iScience, Volume *23*

Supplemental Information

Microevolutionary Dynamics of Chicken Genomes under Divergent Selection for Adiposity

Hui Zhang, Qiqi Liang, Ning Wang, Qigui Wang, Li Leng, Jie Mao, Yuxiang Wang, Shouzhi Wang, Jiyang Zhang, Hao Liang, Xun Zhou, Yumao Li, Zhiping Cao, Peng Luan, Zhipeng Wang, Hui Yuan, Zhiquan Wang, Xuming Zhou, Susan J. Lamont, Yang Da, Ruiqiang Li, Shilin Tian, Zhiqiang Du, and Hui Li

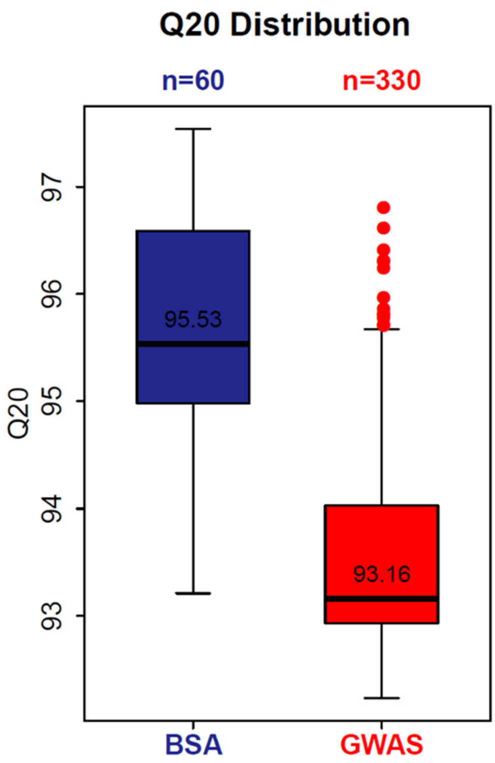


Figure S1. Distribution of Q20 values for genome resequencing data. Related to Figures 1 and 3.

BSA, pooled-DNA sequencing; GWAS, genome-wide association study.

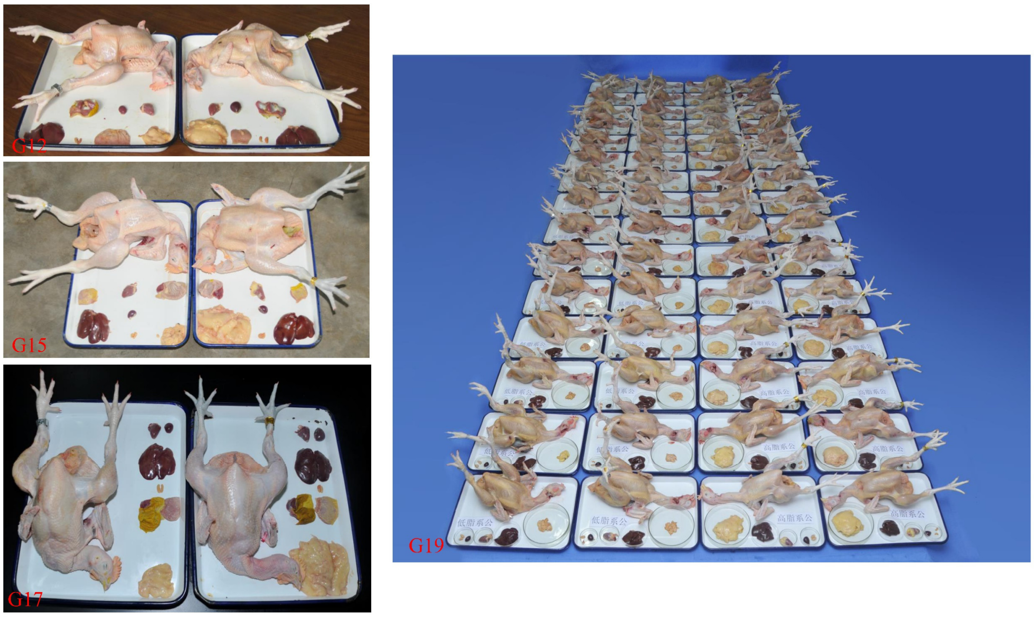


Figure S2. Difference in abdominal fat content between fat (right) and lean (left) lines. Related to Figure 1.

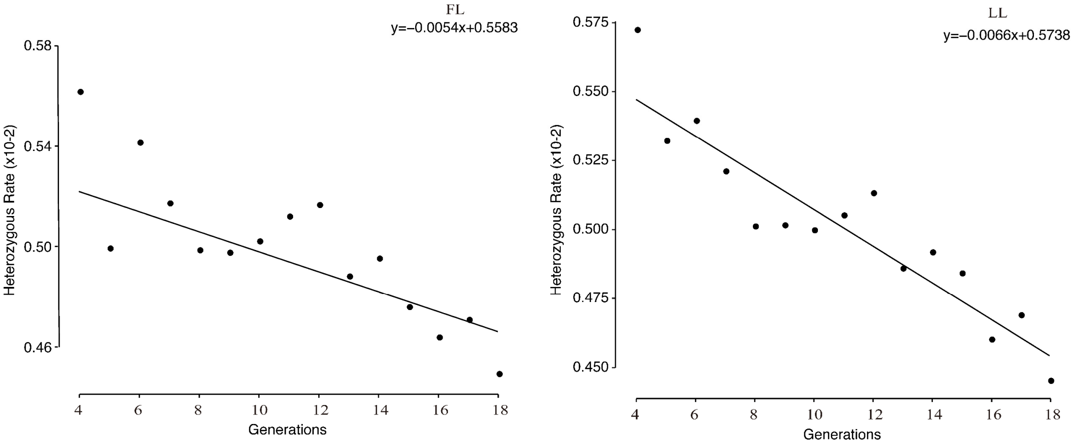


Figure S3. The significant reduction in heterozygosity rates calculated by a linear regression method. Related to Figure 1.

FL and LL, fat and lean chicken lines.

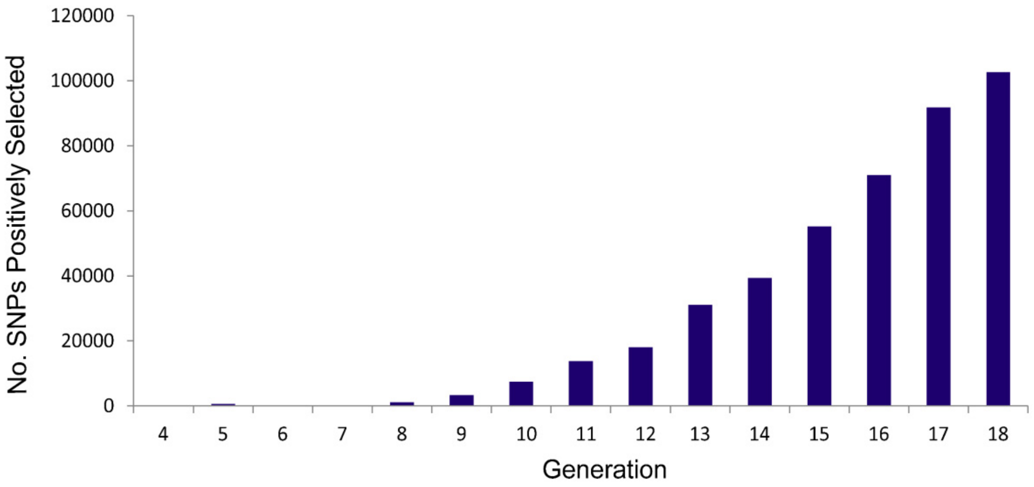


Figure S4. Dynamics of the number of SNPs detected by the △AF method in each generation. Related to Figure 1.

Number of SNPs associated with population differentiation continued to increase from G9 to G18. The increased number of SNPs subjected to selection paralleled the pattern of phenotypic changes.

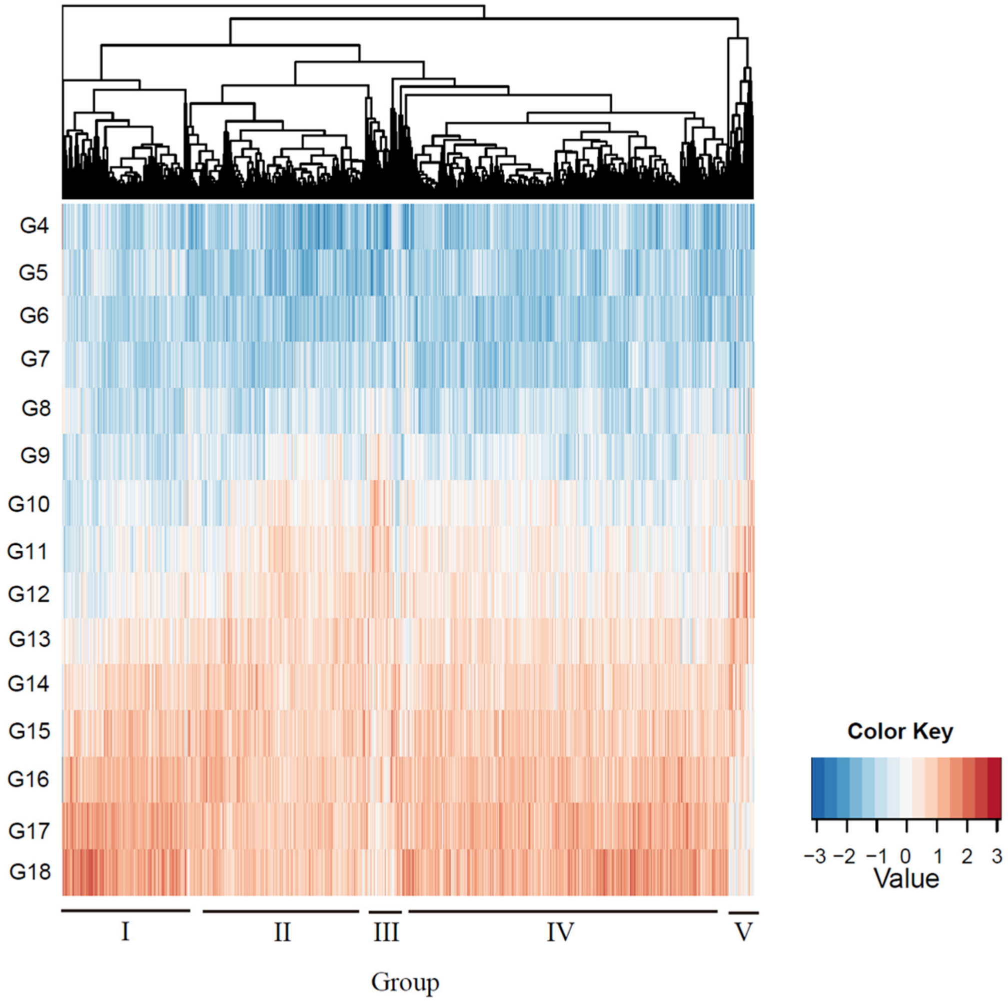


Figure S5. Clustering analysis on SNPs detected by the △AF method for all generations. Related to Figure 1.

Heatmap for regions detected to be under selection can be clustered into five groups, indicating their differently changing pattern of allele frequencies.

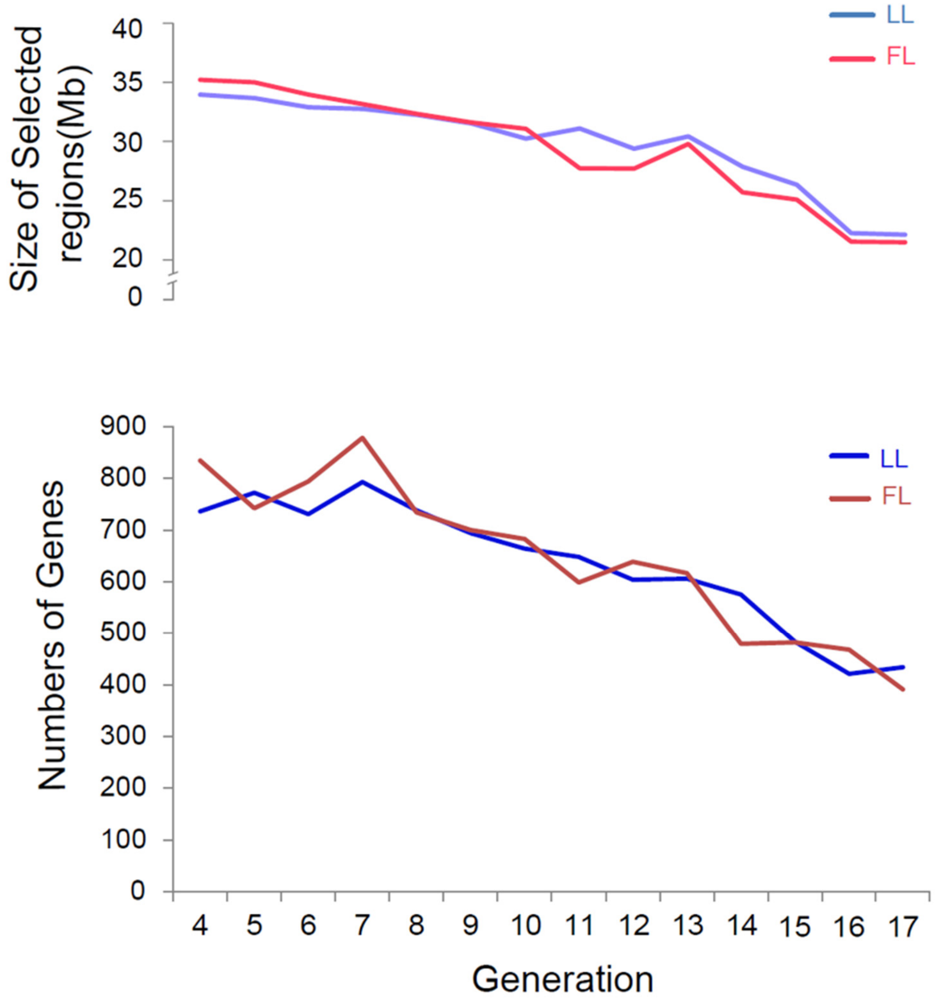


Figure S6. Number of genome windows and genes under selection within the fat and lean chicken lines across generations. Related to Figure 1.

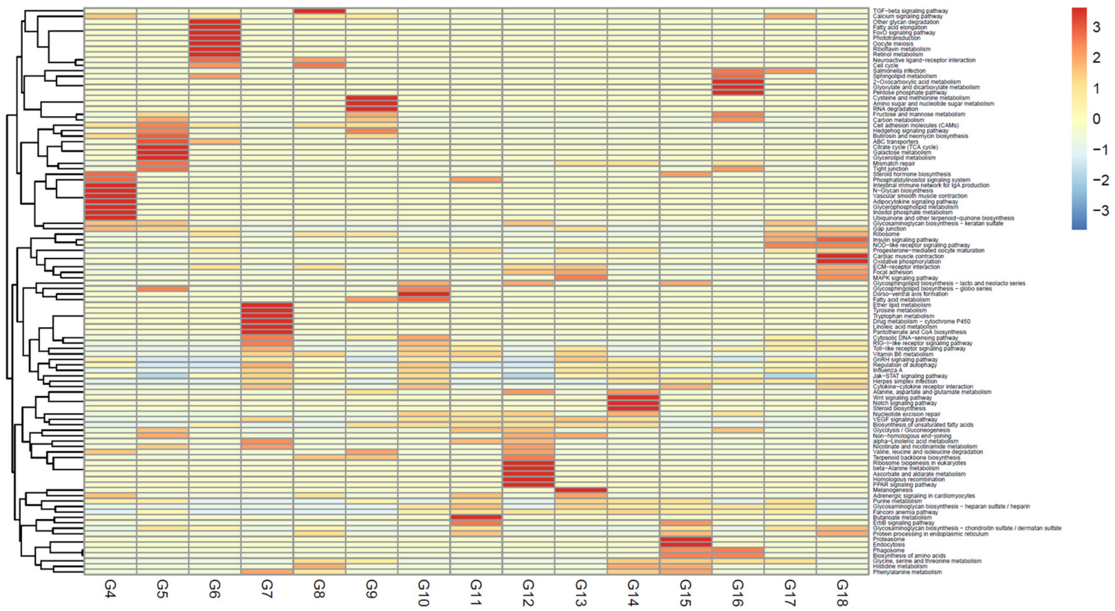


Figure S7. SNPs under positive selection across generation gradually recruited and enriched in pathways related to adipose tissue growth (such as fatty acid and ribosome biosynthesis and metabolism). Related to Figure 1.

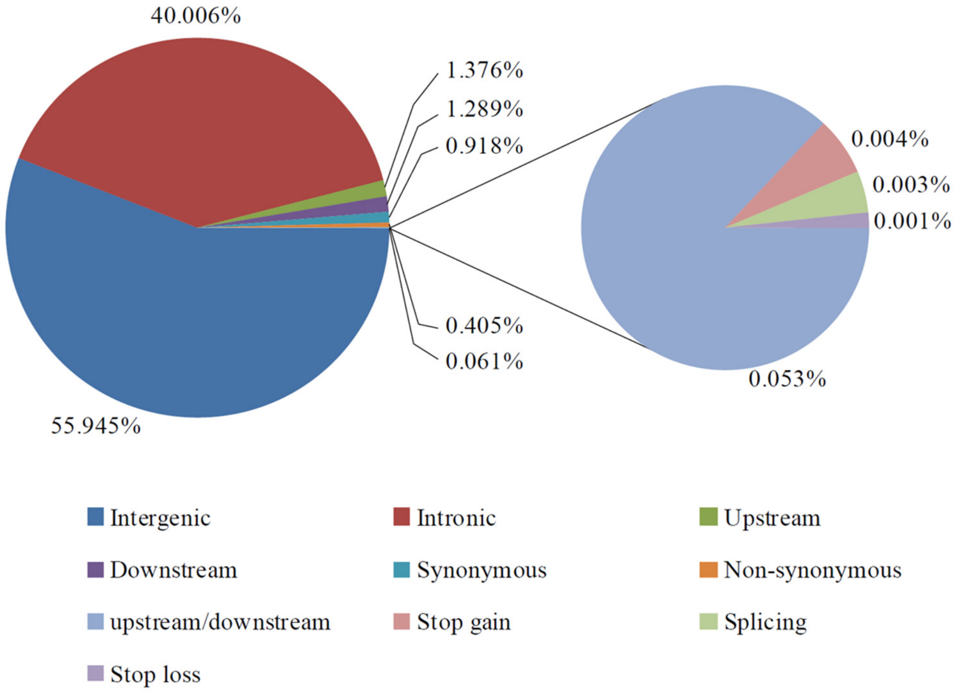


Figure S8. Genomic distribution of SNPs for generation 19. Related to Figure 3.

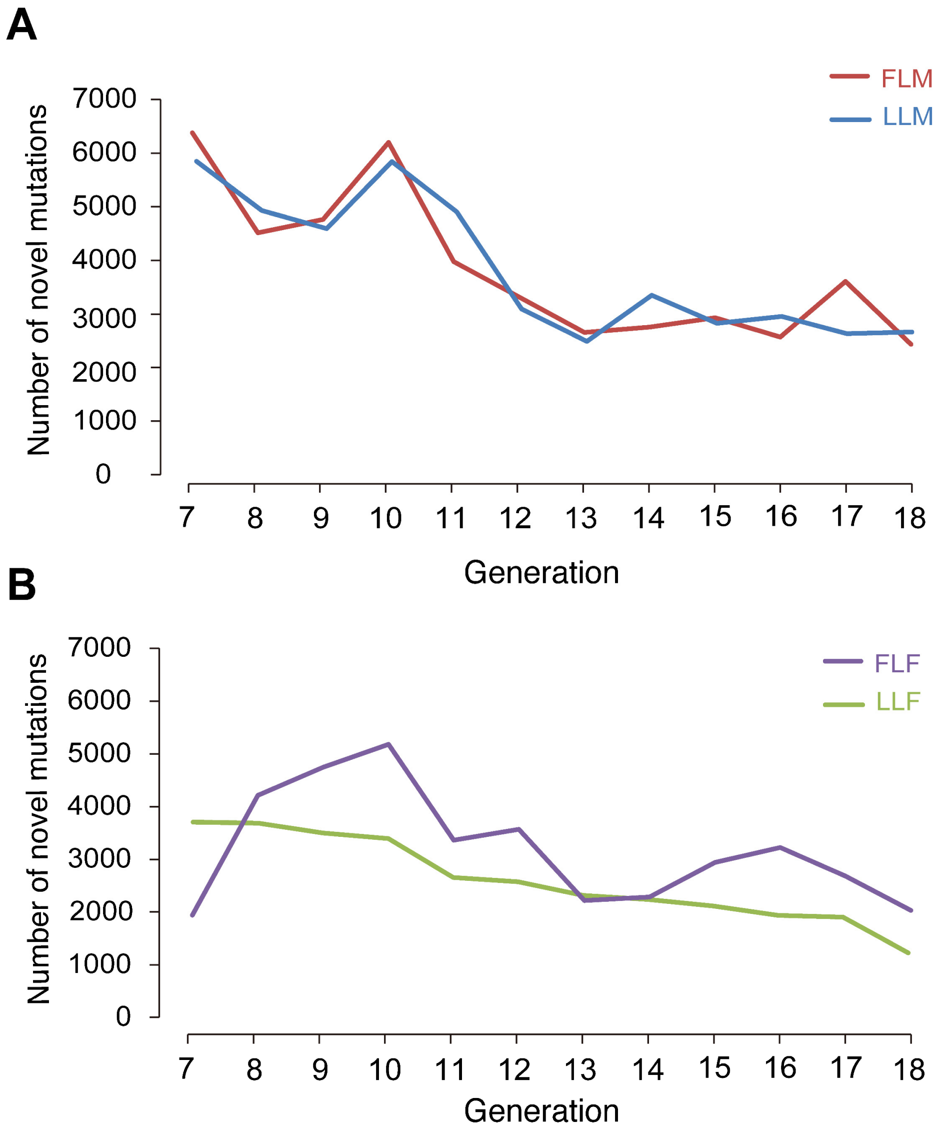


Figure S9. Number of novel mutations distributed in the chicken genome. Related to Figure 4. Novel Mutations as distributed in cocks (FLM, LLM) and hens (FLF, LLF) for both chicken lines. FLM and LLM, male birds of the fat and lean lines; FLF and LLF, female birds of the fat and lean lines.

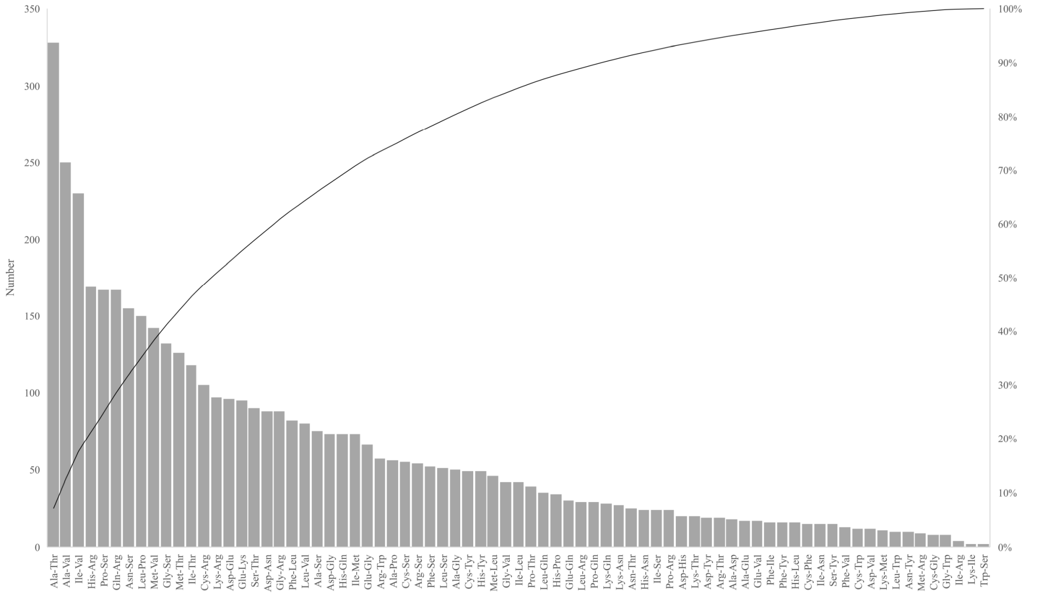


Figure S10. The number of amino acid changes caused by novel mutations. Related to Figure 4.

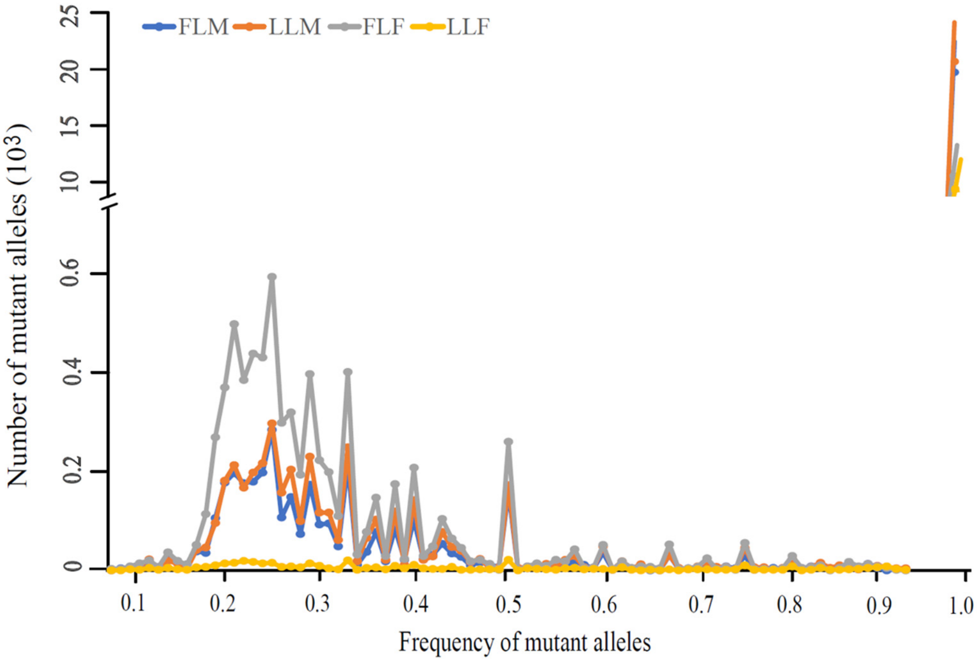


Figure S11. Distribution of frequency and number of novel mutations discovered at G18. Related to Figure 4.

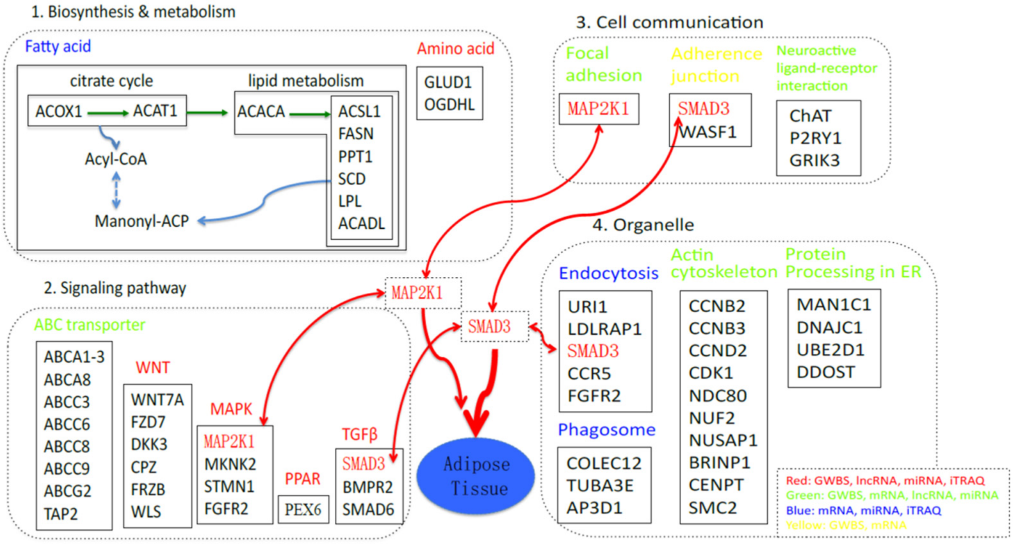


Figure S12. Signalling pathways found by functional genomics analysis at G19. Related to Figures 1, 2 and 3.

Molecular pathways related to adipose tissue growth and development were found, and were common to the results obtained by time-series analysis.

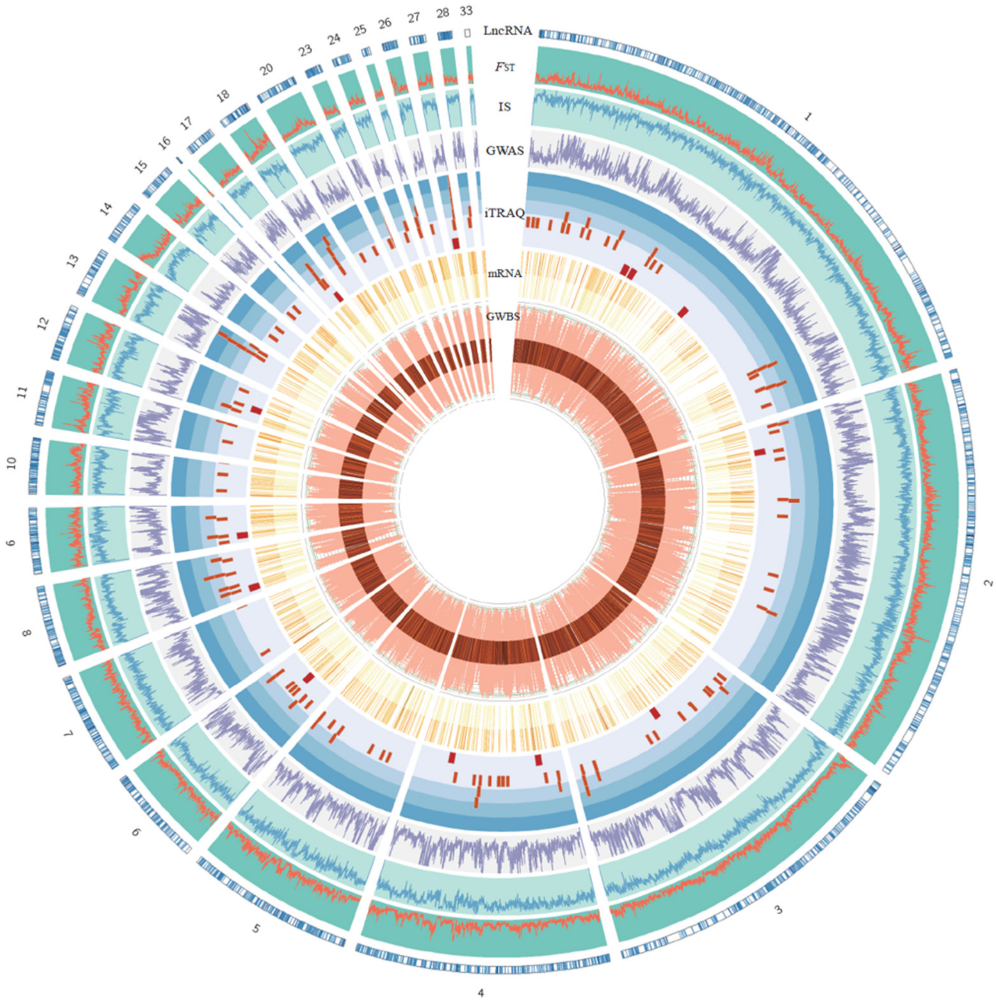
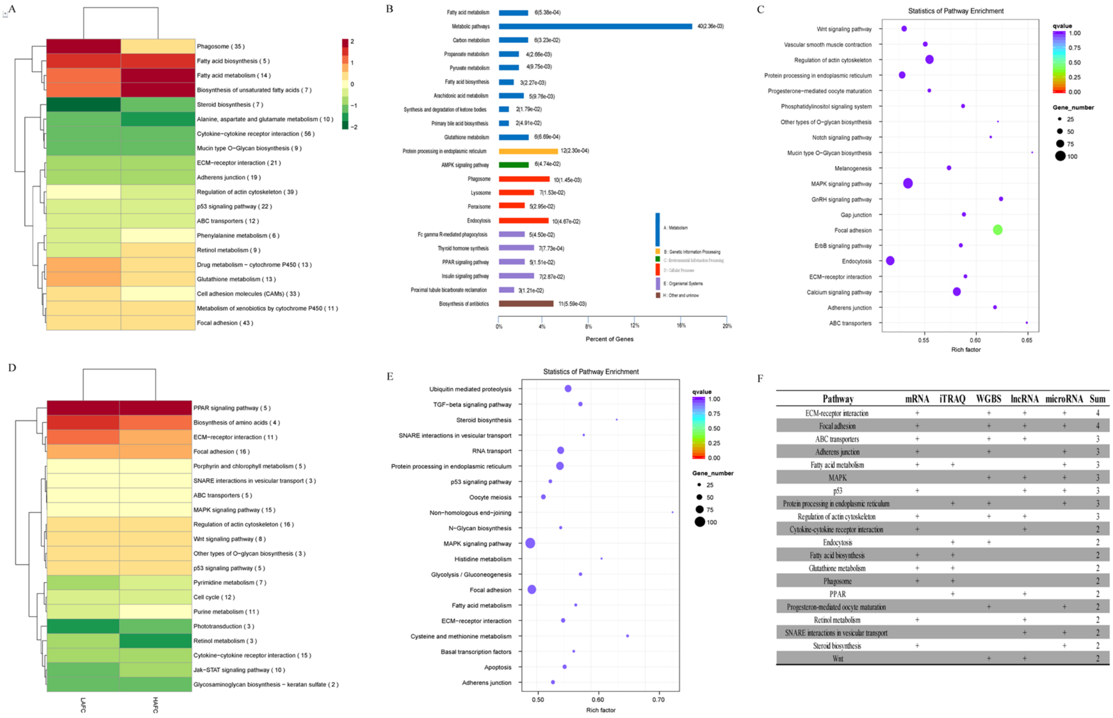


Figure S13. Integrative genomics analysis. Related to Figures 1, 2, 3 and 4.

An integrated view on functional genomics study. From outer to inner circle, F_{ST} (red) and IS (blue) at G18; GWAS P-values (violet); differentially expressed proteins by iTRAQ (brown); RNA expression (FPKM) (yellow) and DNA methylation levels (pink) of fat and lean birds.



Pathway	mRNA	iTRAQ	WGBS	lncRNA	microRNA	Sum
ECM-receptor interaction	+		+	+	+	4
Focal adhesion	+		+	+	+	4
ABC transporters	+		+	+		3
Adherens junction	+		+		+	3
Fatty acid metabolism	+	+			+	3
MAPK			+	+	+	3
p53	+			+	+	3
Protein processing in endoplasmic reticulum		+	+		+	3
Regulation of actin cytoskeleton	+		+	+		3
Cytokine-cytokine receptor interaction	+			+		2
Endocytosis		+	+			2
Fatty acid biosynthesis	+	+				2
Glutathione metabolism	+	+				2
Phagosome	+	+				2
PPAR		+		+		2
Progesteron-mediated oocyte maturation			+		+	2
Retinol metabolism	+			+		2
SNARE interactions in vesicular transport				+	+	2
Steroid biosynthesis	+				+	2
Wnt			+	+		2

Figure S14. Comprehensive functional genomics reveal the molecular consequences of genomic variants under selection. Related to Figures 1, 2, 3 and 4.

Common signalling pathways important for adipose tissue growth and development were found, such as fatty acid metabolism, ECM-receptor interaction, autophagy and lysosome. However, signalling pathways specific to each functional genomics data type indicated that regulatory mechanisms at the genome, transcriptome and proteome levels could be distinct. (A) 2,325 mRNA significantly differentially expressed. (B) 195 significantly differentially expressed proteins (iTRAQ proteomics). (C) 678 genes were highly methylated in the fat line (GWBS). (D) and (E) 111 lncRNAs and 120 miRNAs significantly differentially expressed. (F) Table showing common signalling pathways (ECM-receptor interaction, focal adhesion, ABC transporters).

Transparent Methods

1. Study samples

Broilers used in this study were from two Northeast Agricultural University (NEAU) broiler lines divergently selected for abdominal fat content (NEAUHLF). The two broiler lines have been selected since 1996 using abdominal fat percentage (AFP = AFW/body weight) and plasma very low-density lipoprotein (VLDL) concentration as selection criteria (Guo et al., 2011). The G0 generation of the two lines came from the same grandsire line originating from the Arbor Acres broiler, which was then divided into two lines according to their plasma VLDL concentration at 7 weeks of age. The G0 birds were mated (one sire: four dams) to produce 25 half-sib families for each line, with an average of 70 G1 offspring per family in two hatches. From G1 to G19, the birds of each line were raised in two hatches and housed in pens with five birds per cage. Plasma VLDL concentrations were measured for all male birds, which had free access to feed and water, and the AFP of the male birds in the first hatch was measured after slaughter at 7 weeks of age. Sib birds from the families with lower (lean line) or higher (fat line) AFP than the average value for the population were selected as candidates for breeding, considering the plasma VLDL concentration and the body weights of male birds in the second hatch and the egg production of female birds in both hatches. These birds were kept under the same environmental conditions and had free access to feed and water. Commercial corn-soybean-based diets that met all National Research Council (NRC) requirements were provided. From hatch to 3 weeks of age, the birds received a starter feed (3,000 kal ME= kg and 210 g =kg CP) and from 4 weeks of age to slaughter the birds were fed a grower diet (3,100 kal ME= kg and 190 g= kg CP).

For the 15 generations genome sequencing experiment, 60 pooled-DNA samples were produced from G4-G18, including all roosters and hens used as parents for the next generation (3,642 birds in total, and number of birds in each pool was given in Table S1). Between the divergent lines, striking phenotypic differences in fatness could be seen starting from G4, and those samples were used to explore the allele frequency changes. GWAS was conducted on 330 male birds from G19 (160 and 170 from the fat and lean lines, respectively) (Table S11). DNA samples for 15 generations genome sequencing and GWAS experiments were prepared from blood samples, and their quality were checked for preparing sequencing libraries according to the standard protocols. Abdominal adipose tissue samples from 10 male birds (5 from each line) (selected based on similar body weights but different abdominal fat weights) were immediately frozen in liquid nitrogen, and stored at -80℃. Then DNA, total RNA, and protein samples from abdominal fat tissue were prepared for functional genomics studies (Figure S13).

2. DNA library preparation and sequencing

A total amount of 1.5 μg DNA per sample was used as the input material for the DNA sample preparations. Sequencing libraries were generated using the Truseq Nano DNA HT Sample Preparation Kit (Illumina USA) following the manufacturer's recommendations and index codes were added to attribute sequences to each sample. The DNA sample was fragmented by sonication to an average size of 350bp, and then DNA fragments were end-polished, A-tailed, and ligated with the full-length adaptor for Illumina sequencing with further PCR amplification. Then PCR products were purified (AMPure XP system) and libraries were analysed for size distribution by Agilent2100 Bioanalyzer and quantified using real-time PCR.

The 15 generations genome sequencing of 60 pooled DNA libraries generated ~2.35 Tb of sequences on the Illumina HiSeq 4000 platform. We also constructed individual 350bp DNA libraries for 330 male birds of G19, and carried out whole-genome re-sequencing with approximately 2.19 Tb for GWAS analysis.

3. Whole-genome DNA bisulfite sequencing

A total amount of 5.2 μg genomic DNA spiked with 26 ng lambda DNA was fragmented by sonication to 200-300 bp with Covaris S220, followed by end repair and adenylation. Cytosine-methylated barcodes were ligated to sonicated DNA as per the manufacturer's instructions. Then these DNA fragments were treated with bisulfite using the EZ DNA Methylation-Gold Kit (Zymo Research), before the resulting single-strand DNA fragments were PCR amplified using KAPA HiFi HotStart Uracil + ReadyMix (2X).

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on Illumina Hiseq 2000/2500 platform and 125bp paired-end reads were generated. Image analysis and base calling were performed with Illumina CASAVA pipeline, and finally 125bp paired-end reads were generated.

4. RNA library preparation and sequencing

A total of 3 μg RNA per sample was used as the input material for the RNA sample preparations. Ribosomal RNA was removed by using the Epicentre Ribo-zero™ rRNA Removal Kit (Epicentre, USA), and the rRNA-free residue was cleaned up by ethanol precipitation. Sequencing libraries were then generated using the rRNA-depleted RNA by the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (NEB, USA) following the manufacturer's recommendations. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X).

First-strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H$^-$). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. In the reaction buffer, dNTPs with dTTP were replaced by dUTP. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure was ligated to prepare for hybridization. To select cDNA fragments of preferentially 350bp in length, the library fragments were purified with the AMPure XP system (Beckman Coulter, Beverly, USA). Then 3 μL USER Enzyme (NEB，USA) was used with size-selected, adaptor-ligated cDNA at 37°C for 15 min followed by 5 min at 95°C before PCR. Then PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. The products were purified (AMPure XP system) and library quality was assessed on the Agilent Bioanalyzer 2100 system.

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq PE Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the libraries were sequenced on the Illumina Hiseq 4000 platform and 150bp paired-end reads were generated.

5. Small RNA sequencing

Total RNA per sample in a total amount of 3 μg was used as input material for the small RNA library. Sequencing libraries were generated using the NEBNext® Multiplex Small RNA Library Prep Set for Illumina® (NEB, USA.) following the manufacturer's recommendations and index codes were added to attribute sequences to each sample. Briefly, an NEB 3' SR Adaptor was directly, and specifically ligated to the 3' end of miRNA, siRNA and piRNA. After the 3' ligation reaction, the SR RT Primer was hybridized to the excess of 3' SR adaptor (that remained free after the 3' ligation reaction) and transformed the single-stranded DNA adaptor into a double-stranded DNA molecule. This step is important to prevent adaptor-dimer formation; dsDNAs are not substrates for ligation mediated by T4 RNA Ligase 1 and therefore do not ligate to the 5' SR adaptor in the subsequent ligation step. A 5' end adaptor was ligated to the 5' end of miRNAs, siRNA and piRNA. Then first strand cDNA was synthesized using M-MuLV Reverse Transcriptase (RNase H$^-$). PCR amplification was performed using LongAmp Taq 2X Master Mix, SR Primer for Illumina and index (X) primer. PCR products were purified on an 8% polyacrylamide gel (100V, 80 min). DNA fragments corresponding to 140~160 bp (the length of small noncoding RNA plus the 3' and 5' adaptors) were recovered and dissolved in 8 μL elution buffer. Library quality was assessed on the Agilent Bioanalyzer 2100 system using DNA High Sensitivity Chips.

Clustering of the index-coded samples was performed on a cBot Cluster Generation System using the TruSeq SR Cluster Kit v3-cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on Illumina Hiseq 2500 platform and 50bp single-end reads were generated.

6. iTRAQ

Protein isolation, quality control, sample preparation and labeling followed the standard procedures (Supplemental Note). A total of 100μg protein was submitted for mass spectrometry (2D-LC-MSMS) analysis on the Agilent 1200 HPLC System (Agilent). After Nano-RPLC treatment on the Eksigent nanoLC-Ultra™ 2D System (AB SCIEX), data acquisition was then performed with a Triple TOF 5600 System (AB SCIEX, USA) fitted with a Nanospray III source (AB SCIEX, USA) and a pulled quartz tip as the emitter (New Objectives, USA).

7. Sequence alignment and variant calling

To ensure reads were reliable and no artificial bias (for example, low quality paired reads, mainly resulting from base-calling duplicates and adaptor contamination) in the following analyses, raw data (raw reads) were first processed through a series of quality control (QC) procedures using in-house C scripts. QC standards were the following:

(1) Removing reads with ≥10% unidentified nucleotides (N);

(2) Removing reads with > 50% bases having Phred quality < 5;

(3) Removing reads with > 10 nt aligned to the adaptor, allowing ≤10% mismatches;

(4) Removing putative PCR duplicates generated by PCR amplification in the library construction process (i.e. identical read 1 and read 2 of two paired-end reads) by SAMtools (Li et al., 2009).

The remaining high-quality paired-end reads were mapped to the *Gallus gallus* (Galgal 6) reference genome using Burrows-Wheeler algorithm (Version: 0.7.8) (Li and Durbin, 2009).

After alignment, variant calling was performed for all samples by using the Unified Genotyper function in GATK 3.3 software (McKenna et al., 2010). SNPs were selected by using the Variant Filtration parameter in GATK (settings: --filterExpression "QD < 4.0 || FS > 60.0 || MQ < 40.0 ", -G_filter "GQ<20", --cluster WindowSize 4), and InDel with the filter expression "FS > 200.0 ||ReadPosRankSum < −20.0 || InbreedingCoeff < −0.8".

8. Selective sweep analysis

We used allele frequencies at variable sites to identify signals of selection in 40 kb windows with a step size of 20 kb using two approaches: for each window we calculated (1) the average fixation index, F_{ST}, between lean and fat lines and also between each of the 14 generations (from G4 to G17) and G18 in lean and fat lines respectively, and (2) Identity scores (IS).

At each detected SNP position, we counted the number of reads corresponding to the most and least frequently observed allele (n_{maj} and n_{min} , respectively) in each pool. We calculated genome-wide distribution of F_{ST}. F_{ST} was calculated from the allele frequencies (not from the allele counts) using the standard equation (Tang et al., 2005; Hartl and Clark, 2007).

Identity scores (IS) were calculated to evaluate the pairwise similarities of the sequences. For each identified SNP, we determined the fraction of reads that corresponded to the reference allele, F, in each sequence. The IS values of pooled SNPs were then calculated as IS = 1 -(|$F_{sequence1}$ - $F_{sequence2}$|), with SNPs being assessed only if at least one read was obtained in each sequence. The IS values for a sequence was the mean of all SNP IS values observed in the sequence for a specific comparison.

To minimize the volatility of F_{ST} and IS when calculated in each generation, the mean of F_{ST} and IS in three consecutive generations was calculated for each bin. Then linear regression was performed to examine the relationship between generation and F_{ST} or IS. The equation was as follows: $Y = aX+b$, where Y was generation, X was F_{ST} or IS, a was slope and b was intercept. Afterwards, the P value of each bin was adjusted by Bonferroni correction (Hochberg, 1988).

9. Mutation analysis

In each generation, the birds were divided into four groups, including males of fat line (FLM), males of lean line (LLM), females of fat line (FLF) and females of lean line (LLF). In order to improve the accuracy of the identification of mutations, reduce the randomness of mutations caused by the comparison between two separate individuals/generations, and exclude mutations that may be inherited from parents to offspring, we compared each group (FLM, LLM, FLF and LLF) in G7 with the groups in G4, G5 and G6 simultaneously. If the mutations are new in G7, which means that they are not detected in G4, G5 or G6, these mutations are defined as "novel mutations". For G8, we compared each group (FLM, LLM, FLF and LLF) in G8 with the groups in G4, G5, G6 and G7 simultaneously and the novel mutations are defined as they appeared in G8 but not detected in G4, G5, G6 or G7, and so on until G18 are compared with G4 to G17. Of course, the novel mutation will be filtered if it is "Miss" genotype in all former generations.

10. Phylogenetic tree and population structure

To avoid statistical bias from low-coverage data, our population genetics inference was based on genotype likelihoods (GL) which can take genotype uncertainty into account. Following reads mapping, high-quality alignments (BAM files) were input to the program ANGSD (Korneliussen et al., 2014). The command was used to identify SNPs with the parameters as './angsd -bam bam.list -GL 1 -doMaf 1 -SNP_pval 0.01 -doMajorMinor 1 -doGeno 5 -dopost 1 -doCounts 1 -doGlf 2 -dumpCounts 2'. Only SNPs with at least 20% varieties genotyped were left for subsequent imputation. A total of 7,070,164 high quality SNPs (coverage depth ≥3 and ≤90 , maf ≥ 0.05) were used in subsequent analysis. To infer genetic structure, we used the package *frappe* (Patterson et al., 2006) to estimate individual ancestry and admixture proportions with estimated individual ancestry and admixture proportions based on 7,070,164 high quality SNPs. We predefined the number of genetic clusters from K = 2 to 5 and 10,000 iterations for each run, with default methods and settings used in *frappe* analysis. To clarify the phylogenetic relationships from a genome-wide perspective, an individual-based neighbour-joining (NJ) tree was constructed based on the p-distance using the software TreeBest (http://treesoft.sourceforge.net/treebest.shtml). The software MEGA5 was used for visualizing the phylogenetic trees (Tamura et al., 2011). We also conducted principal component analysis (PCA) to evaluate genetic structure using the software GCTA, and the significance level of the eigenvectors was determined using the Tracey-Widom test (Yang et al., 2011).

11. Linkage disequilibrium analysis

We compared the pattern of linkage disequilibrium (LD) among 330 chickens using the 7,070,164 high-quality SNPs. To estimate LD decay, the squared correlation coefficient (r^2) between pairwise SNPs was calculated using the software Haploview (Barrett et al., 2005). Parameters in the program were set as: '-n -dprime -minMAF 0.05'. The average r^2 value was calculated for pairwise markers in a 500-kb window and averaged across the whole genome. We found differences in the rate of decay and level of LD value, which reflected variations in population demographic history and effective population size (N_e) among breeds/populations.

12. GWAS analysis

In our association panel containing 330 samples, a total of 7,070,164 SNPs were used in our GWAS for abdominal fat content traits. Association analysis was conducted using GEMMA (genome-wide efficient mixed-model association) (Zhou and Stephens, 2014). The statistical model was $y = X\alpha + S\beta + K\mu + e$, where y represents the phenotype, X represents the genotype, S is the structure matrix, K is the relative kinship matrix, α and β represent fixed effects, Kμ is the random effect, and e is normally distributed residual error. The top three structural components were used to build up the S matrix for population structure correction. The matrix of simple matching coefficients was used to build up the K matrix.

13. Functional annotation of genetic variants

SNP annotation was performed according to the *Gallus gallus* (Galgal6.0) genome using the package ANNOVAR (Version: 2013-05-20) (Wang et al., 2010; Yang and Wang, 2015). Based on the genome annotation, SNPs were categorized into exonic regions (overlapping with a coding exon), intronic regions (overlapping with an intron), splicing sites (within 2 bp of a splicing junction), upstream and downstream regions (within a 1 kb region upstream or downstream from the transcription start site), and intergenic regions. SNPs in coding exons were further grouped into synonymous SNPs (not causing amino acid changes) or nonsynonymous SNPs (causing amino acid changes). Also, mutations causing stop gain and stop loss were also classified into this group.

Supplemental Note. Related to Transparent Methods and Figure 1.

1. Animals. For the pooled-seq analysis, genomic DNA was isolated from the blood of all male and female broilers that have been used as parents for the next generation. After quality control and precise quantification of DNA concentrations, four pools of equal amounts of DNA from each bird were created, for the male and female birds from the fat and lean lines, respectively.

Male birds for GWAS analysis were selected according to their phenotypic records and distribution in the whole population. In order to have a good representation of the whole population, no more than two birds can be selected from one full-sib family.

Birds used for functional genomics study were selected based on their phenotypic records for abdominal fat weight and percentage, which distributed close to the average mean value within each population, rather than extreme values.

2. WGBS data analysis

(1) Quality control. Read sequences produced by the Illumina pipeline in FastQ format were pre-processed through in-house Perl scripts. Firstly, as a subset of reads contained all of part of the 3'adapter oligonucleotide sequence, every read was scanned for the adapter sequence, and if detected the read was filtered out. Then, since some reads had unknown base (N) in their sequences, the percentage of Ns in each read was calculated, and if the percentage of Ns was larger than 10% the read was removed. Finally, reads with low quality (Phred score <= 5, and percentage of the low quality bases >= 50%) were trimmed. In parallel, Q20, Q30 and GC content of the data were calculated. The remaining reads (i.e., clean reads) were used for the subsequent analyses.

(2) Reads mapping to the reference genome. Bismark software (version 0.12.5) was used to perform alignments of bisulfite-treated reads to a reference genome with the default parameters (Krueger and Andrews, 2011). The reference genome was firstly transformed into bisulfite-converted version (C-to-T and G-to-A converted) and then indexed using Bowtie2 (Langmead and Salzberg, 2012). Sequence reads were also transformed into fully bisulfite-converted versions (C-to-T and G-to-A converted). Sequence reads that produce a unique best alignment from the two alignment processes (original top and bottom strand) are then compared to the normal genomic sequence and the methylation state of all cytosine positions in the read is inferred. The same reads that aligned to the same regions of genome were regarded as duplicated ones. The sequencing depth and coverage were summarized using deduplicated reads. The results of methylation extractor were transformed into bigWig format for visualization using IGV browser (Robinson et al., 2011). The sodium bisulfite non-conversion rate was calculated as the percentage of cytosines sequenced at cytosine reference positions in the lambda genome.

(3) Estimating methylation level. To identify the methylation site, we modeled the sum of methylated counts as a binomial (Bin) random variable with methylation rate. We employed a sliding-window approach, which is conceptually similar to approaches that have been used for bulk BS-Seq (http://www.bioconductor.org/packages/2.13/bioc/html/bsseq.html). With window size of 3,000 bp and step size of 600 bp, the sum of methylated and unmethylated read counts in each window were calculated (Smallwood et al., 2014). Methylation level (ML) for each C site shows the fraction of methylated Cs, and is defined as: ML=mC/(mC+umC). Calculated ML was further corrected with the bisulfite non-conversion rate according to previous studies (Lister et al., 2013). Given the bisulfite non-conversion rate r, the corrected ML was estimated as: ML_corrected=(ML-r)/(1-r). Differentially methylated regions (DMRs) were identified using the BS-seq software.

(4) **GO and KEGG enrichment analysis of DMR-related genes.** Gene Ontology (GO) enrichment analysis of genes related to DMRs was implemented by the GOseq R package (Young et al., 2010), in which gene length bias was corrected. GO terms with corrected P-value less than 0.05 were considered significantly enriched by DMR-related genes. We used KOBAS software (Mao et al., 2005) to test the statistical enrichment of DMR-related genes in KEGG pathways.

3. mRNA and lncRNA sequencing data analysis

(1) Quality control. Raw reads in fastq format were firstly processed through in-house Perl scripts. In this step, clean data (clean reads) were obtained by removing reads containing adapter, reads containing poly-N and low quality reads from raw data. At the same time, Q20, Q30 and GC content of the clean data were calculated. All the down stream analyses were based on the clean data with high quality.

(2) Mapping to the reference genome. Reference genome and gene model annotation files were downloaded directly from Ensembl (http://www.ensembl.org/index.html). Index of the reference genome was built using Bowtie2 and paired-end clean reads were aligned to the reference genome using TopHat v2.0.9.

(3) Transcriptome assembly. The mapped reads of each sample were assembled by Cufflinks (v2.1.1) in a reference-based approach (Trapnell et al., 2010). This methods use spliced reads to determine exons connectivity, but with two different approaches. It uses a probabilistic model to simultaneously assemble and quantify the expression level of a minimal set of isoforms that provides a maximum likelihood explanation of the expression data in a given locus (Cabili et al., 2011). Cufflinks was run with ‘min-frags-per-transfrag=0’ and ‘--library-type’, other parameters were set as default.

(4) Coding potential analysis. CPC (Coding Potential Calculator) (0.9-r2) mainly through assess the extent and quality of the ORF in a transcript and search the sequences with known protein sequence database to clarify the coding and non-coding transcripts (Kong et al., 2007). We used the NCBI eukaryotes' protein database and set the e-value ‘1e-10’ in our analysis.

We translated each transcript in all three possible frames and used Pfam-Scan (v1.3) to identify occurrence of any of the known protein family domains documented in the Pfam database (release 27; used both Pfam A and Pfam B) (Finn et al., 2014). Any transcript with a Pfam hit would be excluded in following steps. Pfam searches use default parameters of -E 0.001 --domE 0.001 (Bateman et al., 2002).

PhyloCSF (Lin et al., 2011) (phylogenetic codon substitution frequency) (v20121028) examines evolutionary signatures characteristic to alignments of conserved coding regions, such as the high frequencies of synonymous codon substitutions and conservative amino acid substitutions, and the low frequencies of other missense and non-sense substitutions to distinguish protein-coding and non-coding transcripts. We built multi-species genome sequence alignments and ran phyloCSF with default parameters.

Transcripts predicted with coding potential by either/all of the four tools above were filtered out, and those without coding potential were our candidate set of lncRNAs.

(5) Conservative analysis. Phast (v1.3) is a software package contains many statistical programs, mostly used in phylogenetic analysis (Siepel et al., 2005), and phastCons is a conservation scoring and identification program of conserved elements. We used phyloFit to compute phylogenetic models for conserved and non-conserved regions among species and then gave the model and HMM transition parameters to phastCons to compute a set of conservation scores of lncRNA and coding genes.

(6) Target gene prediction. The cis-acting role of lncRNA is its acting on neighboring target genes. We searched coding genes in the upstream and downstream regions (100 kbp) of lncRNA and then analyzed their function next. The trans-acting lncRNA and its target genes are identified by the expression level. While there were no more than 10 samples, we calculated the expressed correlation between lncRNAs and coding genes with custom scripts; otherwise, we clustered the genes from different samples with WGCNA (Langfelder and Horvath, 2008) to search common expression modules and then analyzed their function through functional enrichment analysis.

(7) Quantification of gene expression level. Cuffdiff (v2.1.1) was used to calculate FPKMs (fragments per kilo-base of exon per million fragments mapped) of both lncRNAs and coding genes in each sample (Trapnell et al., 2010). Gene FPKMs were computed by summing the FPKMs of transcripts in each gene group.

(8) Differential expression analysis. Cuffdiff provides statistical routines for determining differential expression in digital transcript or gene expression data using a model based on the negative binomial distribution. For biological replicates, transcripts or genes with a *P*-adjust <0.05 were assigned as differentially expressed.

(9) GO and KEGG enrichment analysis. Gene Ontology (GO) enrichment analysis of differentially expressed genes or lncRNA target genes was also implemented by the GOseq R package. GO terms with corrected *P* value less than 0.05 were considered significantly enriched by differential expressed genes. We used KOBAS software to test the statistical enrichment of differential expression genes or lncRNA target genes in KEGG pathways.

4. Small RNA sequencing

(1) Quality control. Raw data were firstly processed through custom perl and python scripts. During this step, clean data (clean reads) were obtained by removing reads containing poly-N, with 5' adapter contaminants, without 3' adapter or the insert tag, containing poly A or T or G or C and low quality reads from raw data. At the same time, Q20, Q30 and GC-content of the raw data were calculated. Then, clean reads were used for all the downstream analyses.

(2) Reads mapping to the reference sequence. The small RNA tags were mapped to reference sequence by Bowtie2 without mismatch to analyze their expression and distribution on the reference genome (Langmead and Salzberg, 2012).

(3) Known miRNA alignment. Mapped small RNA tags were used to looking for known miRNA. miRBase20.0 was used as reference, modified software mirdeep2 (Friedländer et al., 2012), and srna-tools-cli were used to obtain the potential miRNA and draw the secondary structures. Custom scripts were used to obtain the miRNA counts as well as base bias on the first position of identified miRNA and on each position of all identified miRNA respectively.

(4) Removing tags. To remove tags originating from protein-coding genes, repeat sequences, rRNA, tRNA, snRNA, snoRNA, and small RNA tags were mapped to RepeatMasker, Rfam database or those types of data from the targeted species itself.

(5) Novel miRNA prediction. The characteristics of hairpin structure of miRNA precursor can be used to predict novel miRNA. The available software miREvo (Wen et al., 2012) and mirdeep2 (Friedländer et al., 2012) were integrated to predict novel miRNA through exploring the secondary structure, the Dicer cleavage site and the minimum free energy of the small RNA tags unannotated in the former steps. At the same time, custom scripts were used to obtain the identified miRNA counts as well as base bias on the first position with certain length and on each position of all identified miRNA respectively.

(6) Small RNA annotation summary. To make every unique small RNA mapped to only one annotation, we follow the following priority rule: known miRNA > rRNA > tRNA > snRNA > snoRNA > repeat > gene > novel miRNA > ta-siRNA. The total rRNA proportion was used a marker as sample quality indicator. Usually it should be less than 40% in animal samples as high quality.

(7) miRNA editing analysis

The seed region (position 2~8) of a mature miRNA was highly conserved. The target of a miRNA might be different with the changing of nucleotides in this region. In our analysis pipeline, miRNA that might have base edit could be detected by aligning all the sRNA tags to mature miRNA, allowing one mismatch.

(8) miRNA family analysis. The occurrence of miRNA families was explored using homology approach. In short, known miRNA were identified by miFam.dat (http://www.mirbase.org/ftp.shtml) and novel miRNA precursor was detected by searching Rfam (http://rfam.sanger.ac.uk/search/).

(9) Target gene prediction. Prediction of the target gene of miRNA was performed by miRanda (Enright et al., 2003).

(10) Quantification of miRNA. The expression levels of miRNA were estimated by TPM (transcript per million) through the following criteria (Zhou et al., 2010): normalized expression = mapped readcount/Total reads*1000000.

(11) Differential expression of miRNA. For the samples with biological replicates: Differential expression analysis of two conditions/groups was performed using the DESeq R package (1.8.3) (Anders and Huber, 2010). *P* values was adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.05 was set as the threshold for significantly differential expression by default.

5. iTRAQ

(1) Protein extraction. Protein extraction was performed according to a standard protocol (Damerval et al., 1986). Briefly, 1) the tissue sample was ground to powder with liquid nitrogen; 2) Add 10mL cooled acetone contained 10% TCA to 1 g sample power at - 20 °C for 1 hour; 3) Centrifuge by 15000g for 15 min at 4 °C, the deposit was collected and then cooled acetone was added at - 20°C for 1 hour; 4) Repeated step 3; 5) Centrifuge by 15000g for 15 min at 4 °C , collect the deposit and dried by vacuum freeze dryer ; 6) The deposit was dissolved in lysis solution at 30°C for 1 hour; 7) Centrifuge the solution by 15000g for 15 min at room temperature, collect the supernatant and centrifuge again; 8) The supernatant was the extracted protein solution. The concentrations of the protein extracts were determined (Bradford, 1976), and store at -80°C for iTRAQ analysis.

(2) SDS-PAGE electrophoresis. A total of 10 μg protein sample was loaded onto 12% SDS-PAGE gel, and visualized by CBB stain (Candiano et al., 2004), and the stained gel was scanned by the Image Scanner (GE Healthcare, USA) at a resolution of 300 dots per inch.

(3) Protein reduction, cysteine block and digest. 1) Take 100μg protein for each sample and add five volume of cold acetone at - 20°C for 1hour. 2) Centrifuge by 12000rpm for 15 mins at 4°C, collect the deposit and dried by vacuum freeze dryer; 3) Add 50μL dissolution buffer for the deposit and add 4μL reducing reagent. Incubate the solution at 60°C for 1hour. 4) Add 2μL cysteine-blocking reagents at room temperature for 10min. Clean the protein solution by using 10 KDa ultrafiltration tube to centrifuge by 12000rpm for 20min. 5) Add 100μL dissolution buffer, centrifuge by 12000rpm for 15 mins and repeat this step three times. 6) Place column in a new tube, add 50μL sequencing-grade trypsin (50ng/μL) and incubate at 37°C for 12 hours. 7) Centrifuge by 12000rpm for 20 mins, collect the peptide. Transfer the filter units to new collection tube and add 50μL dissolution buffers to centrifuge the tube again. Combine the two filtered solution.

(4) Protein labeling and MS analysis. 1) Allow each vial of iTRAQ reagent required to reach room temperature. 2) Centrifuge iTRAQ reagent to the bottom of the tube. 3) Add 150 μL of ethanol to each room- temperature iTRAQ reagent vial. 4) Transfer 50 μL sample (100μg peptide) to one new tube, add iTRAQ reagent and incubate the tube at room temperature for 2 hours; 5) Add 100μL water to stop the labeling reaction; 6) Vortex each tube to mix, then spin and collect the solution; 7) Dry the sample in a vacuum freeze dryer for iTRAQ analysis.

(5) 2D-LC-MSMS. The strong cation exchange (SCX) analysis included the following steps: 1) Dry sample was resuspended with 100μL buffer A; 2) The SCX was employed on the Agilent 1200 HPLC System (Agilent). The HPLC column was from Michrom. The parameter was: Poly-SEA 5μ 300Å 2.0 x 150 mm with 215nm and 280nm UV detection. Separation was performed at 0.3 ml/min using a nonlinear binary gradient starting with buffer A and transitioning to buffer B. 3) Collect the first segment from 0-5 mins, then collect each segment with 4 mins interval for the 6-44 mins, and for the last segment from 45-50 mins, with a total of 12 segments. Dry every segment in a vacuum frozen dryer for LC-MSMS analysis.

(6) RPLC-MSMS analysis. 1) Samples were resuspended with Nano-RPLC buffer A. 2) The online Nano-RPLC was employed on the Eksigent nanoLC-Ultra™ 2D System (AB SCIEX). The samples were loaded on C_{18} nanoLC trap column (100μm× 3cm, C_{18}, 3μm, 150Å) and washed by Nano-RPLC Buffer A(0.1%FA, 2%ACN) at 2μL/min for 10 mins. 3) An elution gradient of 5-35% acetonitrile (0.1% formic acid) in 70 mins gradient was used on an analytical ChromXP C_{18} column (75 μm x 15cm, C_{18}, 3μm 120 Å) with spray tip. 4) Data acquisition was performed with a Triple TOF 5600 System (AB SCIEX, USA) fitted with a Nanospray III source (AB SCIEX, USA) and a pulled quartz tip as the emitter (New Objectives, USA). Data were acquired using an ion spray voltage of 2.5 kV, curtain gas of 30 PSI, nebulizer gas of 5 PSI, and an interface heater temperature of 150℃. For information dependant acquistion (IDA), survey scans were acquired in 250ms and as many as 35 product ion scans were collected if they exceeded a threshold of 150 counts per second (counts/s) with a 2^+ to 5^+ charge-state. The total cycle time was fixed to 2.5s. A rolling collision energy setting was applied to all precursor ions for collision-induced dissociation (CID). Dynamic exclusion was set for ½ of peak width (18s). And the precursor was then refreshed off the exclusion list.

(7) Protein identification and quantification. Data were processed with Protein Pilot Software v.5.0 (AB SCIEX, USA) against *Gallus gallus* database using the Paragon algorithm (Shilov et al., 2007). The experimental data from tandem mass spectrometry (MS) was used to search the database to obtain protein identification. Protein identification was performed with the search option: emphasis on biological modifications. An automatic decoy database search strategy was employed to estimate the false discovery rate (FDR) using the PSPEP (Proteomics System Performance Evaluation Pipeline Software, integrated in the ProteinPilot Software). The FDR was calculated as the false positive matches divided by the total matches. The database search parameters were as follows: The iTRAQ 8-plex was chosen for protein quantification with unique peptides during the search. A total of 2,137 and 1,727 proteins with the value of global FDR less than 1% were considered for further analysis.

6. Integrated analysis of functional genomics data

All genes found by selective sweep analysis, GWAS and functional genomics were searched and compared. Genome-wide DNA methylation sequencing and RNA expression data were combined for the analysis, according to the methylation level of genes and their corresponding gene expression levels. All identified genes were submitted for gene functional enrichment analysis to identify corresponding functional pathways.

Supplemental references:

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. (2002). The Pfam protein families database. Nucleic Acids Res. *30*, 276-80.

Bradford, M.M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. *72*, 248-54.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915-27.

Candiano, G., Bruschi, M., Musante, L., Santucci, L., Ghiggeri, G.M., Carnemolla, B., Orecchia, P., Zardi, L., and Righetti, P.G. (2004). Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. Electrophoresis *25*, 1327-33.

Damerval, C., de Vienne, D., Zivy, M., and Thiellement, H. (1986). Technical improvements in two-dimensional electrophoresis increase the level of genetic-variation detected in wheat-seedling proteins. Electrophoresis *7*, 52-54.

Enright, A.J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D.S. (2003). MicroRNA targets in Drosophila. Genome Biol. *5*, R1.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res. *42*, D222-30.

Friedländer, M.R., Mackowiak, S.D., Li, N., Chen, W., and Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. *40*, 37-52.

Guo, L., Sun, B., Shang, Z., Leng, L., Wang, Y., Wang, N., and Li, H. (2011). Comparison of adipose tissue cellularity in chicken lines divergently selected for fatness. Poult. Sci. *90*, 2024-2034.

Hartl, D.L., and Clark, A.G. (2007). Principles of Population Genetics (Massachusetts, Sinauer Associates).

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. Biometrika. *75*, 800-803.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res. *35*, W345-9.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics *15*, 356.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571-2.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357-9.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lin, M.F., Jungreis, I., and Kellis, M. (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics *27*, i275-82.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.

Mao, X., Cai, T., Olyarchuk, J.G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics *21*, 3787-93.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297-1303.

Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet. *2*, e190.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24-6.

Shilov, I.V., Seymour, S.L., Patel, A.A., Loboda, A., Tang, W.H., Keating, S.P., Hunter, C.L., Nuwaysir, L.M., and Schaeffer, D.A. (2007). The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. Mol. Cell Proteomics *6*, 1638-55.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034-50.

Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods *11*, 817-820.

Tang, H., Peng, J., Wang, P., and Risch, N.J. (2005). Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. *28*, 289-301.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731-2739.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511-5.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wen, M., Shen, Y., Shi, S., and Tang, T. (2012). miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. BMC Bioinformatics *13*, 140.

Yang, H., and Wang, K. (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat. Protoc. *10*, 1556-1566.

Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. *88*, 76-82.

Young, M.D., Wakefield, M.J., Smyth, G.K., and Oshlack, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. *11*, R14.

Zhou, L, Chen, J., Li, Z., Li, X., Hu, X., Huang, Y., Zhao, X., Liang, C., Wang, Y., Sun, L., et al. (2010). Integrated profiling of microRNAs and mRNAs: microRNAs located on Xq27.3 associate with clear cell renal cell carcinoma. PLoS One *5*, e15224.

Zhou, X., and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat. Methods *11*, 407-409.