doi:10.16372/j.issn.1004-6364.2016.17.002

鸡全基因组转录因子预测与分析

王志鹏[1,2,3*],景军红[1,2,3],周 萌[1,2,3],郭媛媛[4],李 婕[3]
(1.农业部鸡遗传育种重点试验室,黑龙江哈尔滨 150030;
2.黑龙江省普通高等学校动物遗传育种与繁殖重点试验室,黑龙江哈尔滨 150030;
3.东北农业大学动物科学技术学院,黑龙江哈尔滨 150030;
4.东北农业大学电气与信息学院,黑龙江哈尔滨 150030)

摘 要:转录因子是真核细胞基因表达调控网络的核心基因。本研究利用隐马尔可夫模型,在鸡基因组上共识别720个转录因子基因,分布在28条常染色体、Z染色体以及7个连锁群。在基因组上呈非均匀分布。发现有554个转录因子位于1 515个QTL内,包括影响外貌性状、健康性状、生理性状和生产性能性状。

关键词:鸡;全基因组;转录因子
中图分类号:S831.2 **文献标识码**:A **文章编号**:1004-6364(2016)17-06-04

Predictive Analysis of Chicken Transcription Factors at Genome-wide Scale

WANG Zhipeng[1,2,3*],JING Junhong[1,2,3],ZHOU Meng[1,2,3],GUO Yuanyuan[4],LI Jie[3]
(1.Key Laboratory of Chicken Genetics and Breeding,
Ministry of Agriculture,Harbin,Heilongjiang 150030;
2.Key Laboratory of Animal Genetics,Breeding and Reproduction,
Education Department of Heilongjiang Province,Harbin,Heilongjiang 150030;
3.College of Animal Science and Technology,
Northeast Agricultural University,Harbin,Heilongjiang 150030;
4.College of Electrical Engineering and Information,
Northeast Agricultural University,Harbin,Heilongjiang 150030)

Abstract:The transcription factor plays important roles in gene regulation networks. In this study,720 transcription factor genes on chicken genome based on HMM model were identified,which distributed non-random on 1 to 28 chromosome,Z chromosome and 7 linkage groups. According to QTL information,it was found that there are 554 transcription factors located on 1 515 QTLs related to exterior traits,health traits,physiology traits and production traits.

Key words:chicken;transcription factor;genome-wide

真核细胞基因表达的调控是一个多级调控系统,其中第一个水平是转录水平的调控。在这一调控水平上,细胞会在特定时间和空间上选择性地合成某些特定蛋白质,以期完成特定的生物学过程。转录水平上的调控是由为数众多的转录因子所主导,转录因子的基因数目占整个基因组基

收稿日期:2016-08-11
基金项目:黑龙江省教育厅面上项目(11551032)
*通讯作者:王志鹏(1979-),男,博士,副教授,研究方向为动物分子数量遗传学与生物信息学,E-mail:wangzhipeng@neau.edu.cn

因总数5%~10%左右[1~3]。这些转录因子特异性结合到所调控的靶基因上,通过增强和抑制靶基因的表达来实现相应的生物学功能。

在转录水平上参与基因表达调控的众多转录因子存在很大差异,且含有复杂的结构域。这些转录因子的序列相似性非常小,但是不同转录因子存在一个共同的特征,即都含有一个DNA结合结构域(DNA-binding domain,DBD),转录因子通过对该结合域特异性地结合到DNA的特异碱基序列上,从而使转录因子执行特定的生物学功能。根据转录因子DNA结合结构域的种类,可将转录因子分成不同的基因家族,如锌指模型、亮氨酸拉链模型、螺旋-转角-螺旋、螺旋-环-螺旋、HMG框等家族。

目前,已经获得多个物种的基因组和蛋白质组数据,以全基因组的视角采用比较基因组学和生物信息学方法揭示生物基因组上的潜在转录因子也多有报道,如在小鼠基因组上预测得到1 485~1 677个转录因子[4,5],在水稻基因组上预测得到2 025个转录因子[6],在拟兰芥基因组上预测得到1 827个转录因子[7],在蓝藻基因组上预测得到1 288个转录因子[8]。但对于具有经济价值的家养动物基因组上的转录因子预测及分析的研究却远远滞后于模式生物基因组学的研究。

在鸡基因组上转录因子的研究,主要聚焦于特定几个转录因子的基因功能研究,如影响体脂生长发育的PPAR、CEBPa等转录因子[9]。转录因子作为转录调控水平的核心,其在基因组水平的分布特点也应是鸡结构基因组学研究的一个重要内容,但目前鲜有这方面的报道。本研究在全基因组水平预测鸡基因组上所有的潜在转录因子,并对所预测得到转录因子的分布特点进行研究,同时结合鸡重要经济性状的QTL定位结果对转录因子进行注释,为相关基因的表达调控研究奠定基础。

1 材料与方法

1.1 数据来源

从ENSEMBL数据库[10]((ftp://ftp.ensembl.org/pub/)下载鸡(*Gallus gallus*)的全基因组序列、基因序列及蛋白质序列。同时获得各基因的结构信息与注释信息。从QTLdb数据库[11](www.animalgenome.org/QTLdb/)收集关于鸡的所有QTL信息(包括影响外貌性状、健康性状、生理性状和生产性能性状的QTLs共4 525个)。

1.2 HMM模型来源

参考DBD数据库[12],从SUPERFAMILY数据库(http://www.supfam.org/SUPERFAMILY/)下载所有DNA结合结构域的HMM模型,包括66个家族,213个HMM模型。

1.3 转录因子预测识别

用HMMER[12]中hmmpscan程序和hmmsearch程序对鸡基因组上所有编码的蛋白质进行分析。设定E-value为1.0E-4。其他参数取默认值。

2 结果与讨论

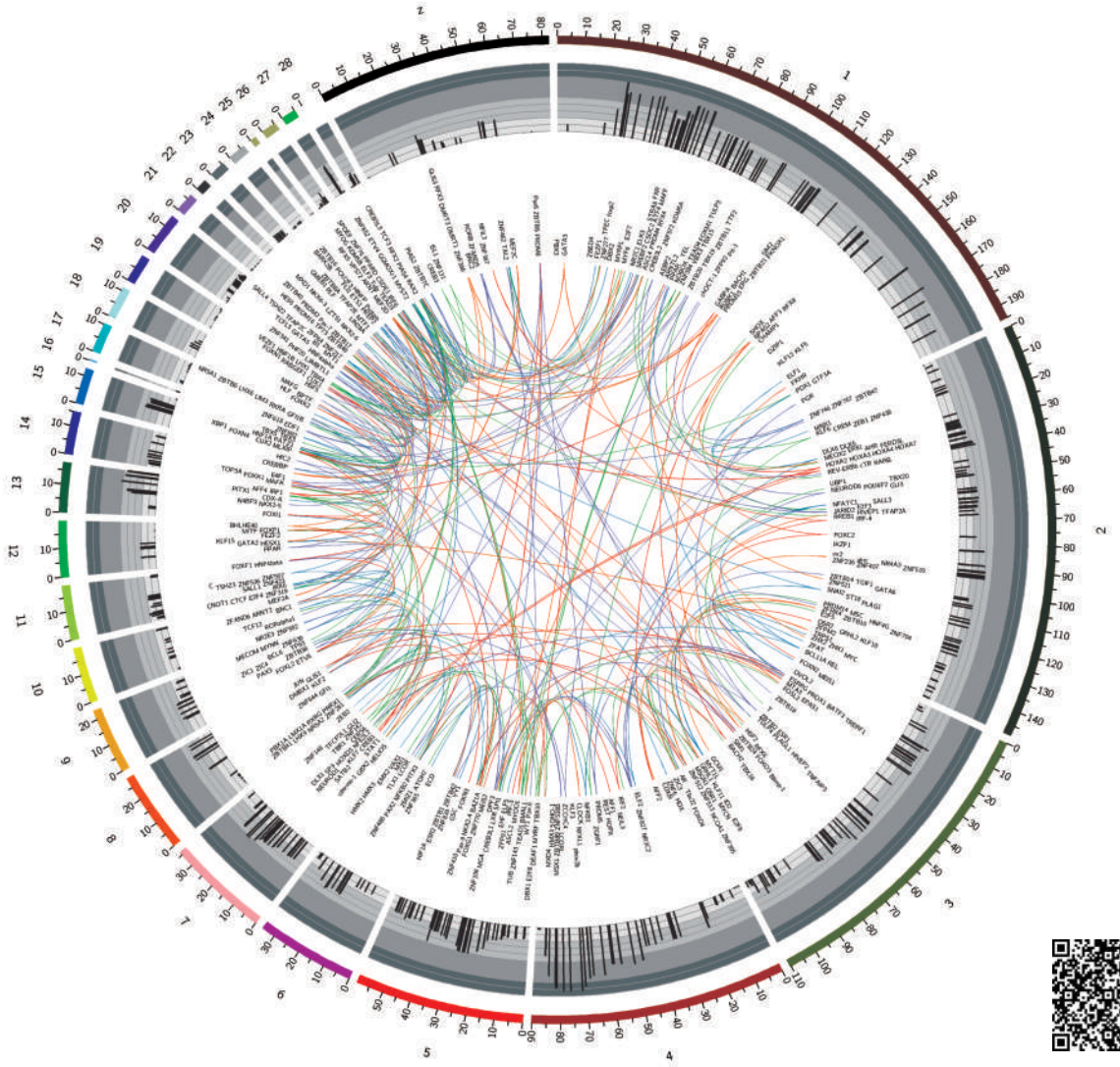
2.1 鸡全基因组转录因子识别

本研究利用隐马尔可夫模型,共识别得到1 084个转录因子蛋白质,占整个蛋白质组6.0%。这些转录因子蛋白质由720个转录因子基因翻译得到,占基因组总基因数的3.2%。这720个转录因子基因中有463个基因只翻译生成1个转录因子蛋白质,其余257个转录因子基因翻译生成2个以上的转录因子蛋白质,ENSGALG00000006013基因翻译生成的转录因子蛋白质最多(10个转录因子蛋白质)。TRANFAC数据库收录了高质量的鸡转录因子,此数据库共收录82个鸡转录因子基因[16](http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/search.cgi),这些转录因子都被本研究识别。除此之外,DBD和AnimalTFDB数据库收录了鸡全基因组上预测的转录因子。在DBD数据库中,共收录1 154个预测转录因子基因[12];AnimalTFDB数据库共收录858个转录因子基因[4]。在其他物种的全基因组转录因子预测工作发现,大部分脊椎动物的转录因子占5.0%~8.9%,非脊椎动物的转录因子占比小于5%[4];拟兰芥的转录因子基因占5.9%[1],大部分植物所含转录因子约占10%[3]。综上,转录因子在不同物种间所占比例存在差异,这可能是不同物种间基因转录调控复杂程度的差异所导致。

2.2 鸡全基因组转录因子分布特点

预测得到的720个鸡转录因子基因分布在28条常染色体、Z染色体以及7个连锁群。其中位于常染色体和Z染色体上的转录因子基因共650个,每条染色体上分布的转录因子基因数目迥异,范围从2个(16号染色体)到83个(2号染色

体)。具体分布情况见图1。对每一条染色体设置区间为500 kb、1 Mb、2 Mb的视窗,计算落入此视窗内转录因子基因的数目,发现转录因子基因在基因组上的分布呈非均匀分布,在染色体上存在转录因子基因的荒漠区与富集区。Oliver等[14]对基因组序列的非随机性做了深刻的探讨,研究发现基因非随机分布于基因组,基因组上相邻基因趋向于相同的基因表达模式。



注:第一层(最外层)为鸡基因组不同染色体;第二层为转录因子位于QTLs的数目的直方图;第三层为转录因子的基因名称;第四层为同一个转录因子家族的连线。

图1 鸡全基因组预测转录因子分布图

2.3 鸡转录因子家族

按照DNA结合结构域序列的特征,可以将转录因子基因分成不同的家族,即具有同样的DBD定义为同一类基因家族。本研究根据DBD对预测的转录因子基因进行家族分类。分类原则如下:若转录因子上仅存在一类DBD,则将该转录因子定义为此类转录因子家族;若转录因子存在多种DBD,则将根据DBD的关系决定该转录因子归属的家族,如果这些DBD都属同一超家族,则将该转录因子归属为该超家族,如果这些DBD不属同一超家族,则将该转录因子归属为"组合型"家族。根据如上定义,本研究对位于常染色体和Z染色体上的转录因子基因(共650个)进行了分析,其中192个转录因子基因为"组合型"基因家族;其余458个基因仅存在一种DBD,归属42个基因家族,这些基因家族的成员数目差异较大,

319个转录因子基因分布在6个基因家族中,其余139个基因分布在36个基因家族中。包含基因数目最多的基因家族主要包括ZF-C2H2、Homeobox家族和HLH家族等,这些家族中分别包含143个、86个和42个基因。有20个基因家族的成员数小于3个转录因子基因,如CBFB_NFYA、HTH_3和Not1家族仅含有1个基因,DM、GCM、HTH_psq家族仅含有2个基因,Fez1、Runt、SRF-TF家族仅含有3个基因。对这些转录因子家族的成员数进行进一步的统计分析,发现其分布服从幂率分布,即只有少数几个家族含有较多的成员数目,在其他的蛋白质家族分类研究中也得到了同样的结果[15]。

2.4 鸡转录因子的QTL注释分析

本研究从QTLdb[11]网站获得目前鸡不同性状QTL定位结果。在鸡基因组上共4 525个QTL,其中影响外貌性状的QTL共190个,影响健康性状的QTL共547个,影响生理性状的QTL共116个,影响生产性能性状的QTL共3 672个。通过对转录因子基因与QTL位置的比较,本研究发现有554个转录因子位于1 515个QTL内(每个转录因子基因与QTLs数目的关系见图1),包括63个影响外貌性状QTL,122个影响健康性状QTL,46个影响生理性状QTL,1 284个影响生产性能性状QTL。

3 结 论

本研究利用隐马尔可夫模型,在鸡基因组上共识别720个转录因子基因,分布在28条常染色体、Z染色体以及7个连锁群。在基因组上呈非均匀分布。发现有554个转录因子位于1 515个QTL内,包括影响外貌性状、健康性状、生理性状和生产性能性状。

参考文献:

[1] RIECHMANN J L,HEARD J,MARTIN G,*et al*. Arabidopsis transcription factors:genome-wide comparative analysis among eukaryotes[J]. Science,2000(290):2105-2110.

[2] MADAN M BABU,SARAH A TEICHMANN. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*[J]. Nucleic Acids Research,2003,31(4):1234-1244.

[3] GUO A,CHEN X,GAO G,*et al*. PlantTFDB:A comprehensive plant transcription factor database[J]. Nucleic Acids Research,2007:1-4

[4] ZHANG H M,LIU T,LIU C J,*et al*. Animal TFDB 2.0:A resource for expression,prediction and functional study of animal transcription factors[J]. Nucleic Acids Res,2015(43):D76-D81.

[5] KANAMORI M,KONNO H,OSATO N,*et al*. A genome-wide and nonredundant mouse transcription factor database[J]. Biochem Biophys Res Commun,2004,24;322(3):787-793.

[6] GAO G,ZHONG Y,GUO A,*et al*. DRTF:A database of rice transcription factors[J]. Bioinformatics,2006(22):1286-1287.

[7] GUO A,HE K,LIU D,*et al*. DATF:A database of Arabidopsis transcription factors[J]. Bioinformatics,2005(21):2568-2569.

[8] WU J,ZHAO F,WANG S,*et al*. cTFbase:A database for comparative genomics of transcription factors in cyanobacteria[J]. BMC Genomics,2007(8):104.

[9] DING N,GAO Y,WANG N,*et al*. Functional analysis of the chicken PPARγ gene 5′-flanking region and C/EBPα-mediated gene regulation[J]. Comp Biochem Physiol B Biochem Mol Biol,2011,158(4):297-303.

[10] KULIKOVA T,AKHTAR R,ALDEBERT P,*et al*. EMBL nucleotide sequence database in 2006[J]. Nucleic Acids Res,2007(35):D16-D20.

[11] HU Z L,PARK C A,WU X L,*et al*. Animal QTLdb:An improved database tool for livestock animal QTL/association data dissemination in the post-genome era[J]. Nucleic Acids Res,2013(41):D871-D879.

[12] SARAH K,KUMMERFELD,SARAH A. Teichmann DBD:A transcription factor prediction database[J]. Nucleic Acids Research 2006(34):D74-D81.

[13] EDDY S R. Profile hidden Markov models[J]. Bioinformatics,1998(14):755-763.

[14] OLIVER B,MISTELI T. A non-random walk through the genome[J]. Genome Biol,2005,6(4):214.

[15] QIAN J,LUSCOMBE N M,GERSTEIN M. Protein family and fold occurrence in genomes:Power-law behaviour and evolutionary model[J]. J Mol Biol,2001,313(4):673-681.

[16] MATYS V,KEL-MARGOULIS O V,FRICKE E,*et al*. TRANSFAC and its module TRANS Compel:Transcriptional gene regulation in eukaryotes[J]. Nucleic Acids Res,2006,34(1):D108-D110.