Supporting Information

Fortes et al. 10.1073/pnas.1002044107

SI Results

The AWM dissects biological information from GWASs in three steps as follows: The first step is to estimate trait correlations from all ~50,000 SNP effects; the second step is to select SNPs for the AWM considering these trait correlations; and the third step is to use the AWM to build a gene network.

Initially, the SNP effect data from the total of available SNPs (50,070) were used to calculate correlations between the 22 traits. As a result, AGECL correlates with WTCL ($R = 0.64$) and with PPAI ($R = 0.31$) (Table S1). Second, the selected set of SNPs included in the AWM (3,159 SNPs) was also used to calculate the correlations between the 22 traits. This second analysis revealed a correlation of $R = 0.82$ when compared with the results obtained with the full 50,070 SNPs.

When selecting SNPs to build the AWM, we considered their distance to the nearest gene (as per Fig. S1). SNPs were considered in four categories according to SNP-to-gene distance: "close," "far," "very far," and "unmapped." A comparison between SNP categories was made to determine if SNP-to-gene distance could influence either the size of the SNP effect or the significance of its association to traits. SNPs from the unmapped category were excluded from the comparison because their distance to the nearest gene is unknown. We compared the close, far, and very far categories within three groups of SNPs (full 50,070 SNPs, AWM SNPs, and top SNPs). First, when the full 50,070 SNPs were examined, the SNP categories were not different in terms of SNP effect. However, we observed decay in the SNP significance with the increasing SNP-to-gene distance, as measured by P value (Fig. S2). Second, among the AWM selected SNPs, the far category had a higher overall SNP effect. This result reflects the selection criteria bias for the AWM SNPs, because far SNPs were included only if they were the top SNPs ($P < 0.05$ in ≥10 traits). Third, considering the group of top SNPs, the close category of SNPs had a higher effect across traits. Furthermore, not one very far SNP could be included in the top group, indicating a negative association between distance to a gene and SNP significance. Overall, the interaction between the SNP group and SNP-to-gene distance influenced SNP effect significantly ($P < 0.0001$).

The AWM is presented in Fig. S3*A*, highlighting a few selected rows from the total of 3,159 SNPs. These selected rows correspond to genes that clustered with the three transcription factors that were further scrutinized in the regulatory sequence analysis: estrogen related receptor γ (ESRRG), prophet of the pituitary-specific transcription factor 1, or prophet of PIT-1 (PROP1), and peroxisome proliferator-activated receptor γ (PPARG). SPOCK1 and ZNF462, genes previously associated with puberty (1, 2), were present in the AWM and are shown in Fig. S3*A*. Fig. S3*A* also shows a pair of highly correlated genes: RUN domain containing 1 (RUNDC1) and breast cancer 1 (BRCA1). The gene–gene interaction between RUNDC1 and BRCA1 is a unique prediction made by the AWM analysis.

Columnwise, the AWM renders itself to the calculation of correlations between the 22 traits under scrutiny. This calculation result is visualized in PermutMatrix as a hierarchical tree where AGECL clusters with WTCL and both are close to PPAI (Fig. S3*B*). On the hierarchical tree cluster, a strong positive correlation is displayed as proximity, whereas a strong negative correlation is displayed as a large distance. To observe negative and positive correlations equally, we developed the quantitative trait network (QTN) from SNP effect data of AWM selected SNPs, which shows a great degree of interaction between all 22 traits (Fig. S3*C*). For visual comparison, we also present a QTN based on the genetic correlations derived from quantitative genetic approaches (3, 4) (Fig. S3*D*).

A formal comparison between published genetic correlations (3, 4) and AWM-derived correlations was completed using the pairwise correlations between 19 traits, AGECL, WTCL, FATCL, and all T1 and T2 traits (ADG, CS, HH, IGF1, SEMA, SP8, SRIB, and WT). The scatter plot of this formal comparison is illustrated in Fig. 1, revealing a moderate agreement between both approaches ($R^2 = 0.6439$). Importantly, when all 50K SNPs were considered, the trait correlations were even closer to the genetic estimates ($R^2 = 0.7034$; Fig. 1). We conclude that SNP effects via the AWM methods can be used to recover the genetic correlations between traits.

The number of SNPs used for the SNP effect-based correlations impacts on the similarity between genetic and SNP-based estimates of trait correlations. Linearly, the higher the number of SNPs analyzed, the higher is the recovery of genetic correlations, even when significance levels are considered. In other words, even if the SNPs with lower P values were selected for the comparisons with genetic estimates, the actual number of SNPs is still important. As mentioned earlier, the trait correlations based on all 50K SNPs were similar to genetic correlations ($R^2 = 0.7034$). But, this similarity decreased when fewer SNPs were used to calculate the trait correlations. This linear relationship between SNP numbers and similarity to genetic correlations is presented in Fig. S4. The equation of best fit presented in Fig. S4 might be used to estimate the number of SNPs required to recover 100% of the genetic correlations, which were REML estimated. Solving the equation results in >200,000 SNPs required to fully recover genetic correlations between traits. This value agrees with the estimated 200,000–300,000 SNPs required to fully exploit GWASs or genomic selection across cattle breeds (5).

Pairwise correlations across AWM rows are used to predict gene–gene (or gene–SNP) interactions and hence build a network. In this network, every gene (or SNP when mapped as very far) is a node and every significant interaction is an edge. Significant interactions were identified according to the PCIT weighted network algorithm. We identified 287,465 significant edges between 3,159 nodes, which were subsequently visualized in Cytoscape. This network is illustrated in Fig. 2*A* where the colors of the nodes follow the MCODE score (6), which indicates network density: Red nodes have high score, yellow a middle score, and green a low score.

Once the network was built, it was subjected to gene ontology and pathway analyses to mine the predicted drivers of puberty. Gene Ontology (GO) analyses performed by BiNGO and GOrilla showed many similar results, but some differences were also noted. These differences likely reflect the different background lists used in each analysis. BiNGO analysis used National Center for Biotechnology Information full *Bos taurus* annotation as a background list, whereas for the GOrilla analysis we created a background list that contained all genes located close to a SNP in our GWAS. These analyses revealed GO term overrepresentation for the molecular functions "binding" ($P = 8.62E$-10), "metal ion binding" ($P = 6.38E$-3), "ATP binding" ($P = 4.28E$-9), and "GABA receptor activity" ($P = 2.51E$-2). Similarly, there were overrepresented GO terms for biological processes including "fatty acid metabolic process," "signal transduction," "protein modification processes," "regulation of epidermal growth factor," and "small GTPases signaling," as well as a number of processes associated with gene transcription (Fig. S5*A*). In addition, there were 539 genes in the AWM associated with the GO term "developmental process," which was highly enriched ($P < 1.00E$-09).

Importantly, these genes along with the genes associated with fatty acid metabolic process would have been missed if the traditional single-trait analysis was performed (Fig. S5*B*). In addition, pathway analyses of the network revealed an enrichment ($P < 0.001$) for "calcium signaling," "axon guidance," and "neuroactive ligand–receptor interaction." This last pathway includes ligands and receptors considered to be involved with pubertal signaling such as GABA receptor activity, glutamate receptor activity, follicular stimulant hormone (FSH) receptor activity, and leptin receptor activity. These pathway analyses also revealed enrichment for cell growth, cell survival, and factors controlling cell cycle progression. This last result supports a theory that implicates a role in puberty for tumor-related genes, which are involved in control of cell proliferation. Both literature-derived theory and these results help to justify the large number of tumor-related genes found using the AWM. These gene ontology and pathway analyses were useful to mine the drivers of puberty from our AWM gene network.

To provide an in silico validation for gene–gene interactions predicted by the AWM, we performed regulatory sequence analysis for predicted targets of selected key TFs. We could select from 34 TFs in the network, which had DNA sequence-binding motif information available in the Genomatix suite of tools. These 34 TFs were BACH2, BSX, ELF2, ESRRG, ETS1, GLI3, GLIS3, GRHL1, GRHL3, HIVEP1, HNF4A, IRX2, JARID2, LHX3, LHX4, LHX8, LHX9, LMX1A, LMX1B, MEOX1, MYT1, PAX2, PBX1, POU2F3, PPARG, PROP1, RORA, RREB1, SATB1, SMAD4, SOX5, SRF, STAT6, and TLX1. From the 34, PROP1, PPARG, and ESRRG were the top 3 TFs according to their reported functional role in the context of reproduction and their position in the gene network.

PROP1, PPARG, ESRRG, and their in silico validated target genes are shown in Fig. 2*B*. Detailed gene lists of the validated target genes, which were 114 for PROP1, 22 for PPARG, and 76 for ESRRG, are as follows: (*i*) List of genes, AWM-predicted partners, with in silico binding site for ESRRG: *UNC5A*, *SDCCAG1*, *GLIS1*, *ACACB*, *ADAMTSL1*, *CCDC86*, *CD47*, *CDH4*, *CEP76*, *COL27A1*, *ETS1*, *GTF3C5*, *HERC1*, *JMJD2B*, *SRP72*, *HORMAD2*, *MSH3*, *SNCAIP*, *TRPM7*, *ACVR2A*, *ACVR2B*, *ADAMTS9*, *AGRN*, *AIG1*, *ALK*, *ARHGAP21*, *BBX*, *CAPN3*, *CNO*, *DCLK2*, *DST*, *EFCAB5*, *EPB41L3*, *EXOC1*, *FBXW8*, *FRMPD4*, *FRYL*, *FSHR*, *FYN*, *GABRA1*, *GLT1D1*, *GRIN2B*, *HIVEP2*, *IKZF1*, *KCNH5*, *KLHL5*, *KRT17*, *MAN2A1*, *MON2*, *NID2*, *NLN*, *PCSK6*, *PDLIM1*, *PGC*, *PPP2R2C*, *RAB35*, *RAB8B*, *RALGDS*, *RNF115*, *RNF122*, *RTL1*, *SCML4*, *SF3B3*, *SI*, *SLC13A1*, *SLC26A8*, *SPG7*, *STXBP1*, *TCF7*, *TMEM163*, *TMEM2*, *TPM1*, *TYRP1*, *WDR66*, *WNT3A*, and *WNT6*. (*ii*) List of genes, AWM-predicted partners, with in silico binding site for PROP1: *ACACA*, *ACTR1A*, *ADAM12*, *ADAMTS3*, *ADD1*, *AKAP10*, *AKAP9*, *AKR1C4*, *ALDH1A1*, *ALDH7A1*, *ANGEL2*, *ANKRD35*, *ANKRD40*, *ARHGAP26*, *ARMC4*, *ASRGL1*, *AZI2*, *BAZ2B*, *BCAS3*, *BRCA1*, *CACHD1*, *CCDC25*, *CDC20B*, *CDH7*, *CDH8*, *CDKAL1*, *CHD2*, *CNTNAP2*, *CRLF3*, *CROP*, *CSMD1*, *CSMD3*, *CSNK1E*, *DIO1*, *DNAH7*, *DYM*, *EDIL3*, *ENOX1*, *EPS15*, *EXT1*, *FANCC*, *FBLN7*, *GARNL1*, *GBF1*, *GNL2*, *HEATR1*, *HRNBP3*, *HTR4*, *HUNK*, *IPO8*, *JAKMIP1*, *KCNIP4*, *LARGE*, *MAP2*, *MAPK10*, *MGAT4A*, *MSRA*, *NAALADL2*, *NBR1*, *NCOA2*, *NEDD4*, *NFATC3*, *NFXL1*, *NLGN1*, *NLK*, *OXR1*, *PARD6G*, *PCMTD1*, *PDE11A*, *PDE3A*, *PDLIM5*, *PHF17*, *PINX1*, *PKD2*, *PLEKHA6*, *PLEKHA7*, *PPM1E*, *PSD3*, *PTGFRN*, *PTPRF*, *PTPRK*, *PTPRM*, *QSER1*, *RANBP17*, *RASAL2*, *RGS6*, *RUNDC1*, *SCARB2*, *SESTD1*, *SETD5*, *SFRS5*, *SH3TC2*, *SLFN14*, *SMG7*, *SOX5*, *SPATA6*, *SPOCD1*, *SPOP*, *SPRY1*, *STIM2*, *SYNE1*, *TCERG1*, *TFB1M*, *TJP1*, *TOX*, *TRIM23*, *TRIM3*, *TRIO*, *TRPS1*, *TYW1*, *UBASH3B*, *UBE2K*, *USO1*, and *WBSCR17*. (*iii*) List of genes, AWM-predicted partners, with in silico binding site for PPARG: *ARHGAP21*, *BAALC*, *BTBD9*, *CHODL*, *COL4A3*, *GABRA2*, *GFM1*, *LBH*, *MPPED2*, *MYOM2*, *OTOF*, *PDZD2*, *PEPD*, *PPP2R2C*, *PRLR*, *SMC2*, *SPINK5*, *SPRY3*, *TYRP1*, *UFD1L*, *WDR70*, and *ZNF592*.

Further evidence for the interactions predicted by the AWM could be found for ESRRG and 19 of its partners. These partners presented a promoter model derived from experimental data. A promoter model consists of various individual regulatory elements such as TFBSs, repeats, hairpins, their strand orientation, their sequential order, and their distance ranges. In our dataset of 76 target genes with ESSRG binding sites, promoter models were found for 19 genes in tandem with other TFBSs including E-box binding factors (EBOX), PAR/bZIP family (PARF), and vertebrate steroidogenic factor (SF1F) (Table S3).

SI Discussion

The AWM is constructed with as many columns as related traits and as many rows as genes selected from GWASs. In our GWAS, the selected genes were the nearest from a selected SNP, which was, in brief, a SNP with $P < 0.05$ in ≥3 related traits. The exact selection criteria and thresholds proposed to include or exclude SNPs from the AWM may vary according to each GWAS under investigation. In our case study, on average, SNPs that were minimally significant ($P < 0.05$) for AGECL were also significant for two more traits. Therefore the AWM selection was expanded by adding all SNPs that were significant for any ≥3 traits. This criterion forces the selection of all SNPs that exceed the minimum significance for AGECL and, at the same time, controls bias by selecting other SNPs with the same overall significance ($P < 0.05$ in ≥3 traits). Thus, none of the related traits are penalized. This is important, considering the connectivity of the QTN presented and because this is the power of a network approach. Our approach is such that up to 22 traits were considered and priority for SNP selection was given to those SNPs associated with >1 trait. Of course, the possibility of a SNP being significant for one trait and again for another trait by chance alone is reduced, minimizing false discovery rate issues of a relaxed threshold ($P < 0.05$). The number of traits available as well as the strength of their joint correlation structure will impact the power to detect the genetic drivers of a complex trait. Also, we proposed a selection criterion focused on gene-based SNPs (SNPs at least 2.5 kb close to a gene) for two reasons: biological interpretation of results (gene-centered inference) and overall SNP significance. SNPs with higher significance across 22 traits tend to be close to a gene. SNPs located very far (1.5 Mb) from genes were selected using the same criteria for the genes that were close as an inbuilt control for the method (i.e., to avoid biasing toward *cis*-acting SNPs). Nonetheless, we acknowledge an indisputable bias on the close, far, and very far numbers of SNPs available to the study. This bias is imposed by the heterogeneous nature of the genome with unevenly spaced genes and that of the BovineSNP50 Bead Chip with equally distant SNPs. Once again, this will vary with the genome under investigation and the genotyping platform used and, accordingly, it will require adaptation of the method to a particular study. Finally, the GWAS results might present some SNPs that are far from any annotated gene but should be included in the AWM on account of overall significance exceeding the selection threshold by 3-fold. In our case study, we selected the SNPs that were far but presented significant associations ($P < 0.05$) with ≥10 traits, rather than 3 traits (the standard threshold for close SNPs). The inclusion of far and very far SNPs might allow the discovery of putative regulatory sites by verifying the correlation between these SNPs and SNPs that are close to genes. In conclusion, all of the steps for constructing an AWM were carefully reasoned beforehand.

SNP effects can recover the correlation existing among traits. However, the negative or positive nature of a correlation estimated by SNP effect must be viewed with caution. Each SNP effect has a positive or negative signal attributed to it as an artifact of allele order computed in the ASREML software (7). In the Permut Matrix display of the AWM (distance on hierarchical tree cluster), a strong negative correlation between traits or genes might be an artifact from the signal of the SNP effect and it does not imply that

the gene products are opposite in function. This signal issue is minimized as it occurs evenly throughout the dataset and we address this difficulty by using the network theory and building a quantitative trait network from the AWM columns and a gene network from the AWM rows.

Here we acknowledge a limitation of our methods. Given the expected relevance of GnRH for the initiation of puberty, the gene encoding the hormone (GNRH1) should be in our network. However, the nearest SNP from GNRH1 is ≈12 kb from it, which means it would classify as far by AWM criteria. Also, this SNP is located within KCDT9 and so it would represent KCDT9 in the AWM and not GNRH1. Therefore, there were no SNPs in GNRH1 accessed by our GWAS effort and this "wet lab" limitation is carried on to our AWM approach.

Despite the absence of GNRH1, aspects of the brain remodeling that changes the input in GnRH neurons and triggers the increase in GnRH release were captured. From our network viewpoint, ESRRG targeting GABRA1 and NMDAR2B indicates a link between estrogen pathways and GABA and glutamate signaling. GABAergic and glutamatergic synaptic inputs are important drivers of GnRH neuron remodeling, known to influence age of puberty (8, 9).

PPARG is an important regulator of energy balance. Among the 124 AWM-predicted targets, 22 presented TFBSs for PPARG. The presence of PPARG and its 23 targets in common with ESRRG in our network is evidence for the AWM capturing the known biology behind puberty. There are demonstrated associations between energy balance and reproduction (10, 11). The genes ARHGAP21, PPP2R2C, and TYRP1 are ESRRG and PPARG targets and present binding sites for both. These three targets might be components of the known metabolic link between estrogen-related receptors and PPARG (12). Additionally, PPARG is a predicted regulator of GABRA2 and so it could also be influencing GnRH neurons remodeling.

SI Methods

Animals, Traits, and Genotypes. We used data from 866 cows representing the genotyped subset of animals from a larger population bred by the Cooperative Research Centre for Beef Genetic Technologies (Beef CRC) previously described in detail (3, 4, 13–15). Briefly, tropical composite animals are 50% taurine and 50% tropically adapted breeds. The tropically adapted component was either zebu (Brahman) or taurine adapted breeds, such as Africander, and the N'Dama nonadapted component was from taurine breeds, which originate from continental Europe and Britain.

In broad terms, there are only two seasons in the tropical regions of Queensland, Australia: the wet and the dry. The wet season begins with the first monsoonal rains, usually around November, and ends around May (http://www.tropicalaustralia.com.au/about_tropical_queensland/climate).

We consider a total of 22 traits from measurements taken on three occasions: (*i*) at the end of the cows' first wet season (T1), when the mean age of animals was 18 mo; (*ii*) at the time of observation of the first corpus luteum (CL), when the mean age of the animals was 22 mos; and (*iii*) at the end of the cows' second dry season (T2), when the mean age was 24 mo. A brief description of the 22 traits along with a summary of descriptive statistics for this population is provided in Table S4.

The cows' first CL was detected through regular ovarian scans (every 4–6 wk), performed when the heifers' average body weight reached ∼200 kg (or ∼12 mo of age). We consider the age at first observed CL (AGECL) as a trait for age of puberty (16), although we recognize that puberty is a developmental process that takes place over a period. The presence (1) or absence (0) of a CL close to the first day of joining, i.e., when the cows are first joined to bulls, was also recorded (CLJOIN). When the first CL was observed, live weight (WTCL, kg) and s.c. fat depth at the rump or P8 site (FATCL, mm) were measured. The P8 site is located over the gluteus muscle on the rump, at the intersection of a line through the ischiatic tuberosity parallel to the spine and its perpendicular through the third sacral crest (17).

At T1 and T2, eight growth and growth-related traits were measured, including live weight (WT, kg), hip height (HH, cm), serum concentration of insulin-like growth factor I (IGF-I, ng/mL), average daily weight gain (ADG, kg/d), body condition score (CS, score 1–10), ultrasound scanned eye muscle, or *longissimus dorsi*, area (SEMA, cm^2), scanned fat depth at the P8 site (SP8, mm), and scanned fat depth measured between the last two ribs (SRIB, mm). A full description of these trait measurements is published elsewhere (3, 4).

Heifers that reached puberty before the first mating season, conceived during this season, and later calved had an additional trait measured: postpartum anoestrus interval (PPAI). PPAI is defined as the interval, in days, between calving and first CL observed after calving. For this study, we also used a related binary trait: PPAI with respect to weaning time (PW), for cows that either had a CL (score 0) or did not have a CL (score 1) recorded before weaning of their calves.

LD between all possible SNP pairs was calculated using two metrics: D' and R^2. For a review on these metrics refer to ref. 18.

SNP effects were calculated via single-trait–single-SNP association analysis. The additive effect of a SNP on each trait, or the allele substitution effect, was calculated by regression analysis, with values in the covariate coded as zero, one, or two copies of the variant allele, and after fitting the following mixed model,

$$y_{ij} = X\beta + Zu + s_{j,k} + e_{ij}, \qquad \textbf{[S1]}$$

with terms defined as follows: y_{ij} represents the vector of observations from the ith cow at the jth trait; X is the incidence matrix relating fixed effects in β with observation in y_{ij}; Z is the incidence matrix relating random additive polygenic effects in u with observation in y_{ij}; $s_{j,k}$ represents the additive association of the kth SNP on the jth trait; and e_{ij} is the vector of random residual effects.

Fixed effects included in β were contemporary groups (i.e., group of cows raised together), herd of origin, sex of calf, month of calving, and sire of calf. Polygenic effects were included to reduce the effect of family structure on family-specific alleles (19).

Standard stochastic assumptions were applied to the random effects in model [**S1**], which were assumed to be distributed as multivariate normal with zero mean and variance, as

$$\mathrm{V}\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} A\sigma_u^2 & 0 \\ 0 & I\sigma_e^2 \end{bmatrix}, \qquad \textbf{[S2]}$$

where A is the numerator relationship matrix across all cows and derived from the pedigree structure (20), σ_u^2 is the additive polygenic component of variance, I is an identity matrix, and σ_e^2 is the residual component of variance.

Solutions to the effects in model [**S1**] as well as variance components [Eq. **S2**] were estimated using the ASREML software (http://www.vsni.co.uk/software/asreml/) (7). The log inverse of the P value of each SNP, for AGECL, was plotted according to genomic positions.

Association Weight Matrix (AWM). Constructing an AWM starts with the selection of relevant SNPs from a GWAS to represent genes. A diagrammatic representation of the selection criteria is shown in Fig. S1. In detail, these criteria were applied in a sequential fashion as follows.

First, the allele substitution effect of the ith SNP on the jth trait was z-score standardized to allow comparison across traits as

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad [S3]$$

where z_{ij} is the standardized effect of the ith SNP on the jth trait, x_{ij} is the effect of the ith SNP on the jth trait in the original units, $\bar{x}_j$ is the mean of x_{ij} over all SNPs, and s_j is the SD of x_{ij} over all SNPs.

Second, regardless of their genomic position, SNPs with $P < 0.05$ in ≥10 traits were included in the AWM. These SNPs represent the top 0.2% SNPs of our GWAS (Fig. S2), which were associated at the 0.05 level with 45% of the traits. Selecting for ≥10 traits means these top SNPs were associated either directly with AGECL or at least with one correlated trait (minimum $R = 0.28$).

Third, we classified each SNP as close, far, very far, or unmapped according to its mapped distance from the nearest annotated gene (BTAU4.0 assembly, ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/). SNPs considered close were located at ≤2.5 kb from the nearest gene (either 5′ or 3′). SNPs considered very far were ≥1.5 Mb distant from the nearest gene. Accordingly, any SNP-to-gene distance that fell between close and very far was annotated as far.

The remaining far or unmapped SNPs were discarded from further analysis and we continued to select from the close and very far groups only. The very far group is intended as an inbuilt control for the AWM associations. As puberty was our primary trait, all SNPs with $P < 0.05$ for AGECL were selected. Then, and given that on average, SNPs that were significant for AGECL were also significant for two more traits, the AWM selection was expanded by adding all SNPs that were significant for any three or more traits.

Finally, the marker density of the BovineSNP50 resulted in some genes being represented in the AWM by more than one SNP. In these cases, the SNP with a $P < 0.05$ in the largest number of traits was chosen to represent that gene. If still more than one SNP in the same gene had a significant P value in the same number of traits, the selection was based on the lower sum of P values.

An AWM was constructed with as many rows as selected SNPs (according to Fig. S1) and as many columns as traits. The rows are identified as genes or SNPs according to each SNP location. The SNPs selected from the close group were identified in the AWM by the official symbol of the nearest gene. The SNPs selected from the far, very far, or unmapped groups were designated by the SNP name (Illumina code). Each $\{i, j\}$ cell value in the AWM corresponds to the z-score normalized additive effect of the ith SNP on the jth trait. The AWM approach explores trait correlations columnwise and gene correlations rowwise.

Pearson correlations between AGECL and the other 21 traits were calculated using the SNP effect values. We call this procedure SNP-based correlations and they were compared with the genetic correlations, estimated via pedigree-based restricted maximum likelihood (REML), established for the same population previously (3, 4). Genetic correlations for PPAI and PW were not available for comparison with SNP-based correlations.

Furthermore, we used the previous genetic correlations across traits (3, 4) and the SNP-based correlations to form QTNs for puberty. In the genetic QTN each of 19 traits is a node and each significant genetic correlation ($r^2 > 2 \times$ SD) is represented by an edge, which is a line linking the correlated traits. In the AWM-derived QTN all 22 traits were considered. Both QTNs were analyzed with MCODE (6) for exploring network density and clustering of traits.

Rowwise AWM explores the correlations between SNP effects to predict gene interactions. We studied the predicted gene interaction using a combination of hierarchical clustering, weighted gene network, and pathway analyses to identify genetic drivers of puberty.

Visualizations of AWM and hierarchical clustering analyses for rows and columns were performed using the PermutMatrix software (21). The significant correlations between rows were identified with the PCIT algorithm (22) and reported as gene–gene or gene–SNP interactions in a network. Cytoscape (23) was used to visualize these networks, where genes and SNPs were nodes and significant correlations were edges, linking the genes. The BiNGO plug-in (24) of Cytoscape was used to test for gene ontology (GO term) enrichment in the network. BiNGO was used in the GO-full mode, retrieving National Center for Biotechnology Information annotations (http://www.ncbi.nlm.nih.gov/Ftp/) for biological processes, molecular functions, and cellular components, all of which were specific for *B. taurus*.

Genes included in the AWM were further analyzed using (*i*) DAVID (25, 26) to review known pathways and (*ii*) GOrilla (27) to infer GO term enrichment for biological processes in a tree-based structure. When applying GOrilla, genes in the AWM were contrasted against a background list that contained all genes located close to a SNP in our GWAS. Thus, the background list considered all genes that could have been selected from our GWAS. When using DAVID, *Homo sapiens* full genome annotation was used as a background list. Using different background lists, *H. sapiens* in DAVID and *B. taurus* in BiNGO, allows input from different databases, not limiting the retrieval of biological information.

To provide an in silico validation for gene–gene interactions predicted by the AWM, we performed regulatory sequence analysis for predicted targets of selected key TFs. Among the 3,159 genes (or SNPs) in the AWM, there were 236 TFs according to the original census of 1,391 TFs (28). Of these 236 TFs, 34 had DNA sequence binding motifs information available in the Genomatix suite of tools. Finally, this list of 34 TFs was further scrutinized on a 2-fold basis: their reported functional role in the context of reproduction and their position in the gene network (i.e., separated enough to guarantee a maximum coverage of the landscape spanned by the network). On the basis of these criteria, three TFs (PROP1, PPARG, and ESRRG) were deemed to be "key" and we focused on them.

1. Liu YZ, et al. (2009) Genome-wide association analyses identify SPOCK as a key novel gene underlying age at menarche. *PLoS Genet* 5:e1000420.
2. Perry JR, et al. (2009) Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat Genet* 41:648–650.
3. Johnston DJ, et al. (2009) Genetics of heifer puberty in two tropical beef genotypes in northern Australia and associations with heifer- and steer-production traits. *Anim Prod Sci* 49:399–412.
4. Barwick SA, et al. (2009) Genetics of heifer performance in 'wet' and 'dry' seasons and their relationships with steer performance in two tropical beef genotypes. *Anim Prod Sci* 49:367–382.
5. de Roos AP, Hayes BJ, Spelman RJ, Goddard ME (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* 179:1503–1512.
6. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2.
7. Gilmour ARCB, Gogel BJ, Welham SJ, Thompson R (2006) *ASReml, User Guide. Release 2.0* (VSN International, Hemel Hempstead, UK).
8. Terasawa E (2005) Role of GABA in the mechanism of the onset of puberty in non-human primates. *Int Rev Neurobiol* 71:113–129.
9. Clarkson J, Herbison AE (2006) Development of GABA and glutamate signaling at the GnRH neuron in relation to puberty. *Mol Cell Endocrinol* 254–255:32–38.
10. Fernandez-Fernandez R, et al. (2006) Novel signals for the integration of energy balance and reproduction. *Mol Cell Endocrinol* 254–255:127–132.
11. Gasser CL, Behlke EJ, Grum DE, Day ML (2006) Effect of timing of feeding a high-concentrate diet on growth and attainment of puberty in early-weaned heifers. *J Anim Sci* 84:3118–3122.
12. Feige JN, Auwerx J (2007) Transcriptional coregulators in the control of energy homeostasis. *Trends Cell Biol* 17:292–301.
13. Prayaga KC, et al. (2009) Genetics of adaptive traits in heifers and their relationship to growth, pubertal and carcass traits in two tropical beef cattle genotypes. *Anim Prod Sci* 49:413–425.
14. Barwick SA, Wolcott ML, Johnston DJ, Burrow HM, Sullivan MT (2009) Genetics of steer daily and residual feed intake in two tropical beef genotypes, and relationships among intake, body composition, growth and other post-weaning measures. *Anim Prod Sci* 49: 351–366.
15. Burrow HM, et al. (2003) Relationships between carcass and beef quality and components of herd profitability in Northern Australia. *50 Years of DNA: Proceedings of the Fifteenth*

Conference, Association for the Advancement of Animal Breeding and Genetics, University of Roseworthy, SA, 5371, Australia 7–11 July 2003, pp 359–362.
16. Romano MA, Barnabe VH, Kastelic JP, de Oliveira CA, Romano RM (2007) Follicular dynamics in heifers during pre-pubertal and pubertal period kept under two levels of dietary energy intake. *Reprod Domest Anim* 42:616–622.
17. Reverter A, Tier B, Johnston DJ, Graser HU (1997) Assessing the efficiency of multiplicative mixed model equations to account for heterogeneous variance across herds in carcass scan traits from beef cattle. *J Anim Sci* 75:1477–1485.
18. Zhao H, Nettleton D, Soller M, Dekkers JCM (2005) Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet Res* 86:77–87.
19. Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 10:381–391.
20. Wright S (1922) Coefficients of inbreeding and relationship. *American Naturalist* 56: 330–338.
21. Caraux G, Pinloche S (2005) PermutMatrix: A graphical environment to arrange gene expression profiles in optimal linear order. *Bioinformatics* 21:1280–1281.
22. Reverter A, Chan EK (2008) Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 24:2491–2497.
23. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.
24. Maere S, Heymans K, Kuiper M (2005) BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21: 3448–3449.
25. Dennis G, Jr, et al. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4:3.
26. Huang W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4:44–57.
27. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.
28. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: Function, expression and evolution. *Nat Rev Genet* 10:252–263.

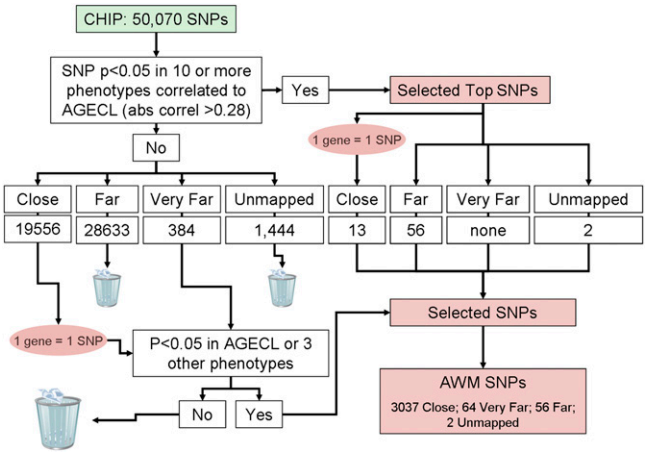

Fig. S1. Method for selecting SNPs from GWASs to perform the AWM. SNPs were selected from the genome-wide study using both their distance to the nearest gene and their overall significance level (smallest *P* value) across 22 traits. SNPs were considered close to a gene (within 2,500 bp up or downstream), far (distance >2,500 bp but <1.5 Mb), very far (distance >1.5 Mb), or unmapped.

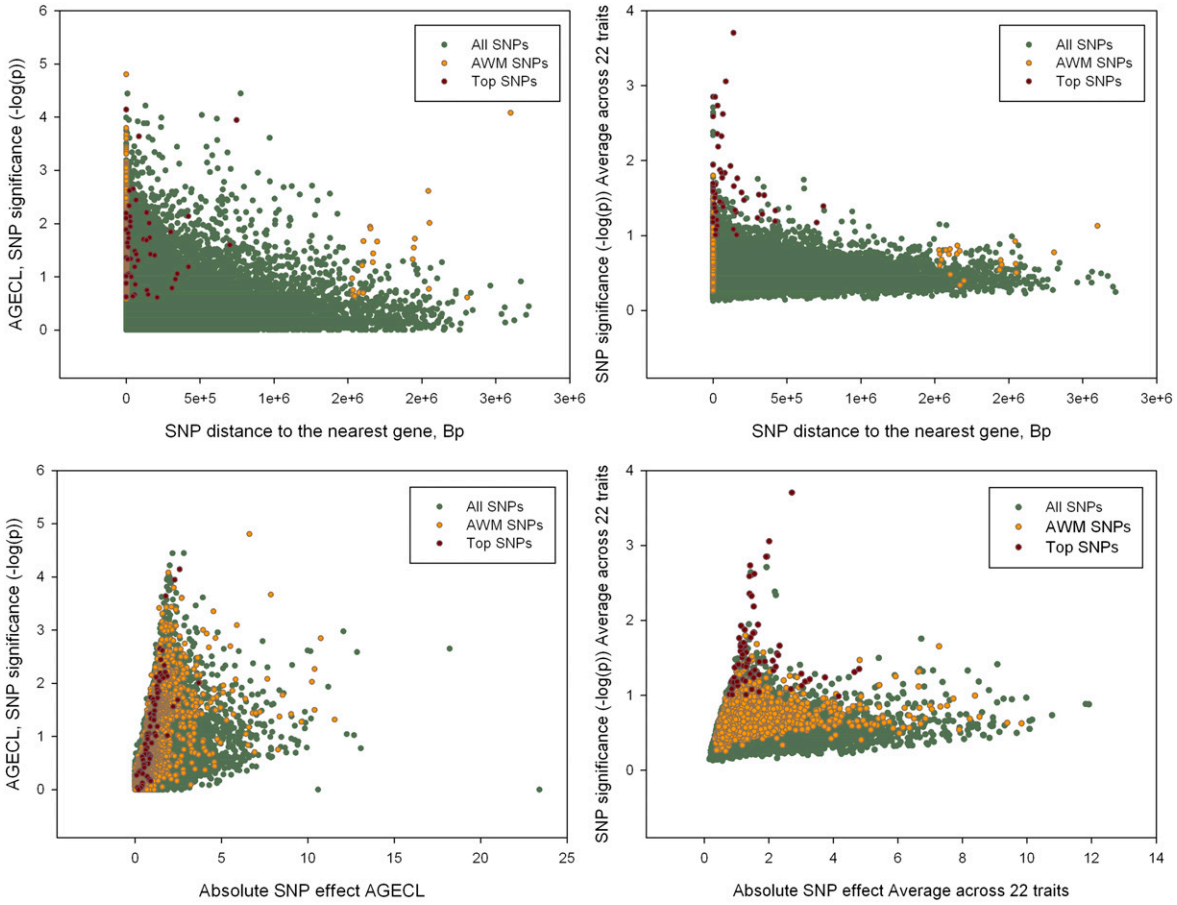

Fig. S2. Significance vs. SNP effect and gene distance for AGECL and across all traits. Clockwise from *Upper Left* the four graphs represent the −log(*P*) for AGECL of each SNP plotted against its distance to the nearest gene, the average −log(*P*) across 22 traits of each SNP plotted against its distance to the nearest gene, the average −log(*P*) across 22 traits of each SNP plotted against its average *z*-normalized effect, and the −log(*P*) for AGECL of each SNP plotted against its effect in days on AGECL. Represented in green are all SNPs, in yellow are the SNPs selected for the AWM, and in red are the top AWM SNPs.

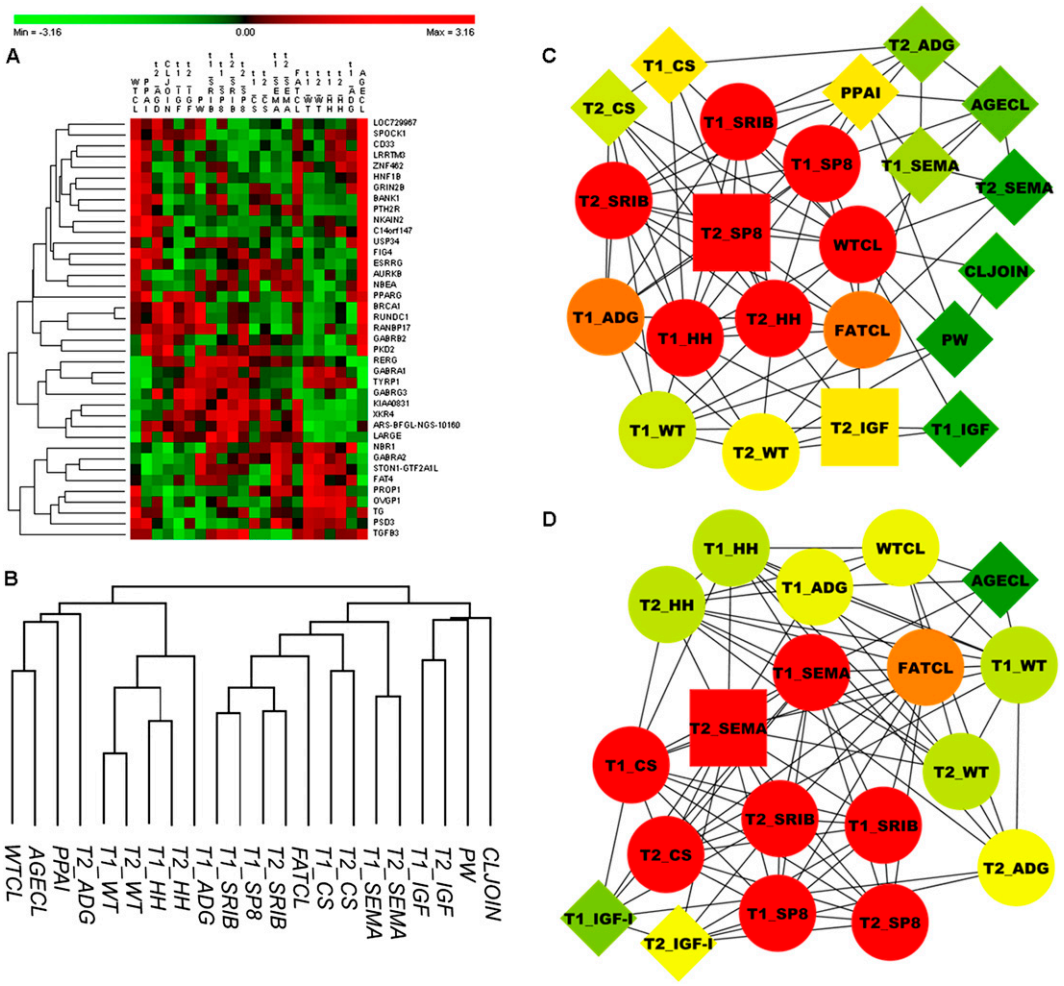

Fig. S3. Quantitative trait networks for puberty-related traits based on REML and SNP effect correlations. Clockwise starting from *Upper Left*: (*A*) subset of the association weight matrix; (*B*) hierarchical tree cluster for the 22 traits as computed by Permut Matrix; (*C*) quantitative trait network analyzed by MCODE representing significant SNP effect correlations between 22 traits, as estimated by AWM-PCIT analyses; and (*D*) quantitative trait network analyzed by MCODE representing the significant genetic correlations for AGECL and 18 other traits, adapted from previously published data (4, 28). In both QTNs, the color code represents MCODE scores related to network density, ranging from green (low density) to red (high density). Density is an expression of network connectivity. Circles, diamonds, and squares represent the clustering status of a trait: Circles are clustered with other traits, diamonds are unclustered, and squares are a cluster seed (MCODE analyses).

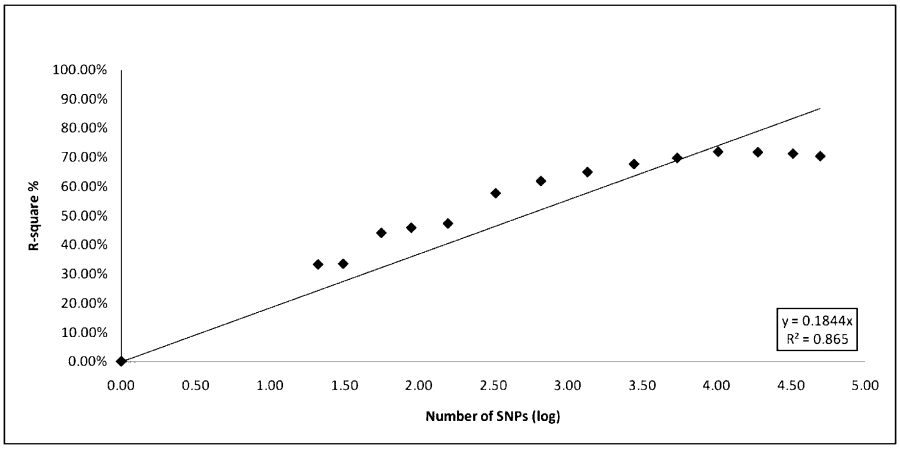

Fig. S4. Number of SNPs analyzed influences the similarity between genetic and SNP effect-driven trait correlations. R^2 (comparison between SNP effect and genetic correlations) is plotted against the number of SNPs used. The correlation of 0.86 between the number of SNPs used and the similarity (R^2) with genetic correlations is explained linearly by the equation in the *Inset*.

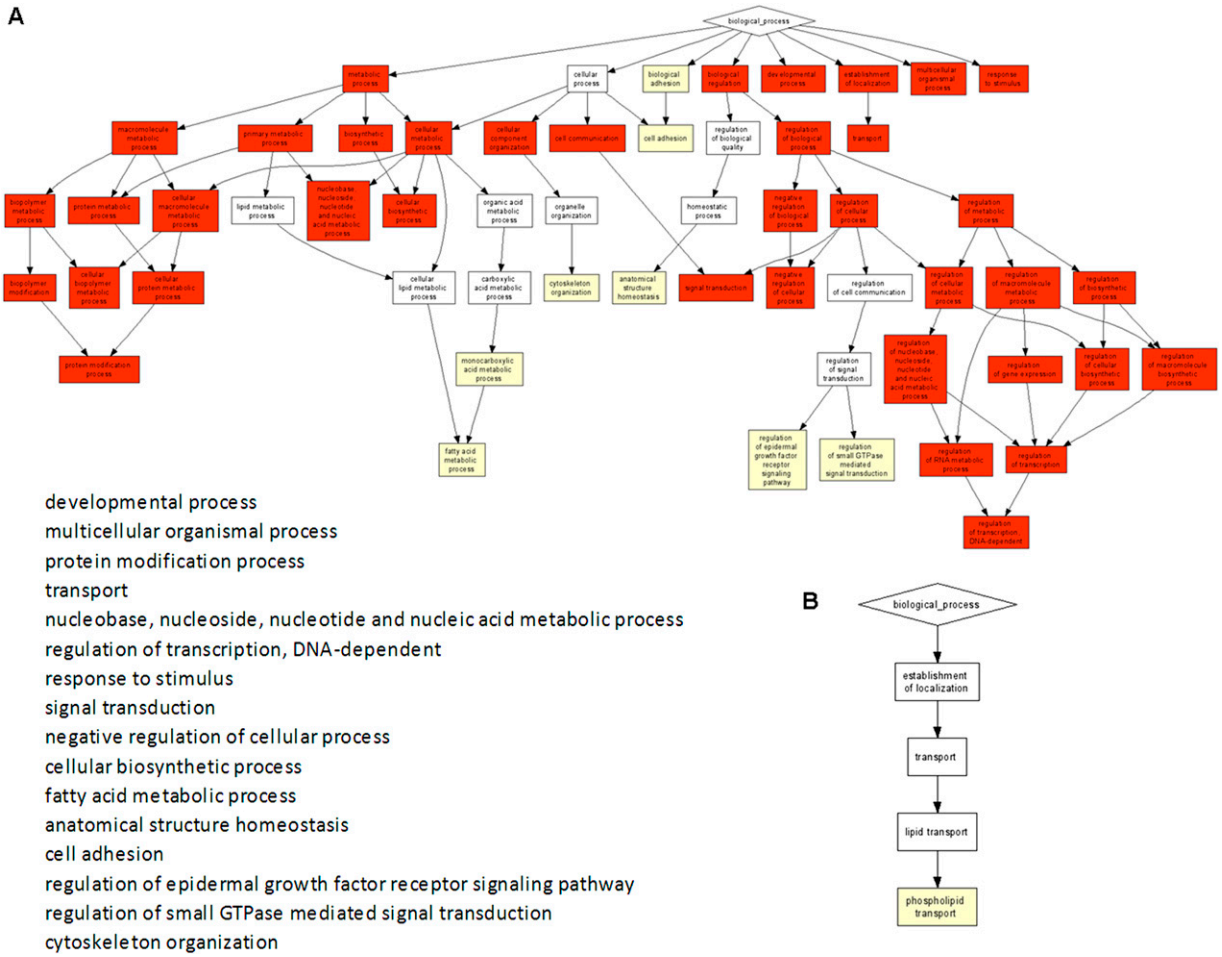

Fig. S5. Comparison of GOrilla results between AWM genes and genes associated only with AGECL. GOrilla results were obtained using the biological process function and two unranked gene lists for analyses: (*A*) AWM genes tested against all GWAS genes (background list) and (*B*) all genes that were significant for AGECL alone ($P < 0.05$) against all GWAS genes (background list). In the two flow charts, the stronger the color is for each GO term box, the higher its significance. The *A* flowchart with a number of significant results was obtained when comparing AWM genes against the background list. The *B* flowchart with only one significant GO term enrichment (phospholipid transport) was the result obtained for AGECL alone. The GO terms found to be significantly enriched in the AWM analysis would have been all missed in the AGECL solo analysis. The list of GO terms shown represents the GO terms that are most relevant, i.e., the ones that were farthest down in the GOrilla hierarchy.

Table S1. All 50K SNP correlations (below diagonal) vs. genetic correlations (above diagonal)

	AGECL	CLJOIN	FATCL	WTCL	ADG	CS	SEMA	HH	IGF-I	SP8	SRIB	WT	T2_ADG	T2_CS	T2_SEMA	T2_HH	T2_IGF-I	T2_SP8	T2_SRIB	T2_WT	PPAI	PW
AGECL		—	0.41	0.68	−0.40	−0.02	−0.33	−0.24	−0.36	−0.18	0.00	−0.38	0.08	0.03	−0.16	−0.27	−0.09	−0.07	−0.01	−0.28	—	—
CLJOIN	−0.27		—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
FATCL	0.13	−0.09		0.28	−0.06	0.74	0.17	−0.35	−0.05	0.91	0.85	−0.09	0.07	0.79	0.30	−0.39	0.15	0.86	0.70	−0.07	—	—
WTCL	0.64	−0.28	0.18		0.38	0.24	0.11	0.35	−0.37	0.01	0.16	0.43	0.49	0.22	0.24	0.42	−0.22	−0.08	0.05	0.47	—	—
T1_ADG	0.02	−0.03	0.05	0.31		−0.01	0.27	0.7	0.04	0.03	0.07	0.89	0.33	0.04	0.34	0.6	0.07	−0.01	0.1	0.9	—	—
T1_CS	−0.16	0.03	0.17	−0.03	0.18		0.31	−0.31	0.19	0.7	0.68	0.15	0.14	1	0.28	−0.38	0.31	0.69	0.61	0.14	—	—
T1_SEMA	−0.23	0.05	0.08	0.06	0.16	0.32		0.42	0.33	0.45	0.54	0.59	0.14	0.41	0.94	0.38	0.32	0.57	0.51	0.41	—	—
T1_HH	−0.01	−0.03	−0.1	0.42	0.31	−0.09	0.17		0.18	−0.11	−0.05	0.73	0.41	0.05	0.46	0.99	−0.01	−0.23	−0.16	0.73	—	—
T1_IGF-I	−0.27	0.09	0.13	−0.27	0.08	0.12	0.12	−0.03		0.24	0.27	0.2	0.57	0.1	0.34	0.13	0.88	0.22	0.31	0.23	—	—
T1_SP8	−0.25	0.12	0.43	−0.13	0.03	0.3	0.24	−0.06	0.24		0.96	0.18	0.12	0.7	0.56	−0.19	0.41	0.95	0.98	0.08	—	—
T1_SRIB	−0.19	0.12	0.34	−0.08	0.02	0.25	0.21	−0.02	0.22	0.72		0.21	0.37	0.71	0.4	−0.09	0.41	0.94	0.86	0.12	—	—
T1_WT	−0.16	0.02	0	0.44	0.53	0.22	0.4	0.69	0.02	0.16	0.14		0.44	0.29	0.59	0.75	0.1	0.11	0.12	0.99	—	—
T2_ADG	0.15	−0.02	−0.03	0.23	0.04	−0.1	−0.15	0.07	−0.15	−0.14	−0.11	−0.01		0.23	0.29	0.49	0.55	0.02	0.13	0.75	—	—
T2_CS	−0.11	−0.02	0.29	−0.03	0.01	0.46	0.2	−0.16	0.17	0.3	0.27	0.05	0.01		0.39	0.06	0.11	0.73	0.58	0.24	—	—
T2_SEMA	−0.2	0.11	0.07	0.05	0.16	0.32	0.64	0.16	0.15	0.26	0.2	0.38	0.01	0.25		0.35	0.4	0.64	0.62	0.6	—	—
T2_HH	0.1	−0.09	−0.08	0.5	0.38	−0.11	0.08	0.76	−0.11	−0.1	−0.11	0.63	0.13	−0.19	0.08		−0.1	−0.21	−0.15	0.77	—	—
T2_IGF-I	−0.16	0.03	0.12	−0.14	0.01	0.06	0.13	−0.05	0.38	0.16	0.14	0.01	0.01	0.09	0.12	−0.11		0.41	0.46	0.06	—	—
T2_SP8	−0.28	0.09	0.53	−0.18	−0.06	0.27	0.2	−0.12	0.23	0.65	0.47	0.05	−0.07	0.4	0.22	−0.2	0.22		0.88	0.24	—	—
T2_SRIB	−0.25	0.08	0.4	−0.15	−0.06	0.27	0.19	−0.06	0.21	0.56	0.6	0.09	−0.05	0.35	0.2	−0.18	0.22	0.72		0.29	—	—
T2_WT	−0.12	0.02	0.02	0.47	0.47	0.17	0.31	0.64	−0.02	0.12	0.08	0.88	0.28	0.09	0.36	0.61	0.01	0.08	0.09		—	—
PPAI	0.31	−0.06	−0.04	0.24	−0.02	−0.06	−0.09	0.03	−0.18	−0.15	−0.12	−0.01	0.06	−0.08	−0.07	0.08	−0.13	−0.16	−0.17	0.01		—
PW	−0.26	0.06	0.03	−0.23	0.01	0.05	0.09	−0.05	0.15	0.15	0.14	−0.01	−0.09	0.06	0.07	−0.09	0.11	0.12	0.13	−0.04	−0.9	

Numbers below the diagonal represent the correlations between 22 traits calculated from SNPs. Each SNP effect, from the total set of 50,070 SNPs, was used as a data point in the calculation of all pairwise correlations between the 22 traits. Numbers above the diagonal represent the genetic correlations (REML) between 19 traits previously published (1).

1. Johnston DJ, et al. (2009) Genetics of heifer puberty in two tropical beef genotypes in northern Australia and associations with heifer- and steer-production traits. *Anim Prod Sci* 49:399–412.

Table S2. Percentage of validated gene–gene interactions for both the AWM gene network and the random gene network: regulatory sequence analysis completed for the predicted targets of ESRRG, PROP1, and PPARG

	AWM gene network			Random gene network		
TF	Predicted targets	Validated targets	% validated	Predicted targets	Validated targets	% validated
ESRRG	211	76	36.02	41	9	21.95
PROP1	320	114	35.63	57	11	19.30
PPARG	124	22	17.74	50	0	0.00

The random gene network predicted a significantly ($P < 0.0001$) smaller proportion of partners that had a binding site when compared with the AWM gene network. The AWM network presented more validated gene–gene interactions.

Table S3. List of genes that are AWM-predicted partners of ESRRG with in silico predicted models

Target	TFBS	ESRRG promoter model
UNC5A	5	EBOX_EREF_01andEREF_SF1F_01
SDCCAG1	4	EREF_AP1F_01
GLIS1	3	EREF_P53F_01,EREF_SF1F_01
ACACB	4	EREF_SF1F_01
ADAMTSL1	3	EREF_SF1F_01
CCDC86	4	EREF_SF1F_01
CD47	3	EREF_SF1F_01
CDH4	5	EREF_SF1F_01
CEP76	4	EREF_SF1F_01
COL27A1	3	EREF_SF1F_01
ETS1	3	EREF_SF1F_01
GTF3C5	1	EREF_SF1F_01
HERC1	1	EREF_SF1F_01
JMJD2B	2	EREF_SF1F_01
SRP72	3	EREF_SF1F_01
HORMAD2	2	PARF_EREF_01
MSH3	1	PARF_EREF_01
SNCAIP	1	PARF_EREF_01
TRPM7	4	PARF_EREF_01

Table S4. Definitions and descriptive statistics for the 22 traits under investigation

Phenotype	Units	Description	Average	SD
CL				
CLJOIN	0 or 1	Presence of corpus luteum when bulls were placed in same paddock as heifers	0.63	0.48
AGECL	d	Age at first detected CL	652.57	117.67
FATCL	mm	Scanned P8 fat at first CL	2.99	1.63
WTCL	kg	Weight at first CL	329.61	45.85
T1				
ADG	kg/d	Average daily gain in live weight	0.58	0.14
CS	Score 1–10	Condition score	7.45	0.91
SEMA	cm^2	Scanned eye muscle area	45.82	6.92
HH	cm	Hip height	125.04	6.02
IGF	kg/mL	Serum IGF level	225.14	76.00
SP8	mm	Scanned s.c. P8 fat	3.14	1.77
SRIB	mm	Scanned rib fat	2.05	1.15
WT	kg	Live weight	313.65	41.05
T2				
ADG	kg/d	Average daily gain in live weight	0.26	0.17
CS	score 1–10	Condition score	7.01	1.09
SEMA	cm^2	Scanned eye muscle area	48.90	6.60
HH	cm	Hip height	130.16	4.79
IGF	ng/mL	Serum IGF level	239.60	71.62
SP8	mm	Scanned s.c. P8 fat	2.93	1.67
SRIB	mm	Scanned rib fat	1.98	1.07
WT	kg	Live weight	354.30	38.95
PPAI				
PPAI	d	Postpartum anoestrus interval following the first calving event	141.47	108.57
PW	0 or 1	CL before (0) or after (1) weaning a calf, following the first calving event	0.82	0.39