maximum likelihood

method with MEGA 6.0 software. Bootstrap values (≥ 50) were indicated by numbers at the nodes. The accession numbers were listed following the species names. Black dots indicated zebrafish NF-κB family genes, while red dots indicated channel catfish NF-κB family genes.

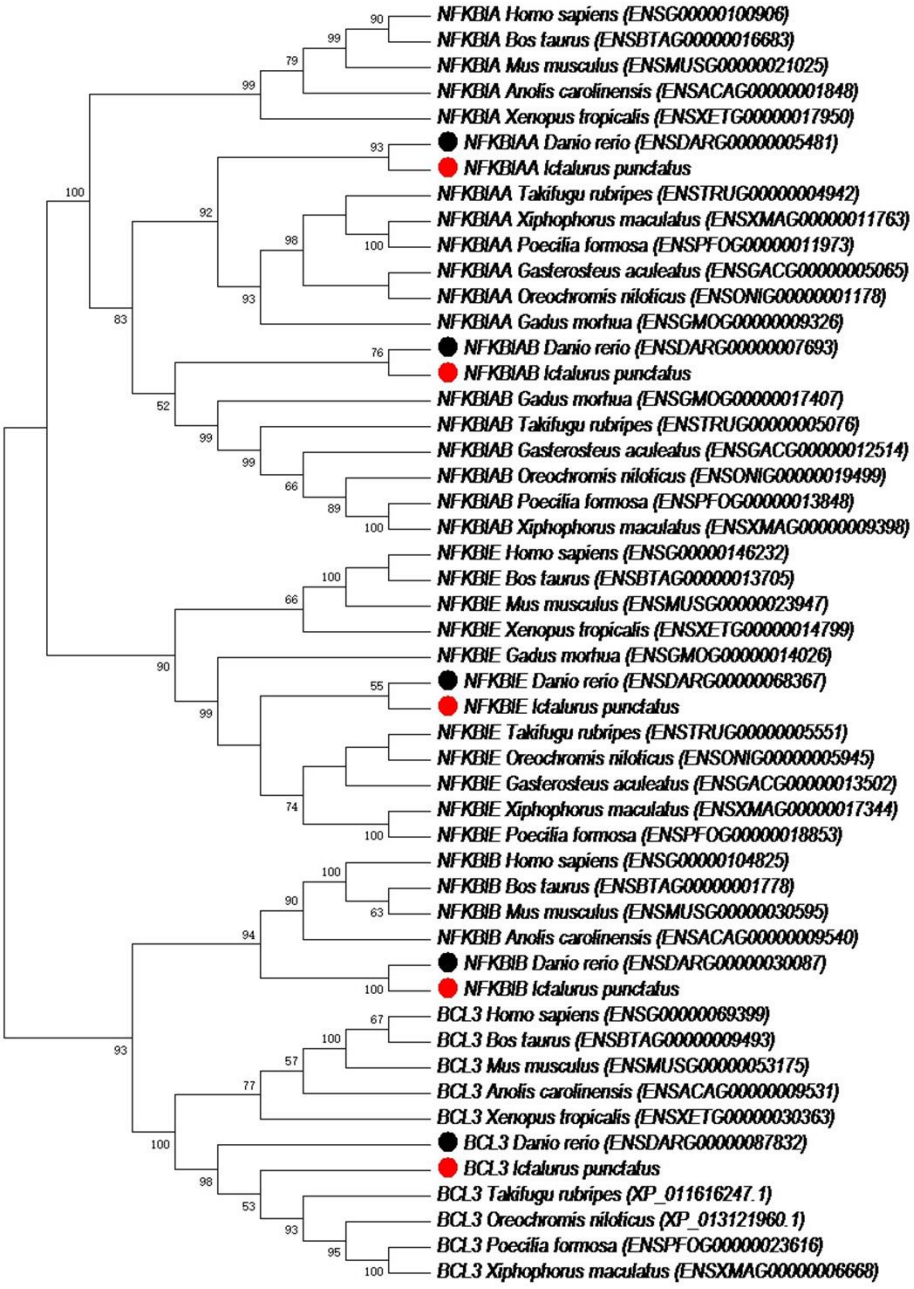

Figure 5. Phylogenetic analysis of IκB family genes. The phylogenetic tree was constructed with amino acid sequences from selected fish and tetrapod species using maximum likelihood method

with MEGA 6.0 software. Bootstrap values (≥ 50) were indicated by numbers at the nodes. The accession numbers were listed following the species names. Black dots indicated zebrafish IκB family genes, while red dots indicated channel catfish IκB family genes.

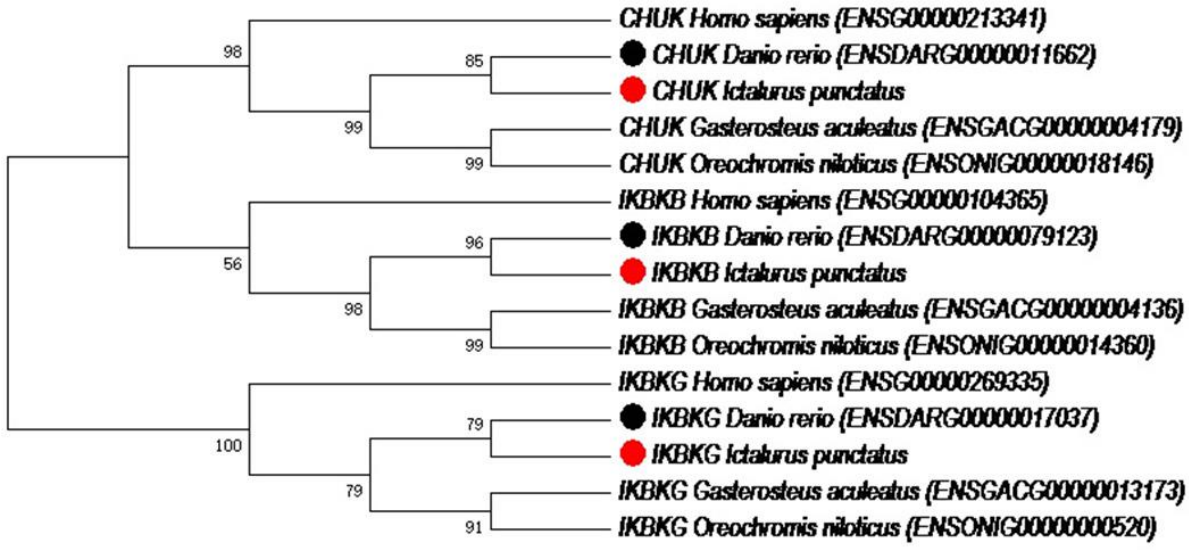

Figure 6. Phylogenetic analysis of IKK family genes. The phylogenetic tree was constructed with amino acid sequences from selected fish and tetrapod species using maximum likelihood method with MEGA 6.0 software. Bootstrap values (≥ 50) were indicated by numbers at the nodes. The accession numbers were listed following the species names. Black dots indicated zebrafish IKK family genes, while red dots indicated channel catfish IKK family genes.

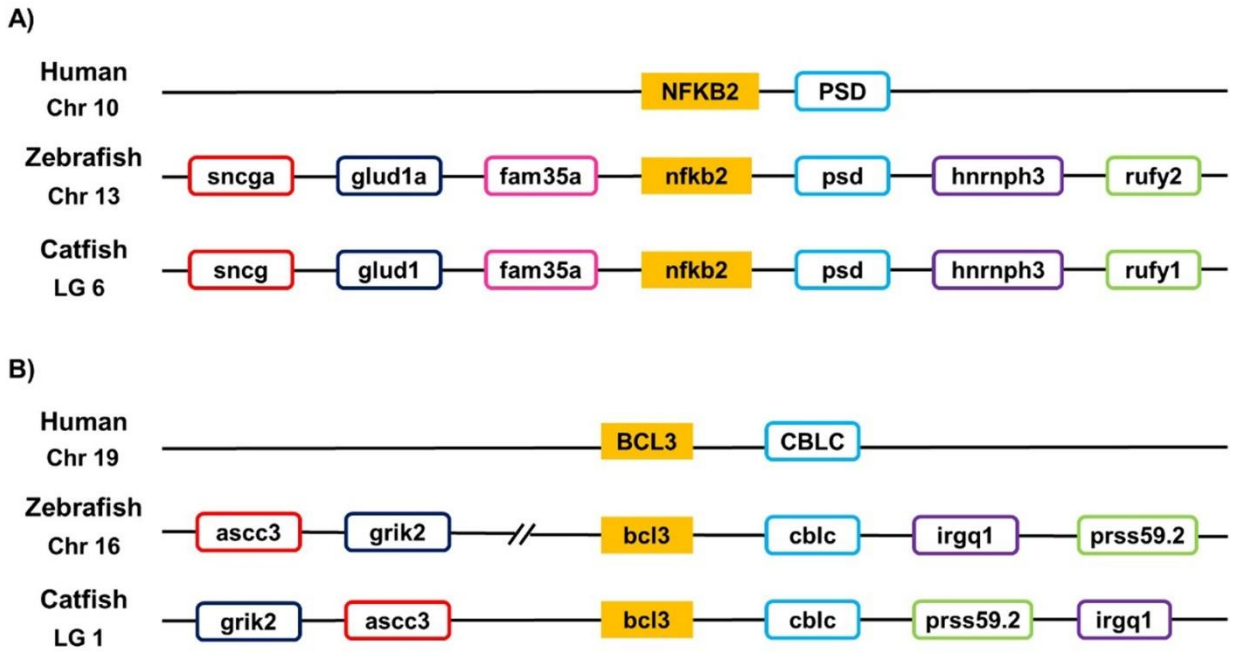

Figure 7. Syntenic analysis of (A) NFKB2 and (B) BCL3 genes in human, zebrafish and channel catfish. Syntenies were constructed based on the genomic information from Genomicus. Gene abbreviations are the following: ascc3 (activating signal cointegrator 1 complex subunit 3), cblc (Cbl proto-oncogene C), fam35a (family with sequence similarity 35, member A), glud1 (glutamate dehydrogenase 1), glud1a (glutamate dehydrogenase 1 a), grik2 (Glutamate Receptor, Ionotropic, Kainate 2), hnrnph3 (heterogeneous nuclear ribonucleoprotein H3), irgq1 (immunity-related GTPase family, q1), prss59.2 (protease, serine, 59, tandem duplicate 2), psd (pleckstrin and Sec7 domain containing), rufy1 (RUN and FYVE domain containing 1), rufy2 (RUN and FYVE domain containing 2), sncg (synuclein, gamma), sncga (synuclein, gamma a).

3.1.3 Expression of NF-κB related genes after columnaris infection

The expression profiles of catfish NF-κB related genes were determined in three mucosal tissues of gill, skin and intestine after *F. columnare* infection. The levels of gene expression were determined of all 13 NF-κB related genes at 0 h and 8 h post-infection (Figure 8). In the gill at 8 h

after *F. columnare* infection, significant down-regulation was observed in three members of NF-κB family (*RELA*, *RELB* and *NFKB1*) and one member of IKK complex (*IKBKB*); while significant up-regulation was observed in three members of IκB family (*NFKBIAA*, *NFKBIAB* and *BCL3*) (Figure 8A). The extent of expression differences of these genes was between 1.2 to 3.1-fold changes. Surprisingly, most of NF-κB related genes were significantly induced in the skin at 8 h after *F. columnare* infection including four members of NF-κB family (*RELA*, *RELB*, *NFKB1* and *NFKB2*), three members of IκB family (*NFKBIB*, *NFKBIE* and *BCL3*), and one member of IKK complex (*IKBKB*) (Figure 8B). Two genes (*BCL3* and *IKBKB*) were observed to be dramatically induced with greater than 5-fold changes at 8 h post-infection, while the remaining genes were induced 2 - 4 folds. In the intestine at 8 h after bacterial infection, four members of NF-κB family (*RELB*, *REL*, *NFKB1* and *NFKB2*), three members of IκB family (*NFKBIAA*, *NFKBIAB* and *NFKBIE*), and two members of IKK complex (*CHUK* and *IKBKG*) were significantly induced with 2- to 5-fold changes (Figure 8C).

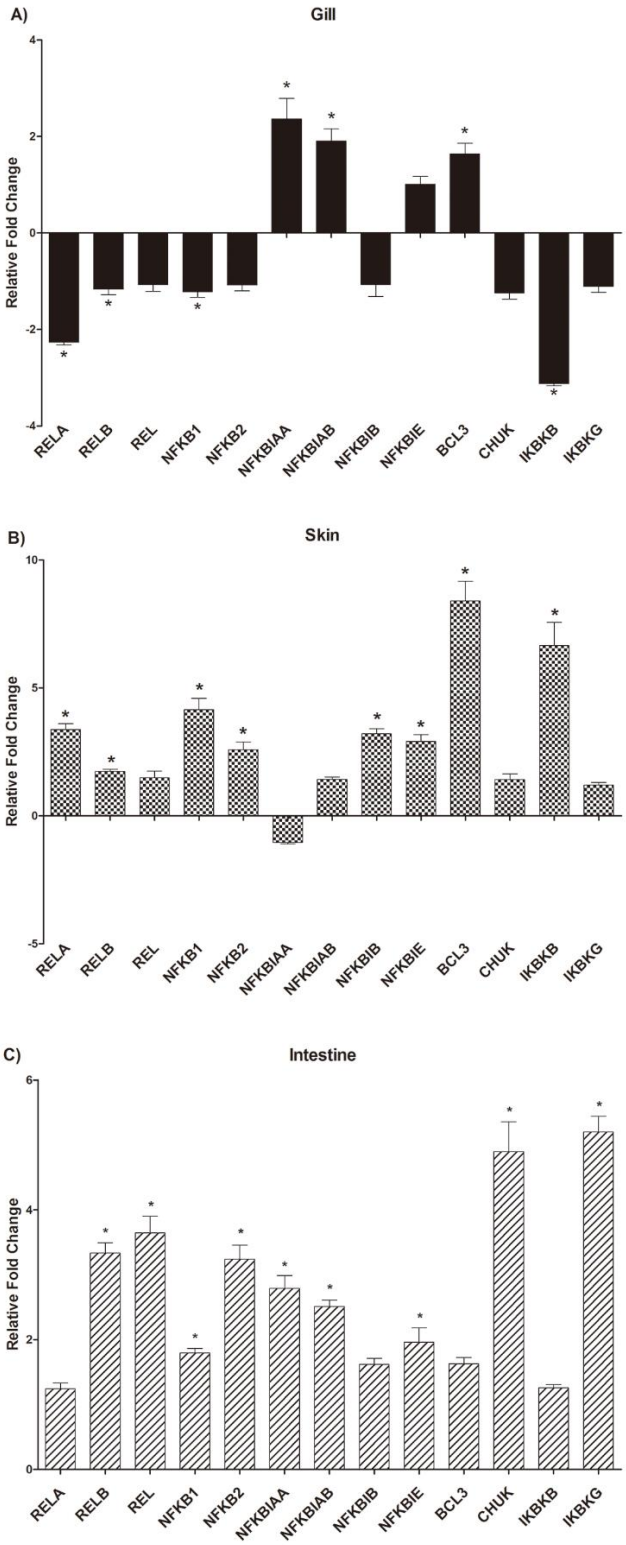

Figure 8. Expression of NF-κB related genes following *Flavobacterium columnare* infection.

Quantitative real-time PCR was used for the expression analysis in three mucosal tissues: (A) gill,

(B) skin and (C) intestine at 8 h post-infection. Fold change of expression as compared with control fish is presented on the Y-axis, 18S rRNA and 28S rRNA were used as internal reference. The results were expressed as mean ± standard error (bars). Significant differences ($p \leq 0.05$) among controls and various treatments were indicated by asterisks (*).

To determine the correlation between expression of NF-κB related genes in response to *F. columnare* infection and catfish resistance against the disease, expression of the 13 NF-κB genes was compared between resistant and susceptible catfish. In general, the NF-κB family and the IKK family of genes were expressed at higher levels in resistant fish than in susceptible fish, whereas the IκB family (except *BCL3*) of genes were expressed at higher levels in susceptible fish than in resistant fish (Figure 9). However, the level and timing of expression differences cannot be generalized. For example, at 1 h post-infection, *NFKB2* and *BCL3* were induced approximately two-fold in resistant catfish; at 2 h post-infection, three members of NF-κB family (*RELA*, *NFKB1* and *NFKB2*) and one member of IκB family (*BCL3*) were induced in resistant catfish; at 8 h post-infection, two members of NF-κB family (*RELA* and *NFKB1*) were induced approximately two-fold in resistant catfish. Taken together, these results suggested that higher expression of NF-κB and IKK genes may be correlated with disease resistance, while higher expression of IκB genes (except *BCL3*) may be correlated with disease susceptibility.

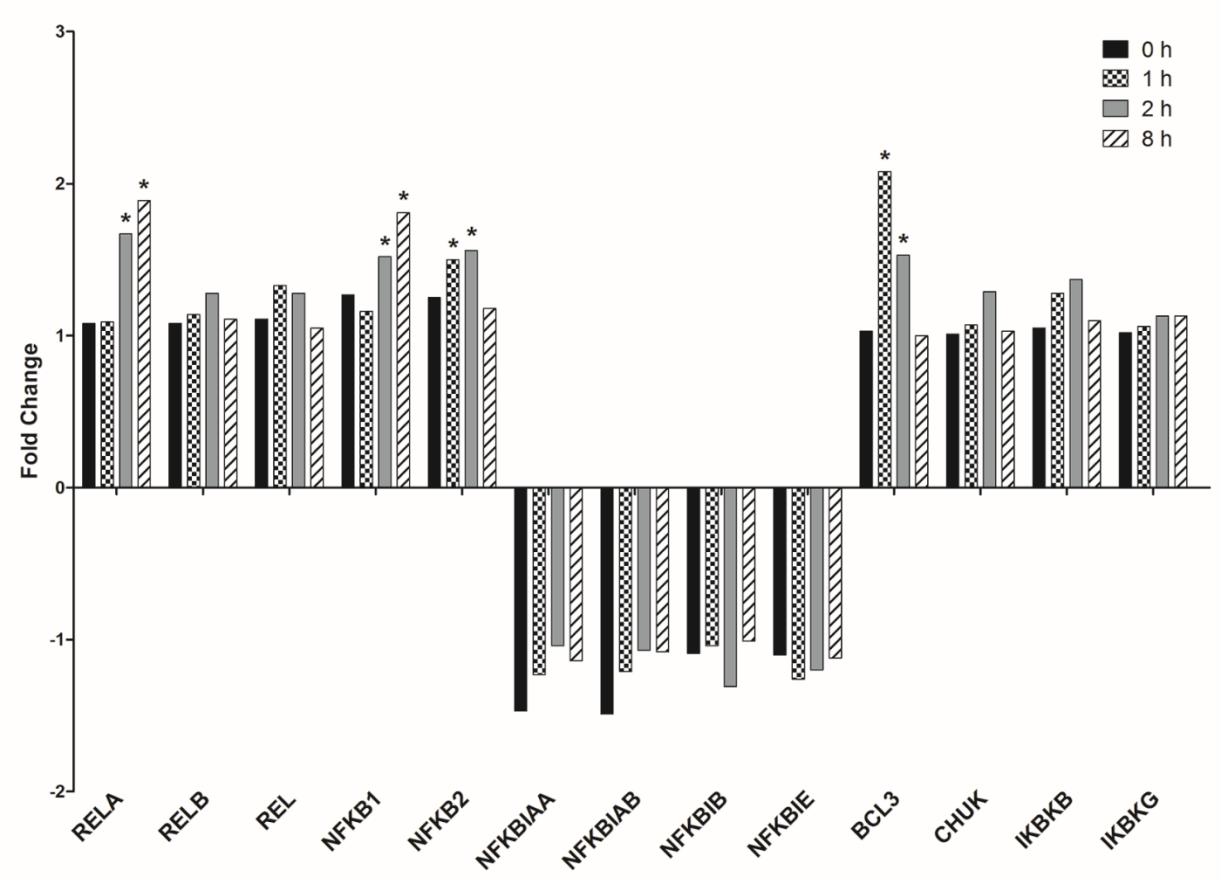

Figure 9. Expression of NF-κB related genes following *Flavobacterium columnare* infection in the gill of channel catfish with different susceptibility (resistant vs susceptible) to *F. columnare* at different time points. Asterisks (*) indicated significant differences (p-value ≤ 0.05, expression fold change ≥ 1.5).

3.2 Discussion

Innate immunity is the first line of defense against invading pathogens. After TLRs detect the microbial components, they recruit adaptor proteins to active the NF-κB pathway (Kawai and Akira, 2007). The interaction of NF-κB, IκB and IKK subunits is a negative feedback-containing signal-transduction pathway, which regulates the expression of an array of inflammatory cytokine genes (Baeuerle and Henkel, 1994; Pahl, 1999). Given the close interaction of NF-κB, IκB and

IKK families, this study included genes from all three families to have a better understanding of their immune functions in catfish. We identified a total of 13 NF-κB related genes in the channel catfish genome including 5 NF-κB family genes, 5 IκB family genes, and 3 IKK complex genes. Similar exon/intron structural features were shared by members within each family. Negative selection patterns were identified in these genes for maintaining their conserved functions. Distinct expression patterns of NF-κB related genes were observed after *F. columnare* infection in three mucosal tissues, as well as in resistant and susceptible catfish, indicating their involvement in immune response to bacterial infection and complex function mechanisms in host defense system. As the first systematic work of analyzing NF-κB related genes in catfish, we provided useful genomic resources, evolutionary patterns and expression profiles for future research in catfish and in other teleost fish.

The annotation of NF-κB related genes was relatively straightforward due to their conservancy during gene evolution. All mammalian NF-κB related genes were identified in channel catfish except for *NFKBID* that is also missing in the zebrafish genome. It is highly likely that *NFKBID* was not present in the channel catfish genome as well (Blomme et al., 2006). Interestingly, both catfish and zebrafish missed *NFKBID*, but they both harbor *NFKBIB* that are not present in any other teleost fish species. This may suggest that *NFKBIB* in catfish and zebrafish may play similar roles as *NFKBID* in other teleost fish species.

Phylogenetic analysis provided sufficient evidence for the identities of most catfish NF-κB related genes as they group with their respective counterparts in higher vertebrates and other fish species. Two genes (*NFKB2* and *BCL3*) required additional syntenic analysis, and their conserved syntenies confirmed our proper annotations.

Copy numbers of NF-κB related genes were compared between catfish and other tetrapods. For the majority of NF-κB related genes, only one copy exists in various tetrapods except for *NFKB1* and *NFKBIE*, which are not found in lizard, and *NFKBIB*, which is not found in frog. All NF-κB related genes in fish species have only a single copy except for *NFKBIA*. Two copies of *NFKBIA* (*NFKBIAA* and *NFKBIAB*) exist in teleost fish, indicating teleost-specific whole genome duplication (Hoegg et al., 2004; Thornton and DeSalle, 2000). It is interesting that most of the NF-κB related genes derived from the whole genome duplication were lost, leading to the current single copy genes. This may suggest that gene dosage of these genes is crucially important such that most of them were lost after whole genome duplication.

Identification of selection pressure of NF-κB related genes during the evolution was conducted using maximum likelihood model to compare nonsynonymous (d_N) and synonymous (d_S) substitutions rates (Yang and Bielawski, 2000). The ratio d_N/d_S of each NF-κB related gene was significantly less than 1 and very close to zero in some cases, indicating that these genes were globally strongly constrained by purifying or negative selection. Negative selection is believed to be crucial to maintain the functional stability and improved structures of gene family by removing deleterious mutations (Charlesworth et al., 1995). Because of the essential roles of NF-κB related genes in various biological processes, negative selection eliminated less-adapted variants of these genes to ensure their proper functions.

Conserved domains were identified for understanding the functional structures of NF-κB related genes. All catfish NF-κB family proteins shared two same domains IPT and RHD, while NFκB1 and NFκB2 contained two extra domains ANK and DD. RHD is crucial for NF-κB family because of its function in DNA-binding, dimerization, nuclear localization signal and interaction with IκBs (Ghosh et al., 1995; Muller et al., 1995). While NFκB1 and NFκB2 can act as inhibitors

like IκBs by combining with other NF-κB proteins due to the presence of ANK repeat domains (Baldwin, 1996). All catfish IκBs contained ANK repeats and their inner helices that were reported to be involved in the interaction of IκBs with NF-κB proteins (Mosavi et al., 2004). For catfish IKK complex, IKKα and IKKβ shared PKc domain for their function as catalytically active kinases, while the regulatory subunit IKKγ contained unique domain NEMO for binding with IKKβ, as reviewed by (Oeckinghaus and Ghosh, 2009).

NF-κB related genes have been widely reported to play key roles in regulating the immune response to bacterial infection in mammals (Johannessen et al., 2013; Le Negrate, 2012). To determine if these genes play a similar role in catfish immune response, we analyzed the expression profiles of all thirteen NF-κB related genes after *F. columnare* infection in three mucosal tissues. Dramatic expression pattern differences of the NF-κB related genes were observed among three mucosal tissues at 8 h post-infection. In gill, all members of NF-κB family and IKK family were generally suppressed, while most members of IκB family were generally induced after bacterial infection (Figure 8A). This may suggest the inhibition of NF-κB pathway in gill by *F. columnare* invasion, which validates our finding in previous RNA-Seq study (Sun et al., 2012). In skin, *IKBKB* and *BCL3* were extremely induced along with the upregulation of several members of NF-κB family and IκB family (Figure 8B). As the significant upregulation of most members of NF-κB family and IKK family were detected, it seems that NF-κB pathway was activated in the skin after *F. columnare* infection. Although some members of IκB family were also induced, this may be caused by the NF-κB-induced feedback regulation (Vallabhapurapu and Karin, 2009). In intestine, most members of NF-κB family and IKK family were extremely induced after *F. columnare* infection, while members of IκB family were also observed to be induced due to the NF-κB-induced feedback regulation as the situation in skin (Figure 8C). Notably, different NF-κB

related genes were significantly regulated in three mucosal tissues, which may suggest potential tissue-specific regulation preference of different NF-κB related genes in response to bacterial infection.

The difference in expression patterns of NF-κB related genes between resistant and susceptible catfish was determined by analyzing an existing RNA-Seq dataset after *F. columnare* infection. *RELA*, *NFKB1* and *NFKB2* from NF-κB family were identified to be significantly induced in resistant catfish, but not so in susceptible catfish. These genes may contribute to the successful defense against *F. columnare* by activing NF-κB pathway and its downstream TLR pathway in the resistant catfish (Yamamoto and Takeda, 2008). Notably, *BCL3* from IκB family was also significantly induced in resistant catfish. Unlike other members of IκB family, *BCL3* can either positively or negatively modulate nuclear NF-kB activity through the exclusive interaction with p50 and p52 homodimers (Palmer and Chen, 2008; Zhang et al., 1994). For instance, *BCL3* was reported to positively regulate the NF-κB activity in response to HIV in human (Franzoso et al., 1992) and *Klebsiella pneumoniae* infection in mice (Pène et al., 2011). On the other hand, *BCL3* was considered to be a negative regulator to NF-κB (Carmody et al., 2007). Given the success of the immune response of the resistant catfish against *F. columnare* invasion, *BCL3* may respond positively by promoting the NF-κB activity to defend the host against *F. columnare* in resistant catfish.

In most cases, regulation of NF-κB, IκB and IKK family members has been believed at the translational levels. Although expressed in most cells, NF-κB dimers are kept inactivated by binding with IκB proteins until the activation of IKKs (Oeckinghaus and Ghosh, 2009). However, RelB protein level was reported to be transcriptionally regulated in a RelA dependent manner, leading to its delay in response to stimuli in human cells (Bren et al., 2001). Our results suggest

that NF-κB related genes are also regulated at the transcriptional level with the continued bacterial infection. Clearly, the regulation of NF-κB related genes could be quite complex regarding the tissue specificity, species susceptibility to diseases, and specific genes under study.

Chapter 4: Conclusion

In conclusion, 5 NF-κB family genes, 5 IκB family genes, and 3 IKK complex genes were identified and characterized in channel catfish genome. The annotation of these genes was verified by conducting phylogenetic analysis with other species using maximum likelihood methods and the bootstrap hypothesis testing. Based on the results of likelihood ratio test, negative selection was identified to play a key role in maintaining the conserved structure and function of these genes. Taken together, genes from NF-κB family, IκB family and IKK complex have high evolutional conservancy in their sequences, functional domains, and copy numbers. This high conservancy from fish to human indicated functional importance of these NF-κB related genes in biological processes.

Randomization test was conducted to compare the expression levels of three mucosal tissues (gill, intestine and skin) at 0h and 8h after columnaris infection. Expression of most of NF-κB related genes were significantly regulated in these three mucosal tissues after *F. columnare* infection, indicating their involvement in immune response to bacterial infection. Kal's Z test was conducted to analyze the previous RNA-Seq dataset with FDR correction. The results indicated a dramatical variation in the expression patterns of NF-κB related genes between susceptible catfish and resistant catfish after *F. columnare* infection, further suggesting their contribution to the variations in disease resistance/susceptibility of catfish.

This thesis applied various statistical methods such as maximum likelihood methods and hypothesis testing to identify the NF-κB related genes in the catfish genome and their regulation in expression patterns after bacterial infection. The results of this study provide insight into the evolutional conservancy of biological function of these NF-κB related genes among different species, and suggesting their important roles in host defense responses. In particular, the significant variation of expression levels of NF-κB related genes existing between susceptible catfish and resistant catfish may provide a foundation for further genetic improvement of disease resistance in catfish.

As a core component of NF-κB pathway, the interaction of NF-κB proteins, IκB proteins, and IKK complexes still require further statistical analysis for a more detailed understanding of how they respond to bacterial infections and their roles in the host defense.

References

Anisimova, M., Bielawski, J.P., Yang, Z., 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution 18, 1585-1592.

Baeuerle, P.A., Baltimore, D., 1988. Activation of DNA-Binding Activity in an Apparently Cytoplasmic Precursor of the Nf-Kappa-B Transcription Factor. Cell 53, 211-217.

Baeuerle, P.A., Henkel, T., 1994. Function and activation of NF-kappaB in the immune system. Annual Review of Immunology 12, 141-179.

Baldwin, A.S., 1996. The NF-kappa B and I kappa B proteins: New discoveries and insights. Annual Review of Immunology 14, 649-683.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Journal of the Royal Statistical Society: Series B (Methodological) 57, 289-300.

Bielawski, J.P., Yang, Z., 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. Journal of Structural and Functional Genomics 3, 201-212.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., Van de Peer, Y., 2006. The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biologyogy 7, R43.

Bren, G.D., Solan, N.J., Miyoshi, H., Pennington, K.N., Pobst, L.J., Paya, C.V., 2001. Transcription of the RelB gene is regulated by NF-kappa B. Oncogene 20, 7722-7733.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., Jaffe, D.B., 2008. Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Research 18, 763-770.

Bush, R.M., 2001. Predicting adaptive evolution. Nature Reviews Genetics 2, 387-392.

Cabin, R.J., Mitchell, R.J., 2000. To Bonferroni or not to Bonferroni: when and how are the questions. Bulletin of the Ecological Society of America 81, 246-248.

Carmody, R.J., Ruan, Q., Palmer, S., Hilliard, B., Chen, Y.H., 2007. Negative regulation of toll-like receptor signaling by NF-κB p50 ubiquitination blockade. Science 317, 675-678.

Casella, G., Berger, R.L., 2002. Statistical inference. Duxbury Pacific Grove, CA.

Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. Models and estimation procedures. American Journal of Human Genetics 19, 233-257.

Charlesworth, D., Charlesworth, B., Morgan, M.T., 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141, 1619-1632.

Chen, Z.J., Parent, L., Maniatis, T., 1996. Site-specific phosphorylation of IκBα by a novel ubiquitination-dependent protein kinase activity. Cell 84, 853-862.

Comeron, J.M., 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. Journal of Molecular Evolution 41, 1152-1159.

Correa, R.G., Tergaonkar, V., Ng, J.K., Dubova, I., Izpisua-Belmonte, J.C., Verma, I.M., 2004. Characterization of NF-κB/IκB proteins in zebra fish and their involvement in notochord development. Molecular and Cellular Biology 24, 5257-5268.

Czelusniak, J., Goodman, M., Moncrief, N.D., Kehoe, S.M., 1990. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. Methods in Enzymology 183, 601-615.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797.

Edwards, A.W., Sforza, C.L., 1963. The reconstruction of evolution. Heredity 18, 553.

Efron, B., 1982. The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadephia, PA.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17, 368-376.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783-791.

Felsenstein, J., 1988. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22, 521-565.

Franzoso, G., Bours, V., Park, S., Tomfta-Yamaguchi, M., Kelly, K., Siebenlist, U., 1992. The candidate oncoprotein Bcl-3 is an antagonist of pSO/NF-κB-mediated inhibition. Nature 359, 339-342.

Gerondakis, S., Grossmann, M., Nakamura, Y., Pohl, T., Grumont, R., 1999. Genetic approaches in mice to understand Rel/NF-κB and IκB function: transgenics and knockouts. Oncogene 18, 6888-6895.

Ghosh, G., Vanduyne, G., Ghosh, S., Sigler, P.B., 1995. Structure of Nf-Kappa-B P50 homodimer bound to a Kappa-B site. Nature 373, 303-310.

Goldman, N., Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11, 725-736.

Grilli, M., Chiu, J.J.-S., Lenardo, M.J., 1993. IMF-κB and Rel: Participants in a Multiform Transcriptional Regulatory System. International Review of Cytology 143, 1-62.

Hawke, J.P., Mcwhorter, A.C., Steigerwalt, A.G., Brenner, D.J., 1981. Edwardsiella-Ictaluri Sp-Nov, the Causative Agent of Enteric Septicemia of Catfish. International Journal of Systematic and Evolutionary Microbiology 31, 396-400.

Hoegg, S., Brinkmann, H., Taylor, J.S., Meyer, A., 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. Journal of Molecular Evolution 59, 190-203.

Hoffmann, A., Levchenko, A., Scott, M.L., Baltimore, D., 2002. The IκB-NF-κB signaling module: temporal control and selective gene activation. Science 298, 1241-1245.

Israel, A., 2010. The IKK Complex, a Central Regulator of NF-kappa B Activation. Cold Spring Harbor Perspectives in Biology 2, a000158.

Johannessen, M., Askarian, F., Sangvik, M., Sollid, J.E., 2013. Bacterial interference with canonical NFκB signalling. Microbiology 159, 2001-2013.

Kal, A.J., van Zonneveld, A.J., Benes, V., van den Berg, M., Koerkamp, M.G., Albermann, K., Strack, N., Ruijter, J.M., Richter, A., Dujon, B., 1999. Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. Molecular Biology of the Cell 10, 1859-1872.

Karin, M., 1999. How NF-κB is activated: the role of the IκB kinase (IKK) complex. Oncogene 18, 6867-6874.

Kishino, H., Hasegawa, M., 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. Journal of Molecular Evolution 29, 170-179.

Kishino, H., Miyata, T., Hasegawa, M., 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution 31, 151-160.

LaFrentz, B.R., Shoemaker, C.A., Booth, N.J., Peterson, B.C., Ourth, D.D., 2012. Spleen index and mannose-binding lectin levels in four channel catfish families exhibiting different susceptibilities to *Flavobacterium columnare* and *Edwardsiella ictaluri*. Journal of Aquatic Animal Health 24, 141-147.

Le Negrate, G., 2012. Subversion of innate immune responses by bacterial hindrance of NF-κB pathway. Cellular Microbiology 14, 155-167.

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.

Li, W.-H., 1989. A statistical test of phylogenies estimated from sequence data. Molecular Biology and Evolution 6, 424-435.

Li, Y., Liu, S.K., Qin, Z.K., Yao, J., Jiang, C., Song, L., Dunham, R., Liu, Z.J., 2015. The serpin superfamily in channel catfish: Identification, phylogenetic analysis and expression profiling in mucosal tissues after bacterial infections. Developmental & Comparative Immunology 49, 267-277.

Liu, S., Zhang, Y., Zhou, Z., Waldbieser, G., Sun, F., Lu, J., Zhang, J., Jiang, Y., Zhang, H., Wang, X., 2012. Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. BMC Genomics 13, 595.

Liu, Z.J., Liu, S.K., Yao, J., Bao, L.S., Zhang, J.R., Li, Y., Jiang, C., Sun, L.Y., Wang, R.J., Zhang, Y., Zhou, T., Zeng, Q.F., Fu, Q., Gao, S., Li, N., Koren, S., Jiang, Y.L., Zimin, A., Xu, P., Phillippy, A.M., Geng, X., Song, L., Sun, F.Y., Li, C., Wang, X.Z., Chen, A.L., Jin, Y.L., Yuan, Z.H., Yang, Y.J., Tan, S.X., Peatman, E., Lu, J.G., Qin, Z.K., Dunham, R., Li, Z.X., Sonstegard, T., Feng, J.B., Danzmann, R.G., Schroeder, S., Scheffler, B., Duke, M.V., Ballard, L., Kucuktas, H., Kaltenboeck, L., Liu, H.X., Armbruster, J., Xie, Y.J., Kirby, M.L., Tian, Y., Flanagan, M.E., Mu, W.J., Waldbieser, G.C., 2016. The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. Nature Communications 7, 11757.

Louis, A., Nguyen, N.T.T., Muffato, M., Crollius, H.R., 2015. Genomicus update 2015: KaryoView and MatrixView provide a genome-wide perspective to multispecies comparative genomics. Nucleic Acids Research 43, 682-689.

Miyata, T., Yasunaga, T., 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. Journal of Molecular Evolution 16, 23-36.

Mosavi, L.K., Cammett, T.J., Desrosiers, D.C., Peng, Z.Y., 2004. The ankyrin repeat as molecular architecture for protein recognition. Protein Science 13, 1435-1448.

Muffato, M., Louis, A., Poisnel, C.E., Crollius, H.R., 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26, 1119-1121.

Muller, C.W., Rey, F.A., Sodeoka, M., Verdine, G.L., Harrison, S.C., 1995. Structure of the Nf-Kappa-B P50 homodimer bound to DNA. Nature 373, 311-317.

Muse, S.V., 1996. Estimating synonymous and nonsynonymous substitution rates. Molecular Biology and Evolution 13, 105-114.

Nei, M., 1987. Molecular evolutionary genetics. Columbia university press.

Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3, 418-426.

Nei, M., Kumar, S., 2000. Molecular evolution and phylogenetics. Oxford university press.

Nei, M., Stephens, J.C., Saitou, N., 1985. Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes.

Molecular Biology and Evolution 2, 66-85.

Oeckinghaus, A., Ghosh, S., 2009. The NF-kappa B family of transcription factors and its regulation. Cold Spring Harbor Perspectives in Biology 1, a000034.

Oeckinghaus, A., Hayden, M.S., Ghosh, S., 2011. Crosstalk in NF-[kappa] B signaling pathways. Nature immunology 12, 695-708.

Pahl, H.L., 1999. Activators and target genes of Rel/NF-κB transcription factors. Oncogene 18, 6853-6866.

Palmer, S., Chen, Y.H.H., 2008. Bcl-3, a multifaceted modulator of NF-kappa B-mediated gene transcription. Immunology Research 42, 210-218.

Peatman, E., Baoprasertkul, P., Terhune, J., Xu, P., Nandi, S., Kucuktas, H., Li, P., Wang, S., Somridhivej, B., Dunham, R., Liu, Z., 2007. Expression analysis of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with a Gram-negative bacterium. Developmental & Comparative Immunology 31, 1183-1196.

Peatman, E., Li, C., Peterson, B.C., Straus, D.L., Farmer, B.D., Beck, B.H., 2013. Basal polarization of the mucosal compartment in *Flavobacterium columnare* susceptible and resistant channel catfish (*Ictalurus punctatus*). Molecular Immunology 56, 317-327.

Pène, F., Paun, A., Sønder, S.U., Rikhi, N., Wang, H., Claudio, E., Siebenlist, U., 2011. The IκB family member Bcl-3 coordinates the pulmonary defense against *Klebsiella pneumoniae* infection. The Journal of Immunology 186, 2412-2421.

Perkins, N.D., 2006. Post-translational modifications regulating the activity and function of the nuclear factor kappa B pathway. Oncogene 25, 6717-6730.

Pfaffl, M.W., Horgan, G.W., Dempfle, L., 2002. Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Research 30, e36.

Pomerantz, J.L., Baltimore, D., 2002. Two pathways to NF-kappa B. Molecular Cell 10, 693-695.

Pradhan, A., Khalaf, H., Ochsner, S.A., Sreenivasan, R., Koskinen, J., Karlsson, M., Karlsson, J., McKenna, N.J., Orbán, L., Olsson, P.-E., 2012. Activation of NF-κB protein prevents the transition from juvenile ovary to testis and promotes ovarian development in zebrafish. Journal of Biological Chemistry 287, 37926-37938.

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19, 368-375.

Rzhetsky, A., Nei, M., 1992a. A simple method for estimating and testing minimum-evolution trees. Molecular Biology and Evolution 9, 945-967.

Rzhetsky, A., Nei, M., 1992b. Statistical properties of the ordinary least-squares, generalized least-

squares, and minimum-evolution methods of phylogenetic inference. Journal of Molecular Evolution 35, 367-375.

Scheidereit, C., 2006. IκB kinase complexes: gateways to NF-κB activation and transcription. Oncogene 25, 6685-6705.

Schlezinger, J.J., Blickarz, C.E., Mann, K.K., Doerre, S., Stegeman, J.J., 2000. Identification of NF-κB in the marine fish *Stenotomus chrysops* and examination of its activation by aryl hydrocarbon receptor agonists. Chemico-Biological Interactions 126, 137-157.

Senftleben, U., Cao, Y., Xiao, G., Greten, F.R., Krähn, G., Bonizzi, G., Chen, Y., Hu, Y., Fong, A., Sun, S.-C., 2001. Activation by IKKα of a second, evolutionary conserved, NF-κB signaling pathway. Science 293, 1495-1499.

Shoemaker, C.A., Olivares-Fuster, O., Arias, C.R., Klesius, P.H., 2008. Flavobacterium columnare genomovar influences mortality in channel catfish (*Ictalurus punctatus*). Veterinary Microbiology 127, 353-359.

Sitnikova, T., Rzhetsky, A., Nei, M., 1995. Interior-branch and bootstrap tests of phylogenetic trees. Molecular Biology and Evolution 12, 319-333.

Solovyev, V., Kosarev, P., Seledsov, I., Vorobyev, D., 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biology 7, S10.

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

Sun, L.Y., Liu, S.K., Bao, L.S., Li, Y., Feng, J.B., Liu, Z.J., 2015. Claudin multigene family in channel catfish and their expression profiles in response to bacterial infection and hypoxia as revealed by meta-analysis of RNA-Seq datasets. Comparative Biochemistry and Physiology Part D 13, 60-69.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution 28, 2731-2739.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular Biology and Evolution, 30, 2725-2729.

Thornton, J.W., DeSalle, R., 2000. Gene family evolution and homology: genomics meets phylogenetics. Annual Review of Genomics and Human Genetics 1, 41-73.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562-578.

Vallabhapurapu, S., Karin, M., 2009. Regulation and function of NF-kappa B transcription factors in the immune system. Annual Review of Immunology 27, 693-733.

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.

Xu, B., Yang, Z., 2013. PAMLX: a graphical user interface for PAML. Molecular Biology and Evolution 30, 2723-2724.

Yamamoto, M., Takeda, K., 2008. Role of nuclear IκB proteins in the regulation of host immune responses. Journal of Infection and Chemotherapy 14, 265-269.

Yamamoto, Y., Gaynor, R.B., 2001. Role of the NF-kB pathway in the pathogenesis of human disease states. Current Molecular Medicine 1, 287-296.

Yang, Y., Fu, Q., Zhou, T., Li, Y., Liu, S., Zeng, Q., Wang, X., Jin, Y., Tian, C., Qin, Z., 2017. Analysis of apolipoprotein genes and their involvement in disease response of channel catfish after bacterial infection. Developmental & Comparative Immunology 67, 464-470.

Yang, Z., 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15, 568-573.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24, 1586-1591.

Yang, Z., Bielawski, J.P., 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution 15, 496-503.

Yang, Z., Goldman, N., Friday, A., 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Systematic Biology 44, 384-399.

Yang, Z., Nielsen, R., 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17, 32-43.

Zhang, Q., Didonato, J.A., Karin, M., McKeithan, T.W., 1994. BCL3 encodes a nuclear protein which can alter the subcellular location of NF-kappa B proteins. Molecular and Cellular Biology 14, 3915-3926.