RESEARCH ARTICLE

RUNNING HEAD: Using GIFT to map complex traits in small samples

Genomic Informational Field Theory (GIFT) to map complex traits in small sample sizes

Panagiota Kyratzi[1,2], Scott Gadsby[1,5], Edd Knowles[3,4], Patricia Harris[4], Nicola Menzies-Gow[3], Jonathan Elliott[5], Andras Paldi[2], Jonathan Wattis[6] and, Cyril Rauch[1]

[1]School of Veterinary Medicine and Science, University of Nottingham, College Road, Sutton Bonington, LE12 5RD, UK.
[2]École Pratique des Hautes Études, PSL Research University, St-Antoine Research Center, Inserm U938, 34 rue Crozatier, 75012 Paris, France.
[3]The Royal Veterinary College, Hawkshead Lane, Hatfield, Hertfordshire AL97TA, UK.
[4]Equine Studies Group, Waltham Centre for Pet Nutrition, Melton Mowbray, Leicester LE14 4RT, UK.
[5]The Royal Veterinary College, Royal College Street, London NW1 0TU, UK
[6]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
Correspondence: *Cyril Rauch (cyril.rauch@nottingham.ac.uk)*

ABSTRACT

Genome-wide association studies (GWAS) are commonly used to investigate the genetic basis of complex traits. However, to be adequately powered they typically require large sample sizes to provide precise inferences. To address this challenge, this paper introduces GIFT, a novel data analytic method that enhances the power of genetic analyses, enabling the use of smaller datasets without compromising precision. In a small cohort of 157 ponies, GIFT was applied to examine the complex trait of "height at withers", comparing its performance to traditional GWAS. Only GIFT identified genetic loci linked to insulin physiology validating, in turn, a long-standing hypothesis that "height at withers" is associated with insulin physiology in equids, potentially promoting equine metabolic syndrome (EMS). Additionally, by redefining correlations between single nucleotide polymorphisms (SNPs), GIFT offers new insights into linkage disequilibrium and epistatic interactions that, through the development of a network centrality measure, distinguishes core from peripheral genes within gene networks involved in "height at withers". By reducing the time and cost associated with large-scale genotype–phenotype mapping studies, GIFT expands access to quantitative genetic research, enabling smaller-scale investigations to explore the genetic architecture of complex traits with greater resolution.

NEW & NOTEWORTHY

Inferring genotype–phenotype associations typically requires large sample sizes, limiting many genetic studies. GIFT, a novel data analytics tool, overcomes this by enabling accurate association mapping in small datasets. We further show how GIFT's extended framework infers linkage disequilibrium, epistasis, and gene networks, distinguishing core from peripheral genes involved in complex traits. This advancement enhances understanding of biological architecture and enables high-resolution genetic research in limited cohorts, offering a powerful, cost-effective alternative to traditional large-scale approaches.

Keywords: Equids, Epistasis, Gene networks, Linkage Disequilibrium.

INTRODUCTION

Identifying associations between phenotype and genotype is the fundamental basis of genetic analysis and GWAS is currently the method of choice for mapping genotype to phenotype in large populations. It is expected that large-scale GWASs will elucidate the genetic basis of complex traits. The need for "large-scale" GWAS arises from the statistical requirements of the method requiring very large populations to produce valid inferences when effect sizes are small. A prime example of large-scale GWAS applied to a complex trait is the genetic analysis of the phenotype "standing height" in humans. A recent international GWAS involving an unprecedented sample size of 5.4 million individuals successfully produced a saturated map of genetic variants associated with human height (1). Although this represents a landmark achievement in genetics, it also highlights the limitations of GWAS, whose success relies heavily on data acquisition to generate adequately powered studies, an activity that is time-intensive and costly. While human genetic studies benefit from substantial funding and large-scale GWAS capabilities, research on non-human species, especially those facing extinction, often lacks the necessary resources to conduct similar studies (2). As a result, there is a need to explore alternative methods, increasing the investigative power of biological datasets when sample sizes are small.

GWAS is based on the statistical framework developed by Fisher more than 100 years ago. It partitions genotypic values and performs linear regression of phenotype on marker allelic dosage. Regression coefficients estimate the average allele effect sizes, and the regression variance is the additive genetic variance due to the locus. However, there is ongoing debates about whether the statistical method underpinning GWAS is optimal for understanding complex traits (3, 4).

In GWAS, statistical associations rely on frequency or count plots like bar charts or histograms, which categorize data into bins. These plots, known as distribution density functions (DDFs), form the basis of frequentist probability, determining key statistical measures such as averages and variance that are used to infer genotype-phenotype associations. However, this approach involves the loss of significant amounts of information from datasets. To comprehend this point, consider a population where phenotypic data are measured with high precision, making each data point unique. When grouped into bins to form DDFs, individual distinctions are lost since it is not possible to differentiate data within a given category, even though the data were initially differentiated. While averages and variances derived from DDFs appear informative, they are inherently based on this reduced representation limiting access to fine-grained information. Increasing the sample size might improve the precision of statistical summaries, i.e., average and variance, but it does not recover the lost individual details.

This raises a fundamental question: Is more data always beneficial if the method inherently disregards fine-grained information?

GIFT is a novel data analytics method designed to overcome this conceptual limitation by analyzing the full biodiversity of data without requiring data categorization. Through theoretical simulations (5, 6) and analysis of various datasets with sample sizes comparable to those used in GWASs (7), GIFT has been shown to enhance both investigative and discriminative powers of genetic studies. However, the ability of GIFT to operate effectively with smaller sample sizes and provide essential genetic information has not yet been shown.

Herein, using a small cohort of 157 ponies and focusing on the complex trait "height at withers" the investigative powers of GWAS and GIFT are compared. Several reasons motivated the focus on "height at withers".

First, like "standing height" in humans, “height at withers” in equids is a classic complex trait positively selected during domestication and correlated with other morphometric traits (8–10). With a well-established heritability, GWASs have investigated its genetic background (9, 11–24). Thus, GIFT-derived variants can be benchmarked against GWAS results in the equine species and compared to similar phenotype, i.e., “body size” or “body height”, across different species.

Second, "height at withers" is suspected to be associated with equine metabolic syndrome (EMS) (11). EMS describes a cluster of metabolic disturbances predisposing horses to laminitis, a condition where insulin dysregulation plays a central role (11, 16, 25, 26). While a GWAS identified HMGA2 (High mobility group AT-hook protein 2) and IRAK3 (Interleukin-1 receptor-associated kinase 3) as associated with "height at withers" in a cohort of 236 ponies, no genes were linked to insulin physiology (11). However, since a retrospective correlation was identified between blood insulin concentrations and HMGA2 allelic dosages (11), it was hypothesized that GIFT could reveal further genetic insights into the relationship between insulin physiology and "height at withers".

Third, it is postulated that complex traits can be omnigenic involving core and peripheral genes (27). Validating this new paradigm using GWAS would necessitate large sample sizes to detail any information on gene networks. However, since GIFT provides a more detailed genetic analysis, it was hypothesized that redefining SNP and gene pairwise correlations using GIFT could offer a fresh perspective on linkage disequilibrium (LD) and epistasis informing, in turn, on the potential validity of the omnigenic paradigm.

The results herein demonstrate that GIFT's approach is adequate when small sample sizes are used, yielding new insights into the genetic architecture of complex traits.

MATERIALS AND METHODS

Data collection

Data from a study performed by Knowles *et al.* were re-used in which blood collection was performed and the "height at withers" measured in a number of different breeds (25). In order to remain coherent with the study by Norton *et al* (11), Shetland ponies were removed from the current study resulting in a final sample size of 157 ponies

Equine genome sequencing

Blood samples collected (buffy coats) were used for DNA extractions (LGC Biosciences) and genotyping individuals using the Axiom™ Equine Genotyping Array (MNEC670, Cambridge Genomic Services). Further to Axiom Analysis Suite following best practices pathway, a total of 478498 quality controlled single nucleotide polymorphisms were identified.

Phenotype 'height at withers'

Measurements of "height at withers" obtained in (25) were pre-analysed. A Kolmogorov-Smirnov test performed against raw data for the height returning a p-value of 0.05101 suggested that raw data are not normally distributed, requiring the need to pre-correct data for fixed effects. Phenotypic data were then pre-corrected for the fixed effects of Breed (B) and sex (S) in R-studio (GMMAT package) using the following linear model, $\mathrm{H_{ij}} = \mu + \mathrm{B_i} + \mathrm{S_j} + \mathrm{e_{ij}}$, where $\mathrm{H_{ij}}$ is the "height at withers" resulting from the ith Breed and j^{th} Sex (Male vs Female); μ is the overall mean; $\mathrm{B_i}$ is the effect of the ith Breed (i =1,..,7); $\mathrm{S_j}$ the effect of jth Sex (Male vs Female) and, $\mathrm{e_{ij}}$ is the residual. With this correction, a new Kolmogorov-Smirnov test was performed against the residual values for the height, i.e., the $\mathrm{e_{ij}}$-values, providing an acceptable p-value of 0.8357 demonstrating, in turn, that the $\mathrm{e_{ij}}$-values are quasi-normally distributed and usable by GWAS.

GWAS

The genotype dataset was filtered using PLINK (HWE p-value threshold of 10^{-6}, call rate for genotypes of 10% and a MAF of 1%), the number of independent SNPs ($\mathrm{N_{Ind-SNPs}} = 73792$) was determined using BCFTOOLS (r^2-threshold=0.1) and the GWAS Manhattan plots, linked to the determination of the GWAS p-values ($\mathrm{p_{GWAS}}$), were obtained using GEMMA. The threshold of significance for GWAS was determined using 99% and 95% Bonferroni corrections, numerically given by, $-\mathrm{Log}(0.01/\mathrm{N_{Ind-SNPs}}) = 6.8680$ and $-\mathrm{Log}(0.05/\mathrm{N_{Ind-SNPs}}) = 6.1690$, respectively. The code to determine the Manhattan plot for GWAS in Fig.1A in the text is given in supplemental code S14.

GIFT

The same genotype and phenotypic residuals given by the $\mathrm{e_{ij}}$-values, as filtered by GWAS, were used for GIFT. Regarding the representation of GIFT, upon selecting a SNP the different genotypes were assigned arbitrary values of +1, 0, and -1, respectively. More specifically, the assignment of values +1, 0, and -1 were done as a function of the base pairs as follows: AA=TT=+1, GG=CC=-1, and 0 otherwise. The microstates +1, -1 and zero are respectively color coded in red, blue and grey in the figures to facilitate the representation of genetic

paths. With this convention, any barcode can be represented by a string of numbers from which a GIFT analysis can be inferred (Supplemental_Text_S1.doc). As GIFT is not amenable to Bonferroni correction, a different method was used to set a threshold of significance. The information captured by GIFT for a given SNP concerns the relative ordering of microstates in their string to which a p-value, noted $\hat{p}_{GIFT}$, can be numerically determined as explained in (7) or in the Supplemental_Text_S1.doc. To consider the false discovery rate (FDR) Benjamini-Yekutieli procedure (BYp) was applied to the different $\hat{p}_{GIFT}$-values, in turn translating the different $\hat{p}_{GIFT}$-values toward smaller values noted, p_{GIFT}, in the text. To set a threshold of significance a null hypothesis was required at the population level. To capture this global null hypothesis for each string of microstates, i.e., each SNP, all the microstates were randomly permutated once and the corresponding $\hat{p}_{GIFT}$ value resulting from the permutation determined. This operation provided a probability density function of $\hat{p}_{GIFT}$-values. By determining the 99.9% and 99% quantiles of this distribution two thresholds of significance were determined and used in the present study. The code to determine the Manhattan for GIFT in Fig.1B in the main text is given in Supplemental_Code_S15.html. The code to determine the 1,000 random permutations in Fig.1E in the main text is given in supplemental_Code_S16.html.

Enrichment and correlation analyses

Enrichment analyses were performed using STRING, Uniprot, Ensembl, Gephi and the GWAS catalog (31, 41, 42, 62, 65) as cited in the text. Pearson correlations between genetic paths were performed using R-studio and the measure of centrality using Matlab.

RESULTS

Manhattan Plots using GWAS or GIFT

GIFT is a method designed to decouple the notion of precision in phenotypic measurements from that of sample size, providing genotype-phenotype associations without relying on data categorization or grouping (Supplemental_Text_S1.doc). Measurements of "height at withers" obtained by Knowles et al. (2023) were pre-corrected for the fixed effects of breed and sex (25). Further to this correction, phenotypic residual values (Supplemental_Data_S2.xls) were demonstrated to be adequately normally distributed using Kolmogorov-Smirnov test and usable by GWAS and GIFT (Supplemental_Figure_S3.doc). Association analyses using GWAS and GIFT based on the phenotypic residual values were then performed and plotted under the form of Manhattan plots. Fig.1 presents the Manhattan plots for both GWAS (Fig.1A) and GIFT (Fig.1B). Concentrating on the highest threshold levels for GWAS and GIFT, the significant SNPs were then mapped to the reference horse genome assembly (EquCab3.0) available on Ensembl to retrieve the corresponding gene Id and gene name when available. The results summarized in Table 1 and Table 2 clearly demonstrate that SNPs located on chromosome 6 are highly significant. However, while GWAS identifies 5 significant SNPs associated with the HMGA2 gene, GIFT detects 28 significant SNPs for the same gene on chromosome 6. Notably, GIFT also highlights significant SNPs on chromosome 2 that are not detected by GWAS. Figures 1C and 1D provide a comparative visualization of two representative SNPs using both GWAS and GIFT. Figure 1C shows the most significant SNP on chromosome 6, identified by both methods, whereas Figure 1D presents the most significant SNP on chromosome 2, uniquely identified by GIFT. Interestingly, the SNP on chromosome 2 exhibiting a sigmoidal genetic path (Figure 1D, right panels) lacks a traditionally defined gene effect size (Figure 1D, left panel).

To confirm that the significant SNPs identified by GIFT result from the unique organization of their microstates rather than random chance, the microstates of each significant SNP were randomly permuted 1,000 times. After each permutation, the significance level, i.e., the value of $-\mathrm{Log}_{10}(p_{GIFT})$, was determined. The statistical summary of these significance values after 1,000 random permutations was summarized as a boxplot (Fig. 1E). These values were then compared to the original significance value of the same SNP, as determined in Fig. 1b, and represented as a blue dot in Fig.1E. Fig.1E confirms that GIFT's identification of genotype–phenotype associations is not due to random chance. Additionally, using the identified gene names in Table 1 and Table 2, STRING analyses were conducted employing a minimum required interaction score of 0.4 using the *Equus caballus* species. The enrichment p-values for GWAS (Table 1) and GIFT (Table 2) are 1 and 0.0303, respectively. These findings further confirm that the set of genes identified by GIFT is not random, highlighting that GIFT enhances the investigative power of biological datasets.

Finally, to assess whether GIFT can operate effectively with smaller sample sizes, the original dataset (N=157) was randomly subsampled to N=140, 120, and 100. For each subsample, the values of $-\mathrm{Log}_{10}(\mathrm{p}_{\mathrm{GIFT}})$ corresponding to the most significant SNPs (Fig.1E) were recalculated, along with updated quantile thresholds. This procedure was repeated 10 times for each sample size to compute the mean and standard deviation of $-\mathrm{Log}_{10}(\mathrm{p}_{\mathrm{GIFT}})$ for the top SNPs (Fig.1E). As shown in Fig.1F, although a decrease in the magnitude of $-\mathrm{Log}_{10}(\mathrm{p}_{\mathrm{GIFT}})$ is observed with smaller sample sizes, the majority of the average values remain above the 99% quantile threshold when N=120. Notably, N=100 appears to represent the lower limit for reliable GIFT performance, as a substantial portion of the average $-\mathrm{Log}_{10}(\mathrm{p}_{\mathrm{GIFT}})$ values approach the 99% quantile threshold. Nevertheless, the most significant SNPs, typically exhibiting $-\mathrm{Log}_{10}(\mathrm{p}_{\mathrm{GIFT}}) > 20$ at N=157 (Fig.1E), remain above the 99% quantile threshold, even at this reduced sample size.

Biological relevance of genes determined by GIFT

While GWAS previously identified HMGA2 and IRAK3 as the only significant associations with "height at withers" in a cohort of 249 ponies (11), using a different and smaller cohort of 157 ponies the GIFT method expands these findings to other genes also known to be associated to body size but in other species. For example, genes associated to the trait "body size" are HMGA2 and MSRB3 (Methionine-R-sulfoxide reductase B3) in sheep and dogs (28, 29) and, HMGA2 and IRAK3 in cattle (30). Interestingly, HMGA2, MSRB3, IRAK3 and TMBIM4 (Transmembrane BAX Inhibitor Motif Containing 4) are located in close proximity within the equid genome (11). Although the absence of SNPs associated with the insulin gene in the GIFT or GWAS results may appear unexpected, it is explained by the gene array used, which lacked coverage at the genomic locations of the insulin gene.

To assess whether any of the gene names listed in Table 1 and Table 2 have previously been associated with the phenotype "body height" in humans, the GWAS Catalog (31) was used. Table 3 provides a summary of this cross-species comparison, connecting each gene name from Table 1 and Table 2 to the study accession Ids from the GWAS catalog when found. Of the 20 gene names listed in Table 1 only 6, namely, OXCT1 (3-Oxoacid CoA-Transferase 1), IRAK3, FUT1 (fucosyltranferase 1), EN1 (engrailed Homeobox 1), IRS2 (insulin receptor substrate 2) and BECN2 (beclin 2), have no equivalent in the GWAS catalog for the phenotype selected.

Regarding genes associated with insulin physiology, a selection of genes was examined for their roles in insulin action, insulin secretion, glucose uptake and homeostasis, or energy metabolism. Given the limited molecular biology studies in equids, this investigation was extended to other species.

Insulin Action: IRS2 plays a key role in initiating the signalling pathway upon insulin binding to its receptor (32), facilitating glucose uptake (33). Knockout studies of IRS2 have demonstrated phenotypes characterized by glucose intolerance (34). Additionally, PALLD (Palladin), a cytoskeletal protein, is essential for active actin remodelling and the formation of membrane ruffles (35), a process integral to insulin signalling due to its role in the membrane translocation of specific GLUT family members (36).

Insulin Secretion: The KCNB1 (Kv2.1) gene encodes a critical potassium channel expressed in the β-cells of pancreatic cell lines (37). This channel influences insulin secretion levels from pancreatic β-cells (38). Mutations in KCNB1 have been implicated in insulin resistance in humans (39), and its expression is notably lower in human patients with type 2 diabetes (40).

Glucose uptake and homeostasis: Information on MFSDB4 (Major Facilitator Superfamily Domain Containing 4B) is limited. However, a Uniprot search predicts this gene functions as a sodium-dependent glucose transporter, derived from an ancestral gene identified in rats that shares the glucose uptake function (41, 42). According to Ensembl gene tree analyses, MFSDB4 and its ancestral gene share a common origin approximately 96 million years ago (43). Additionally, in mice the SLIT2 (Slit Guidance Ligand 2) gene has been shown to improves glucose homeostasis (44), while BECL2 was shown to be a regulator of body weight and glucose homeostasis in human (45). Interestingly, the genetic locus that contains human BECL2 on chromosome 1q43 in human, has been linked to obesity and diabetes-related traits in multiple ethnic populations (46–50).

Energy Metabolism: Among metabolically involved genes, CRB4 (Carbonyl Reductase Family Member 4) is a mitochondrial NADPH-dependent reductase involved in quinone reduction and fatty acid biosynthesis, playing a vital role in mitochondrial lipid metabolism and overall function (51). Impaired CBR4 functionality can disrupt lipid metabolism in human, leading to lipid accumulation and insulin resistance (52). HIBCH (3-

Hydroxyisobutyryl-CoA Hydrolase) catalyzes the conversion of 3-hydroxyisobutyryl-CoA to 3-hydroxyisobutyrate (3-HIB) as part of the valine catabolic pathway. Reduced valine and elevated 3-HIB blood concentration levels have been associated with the heightened energy demands of endurance racing in horses (53). Elevated blood level of 3-HIB likely reflects its role as an inter-organ metabolite that mediates the gluconeogenic potential of valine (54). Interestingly, 3-HIB blood concentration is also a marker of insulin resistance in human (55). SLIT2 involved in promoting thermogenesis of adipose tissue and energy expenditure in mice (44), is also associated with backfat thickness in pigs (56). In addition, EN1 has been shown to stimulate brown adipogenesis and thermogenesis in mice (57), IRAK3 has been demonstrated to be related to the early onset of obesity-related insulin resistance and metabolic syndrome in human (58) and, OXCT1, a rate limiting enzyme of ketone oxidation, has been found to reverse obesity-induced hyperglycemia in mice when inhibited using gene knock-out or pharmacological intervention (59, 60).

Altogether, GIFT's cross-species comparative results strongly suggest that insulin physiology is associated with the phenotype "height at withers" in the equine species. Notably, the genes listed in Table 2 are located on different chromosomes, indicating that GIFT's pairwise SNP correlations may offer novel insights into linkage disequilibrium (LD) and epistasis contributing to this phenotype.

GIFT's determination of LD

At the core of GIFT lies the concept of the genetic path, represented by $\Delta\theta$ in Fig.1C and Fig.1D (right panels) for each SNP in the genome. The overall pattern of $\Delta\theta$s reflects the ordering of genetic microstates when the sample is ordered by phenotypic residual values (Supplemental_Text_S1.doc). Consequently, SNPs similarly related to the phenotype exhibit comparable large $\Delta\theta$ patterns. Conversely, SNPs that are unrelated to the phenotype give small and noisy $\Delta\theta$ curves. This property is essential to determine precise correlations between loci, which can be compared to that of usual LD definition. An example will illustrate this point. Using GIFT, HMGA2 contains 28 significant SNPs (Table 2) and it is therefore possible to determine the local LD in the portion of genome that defines this protein. The measure of LD commonly used, a.k.a. the r^2 measure of LD, is given by (61):

$$r^2 = \frac{(p_{ab} - p_a p_b)^2}{p_a(1-p_a)p_b(1-p_b)} \quad (1)$$

where p_{ab} is the frequency of haplotypes having allele 'a' at locus 1 and allele 'b' at locus 2. As the square of a correlation, r^2 can range from 0 to 1 as p_a, p_b and p_{ab} vary. In Fig.2A, the genetic paths $\theta(i)$ and $\theta_0(i)$ pertaining to the first 16 SNPs belonging to HMGA2, enumerated from 1 to 16 in the figures, are represented. At the exception of SNPs 5, 7 and 8 on one hand, and SNPs 11, 12, 15 and 16 on another, the curves determined by, $\theta(i)$ and $\theta_0(i)$, for each SNP are clearly different. In Fig.2B the r^2 measure of LD for HMGA2 is determined where the enumeration of SNPs (from 1 to 16) is given next to their SNP-Id on the diagonal. Fig.2B demonstrates that while the r^2 measure of LD for HMGA2 (detailed in Supplemental_Data_S7.xls) is very patchy, strong correlations exist between some SNPs. It is interesting to note from Fig.2B that while the correlation between SNPs 6 and 10, marked by a white star in Fig.2B is strong, the paths are anti-symmetrical. This anti-symmetry arises because, for SNP6, the path begins with '-1' (blue) microstates and ends with '+1' (red) microstates, whereas the opposite is observed for SNP10. In contrast, this effect is not reflected in equation (1), as it considers the entire population to compute bulk probabilities (p_a, p_b, p_{ab}), meaning that the notion of ordering does not influence the result from equation (1). This highlights that the initial and arbitrary assignment of microstates ('+1' and '-1') is not fundamental for the r^2 measure of LD.

To develop a novel measure of LD based on genetic paths, while maintaining the core principles of equation (1), these paths can be rectified and overlapped to reveal their underlying patterns. This is possible by swapping the initial assignment of microstates ('+1' and '-1'). Note that as the level of significance calculated by GIFT relies on the amplitude of $\Delta\theta$ (see supplemental text S1 or (7)), changing the initial assignment of microstates, e.g., swapping the microstates '+1' and '-1', does not change the level of significance calculated. In the left panel of Fig.2C, the genetic paths ($\Delta\theta$s) of all SNPs within HMGA2 are plotted. In the right panel of Fig.2C, these paths are rectified and overlapped, revealing the presence of two distinct groups of genetic paths (light grey and dark grey). Since the key to inferring associations with the phenotype lies in the overall pattern of these genetic paths, a refined measure of LD as described by GIFT can be formulated simply using Pearson's correlation formula in the following form:

$$r^2_{GIFT} = \frac{[\mathrm{cov}(\Delta\theta_1,\Delta\theta_2)]^2}{\mathrm{var}(\Delta\theta_1)\mathrm{var}(\Delta\theta_2)} \quad (2)$$

Where $\Delta\theta_1$ and $\Delta\theta_2$ represent two genetic paths with, $\mathrm{var}(\Delta\theta_p) = \frac{1}{N-1}\sum_{i=1}^{N}(\Delta\theta_p(i) - \langle\Delta\theta_p(i)\rangle)^2$ where $p \in \{1,2\}$ and, $\mathrm{cov}(\Delta\theta_1,\Delta\theta_2) = \frac{1}{N-1}\sum_{i=1}^{N}(\Delta\theta_1(i) - \langle\Delta\theta_1(i)\rangle)(\Delta\theta_2(i) - \langle\Delta\theta_2(i)\rangle)$. Note that with equation (2) the initial and arbitrary assignment of microstates is not fundamental. This can be demonstrated trivially as swapping '+1' and '-1' in $\Delta\theta_1(i)$ transforms it in $-\Delta\theta_1(i)$ with a resulting $r^2_{GIFT} = \frac{[\mathrm{cov}(-\Delta\theta_1,\Delta\theta_2)]^2}{\mathrm{var}(-\Delta\theta_1)\mathrm{var}(\Delta\theta_2)} = \frac{[-\mathrm{cov}(\Delta\theta_1,\Delta\theta_2)]^2}{\mathrm{var}(\Delta\theta_1)\mathrm{var}(\Delta\theta_2)} = \frac{[\mathrm{cov}(\Delta\theta_1,\Delta\theta_2)]^2}{\mathrm{var}(\Delta\theta_1)\mathrm{var}(\Delta\theta_2)}$.

Using equation (2), Fig.2D presents a new heatmap of SNP pairwise correlations for HMGA2 (detailed in Supplemental_Data_S8.xls). This result reveals stronger correlations between SNPs that are not apparent in Fig.2B. Notably, examining the microstate composition of the two groups of genetic paths in Fig.2E clarifies why they can be distinguished in Fig.2C, given the possible reassignment of microstates '+1' and '-1'. Each group exhibits highly similar microstate profiles, resulting in slightly different levels of significance (Fig.2F).

GIFT's determination of epistasis

Returning to the set of significant SNPs identified by GIFT in Table 2, heatmaps using equation (1) (Fig.3A) and equation (2) (Fig.3B) were generated to assess pairwise correlations between SNPs. The detail of pairwise correlations between SNPs using equations (1) and (2) are reported in Supplemental_Data_S9.xls and Supplemental_Data_S10.xls, respectively. Notably, equation (1) does not suggest strong correlations between SNPs across different chromosomes (Fig.3A), whereas equation (2) indicates the opposite (Fig.3B). Since all SNPs represented in Fig.3B are significant, the absence of correlation between two SNPs implies that their genetic paths, i.e., their $\Delta\theta$ curves, differ in nature, for instance one following a parabolic trajectory and the other a sigmoidal one. To assess the presence of strong correlations using Eq.2 a threshold of $1 \geq r^2_{GIFT} \geq 0.90$ was set generating a new heatmap (Supplemental_Figure_S11.doc). To determine strongly pairwise correlated gene names the average of pairwise correlation between SNPs pertaining to any two genes and noted $\langle r^2_{GIFT}\rangle$ was used as numerical representative of pairwise correlation between genes (mean and variance of pairwise correlations are given in Supplemental_Data_S12.xls). As relatively large haplotype blocks are expected in the equine species due to breeding practices, only strongly correlated gene names across different chromosomes were selected and represented under the form of networks in Fig.3C, with the edge colors representing the strength of correlation resulting from $\langle r^2_{GIFT}\rangle$. For information, an enrichment p-value of 0.00887 for both networks was obtained conducting a STRING analysis employing a minimum required interaction score of 0.4 for the *Equus caballus* species. Since the connecting edges exhibited varying correlation values, some rectified genetic paths were grouped and plotted. This approach provided insight into the qualitative differences between genetic paths from different chromosomes with varying correlation strengths. Fig.3D demonstrates that the paraboloid genetic paths (top panel) pertaining to TMBIM4, IRAK3 and OXCT1 with a resulting average correlation between edges of $\langle r^2_{GIFT}\rangle \sim 0.97$ are visibly overlapped. A similar procedure was applied for sigmoidal paths (Fig.3D bottom panel) using CBR4, PALLD and IRS2 with an average correlation between edges of $\langle r^2_{GIFT}\rangle \sim 0.91$. These results demonstrate that the methodology underscoring GIFT can underline correlations between genes located on the same chromosome or across different chromosomes.

While GIFT supports the identification of gene networks it also enables the detection of more nuanced genetic structures. For instance, its analysis of the association between HMGA2 and the phenotype "height at withers" revealed two distinct clusters of genetic pathways, i.e., two separate groups of SNPs (Fig.2C). To explore whether these groups exhibit preferential correlations with specific gene names, they were analysed individually. As summarized in Table 4, SNPs from Group 1 are more frequently associated with other gene names than those from Group 2, highlighting GIFT's capacity to uncover fine-grained patterns of genetic interaction.

The key result obtained by Fig.3 concerns the possibility to extract gene networks based on similarity between genetic paths, underlining the possibility to detect what may be thought of as epistasis for complex traits.

GIFT's determination of gene-networks

Before applying correction measures for the false discovery rate, Bonferroni for GWAS and the Benjamini-Yekutieli procedure for GIFT, nominal significance levels were set at 1% (99% quantile threshold) for GWAS

and 0.1% (99.9% quantile threshold) for GIFT. However, this constraint can be relaxed for GIFT by using a nominal significance level similar to that of GWAS (i.e., 1%). Referring to Fig.1A and Fig.1B, this adjustment corresponds to maintaining the threshold indicated by the dashed line for GWAS while adopting the threshold represented by the continuous line for GIFT. Doing so allowed GIFT to identify 940 significant SNP-Ids, corresponding to 421 gene-Ids and 282 final known gene names. To assess the presence of strong pairwise correlations between SNPs using Eq.2, only r^2_{GIFT} values within the range $1 \geq r^2_{GIFT} \geq 0.90$ were considered. The numerical details of SNP pairwise correlations are given in Supplemental_Data_S13.xls.

This approach resulted in a binary mapping of pairwise correlations, as shown in Fig.4A, highlighting what may be termed as epistatic interactions. Next, using gene names and a mean correlation value for correlations also within the range $1 \geq \langle r^2_{GIFT} \rangle \geq 0.9$, the $N \times N$ adjacency matrix defined as $\boldsymbol{A} = (A_{i,j})$, where $N = 282$ corresponds to known gene names, $A_{i,j} = 1$ if genes 'i' and 'j' are connected or zero otherwise, was computed and the gene networks visualized using the Fruchterman-Reingold layout in Gephi (62). With this layout genes are represented with their level of connectivity shown by a circle whose radius increases as the level of connectivity increases. As shown in Fig.4B, this representation differentiates genes with a high number of connections, which are typically positioned near the center of the network, from those with fewer connections, located at the periphery highlighting, in turn, a distinction between core and peripheral genes. To assess the information from Fig.4B, the STRING database of known and predicted protein-protein interactions was utilized. By analyzing all gene names presented in Fig. 4B, a STRING analysis was performed with a minimum required interaction score of 0.4 for the *Equus caballus* species, yielding an enrichment p-value of 0.000966. Fundamental to gene networks is the notion of gene centrality that determines a measure of a gene's importance or influence within the network. One way to quantify centrality is through eigenvector centrality using the adjacency matrix previously computed. As the gene networks from Fig.4B are undirected, the matrix $\boldsymbol{A}$ is symmetric for each network, and the centrality measure for the gene 'i', denoted g_i, is related to the centralities of i's neighboring genes in the network as follows:

$$g_i = \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} \cdot g_j \quad (3)$$

where λ is the largest eigenvalue of matrix $\boldsymbol{A}$. Defining the vector of centralities as $\vec{g} = (g_1, g_2, \ldots, g_N)$, Eq.3 may be rewritten in matrix form:

$$\lambda \vec{g} = \boldsymbol{A} \cdot \vec{g} \quad (4)$$

The vector $\vec{g}$ may be normalized so that the most important gene in the network has a weight of 1, that is, $max(g_i) = 1$. Table 5 provides the centrality measures of the first ten genes determined by GIFT.

To assess whether the information in Table 5 has some relevance for the human "height", the GWAS catalog was used to determine whether the central genes found in equine were also found in previous GWASs. Of all ten central genes from Table 5, TECPR1 (Tectonin beta-propeller repeat-containing protein 1) and IRAK3, are the only ones showing no equivalence to humans from the GWAS catalog. However, it is worth noting that TECPR1 is central in promoting the fusion of autophagosomes with lysosomes to promote autophagy (63), a process up-regulated in the human adipose tissue of patients suffering of type 2 diabetes (64).

DISCUSSION

As statistical methods, GWAS and GIFT enable the determination of genotype-phenotype associations. However, they differ in the level of genetic detail captured. GWAS employs a coarse-grained approach using DDFs, whereas GIFT adopts a fine-grained strategy based on genetic paths that appear naturally upon deconstructing DDFs. Since statistical inferences are always underpinned by probability, it may seem counterintuitive that GIFT, while outperforming GWAS within such as small dataset, also relies on probability to determine significance levels, a field fundamentally based on a lack of complete information. However, GIFT and GWAS use distinct types of probabilities. GWAS employs hypothesis-free probability in the form of frequentist probability, typically represented by DDFs such as the normal distribution. In contrast, GIFT utilizes combinatorial probability, which calculates the likelihood of specific outcomes based on counting methods. Indeed, GIFT focuses on enumerating the number of ways genetic paths can occur within a defined sample space. By considering all possible configurations of microstates as their number and type are predefined, GIFT delivers precise insights without requiring extensive data collection. This way of extracting information from complex systems has parallels in statistical physics, where all complex system's states are predefined. On the

other hand, GWAS lets patterns emerge under the form of DDFs directly from data through long-run frequencies. Consequently, in the absence of sufficient data, these DDFs may become unreliable, exposing a significant limitation of GWAS when compared to GIFT's approach.

This limitation is exemplified by the study from Norton *et al.* that hypothesized that "height at withers" and insulin physiology are genetically linked in equids (11). Using GWAS with a sample size of 236 ponies, this study failed to uncover this association. In contrast, by applying GIFT to a smaller cohort, the present study extracted such genetic insights validating their initial hypothesis. Whilst this marks the first application of GIFT in equids with potential relevance to EMS, GIFT raises questions in relation to how genetic data must be interpreted.

The specific representation of scientific datasets shapes the foundation and structure of scientific knowledge bases. However, when the representation shifts, the underlying scientific concepts must also be redefined. When DDFs are used, their statistical summaries often based on their leading-order moments such as averages and variances, are expected to capture key scientific insights. In genetics, these statistical measures are closely tied to quantifiable concepts such as gene effect size and dominance for the average, and heritability for the variance. However, as the conceptual framework introduced by GIFT establishes the genetic path as a fundamental entity, one that cannot be reduced to conventional statistical summaries, this shift has important implications. Indeed, with GIFT a SNP can be significantly associated with a phenotype even when its effect size cannot be defined (see Fig.1D). Accordingly, if heredity is defined solely through variance, which itself is determined in relation to a mean or effect size, how can heredity be conceptualized for a significant SNP that lacks a measurable effect size (see Fig.1D)? These observations do not dispute the potential for any SNP to be associated with a trait or to exhibit heritability. Instead, they highlight the need to consider SNPs collectively rather than in isolation, positioning them within intricate dependencies and higher-order relationships that may elude traditional variance- or average-based approaches.

Traditionally, GWAS relies on genotyping a subset of SNPs, known as tag SNPs, which are in strong LD with many nearby SNPs. When a tag SNP shows a strong association with a trait, it suggests that a variant in LD with it may be associated with the trait. However, LD is typically determined based on bulk allele (or microstate) frequency differences calculated across the entire population. By using the detailed genetic paths provided by GIFT, it becomes now possible to refine the correlations between adjacent SNPs along a given chromosome, thereby offering a new perspective on LD. Note that, given the comprehensive information provided by GIFT, no base pair extensions were applied around tag SNPs in this present work. Due to the strength of correlations captured by GIFT, the method was extended to detect associations between SNPs across different chromosomes, commonly referred to as epistasis. Accordingly, the results demonstrate that the emergence of high-order relationships becomes visible when the shape of genetic paths is considered (see Fig.2C, Fig.3B and Fig.4A) demonstrating, in turn, that the genome is far more organized than initially thought.

GIFT's ability to infer potential epistatic interactions, in turn, enables discussion of a recently proposed concept in genetics: the omnigenic paradigm. The omnigenic paradigm for complex traits in genetics is a model that suggests that most genes in the genome contribute, directly or indirectly, to complex traits. This concept was first proposed by Boyle *et al.* and challenges the traditional polygenic model based on GWAS, which focuses on a limited set of key, often non-epistatic, genes that interact additively to influence a phenotypic trait (27). The key concepts of the omnigenic paradigm suggest that a small subset of core genes directly regulates the complex trait, while peripheral genes influence it indirectly through interactions with core genes within regulatory networks. While the theoretical and analytical frameworks, along with the present results from GIFT, support an omnigenic-like representation of complex traits, where an increasing number of SNPs are associated, it is important to acknowledge certain limitations. GIFT reinforces the concept of core versus peripheral genes, as illustrated in Fig.4b. However, contrary to the initial assumption that *most* genes in the genome contribute directly or indirectly to complex traits, GIFT suggests that only statistically significant genes are involved. In other words, it is unlikely that *most* genes contribute by default. In this context, GIFT can be seen as a hybrid conceptual framework for complex traits positioned between GWAS and the omnigenic paradigm.

Finally, it is essential to acknowledge that, while GIFT is a powerful tool for identifying genes associated with complex traits and gene networks, true validation of their functional roles requires experimental confirmation in live biological systems. Statistical associations alone do not establish causality; rather, they highlight

potential candidates that may influence a trait. Without such validation, associations remain putative, emphasizing the need for integrative approaches that combine computational methods like GIFT with experimental genetics to fully understand the molecular mechanisms underlying complex traits.

DATA AVAILABILITY

Source data for this study are not publicly available due to commercial privacy or ethical restrictions. However, the source data are available to verified researchers upon request following instructions as explained in (25).

SUPPLEMENTAL MATERIAL

XXXXXXXXXX

ACKNOWLEDGMENTS

We would like to thank Dr Oswald Matika and Dr Sian Bray for useful discussions concerning data analysis.

GRANTS

The work was funded by Mars Horsecare and WALTHAM™ Equine Studies Group (grant R01958), and the École Pratique des Hautes Études (grant EPHE1021).

DISCLOSURES

Dr Patricia Harris is an advisor to Mars Horsecare & WALTHAM ™ Equine Studies Group at Waltham Petcare Science Institute.

AUTHOR CONTRIBUTIONS

CR conceptualized GIFT; CR, JW formalized GIFT; PK coded GIFT simulations; SG computed network centrality results; the dataset used was designed and obtained by EK, PH, NMG, JE; paper was written by CR; paper was proofread by EK, PH, NMG, JE, AP, CR.

REFERENCES

1. Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, Graff M, Eliasen AU, Jiang Y, Raghavan S, Miao J, Arias JD, Graham SE, Mukamel RE, Spracklen CN, Yin X, Chen S-H, Ferreira T, Highland HH, Ji Y, Karaderi T, Lin K, Lüll K, Malden DE, Medina-Gomez C, Machado M, Moore A, Rüeger S, Sim X, Vrieze S, Ahluwalia TS, Akiyama M, Allison MA, Alvarez M, Andersen MK, Ani A, Appadurai V, Arbeeva L, Bhaskar S, Bielak LF, Bollepalli S, Bonnycastle LL, Bork-Jensen J, Bradfield JP, Bradford Y, Braund PS, Brody JA, Burgdorf KS, Cade BE, Cai H, Cai Q, Campbell A, Cañadas-Garre M, Catamo E, Chai J-F, Chai X, Chang L-C, Chang Y-C, Chen C-H, Chesi A, Choi SH, Chung R-H, Cocca M, Concas MP, Couture C, Cuellar-Partida G, Danning R, Daw EW, Degenhard F, Delgado GE, Delitala A, Demirkan A, Deng X, Devineni P, Dietl A, Dimitriou M, Dimitrov L, Dorajoo R, Ekici AB, Engmann JE, Fairhurst-Hunter Z, Farmaki A-E, Faul JD, Fernandez-Lopez J-C, Forer L, Francescatto M, Freitag-Wolf S, Fuchsberger C, Galesloot TE, Gao Y, Gao Z, Geller F, Giannakopoulou O, Giulianini F, Gjesing AP, Goel A, Gordon SD, Gorski M, Grove J, Guo X, Gustafsson S, Haessler J, Hansen TF, Havulinna AS, Haworth SJ, He J, Heard-Costa N, Hebbar P, Hindy G, Ho Y-LA, Hofer E, Holliday E, Horn K, Hornsby WE, Hottenga J-J, Huang H, Huang J, Huerta-Chagoya A, Huffman JE, Hung Y-J, Huo S, Hwang MY, Iha H, Ikeda DD, Isono M, Jackson AU, Jäger S, Jansen IE, Johansson I, Jonas JB, Jonsson A, Jørgensen T, Kalafati I-P, Kanai M, Kanoni S, Kårhus LL, Kasturiratne A, Katsuya T, Kawaguchi T, Kember RL, Kentistou KA, Kim H-N, Kim YJ, Kleber ME, Knol MJ, Kurbasic A, Lauzon M, Le P, Lea R, Lee J-Y, Leonard HL, Li SA, Li X, Li X, Liang J, Lin H, Lin S-Y, Liu J, Liu X, Lo KS, Long J, Lores-Motta L, Luan J, Lyssenko V, Lyytikäinen L-P, Mahajan A, Mamakou V, Mangino M, Manichaikul A, Marten J, Mattheisen M, Mavarani L, McDaid AF, Meidtner K, Melendez TL, Mercader JM,

Milaneschi Y, Miller JE, Millwood IY, Mishra PP, Mitchell RE, Møllehave LT, Morgan A, Mucha S, Munz M, Nakatochi M, Nelson CP, Nethander M, Nho CW, Nielsen AA, Nolte IM, Nongmaithem SS, Noordam R, Ntalla I, Nutile T, Pandit A, Christofidou P, Pärna K, Pauper M, Petersen ERB, Petersen LV, Pitkänen N, Polašek O, Poveda A, Preuss MH, Pyarajan S, Raffield LM, Rakugi H, Ramirez J, Rasheed A, Raven D, Rayner NW, Riveros C, Rohde R, Ruggiero D, Ruotsalainen SE, Ryan KA, Sabater-Lleal M, Saxena R, Scholz M, Sendamarai A, Shen B, Shi J, Shin JH, Sidore C, Sitlani CM, Slieker RC, Smit RAJ, Smith AV, Smith JA, Smyth LJ, Southam L, Steinthorsdottir V, Sun L, Takeuchi F, Tallapragada DSP, Taylor KD, Tayo BO, Tcheandjieu C, Terzikhan N, Tesolin P, Teumer A, Theusch E, Thompson DJ, Thorleifsson G, Timmers PRHJ, Trompet S, Turman C, Vaccargiu S, van der Laan SW, van der Most PJ, van Klinken JB, van Setten J, Verma SS, Verweij N, Veturi Y, Wang CA, Wang C, Wang L, Wang Z, Warren HR, Bin Wei W, Wickremasinghe AR, Wielscher M, Wiggins KL, Winsvold BS, Wong A, Wu Y, Wuttke M, Xia R, Xie T, Yamamoto K, Yang J, Yao J, Young H, Yousri NA, Yu L, Zeng L, Zhang W, Zhang X, Zhao J-H, Zhao W, Zhou W, Zimmermann ME, Zoledziewska M, Adair LS, Adams HHH, Aguilar-Salinas CA, Al-Mulla F, Arnett DK, Asselbergs FW, Åsvold BO, Attia J, Banas B, Bandinelli S, Bennett DA, Bergler T, Bharadwaj D, Biino G, Bisgaard H. A saturated map of common genetic variants associated with human height. *Nature* 610: 704–712, 2022. doi: 10.1038/s41586-022-05275-y.

2. Steiner CC, Putnam AS, Hoeck PEA, Ryder OA. Conservation genomics of threatened animal species. *Annu Rev Anim Biosci* 1: 261–281, 2013. doi: 10.1146/annurev-animal-031412-103636.

3. Rauch C, Wattis J, Bray S. On the Meaning of Averages in Genome-wide Association Studies: What Should Come Next? *Organisms Journal of Biological Sciences* 6: 7–22, 2023. doi: 10.13133/2532-5876/17811.

4. Nelson RM, Pettersson ME, Carlborg Ö. A century after Fisher: time for a new paradigm in quantitative genetics. *Trends in genetics : TIG* 29: 669–676, 2013. doi: 10.1016/j.tig.2013.09.006.

5. Wattis JAD, Bray SM, Kyratzi P, Rauch C. Analysis of phenotype-genotype associations using genomic informational field theory (GIFT). *J Theor Biol* 548: 111198, 2022. doi: 10.1016/j.jtbi.2022.111198.

6. Rauch C, Kyratzi P, Blott S, Bray S, Wattis J. GIFT: new method for the genetic analysis of small gene effects involving small sample sizes. *Phys Biol* 20, 2022. doi: 10.1088/1478-3975/ac99b3.

7. Kyratzi P, Matika O, Brassington AH, Clare CE, Xu J, Barrett DA, Emes RD, Archibald AL, Paldi A, Sinclair KD, Wattis J, Rauch C. Investigative power of genomic informational field theory relative to genome-wide association studies for genotype-phenotype mapping. *Physiological Genomics* 56: 791–806, 2024. doi: 10.1152/physiolgenomics.00049.2024.

8. Sadek MH, Al-Aboud AZ, Ashmawy AA. Factor analysis of body measurements in Arabian horses. *J Anim Breed Genet* 123: 369–377, 2006. doi: 10.1111/j.1439-0388.2006.00618.x.

9. Molina A, Valera M, Santos RD, Rodero A. Genetic parameters of morphofunctional traits in Andalusian horse. *Livestock Production Science* 60: 295–303, 1999. doi: https://doi.org/10.1016/S0301-6226(99)00101-3.

10. Outram AK, Stear NA, Bendrey R, Olsen S, Kasparov A, Zaibert V, Thorpe N, Evershed RP. The earliest horse harnessing and milking. *Science* 323: 1332–1335, 2009. doi: 10.1126/science.1168594.

11. Norton EM, Avila F, Schultz NE, Mickelson JR, Geor RJ, McCue ME. Evaluation of an HMGA2 variant for pleiotropic effects on height and metabolic traits in ponies. *J Vet Intern Med* 33: 942–952, 2019. doi: 10.1111/jvim.15403.

12. Krebs LC, Santos MM de M, Siqueira MC, Araujo BPG de, Solar Diaz IDP, Costa RB, Oliveira CA de A, Barbero MMD, de Camargo GMF, Godoi FN de. Candidate genes for height measurements in Campolina horses [Online]. *Anim Prod Sci* 64, 2024. https://doi.org/10.1071/AN23071.

13. Junior AB, Quirino CR, Vega WHO, Rua MAS, David CMG, Jardim JG. Polymorphisms in the LASP1 gene allow selection for smaller stature in ponies. *Livestock Science* 216: 160–164, 2018. doi: https://doi.org/10.1016/j.livsci.2018.07.015.

14. Makvandi-Nejad S, Hoffman GE, Allen JJ, Chu E, Gu E, Chandler AM, Loredo AI, Bellone RR, Mezey JG, Brooks SA, Sutter NB. Four loci explain 83% of size variation in the horse. *PLoS One* 7: e39929, 2012. doi: 10.1371/journal.pone.0039929.

15. Frischknecht M, Jagannathan V, Plattet P, Neuditschko M, Signer-Hasler H, Bachmann I, Pacholewska A, Drögemüller C, Dietschi E, Flury C, Rieder S, Leeb T. A Non-Synonymous HMGA2 Variant Decreases Height in Shetland Ponies and Other Small Horses. *PLoS One* 10: e0140749, 2015. doi: 10.1371/journal.pone.0140749.

16. Norton EM, Schultz NE, Rendahl AK, Mcfarlane D, Geor RJ, Mickelson JR, McCue ME. Heritability of metabolic traits associated with equine metabolic syndrome in Welsh ponies and Morgan horses. *Equine Vet J* 51: 475–480, 2019. doi: 10.1111/evj.13053.

17. Tozaki T, Kikuchi M, Kakoi H, Hirota K-I, Nagata S-I. A genome-wide association study for body weight in Japanese Thoroughbred racehorses clarifies candidate regions on chromosomes 3, 9, 15, and 18. *J Equine Sci* 28: 127–134, 2017. doi: 10.1294/jes.28.127.

18. Metzger J, Schrimpf R, Philipp U, Distl O. Expression levels of LCORL are associated with body size in horses. *PLoS One* 8: e56497, 2013. doi: 10.1371/journal.pone.0056497.

19. Reich P, Möller S, Stock KF, Nolte W, von Depka Prondzinski M, Reents R, Kalm E, Kühn C, Thaller G, Falker-Gieske C, Tetens J. Genomic analyses of withers height and linear conformation traits in German Warmblood horses using imputed sequence-level genotypes. *Genet Sel Evol* 56: 45, 2024. doi: 10.1186/s12711-024-00914-6.

20. Signer-Hasler H, Flury C, Haase B, Burger D, Simianer H, Leeb T, Rieder S. A genome-wide association study reveals loci influencing height and other conformation traits in horses. *PLoS One* 7: e37282, 2012. doi: 10.1371/journal.pone.0037282.

21. Tetens J, Widmann P, Kühn C, Thaller G. A genome-wide association study indicates LCORL/NCAPG as a candidate locus for withers height in German Warmblood horses. *Anim Genet* 44: 467–471, 2013. doi: 10.1111/age.12031.

22. Staiger EA, Al Abri MA, Pflug KM, Kalla SE, Ainsworth DM, Miller D, Raudsepp T, Sutter NB, Brooks SA. Skeletal variation in Tennessee Walking Horses maps to the LCORL/NCAPG gene region. *Physiol Genomics* 48: 325–335, 2016. doi: 10.1152/physiolgenomics.00100.2015.

23. Suontama M, Saastamoinen MT, Ojala M. Estimates of non-genetic effects and genetic parameters for body measures and subjectively scored traits in Finnhorse trotters. *Livestock Science* 124: 205–209, 2009. doi: https://doi.org/10.1016/j.livsci.2009.01.017.

24. Zechner P, Zohman F, Sölkner J, Bodo I, Habe F, Marti E, Brem G. Morphological description of the Lipizzan horse population. *Livestock Production Science* 69: 163–177, 2001. doi: https://doi.org/10.1016/S0301-6226(00)00254-2.

564 25. Knowles EJ, Elliott J, Harris PA, Chang Y-M, Menzies-Gow NJ. Predictors of laminitis development in a
565 cohort of nonlaminitic ponies. *Equine Vet J* 55: 12–23, 2023. doi: 10.1111/evj.13572.

566 26. Clark BL, Bamford NJ, Stewart AJ, McCue ME, Rendahl A, Bailey SR, Bertin F-R, Norton EM. Evaluation
567 of an HMGA2 variant contribution to height and basal insulin concentrations in ponies. *J Vet Intern*
568 *Med* 37: 1186–1192, 2023. doi: 10.1111/jvim.16723.

569 27. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*
570 169: 1177–1186, 2017. doi: 10.1016/j.cell.2017.05.038.

571 28. Webster MT, Kamgari N, Perloski M, Hoeppner MP, Axelsson E, Hedhammar Å, Pielberg G, Lindblad-
572 Toh K. Linked genetic variants on chromosome 10 control ear morphology and body mass among dog
573 breeds. *BMC Genomics* 16: 474, 2015. doi: 10.1186/s12864-015-1702-2.

574 29. Posbergh CJ, Huson HJ. All sheeps and sizes: a genetic investigation of mature body size across sheep
575 breeds reveals a polygenic nature. *Anim Genet* 52: 99–107, 2021. doi: 10.1111/age.13016.

576 30. Naval-Sánchez M, Porto-Neto LR, Cardoso DF, Hayes BJ, Daetwyler HD, Kijas J, Reverter A. Selection
577 signatures in tropical cattle are enriched for promoter and coding regions and reveal missense
578 mutations in the damage response gene HELB. *Genet Sel Evol* 52: 27, 2020. doi: 10.1186/s12711-020-
579 00546-6.

580 31. Cerezo M, Sollis E, Ji Y, Lewis E, Abid A, Bircan KO, Hall P, Hayhurst J, John S, Mosaku A, Ramachandran
581 S, Foreman A, Ibrahim A, McLaughlin J, Pendlington Z, Stefancsik R, Lambert SA, McMahon A, Morales
582 J, Keane T, Inouye M, Parkinson H, Harris LW. The NHGRI-EBI GWAS Catalog: standards for reusability,
583 sustainability and diversity. *Nucleic Acids Research* 53: D998–D1005, 2025. doi: 10.1093/nar/gkae1070.

584 32. Copps KD, White MF. Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin
585 receptor substrate proteins IRS1 and IRS2. *Diabetologia* 55: 2565–2582, 2012. doi: 10.1007/s00125-
586 012-2644-8.

587 33. Fasshauer M, Klein J, Ueki K, Kriauciunas KM, Benito M, White MF, Kahn CR. Essential role of insulin
588 receptor substrate-2 in insulin stimulation of Glut4 translocation and glucose uptake in brown
589 adipocytes. *J Biol Chem* 275: 25494–25501, 2000. doi: 10.1074/jbc.M004046200.

590 34. Dong X, Park S, Lin X, Copps K, Yi X, White MF. Irs1 and Irs2 signaling is essential for hepatic glucose
591 homeostasis and systemic growth. *J Clin Invest* 116: 101–114, 2006. doi: 10.1172/JCI25735.

592 35. Jin L. The actin associated protein palladin in smooth muscle and in the development of diseases of the
593 cardiovasculature and in cancer. *J Muscle Res Cell Motil* 32: 7–17, 2011. doi: 10.1007/s10974-011-
594 9246-9.

595 36. Tsakiridis T, Tong P, Matthews B, Tsiani E, Bilan PJ, Klip A, Downey GP. Role of the actin cytoskeleton in
596 insulin action. *Microsc Res Tech* 47: 79–92, 1999. doi: 10.1002/(SICI)1097-
597 0029(19991015)47:2<79::AID-JEMT1>3.0.CO;2-S.

598 37. MacDonald PE, Sewing S, Wang J, Joseph JW, Smukler SR, Sakellaropoulos G, Wang J, Saleh MC, Chan
599 CB, Tsushima RG, Salapatek AMF, Wheeler MB. Inhibition of Kv2.1 voltage-dependent K+ channels in
600 pancreatic beta-cells enhances glucose-dependent insulin secretion. *J Biol Chem* 277: 44938–44945,
601 2002. doi: 10.1074/jbc.M205532200.

602 38. Lu J, Zhao R-X, Xiong F-R, Zhu J-J, Shi T-T, Zhang Y-C, Peng G-X, Yang J-K. All-potassium channel CRISPR
603 screening reveals a lysine-specific pathway of insulin secretion. *Mol Metab* 80: 101885, 2024. doi:
604 10.1016/j.molmet.2024.101885.

605 39. Yu Y, Wang J, Kang R, Dong J, Zhang Y, Liu F, Yan Y, Zhu R, Xia L, Peng X, Zhang L, He D, Gaisano HY,
606 Chen Z, He Y. Association of KCNB1 polymorphisms with lipid metabolisms and insulin resistance: a
607 case-control design of population-based cross-sectional study in Chinese Han population. *Lipids Health*
608 *Dis* 14: 112, 2015. doi: 10.1186/s12944-015-0115-1.

609 40. Fu J, Dai X, Plummer G, Suzuki K, Bautista A, Githaka JM, Senior L, Jensen M, Greitzer-Antes D, Manning
610 Fox JE, Gaisano HY, Newgard CB, Touret N, MacDonald PE. Kv2.1 Clustering Contributes to Insulin
611 Exocytosis and Rescues Human β-Cell Dysfunction. *Diabetes* 66: 1890–1900, 2017. doi: 10.2337/db16-
612 1170.

613 41. Horiba N, Masuda S, Takeuchi A, Takeuchi D, Okuda M, Inui K. Cloning and characterization of a novel
614 Na+-dependent glucose transporter (NaGLT1) in rat kidney. *J Biol Chem* 278: 14669–14676, 2003. doi:
615 10.1074/jbc.M212240200.

616 42. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 53: D609–D617, 2025. doi:
617 10.1093/nar/gkae1010.

618 43. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I, Becker A, Bennett R, Berry
619 A, Bhai J, Bhurji SK, Boddu S, Branco Lins PR, Brooks L, Ramaraju SB, Campbell LI, Martinez MC,
620 Charkhchi M, Chougule K, Cockburn A, Davidson C, De Silva NH, Dodiya K, Donaldson S, El Houdaigui B,
621 Naboulsi TE, Fatima R, Giron CG, Genez T, Grigoriadis D, Ghattaoraya GS, Martinez JG, Gurbich TA,
622 Hardy M, Hollis Z, Hourlier T, Hunt T, Kay M, Kaykala V, Le T, Lemos D, Lodha D, Marques-Coelho D,
623 Maslen G, Merino GA, Mirabueno LP, Mushtaq A, Hossain SN, Ogeh DN, Sakthivel MP, Parker A, Perry
624 M, Piližota I, Poppleton D, Prosovetskaia I, Raj S, Pérez-Silva JG, Salam AIA, Saraf S, Saraiva-Agostinho
625 N, Sheppard D, Sinha S, Sipos B, Sitnik V, Stark W, Steed E, Suner M-M, Surapaneni L, Sutinen K, Tricomi
626 FF, Urbina-Gómez D, Veidenberg A, Walsh TA, Ware D, Wass E, Willhoft NL, Allen J, Alvarez-Jarreta J,
627 Chakiachvili M, Flint B, Giorgetti S, Haggerty L, Ilsley GR, Keatley J, Loveland JE, Moore B, Mudge JM,
628 Naamati G, Tate J, Trevanion SJ, Winterbottom A, Frankish A, Hunt SE, Cunningham F, Dyer S, Finn RD,
629 Martin FJ, Yates AD. Ensembl 2024. *Nucleic Acids Research* 52: D891–D899, 2024. doi:
630 10.1093/nar/gkad1049.

631 44. Svensson KJ, Long JZ, Jedrychowski MP, Cohen P, Lo JC, Serag S, Kir S, Shinoda K, Tartaglia JA, Rao RR,
632 Chédotal A, Kajimura S, Gygi SP, Spiegelman BM. A Secreted Slit2 Fragment Regulates Adipose Tissue
633 Thermogenesis and Metabolic Function. *Cell Metab* 23: 454–466, 2016. doi:
634 10.1016/j.cmet.2016.01.008.

635 45. He C, Wei Y, Sun K, Li B, Dong X, Zou Z, Liu Y, Kinch LN, Khan S, Sinha S, Xavier RJ, Grishin NV, Xiao G,
636 Eskelinen E-L, Scherer PE, Whistler JL, Levine B. Beclin 2 functions in autophagy, degradation of G
637 protein-coupled receptors, and metabolism. *Cell* 154: 1085–1099, 2013. doi:
638 10.1016/j.cell.2013.07.035.

639 46. Shmulewitz D, Heath SC, Blundell ML, Han Z, Sharma R, Salit J, Auerbach SB, Signorini S, Breslow JL,
640 Stoffel M, Friedman JM. Linkage analysis of quantitative traits for obesity, diabetes, hypertension, and
641 dyslipidemia on the island of Kosrae, Federated States of Micronesia. *Proc Natl Acad Sci U S A* 103:
642 3502–3509, 2006. doi: 10.1073/pnas.0510156103.

643 47. Aissani B, Perusse L, Lapointe G, Chagnon YC, Bouchard L, Walts B, Bouchard C. A quantitative trait
644 locus for body fat on chromosome 1q43 in French Canadians: linkage and association studies. *Obesity*
645 *(Silver Spring)* 14: 1605–1615, 2006. doi: 10.1038/oby.2006.185.

646 48. Dong C, Beecham A, Slifer S, Wang L, McClendon MS, Blanton SH, Rundek T, Sacco RL. Genome-wide
647 linkage and peak-wide association study of obesity-related quantitative traits in Caribbean Hispanics.
648 *Hum Genet* 129: 209–219, 2011. doi: 10.1007/s00439-010-0916-2.

649 49. Bossé Y, Chagnon YC, Després J-P, Rice T, Rao DC, Bouchard C, Pérusse L, Vohl M-C. Genome-wide
650 linkage scan reveals multiple susceptibility loci influencing lipid and lipoprotein levels in the Quebec
651 Family Study. *J Lipid Res* 45: 419–426, 2004. doi: 10.1194/jlr.M300401-JLR200.

652 50. Schelling JR, Abboud HE, Nicholas SB, Pahl MV, Sedor JR, Adler SG, Arar NH, Bowden DW, Elston RC,
653 Freedman BI, Goddard KAB, Guo X, Hanson RL, Ipp E, Iyengar SK, Jun G, Kao WHL, Kasinath BS, Kimmel
654 PL, Klag MJ, Knowler WC, Nelson RG, Parekh RS, Quade SR, Rich SS, Saad MF, Scavini M, Smith MW,
655 Taylor K, Winkler CA, Zager PG, Shah VO. Genome-wide scan for estimated glomerular filtration rate in
656 multi-ethnic diabetic populations: the Family Investigation of Nephropathy and Diabetes (FIND).
657 *Diabetes* 57: 235–243, 2008. doi: 10.2337/db07-0313.

658 51. Endo S, Matsunaga T, Kitade Y, Ohno S, Tajima K, El-Kabbani O, Hara A. Human carbonyl reductase 4 is
659 a mitochondrial NADPH-dependent quinone reductase. *Biochem Biophys Res Commun* 377: 1326–
660 1330, 2008. doi: 10.1016/j.bbrc.2008.11.003.

661 52. Savage DB, Petersen KF, Shulman GI. Disordered lipid metabolism and the pathogenesis of insulin
662 resistance. *Physiol Rev* 87: 507–520, 2007. doi: 10.1152/physrev.00024.2006.

663 53. Halama A, Oliveira JM, Filho SA, Qasim M, Achkar IW, Johnson S, Suhre K, Vinardell T. Metabolic
664 Predictors of Equine Performance in Endurance Racing. *Metabolites* 11, 2021. doi:
665 10.3390/metabo11020082.

666 54. Letto J, Brosnan ME, Brosnan JT. Valine metabolism. Gluconeogenesis from 3-hydroxyisobutyrate.
667 *Biochem J* 240: 909–912, 1986. doi: 10.1042/bj2400909.

668 55. Nilsen MS, Jersin RÅ, Ulvik A, Madsen A, McCann A, Svensson P-A, Svensson MK, Nedrebø BG,
669 Gudbrandsen OA, Tell GS, Kahn CR, Ueland PM, Mellgren G, Dankel SN. 3-Hydroxyisobutyrate, A Strong
670 Marker of Insulin Resistance in Type 2 Diabetes and Obesity That Modulates White and Brown
671 Adipocyte Metabolism. *Diabetes* 69: 1903–1916, 2020. doi: 10.2337/db19-1174.

672 56. Lee Y-S, Shin D. Genome-Wide Association Studies Associated with Backfat Thickness in Landrace and
673 Yorkshire Pigs. *Genomics Inform* 16: 59–64, 2018. doi: 10.5808/GI.2018.16.3.59.

674 57. Zhang C, Weng Y, Shi F, Jin W. The Engrailed-1 Gene Stimulates Brown Adipogenesis. *Stem Cells Int*
675 2016: 7369491, 2016. doi: 10.1155/2016/7369491.

676 58. Hulsmans M, Geeraert B, De Keyzer D, Mertens A, Lannoo M, Vanaudenaerde B, Hoylaerts M,
677 Benhabilès N, Tsatsanis C, Mathieu C, Holvoet P. Interleukin-1 receptor-associated kinase-3 is a key
678 inhibitor of inflammation in obesity and metabolic syndrome. *PLoS One* 7: e30414, 2012. doi:
679 10.1371/journal.pone.0030414.

680 59. Tabatabaei Dakhili SA, Greenwell AA, Yang K, Abou Farraj R, Saed CT, Gopal K, Chan JSF, Chahade JJ,
681 Eaton F, Lee C, Velázquez-Martínez CA, Crawford PA, Glover JNM, Al Batran R, Ussher JR. The
682 Antipsychotic Dopamine 2 Receptor Antagonist Diphenylbutylpiperidines Improve Glycemia in

683 Experimental Obesity by Inhibiting Succinyl-CoA:3-Ketoacid CoA Transferase. *Diabetes* 72: 126–134,
684 2023. doi: 10.2337/db22-0221.

685 60. Al Batran R, Gopal K, Capozzi ME, Chahade JJ, Saleme B, Tabatabaei-Dakhili SA, Greenwell AA, Niu J,
686 Almutairi M, Byrne NJ, Masson G, Kim R, Eaton F, Mulvihill EE, Garneau L, Masters AR, Desta Z,
687 Velázquez-Martínez CA, Aguer C, Crawford PA, Sutendra G, Campbell JE, Dyck JRB, Ussher JR. Pimozide
688 Alleviates Hyperglycemia in Diet-Induced Obesity by Inhibiting Skeletal Muscle Ketone Oxidation. *Cell*
689 *Metab* 31: 909-919.e8, 2020. doi: 10.1016/j.cmet.2020.03.017.

690 61. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics* 38:
691 226–231, 1968. doi: 10.1007/BF01245622.

692 62. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating
693 Networks. *ICWSM* 3: 361–362, 2009. doi: 10.1609/icwsm.v3i1.13937.

694 63. Wetzel L, Blanchard S, Rama S, Beier V, Kaufmann A, Wollert T. TECPR1 promotes aggrephagy by direct
695 recruitment of LC3C autophagosomes to lysosomes. *Nature Communications* 11: 2993, 2020. doi:
696 10.1038/s41467-020-16689-5.

697 64. Kosacka J, Kern M, Klöting N, Paeschke S, Rudich A, Haim Y, Gericke M, Serke H, Stumvoll M, Bechmann
698 I, Nowicki M, Blüher M. Autophagy in adipose tissue of patients with obesity and type 2 diabetes.
699 *Molecular and Cellular Endocrinology* 409: 21–32, 2015. doi: 10.1016/j.mce.2015.03.015.

700 65. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT,
701 Pyysalo S, Bork P, Jensen LJ, von Mering C. The STRING database in 2023: protein-protein association
702 networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*
703 51: D638–D646, 2023. doi: 10.1093/nar/gkac1000.

704

FIGURE LEGENDS

Figure 1: Comparison of GWAS and GIFT results for "height at withers". (A) GWAS Manhattan plot. The vertical dashed and solid lines represent the 99% and 95% Bonferroni correction thresholds, respectively. The horizontal blue arrow highlights the most significant SNP (SNP-Id: AX-104415011, Chr6, position: 82439707, $-\log(\mathrm{p_{GWAS}}) = 8.243$ ($\mathrm{a} = -2.18 \cdot 10^{-5}$ a.u., $\mathrm{d} = -5.58 \cdot 10^{-7}$ a.u., $r^2 = 0.173$). **(B)** GIFT Manhattan plot. The vertical dashed and solid line represents the 99.9% and 99% quantiles of the null hypothesis distribution by GIFT, respectively. The horizontal blue arrows point to the most significant SNP on chromosome 6 found by GIFT ($-\log(\mathrm{p_{GIFT}}) = 27.04$), that happens to be the same as determined by GWAS. The horizontal purple arrow points to the most significant SNP on chromosome 2 only determined by GIFT (SNP-Id: AX-103149967, Chr2, position: 82564015, $-\log(\mathrm{p_{GIFT}}) = 24.30$). **(C)** GWAS and GIFT representations of the SNP pointed by the horizontal blue arrows in (A) and (B). **(D)** GWAS and GIFT representations of the SNP pointed by the horizontal purple arrow in (B). **(E)** The microstates of each significant SNP were permutated randomly 1000 times and the statistics of their $-Log_{10}(p_{GIFT})$ values displayed using a box plot and compared to the $-Log_{10}(p_{GIFT})$ determined from the dataset (blue disk). The red dashed and continuous lines correspond, respectively, to the 99.9% and 99% quantile thresholds; they are the same threshold given by the black dashed and continuous lines in (B). **(F)** The original dataset (N=157) was randomly subsampled to N=140, 120, and 100. This procedure was repeated 10 times for each sample size to compute the mean and standard deviation of $-\mathrm{Log_{10}}(\mathrm{p_{GIFT}})$ for the most significant SNPs. The ordering of SNPs on the x-axis is the same as in (B). Supplemental_Data_S4.xls and Supplemental_Data_S5.xls provide the correspondence between the Axiom™ SNP annotations, typically 'AX-…', and the chromosomal positions of the significant SNPs for GWAS and GIFT, respectively) (a.u.: arbitrary unit, Chr: chromosome).

Figure 2: Genetic Paths and Linkage Disequilibrium (LD) for HMGA2. (A) Plot of genetic paths, $\theta(\mathrm{i})$ and $\theta_0(\mathrm{i})$, for the first 16 SNPs associated with HMGA2, numbered from 1 to 16. **(B)** r^2 measure of linkage disequilibrium (LD) for HMGA2. The first 16 SNPs (as shown in (A)) are enumerated from 1 to 16, with their SNP-Ids displayed along the diagonal. **(C)** Genetic trajectories ($\Delta\theta$s) of all SNPs within HMGA2 are plotted (left panel) and rectified (right panel), revealing the presence of two distinct SNP groups (coloured in light and dark grey). **(D)** $\mathrm{r^2_{GIFT}}$ measure quantifies pairwise correlations between SNPs within HMGA2. **(E)** The two SNP groups identified in **c** exhibit strong symmetries when analysed in terms of microstate composition. **(F)** The statistical significances of the two SNP groups from (C) are presented. Supplemental_Data_S5.xls provides the correspondence between the Axiom™ annotations and the chromosomal positions of the significant SNPs for GIFT.

Figure 3: $\mathbf{r^2_{GIFT}}$ measure of SNP pairwise correlations to determine local gene networks. (A) r^2-measure of pairwise correlation between significant SNPs (Table 2). The alternate red and blue colors of SNP Ids along the diagonal reflects that SNPs are located on different chromosomes. **(B)** $\mathrm{r^2_{GIFT}}$-measure of pairwise correlations between SNPs (Table 2). **(C)** Networks of correlated gene names across different chromosomes for which $1 \geq \mathrm{r^2_{GIFT}} \geq 0.90$. **(D)** Resulting representation of correlated paraboloid (left panel) and sigmoidal (right panel) genetic paths. The variable 'n' represents the number of SNPs linked to a given gene. The green triangular line delineates SNP-Ids (Table 2) associated with specific chromosomes. Supplemental_Data_S5.xls provides the correspondence between the Axiom™ annotations and the chromosomal positions of the significant SNPs for GIFT.

Figure 4: $\mathbf{r^2_{GIFT}}$ measure of gene names pairwise correlations to determine the gene networks associated with "height at withers". (A) Binary map of pairwise correlations between SNPs when adjusting the quantile threshold from 99.9% to 99% for GIFT, the green triangles separate the chromosomes. All chromosomes are present but in different proportions. **(B)** Resulting network of gene names associated with (A), where blue edges represent connections within the same chromosome, while red edges represent connections across different chromosomes.

TABLES

Table 1: List of significant genes detected by GWAS.

GWAS (99% Bonferroni correction)						
Chr.	**Gene-Id**	**Gene name**	**Biotype**	**#SNPs detected**	$\langle -\log(p_{GWAS})\rangle$	$\sigma_{-\log(p_{GWAS})}$
6	ENSECAG00000040252	HMGA2	protein	5	7.548	0.298
6	ENSECAG00000057296	NA	lncRNA	1	7.308	NA
GWAS (95% Bonferroni correction)						
Chr.	**Gene-Id**	**Gene name**	**Biotype**	**#SNPs detected**	$\langle -\log(p_{GWAS})\rangle$	$\sigma_{-\log(p_{GWAS})}$
6	ENSECAG00000040252	HMGA2	protein	5	7.548	0.298
6	ENSECAG00000057296	NA	lncRNA	2	6.769	0.763
6	ENSECAG00000051733	NA	lncRNA	1	6.229	NA

When multiple significant SNPs were detected on a gene, $\langle -\log(p_{GWAS})\rangle$ represents the average p-value calculated and, $\sigma_{-\log(p_{GWAS})}$ represents the standard deviation of p-values calculated. Supplemental_Data_S4.xls provides detailed information on the SNPs involved. (Chr.: Chromosome).

Table 2: List of significant genes detected by GIFT.

GIFT (99.9% quantile significance including FDR correction)						
Chr.	**Gene-Id**	**Gene name**	**Biotype**	**#SNPs detected**	$\langle -\log(p_{GIFT})\rangle$	$\sigma_{-\log(p_{GIFT})}$
6	ENSECAG00000051733	NA	lncRNA	1	24.633	NA
2	ENSECAG00000022089	CBR4	protein	1	22.182	NA
6	ENSECAG00000057296	NA	lncRNA	4	20.993	4.759
6	ENSECAG00000040252	HMGA2	protein	28	19.756	2.887
2	ENSECAG00000023122	PALLD	protein	2	19.612	3.634
2	ENSECAG00000050428	NA	lncRNA	3	19.035	0.328
10	ENSECAG00000049140	MFSD4B	protein	1	18.408	NA
23	ENSECAG00000020142	ELAVL2	protein	2	18.005	0.000
6	ENSECAG00000005110	MSRB3	protein	2	17.455	0.026
22	ENSECAG00000016080	KCNB1	protein	1	17.309	NA
6	ENSECAG00000006721	TMBIM4	protein	3	17.273	1.371
3	ENSECAG00000022407	SLIT2	protein	2	17.247	1.116
10	ENSECAG00000009918	MYB	protein	1	17.168	NA
18	ENSECAG00000023666	HIBCH	protein	1	17.122	NA
21	ENSECAG00000018312	OXCT1	protein	1	17.031	NA
6	ENSECAG00000026958	IRAK3	protein	4	16.977	0.573
10	ENSECAG00000006058	FUT1	protein	1	16.832	NA
18	ENSECAG00000022853	CD28	protein	1	16.521	NA
18	ENSECAG00000017122	NCKAP5	protein	1	16.481	NA
18	ENSECAG00000000157	EN1	protein	1	16.436	NA
28	ENSECAG00000009835	ABTB3	protein	1	16.368	NA
14	ENSECAG00000055847	NA	lncRNA	1	16.313	NA
17	ENSECAG00000003315	IRS2	protein	1	16.307	NA
30	ENSECAG00000017996	BECN2	protein	1	16.248	NA

When multiple significant SNPs were detected on a gene, $\langle -\log(p_{GIFT})\rangle$ represents the average p-value calculated for the gene and, $\sigma_{-\log(p_{GIFT})}$ represents the standard deviation of p-values calculated. Supplemental_Data_S5.xls provides detailed information on the SNPs involved. (Chr.: Chromosome).

765 **Table 3:** Comparison of gene names from Table 1 to gene names found in human for the phenotype "body
766 height" from the GWAS catalog.

Gene name	Study Accession Id (GWAS Catalog)
CBR4	GCST90245848
HMGA2	GCST000068, GCST000644, GCST000372, GCST000175, GCST000176, GCST001290, GCST000174, GCST001956, GCST001263, GCST000817, GCST008053, GCST008053, GCST002702, GCST002647, GCST008839, GCST007841, GCST90018959, GCST90435412, GCST008163, GCST90245844, GCST90245843, GCST90131910, GCST90095033, GCST90245848, GCST90267284, GCST90245848, GCST90245845, GCST90245846, GCST90245847, GCST90278637, GCST90479634, GCST90479636, GCST90479637, GCST90479635
PALLD	GCST90245848
MSFD4B	GCST90245848
ELAVL2	GCST90245848
MSRB3	GCST008839, GCST007841, GCST90435412, GCST90245844, GCST90267284
KCNB1	GCST008163, GCST90245848, GCST90479636
TMBIM4	GCST90245848
SLIT2	GCST010988, GCST90245848
MYB	GCST90245848
OXCT1	GCST90245848
CD28	GCST90245848
NCKAP5	GCST90245848
ABTB3	GCST90245848

767 Supplemental_Data_S6.xls provides detailed information of SNPs found from the GWAS Catalog.

768

769 **Table 5:** Equine gene names centrality (eigenvector ranks) and their comparison to gene names found in
770 human for the phenotype "body height" from the GWAS catalog.

Gene name	Eigenvector rank value	Study accession Id (GWAS catalog)
HMGA2	1.0000	See Table 2
MRSB3	0.6914	See Table 2
TMBIM4	0.5727	See Table 2
TECPR1	0.4964	None
IRAK3	0.4870	None
CELF2	0.4406	GCST90245848
OXCT1	0.4268	See Table 2
SH3RF3	0.4166	GCST008839; GCST007841; GCST90018959; GCST90245848
HELB	0.3693	GCST90245848
GMDS	0.3630	GCST008839; GCST007841; GCST90018959; GCST90018739; GCST90435412; GCST90245848; GCST90245844; GCST90245845; GCST90267284; GCST90479634; GCST90479635; GCST90479636; GCST90475368

771 Supplemental_Data_S14.xls provides detailed information of SNPs found from the GWAS Catalog.

772

FIGURES

FIGURE 1

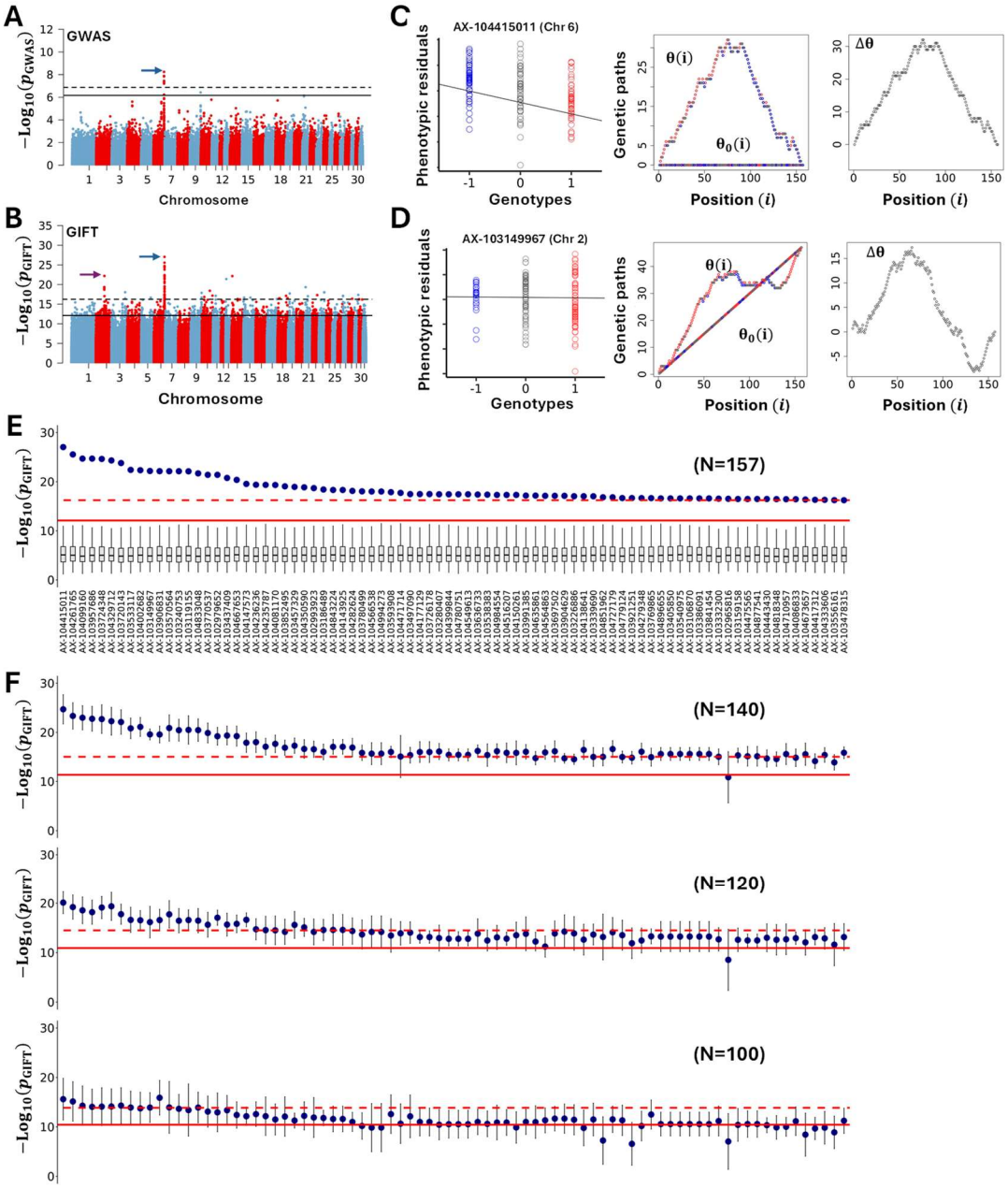
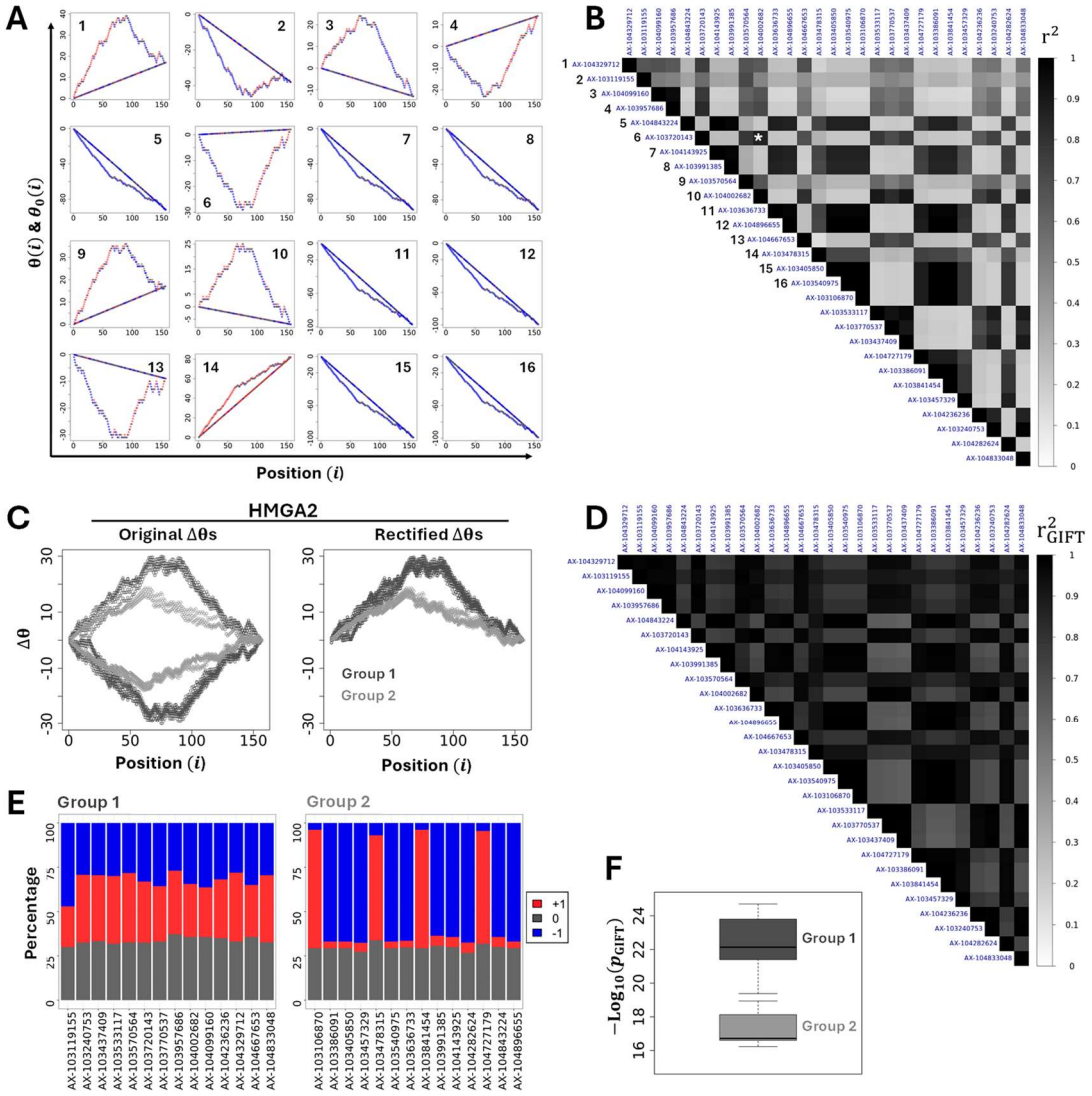
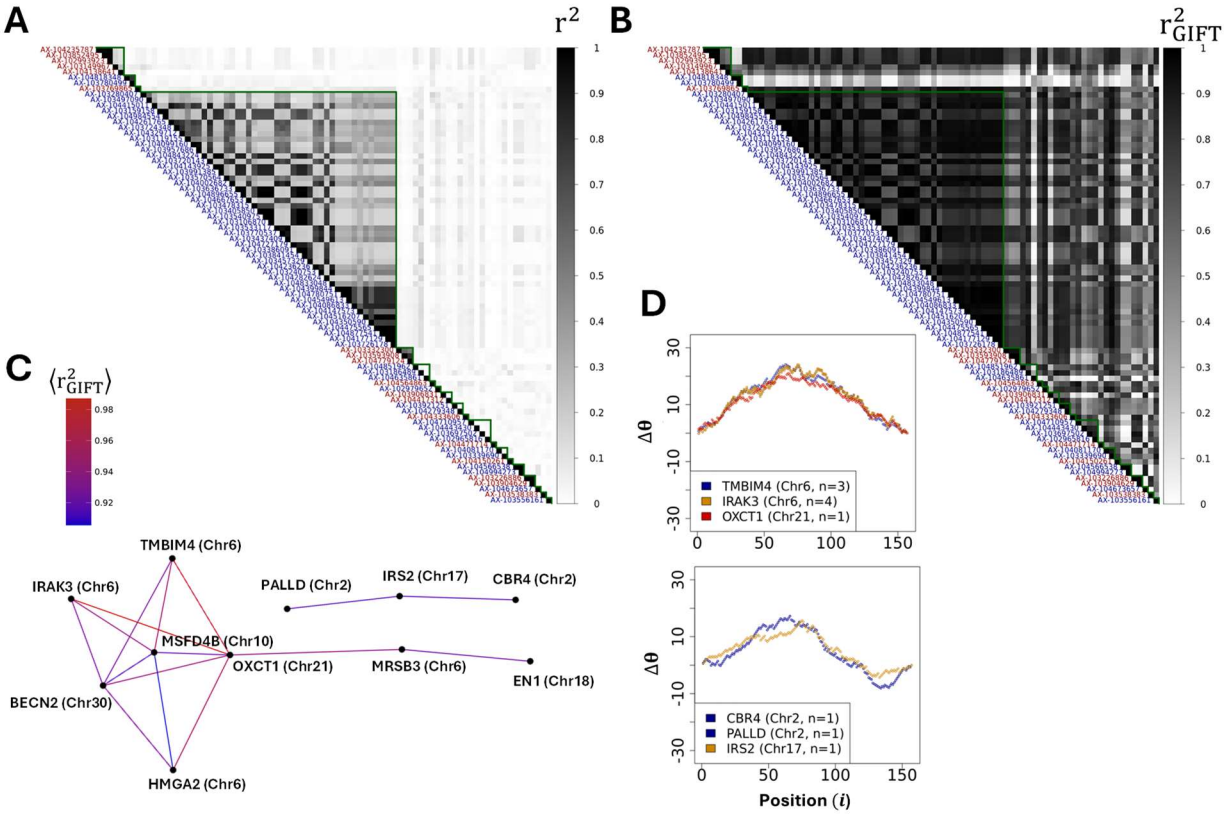


FIGURE 2


A
θ(i) & θ0(i)
Position (i)
B
r2
C
HMGA2
Original Δθs
Rectified Δθs
Δθ
Position (i)
Group 1
Group 2
D
r2GIFT
E
Group 1
Group 2
Percentage
+1
0
-1
F
−Log10(pGIFT)
Group 1
Group 2

783 **FIGURE 3**

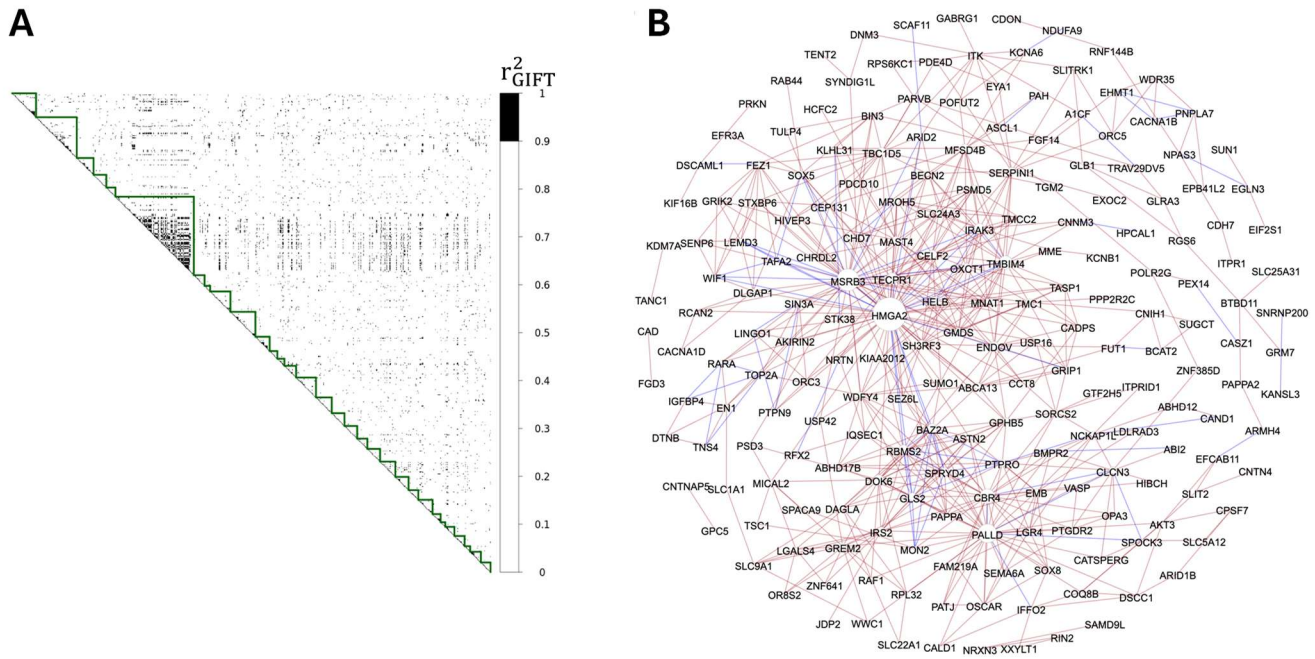
784



785

786

FIGURE 4


A
r^2_{GIFT}
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
B
SCAF11
GABRG1
CDON
NDUFA9
DNM3
TENT2
RPS6KC1
PDE4D
ITK
KCNA6
RNF144B
SLITRK1
RAB44
SYNDIG1L
WDR35
EYA1
EHMT1
PRKN
HCFC2
PARVB
POFUT2
PAH
A1CF
PNPLA7
BIN3
CACNA1B
EFR3A
TULP4
ARID2
ASCL1
ORC5
FGF14
KLHL31
TBC1D5
MFSD4B
NPAS3
SUN1
DSCAML1
FEZ1
SOX5
BECN2
SERPINI1
GLB1
TRAV29DV5
PDCD10
PSMD5
TGM2
EPB41L2
EGLN3
KIF16B
GRIK2
STXBP6
CEP131
MROH5
EXOC2
GLRA3
HIVEP3
SLC24A3
TMCC2
CNNM3
CDH7
EIF2S1
IRAK3
HPCAL1
KDM7A
SENP6
LEMD3
CHD7
MAST4
RGS6
TAFA2
CHRDL2
CELF2
MME
KCNB1
ITPR1
SLC25A31
OXCT1
TMBIM4
POLR2G
WIF1
MSRB3
TECPR1
PEX14
DLGAP1
TASP1
TANC1
SIN3A
HELB
MNAT1
TMC1
PPP2R2C
BTBD11
RCAN2
STK38
HMGA2
CNIH1
SNRNP200
SUGCT
CAD
LINGO1
GMDS
CADPS
AKIRIN2
SH3RF3
ENDOV
USP16
FUT1
BCAT2
CASZ1
GRM7
CACNA1D
RARA
NRTN
KIAA2012
GRIP1
ZNF385D
TOP2A
FGD3
ORC3
SUMO1
ABCA13
CCT8
ITPRID1
PAPPA2
KANSL3
IGFBP4
WDFY4
SEZ6L
GTF2H5
EN1
PTPN9
ABHD12
USP42
SORCS2
CAND1
GPHB5
DTNB
IQSEC1
BAZ2A
ASTN2
NCKAP1L
LDLRAD3
ARMH4
TNS4
PSD3
RBMS2
BMPR2
ABI2
RFX2
PTPRO
EFCAB11
ABHD17B
SPRYD4
CLCN3
CNTN4
CNTNAP5
SLC1A1
MICAL2
DOK6
VASP
HIBCH
GLS2
CBR4
EMB
SLIT2
SPACA9
DAGLA
PAPPA
OPA3
CPSF7
GPC5
TSC1
IRS2
PALLD
LGR4
PTGDR2
AKT3
SPOCK3
SLC5A12
LGALS4
GREM2
MON2
CATSPERG
SLC9A1
FAM219A
SEMA6A
SOX8
ARID1B
ZNF641
RAF1
OR8S2
RPL32
COQ8B
DSCC1
PATJ
OSCAR
IFFO2
JDP2
WWC1
SAMD9L
RIN2
SLC22A1
CALD1
NRXN3
XXYLT1