吉林大学

2019.06.11

比较基因组学揭示家鸡黑白羽色变异的遗传学基础

专业名称 | 动物科学

传统遗传学难以探究，新技术带来突破

毛色研究进展缓慢

1、家鸡驯化历史不明确：至今对于家鸡的起源仍存在争论

2、羽色调控机制复杂：家禽的色素调控机制与哺乳动物存在差异

3、样本缺乏代表性：目前国内研究的多为商品鸡，变异丢失，难以探究毛色的驯化

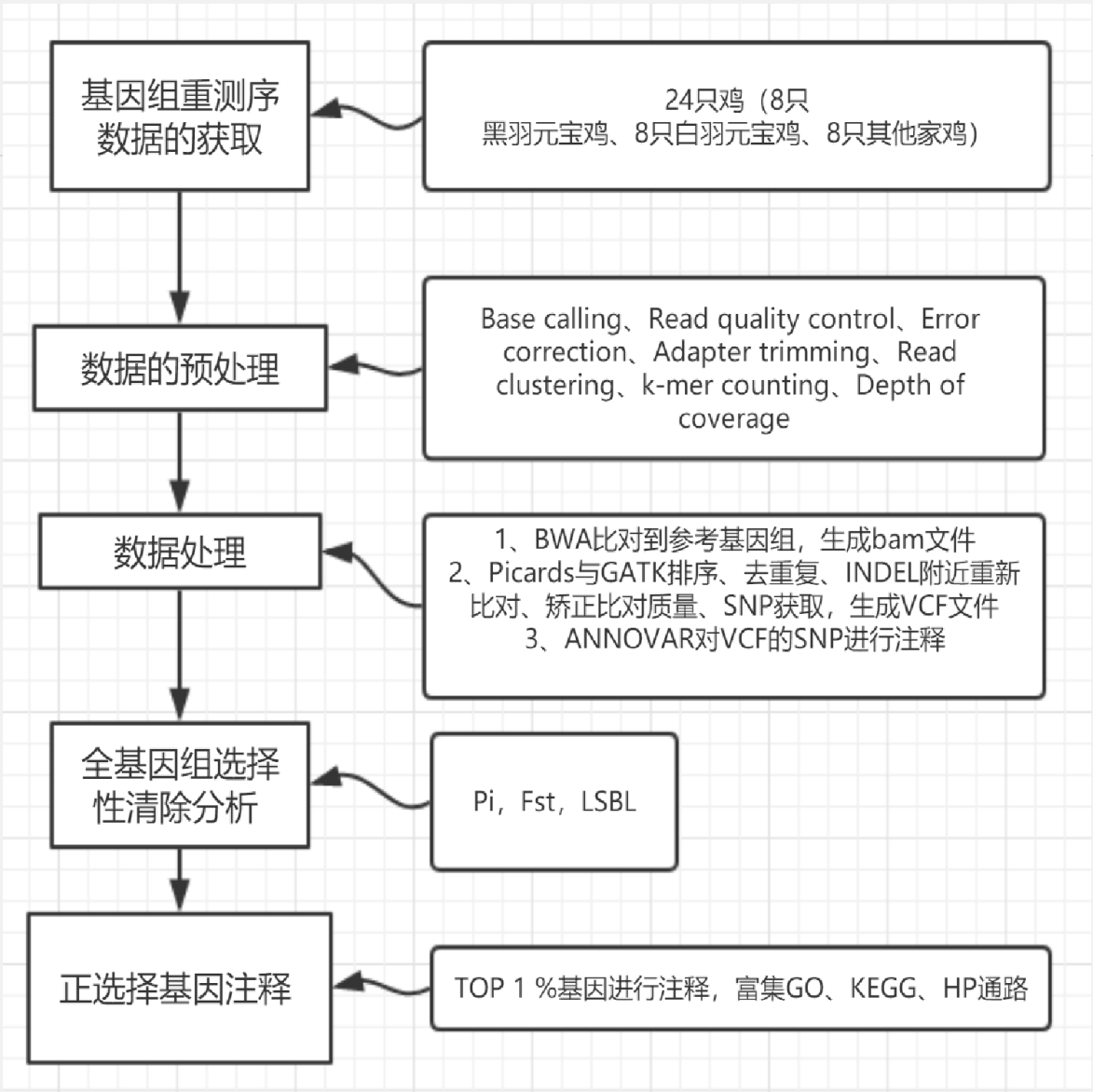
元宝鸡的优异生物学特性

1、驯化历史明确：元宝鸡人工驯化起始于唐代

2、外貌特征：腿短身矮，小腿约6厘米左右，分为黑白两种羽色。

因此我们采用群体遗传学和比较基因组学的结合的技术，探究群体在演化动力的影响下，等位基因的分布和改变，基于NGS基因组图谱，对已知的基因特征比较以了解基因的功能

实验流程主要分为五部分



研究方法

研究要点简述

Linux高性能服务器：生物信息学开发环境构建

实验所撰写的300余行脚本同步上传至 **https://github.com/zhumingyuan123/pipeline-for-fq，配置环境一并上传，实验可复制性**

统计学原理：主要包含群体遗传学相关定理以及随机森林算法

部分公用基础软件
MPICH, Openmpi2，Perl 5.12（包含各种模块），Python 2.6.5（包含各种模块），fftw3.1.2，atlas，lapack,blas, INTEL-COMPLIER, PGI compiler, matlab, confuse, cairo,apache, mysql, imagemagick,? pango, torque,maui, fontconfig, freetype,rrdtool, sar2rrd, pango, pixman, pkgconfig, pigz, mpi4py, Parallel::MPI(mpi implementation on perl), libpng, libxml, glasgow haskell compiler, f95,gfortran, zlib, java， R(包含各种模块),
部分应用软件
Phyml，RAxML-7.2.5， Mrbayes，Parallel_MrBayes, Qiime, GAPipeline, Illumina_Genome_Analyzer，CASAVA, Mireap_0.2, Muscle, Blast, Mpiblast, Abyss，codeml, overlappr, dnadist, dnastar, namd, charm++, blat, mpiblast, tigcl, muscle, DBD-SQLite-1.31, DBD-mysql-4.018, Data-ShowTable-3.3, DBI-1.616, Math-CDF-0.1, prank, paml44, boost1_46_1, cufflinks1.0.3, clustalw, tophat, myrna-1.0.4， amos-3.0.0，bowtie， hapsembler， SOAPdenovo-V1.05，samtools，velvet_1.1.04， Phrap， GapCloser， idba， bwa-0.5.9， MTR， pysam-0.4.1，biopython， bioperl, cd-hit-v4.3, dna_blast, inparanoid, mrbayes, protest, mauve, hmmer， phylip， lastz， Mummer3， Soap2， Velvet， Cap3， T-COFFEE_distribution_Version_8.99

24只元宝鸡测序数据的预处理：质量控制、清除测序误差

JMCB 元宝鸡三个群体的高通量测序数据

JMCB：SRP034930 Comparative population genomics reveals genetic basis underlying body size of domestic chickens；8只元宝黑羽鸡、白羽鸡、其他家鸡

yuanbao_black	140401_Ypt581-700_AGTCAA_L005
yuanbao_black	140401_Ypt582-700_CCGTCC_L005
yuanbao_black	140403_Ypt588-650_GCCAAT_L006
yuanbao_black	140430_Ypt589-650_CAGATC_L004
yuanbao_black	140403_Ypt595-700_GCCAAT_L003
yuanbao_black	140403_Ypt596-700_CAGATC_L004
yuanbao_black	140403_Ypt597-700_ACTTGA_L004
yuanbao_black	140403_Ypt598-700_GATCAG_L004
yuanbao_white	140401_Ypt578-700_GCCAAT_L004
yuanbao_white	140401_Ypt580-700_CTTGTA_L005
yuanbao_white	140403_Ypt583-650_ATCACG_L005
yuanbao_white	140403_Ypt584-650_CGATGT_L005
yuanbao_white	140403_Ypt591-700_CGATGT_L003
yuanbao_white	140403_Ypt593-700_TGACCA_L003
yuanbao_white	140403_Ypt594-700_ACAGTG_L003
yuanbao_white	140430_YPt601-700_CTTGTA_L001
yuanbao_other	140403_Ypt585-650_TTAGGC_L005
yuanbao_other	140403_Ypt586-650_TGACCA_L005
yuanbao_other	140416_Ypt590-650_ATCACG_L004
yuanbao_other	140401_Ypt579-700_CAGATC_L004
yuanbao_other	140403_Ypt587-650_ACAGTG_L005
yuanbao_other	140403_Ypt592-700_TTAGGC_L003
yuanbao_other	140403_Ypt599-700_TAGCTT_L004
yuanbao_other	140403_YPt600-700_GGCTAC_L004

24只元宝鸡测序数据的预处理：质量控制、清除测序误差

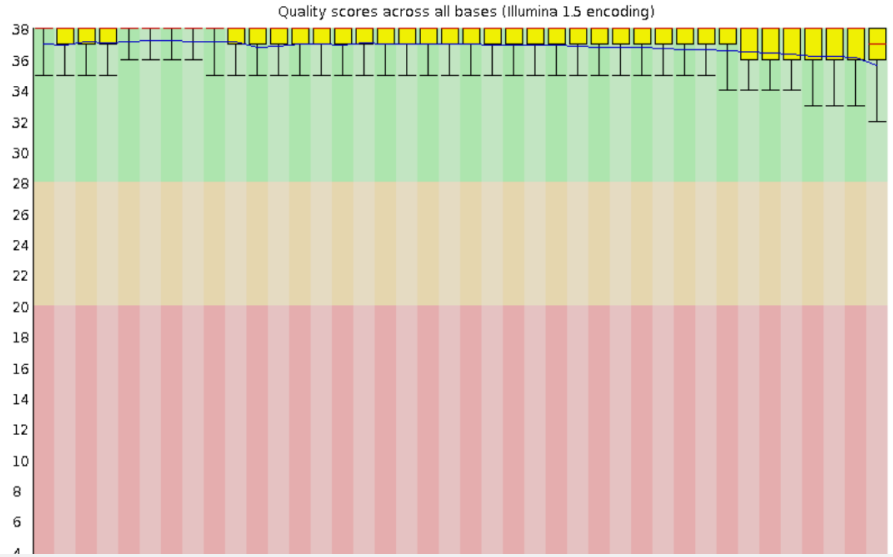
FastQC Report

Summary

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality



数据比对、注释、格式变换 ： FASTQ ->SAM -> VCF

基因组比对、变异位点获取以及注释

1、利用 **BWA** 的 MEM 算法将过滤后的双端测序 reads 比对到家鸡的**参考基因组**上（Galgal4），生成sam（The Sequence Alignment / Map format）文件格式，即序列比对文件的格式

2、**Picards** 和**GATK** 软件包中相应的工具对比对后的bam文件进行一系列的处理，包括**排序、去除其重复序列、INDEL附近的reads重新比对以及矫正其比对质量**等。

3、GATK中的**UnifiedGenotyper**工具获得群体的SNP，并使用**VariantFiltration**对SNP 进行过滤

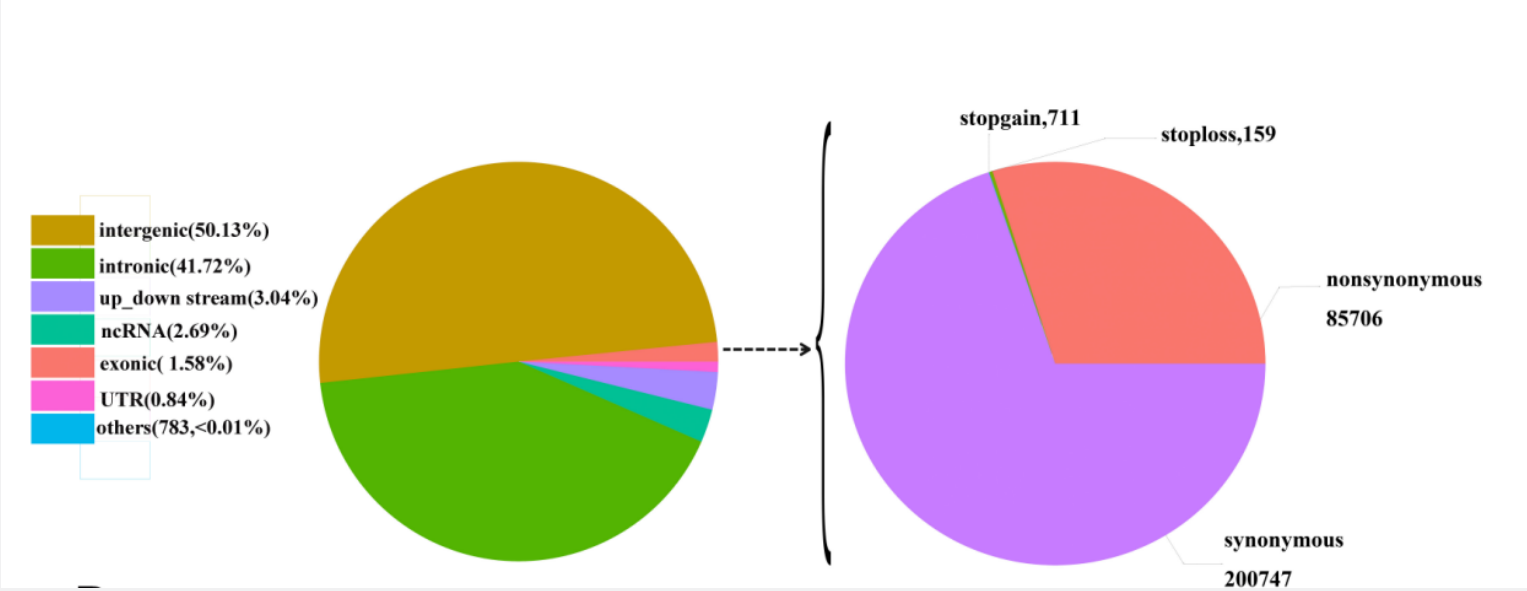
4、分群处理，vcf文件按照群体与染色体进行拆分

```
72497 2  2241723 .  T  C  657.47  PASS  .  GT:AD:DP:GQ:PL  0/0:5,0:5:15:0,15,200  0/0:9,0:9:27:0,27,360  0/0:1,0:1:3:0,3,41  0/0:3,0:3:9:0,9,116 0/0:5,0:5:15
72498 2  2241737 .  C  T  1319.92 PASS  .  GT:AD:DP:GQ:PL  0/0:4,0:4:12:0,12,160  0/0:8,0:8:24:0,24,330  0/0:3,0:3:9:0,9,121 0/0:3,0:3:9:0,9,123 0/0:5,0:5:15
72499 2  2241756 .  C  A  1423.86 PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,286  0/0:8,0:8:24:0,24,331  0/0:3,0:3:9:0,9,124 0/0:4,0:4:12:0,12,167  0/0:7,0:7
72500 2  2241758 .  C  A  486.87  PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,289  0/0:8,0:8:24:0,24,324  0/0:3,0:3:6:0,6,84  0/0:4,0:4:12:0,12,150  0/0:7,0:7
72501 2  2241768 .  G  A  1829.93 PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,287  0/0:7,0:7:21:0,21,283  0/0:3,0:3:9:0,9,118 0/0:4,0:4:12:0,12,158  0/0:7,0:7
72502 2  2241793 .  T  C  9067.82 PASS  .  GT:AD:DP:GQ:PL  0/0:9,0:9:27:0,27,363  0/0:6,0:6:18:0,18,246  0/0:3,0:3:9:0,9,107 1/1:0,1:1:3:41,3,0  0/1:4,2:6:61
72503 2  2241809 .  A  G  1106.14 PASS  .  GT:AD:DP:GQ:PL  0/0:8,0:8:24:0,24,328  0/0:6,0:6:18:0,18,244  0/0:2,0:2:6:0,6,81  0/0:2,0:2:6:0,6,80  0/0:8,0:8:24
72504 2  2241868 .  G  T  624.97  PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,283  0/0:6,0:6:18:0,18,248  0/0:3,0:3:9:0,9,123 0/0:1,0:1:3:0,3,42  0/0:5,0:5:15
72505 2  2241890 .  G  C  2390.96 PASS  .  GT:AD:DP:GQ:PL  0/0:8,0:8:24:0,24,327  0/0:6,0:6:18:0,18,246  0/0:3,0:3:9:0,9,125 0/0:1,0:1:3:0,3,41  0/0:8,0:8:24
72506 2  2241931 .  A  G  270.25  PASS  .  GT:AD:DP:GQ:PL  0/0:6,0:6:18:0,18,233  0/0:6,0:6:18:0,18,229  0/0:3,0:3:9:0,9,122 ./.:.:.:.:. 0/0:7,0:7:21:0,21,27
72507 2  2241954 .  C  A  451.85  PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,274  0/0:5,0:5:15:0,15,193  0/0:1,0:1:3:0,3,30  0/0:1,0:1:3:0,3,40  0/0:7,0:7:21
72508 2  2241964 .  C  A  8414.64 PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,277  0/0:5,0:5:15:0,15,205  ./.:.:.:.:. 0/1:1,1:2:33:36,0,33  0/1:3,4:7:99:141
72509 2  2241974 .  G  A  188.21  PASS  .  GT:AD:DP:GQ:PL  0/0:5,0:5:15:0,15,202  0/0:5,0:5:15:0,15,203  ./.:.:.:.:. 0/0:3,0:3:9:0,9,112 0/0:5,0:5:15:0,15,19
72510 2  2242029 .  C  G  1490.14 PASS  .  GT:AD:DP:GQ:PL  0/0:8,0:8:24:0,24,322  0/0:6,0:6:18:0,18,243  0/0:1,0:1:3:0,3,37  0/0:4,0:4:12:0,12,158  0/0:4,0:4
72511 2  2242047 .  C  T  3149.19 PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,270  0/1:3,4:7:99:130,0,100  0/0:1,0:1:3:0,3,40  0/0:4,0:4:9:0,9,118 0/0:4,0:4:9:
72512 2  2242049 .  G  C  3108.93 PASS  .  GT:AD:DP:GQ:PL  0/0:7,0:7:21:0,21,276  0/1:3,4:7:98:128,0,98  0/0:1,0:1:3:0,3,41  0/0:4,0:4:12:0,12,163  0/0:4,0:4
72513 2  2242063 .  T  C  1649.77 PASS  .  GT:AD:DP:GQ:PL  0/0:6,0:6:18:0,18,243  0/1:1,3:4:28:110,0,28  0/0:1,0:1:3:0,3,33  0/0:3,0:3:9:0,9,110 0/0:7,0:7:21
```

数据比对、注释、格式变换 ： FASTQ ->SAM -> VCF

24只家鸡变异类型的注释结果

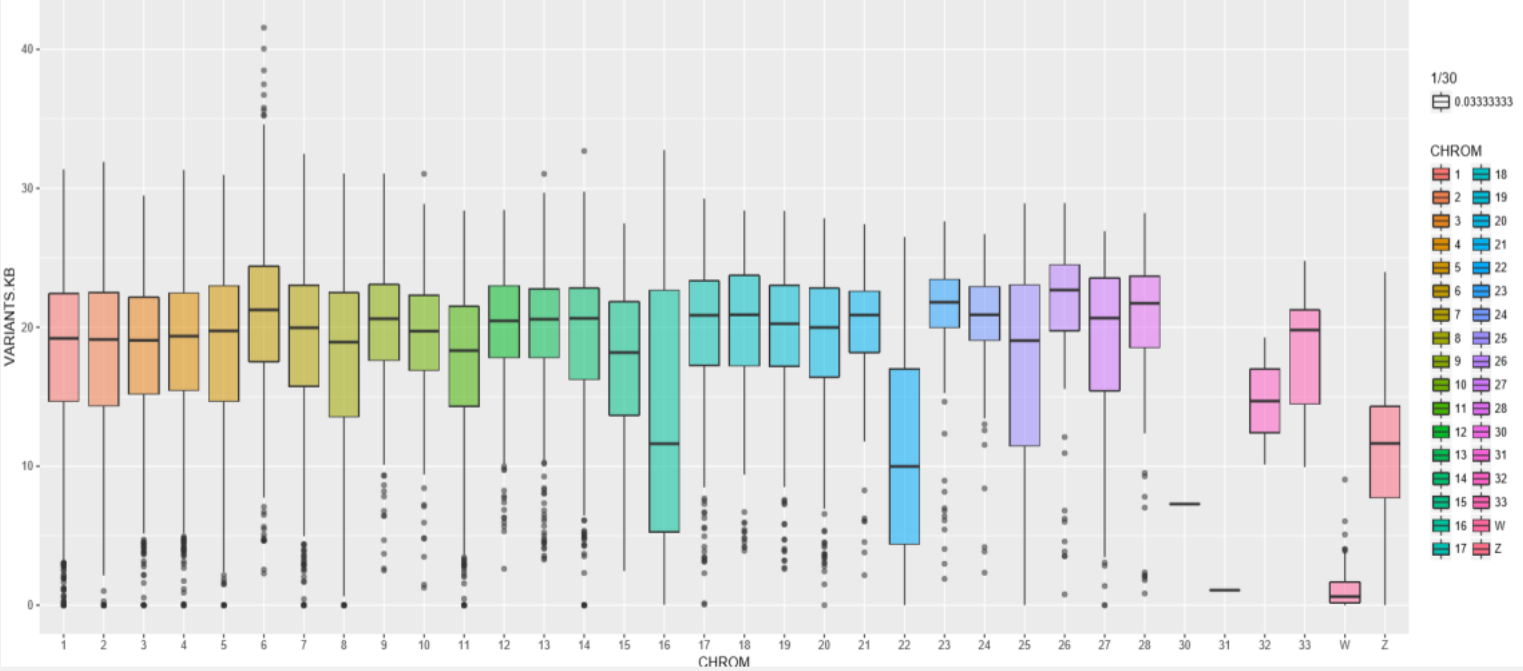
在给定的染色体区域内，基因间突变，内含子，基因上游区域，非编码RNA，外显子和开放阅读框中发生的SNP数目分别为9112159,7584082,552045,488986,287323和52759。SNP的功能注释蛋白质编码区鉴定出非同义替代的85706和同义替代的200747，其中870个SNP导致终止密码子获得或丧失。



数据比对、注释、格式变换 ： FASTQ ->SAM -> VCF

24只家鸡的染色体基因密度变异箱线图

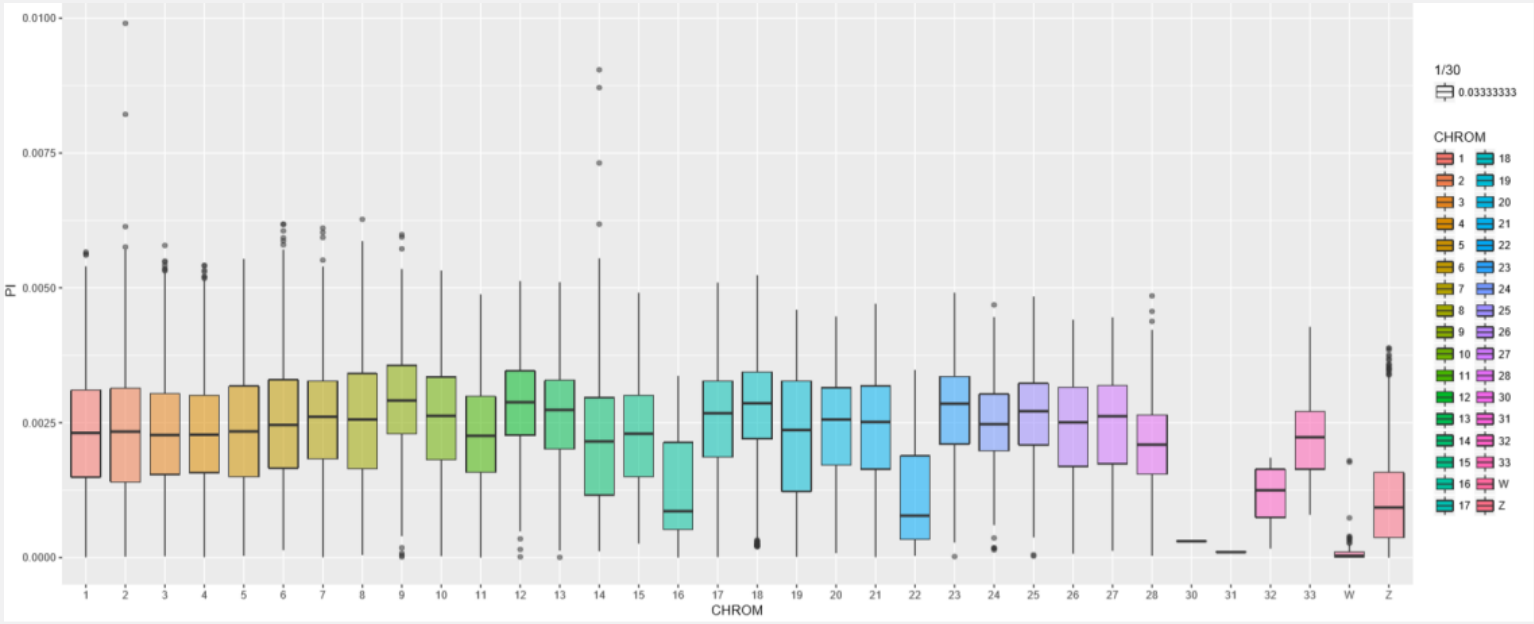
大部分的常染色体都具有较高的变异程度，特别是1~15号染色体；性染色体变异数较少，因此在剩余部分的分析中，我们将考虑这一点。



数据比对、注释、格式变换 ： FASTQ ->SAM -> VCF

24只家鸡的核苷酸多态性箱线图

变异主要集中在常染色体1~15之间，性染色体较少，因此在之后信号分析中我们更多向常染色体偏倚。



全基因组选择性清除分析-群体遗传学相关计算

Fst、LSBL分析

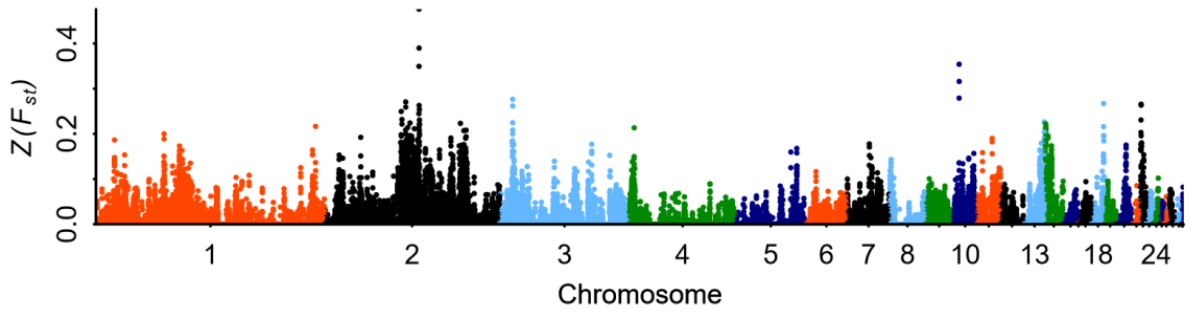
1、Fst的计算: **vcftools**工具的群体间fst功能，以三个群体个体名称为索引，分别计算vcf中三个群体之间位点的fst指数

2、提取三个Fst中相同位置的行，根据公式**LSBL=(FST(AB)+FST(AC)-FST(BC))/2**计算LSBL值：LSBL 进行**滑动窗口**计算，窗口大小为 50Kb，步长为 25Kb

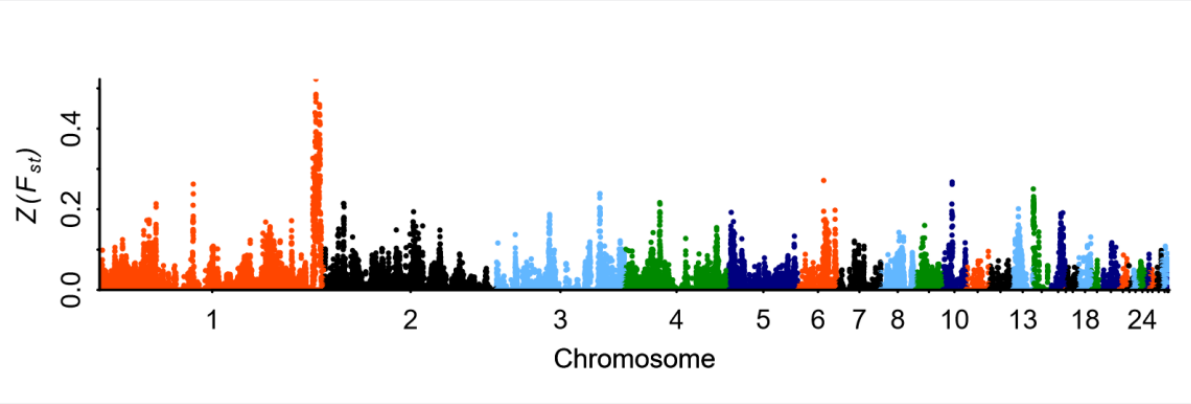
```
#PBS -N chickenfst
#PBS -q small
#PBS -l nodes=1:ppn=1
#PBS -o /home/zhumy/yuanbaochicken  /chickenfst.out
#PBS -e /home/zhumy/yuanbaochicken  /chickenfst.err


cd  /home/zhumy/yuanbaochicken
/home/zhumy/biosoft/vcftools/vcftools_0.1.13/bin/vcftools --gzvcf Frizzle_chicken_filterno3.vcf   --weir-fst-pop yuanbao_black
/home/zhumy/biosoft/vcftools/vcftools_0.1.13/bin/vcftools --gzvcf Frizzle_chicken_filterno3.vcf   --weir-fst-pop yuanbao_black
/home/zhumy/biosoft/vcftools/vcftools_0.1.13/bin/vcftools --gzvcf Frizzle_chicken_filterno3.vcf   --weir-fst-pop yuanbao_white
```

LSBL R语言绘图

B_W_O LSBL分析

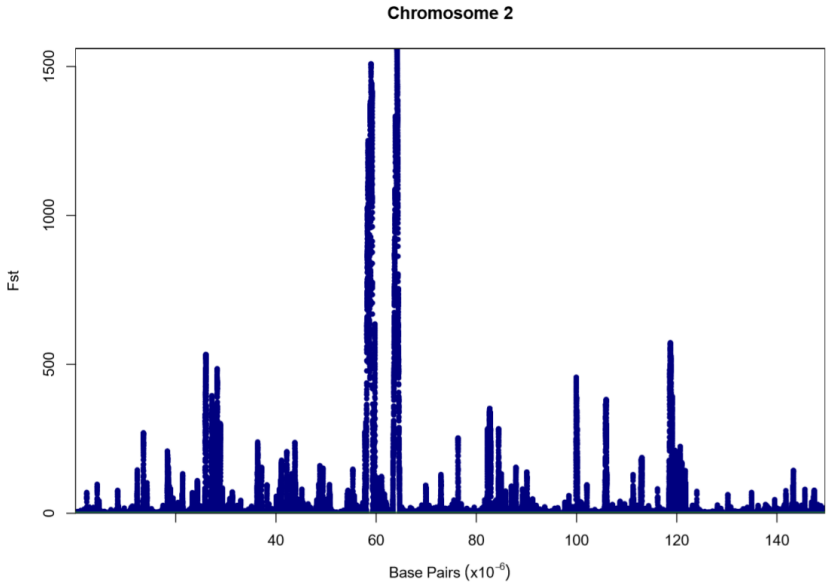


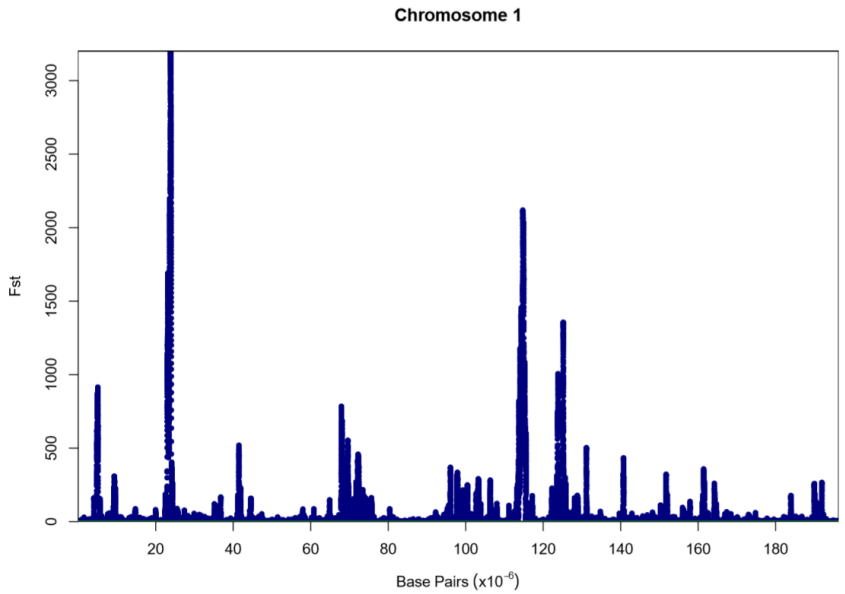
W_B_O LSBL分析



B_W_O LSBL分析的单个染色体区域

二号染色体





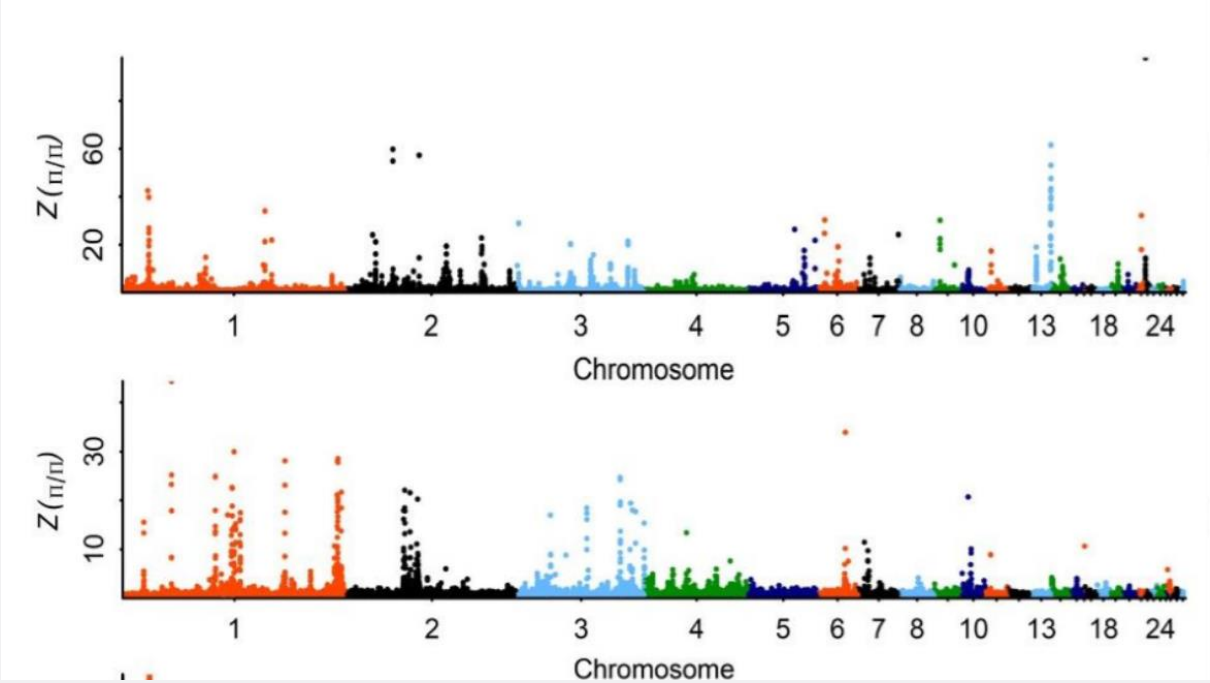
一号染色体

全基因组选择性清除分析-群体遗传学相关计算

PI（核苷酸多态性）

使用**vcftools**，以50Kb 的窗口、25Kb的步长分别计算了两色元宝鸡以及其他家鸡的核苷酸多态性

从黑、白元宝鸡的核苷酸多态性来看，变异主要集中在常染色体1~15之间，性染色体较少



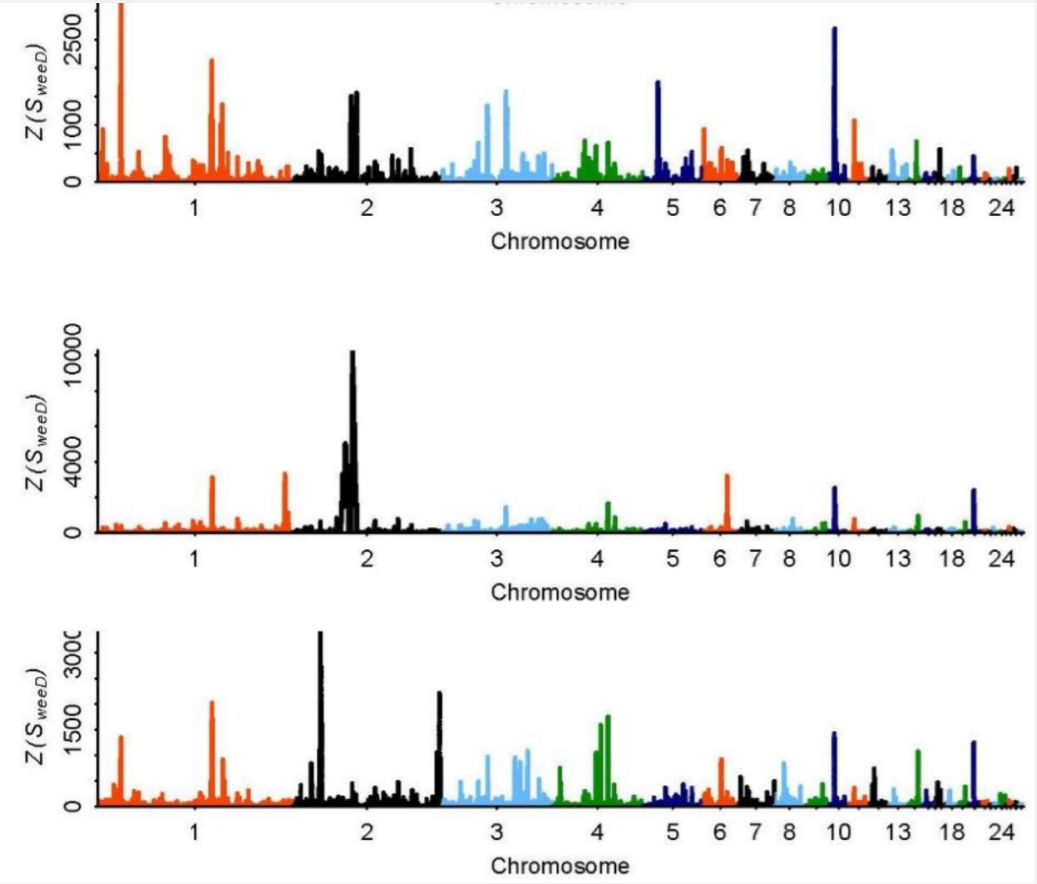
全基因组选择性清除分析-群体遗传学相关计算

SweeD

使用SweeD的**SweeD-P**算法

图依次为元宝黑鸡、白鸡、其他家鸡的SweeD指数

对于该方法，三个亚群的高选择信号区域表现在不同区域。其中**元宝黑鸡**与其他亚群相比，一号染色体信号强烈，而在二号染色体信号较弱；**元宝白鸡**则趋势相反



选择性清除分析结果 -> 基因注释

分析得到大致结论

1、注释结果 ->大部分的突变发生在了非编码区域

2、基因密度 ->性别决定的染色体通常具有有限的遗传多样性

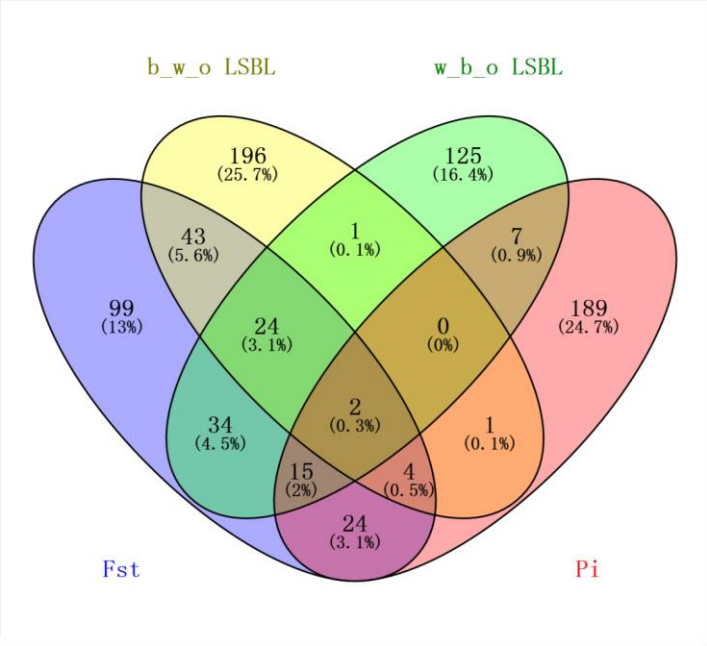
3、四个分析方法 -> 强选择信号不尽相同

对黑、白羽元宝鸡的受选择区域分类讨论

正选择基因的注释

TOP 1% 基因注释

选取 FST、Pi、LSBL、SweeD分析中的前1%的值所相对应的基因组区域为候选的选择性清除的基因座位，通过ensemble、Galaxy、David数据库的变异预测数据库Variant Effect Predictor进行注释。
注释了FST（403基因）、LSBL（B_W_O 369个基因，W_B_O 368个基因）和Pi（403个基因）

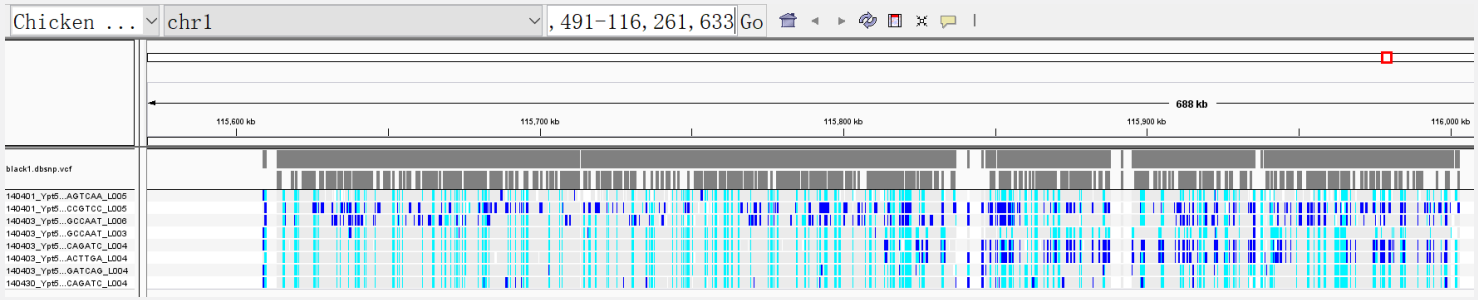


正选择基因的注释

生物学验证

按照数据库中**基因的注释、通路、表达情况，从已发表文献论文、基因功能、连锁上下游通路、已验证的毛色相关通路、和人类与动物疾病**五个角度来统计和评价锁定预测区域

chr1：187.30Mb-188.50M：Tyr和Rab38
ch1：191.45-191.62
ch2：69.1-72.9 Mb



机制复杂，对羽色调控通路提出新的见解

1、驯化过程中家鸡的表型发生了巨大的改变，我们找到了一系列候选区域，可能有助于我们探究家鸡毛色的驯化过程以及作为遗传育种的重要依据。

2、提出了比较群体基因组学的一种新的策略，即利用NGS数据作为前导来推断受选择信号，省时精确

3、毛色研究的复杂性、涉及通路的繁杂以及相关连锁的庞大，我们很难精确定位，但得到结果与前人的研究不谋而合，例如TYR，TYRP1和SLC24A5等并且还发现部分其他新的相关基因

4、提供了一个将家鸡毛色变异与其他动物色素沉着机制相联系的纽带，他们的机制可能在某些方面具有共性。这也为研究者探索新的基因及其功能提供了理论依据。

吉林大学

2019.06.11

请各位老师批评指正！

专业名称 | 动物科学