Genetic analysis of grain protein content and deviation in wheat

Journal of Cereal Science 121 (2025) 104099

Contents lists available at ScienceDirect

Journal of Cereal Science

journal homepage: www.elsevier.com/locate/jcs

Genetic analysis of grain protein content and deviation in wheat

Rohan Richard [a,b], Alison Lovegrove [a], Paola Tosi [b], Richard Casebow [b], Mervin Poole [c], Luzie U. Wingen [d,e], Simon Griffiths [d], Peter R. Shewry [a,*]

[a] *Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK*
[b] *School of Agriculture, Policy and Development, University of Reading, Whiteknights Campus, Early Gate, Reading, RG6 6AR, UK*
[c] *Heygates Ltd., Bugbrooke Mill, Bugbrooke, Northampton, NN7 3QH, UK*
[d] *John Innes Centre, Norwich Research Park, Colney Lane, Norwich, NR4 7UH, UK*
[e] *Population Genetics, School of Life Sciences, Department of Life Sciences, Technical University of Munich, 85354, Freising, Germany*

ARTICLE INFO

Keywords:
Wheat
Nitrogen use efficiency
Grain protein deviation

ABSTRACT

Grain protein content (GPC) is generally inversely correlated with grain yield (GY) but some genotypes consistently have higher or lower grain protein contents than predicted by simple regression analysis: this is called grain protein deviation (GPD). Positive GPD reflects greater nitrogen use efficiency and is an important target for breeders to develop more sustainable types of wheat.

Here, we investigate the genetic architecture of GPC, GY, thousand grain weight (TGW) and GPD using a population of 104 doubled haploid lines derived from a cross between two cultivars with positive (Hereward) and negative (Malacca) GPD and grown in replicated randomised field trials over three years. A total of 9 QTL were detected for all traits, five for GPC, two for GPD and one each for GY and TGW. All of the increasing alleles for GPC and GPD and the single QTL for TGW were contributed by Hereward while Malacca contributed the single increasing allele for GY. The two QTLs for GPD located on chromosomes 3A and 5B explained 23.3% and 16.6% of the variance in the sample sets, respectively. Three QTL for GPC (on chromosomes 3A, 3B, 5B) each explained more than 14% of the variance, with those on chromosomes 3A and 5B having similar locations to the GPD QTLs on the same chromosomes. A survey of the gene content between the markers bordering the confidence intervals for the two GPD QTLs on chromosomes 3A and 5B identified 136 and 704 protein coding genes, respectively, including possible candidate genes.

1. Introduction

Wheat is the most widely grown and consumed staple crop in the world, estimated to provide about 20% of the calories in the human diet. The major uses of wheat are to make breads, other baked goods (including cakes and biscuits), pasta (durum wheat) and noodles (bread wheat), but it is also widely used as an ingredient in processed foods. Furthermore, wheat is widely used as feed for livestock, particularly non-ruminants (pigs and poultry), and as raw material for ethanol production (for alcoholic beverages and bioethanol).

The processing properties of wheat are underpinned by the gluten proteins which form a viscoelastic network in dough. Gluten is a complex mixture of individual proteins and processing quality is determined by variation in both the total protein amount, with loaf volume (a widely used measure of quality) being positively correlated with grain protein content (GPC) (He and Hoseney, 1992), and with allelic variation in some individual components, notably the high molecular weight subunits of glutenin (Payne et al., 1987). Hence, it is possible to compensate, to some extent, for low intrinsic gluten quality by increasing gluten amount (Payne et al., 1987).

The importance of protein content means that grain traders and millers frequently specify minimum protein contents for breadmaking wheat, which are generally about 13% in the UK. This high protein requirement means that farmers often need to apply more nitrogen fertiliser than is optimal for crop yield, typically about 200kgN.ha^{-1} for breadmaking wheat in the UK. This not only adds to the cost of production, but also increases the energy requirement for fertiliser production and the potential environmental footprint. Although it may be possible to reduce the protein requirement for breadmaking by modifying the breadmaking process, this has proved to be difficult to achieve and attention has focused on increasing GPC at lower nitrogen fertilisation.

[*] Corresponding author.
E-mail address: peter.shewry@rothamsted.ac.uk (P.R. Shewry).

https://doi.org/10.1016/j.jcs.2024.104099
Received 22 October 2024; Received in revised form 21 December 2024; Accepted 23 December 2024
Available online 30 December 2024

Many studies have shown that GPC is inversely correlated with grain yield (GY) and hence attempts to increase GPC have generally resulted in decreases in yield (Monaghan et al., 2001; Oury et al., 2003; Oury and Godin, 2007; Bogard et al., 2010). However, Monaghan et al. (2001) compared GY and GPC for a range of cultivars, showing that some deviated positively or negatively from the simple regression line which could be calculated for GY vs GPC, and introduced the term grain protein deviation (GPD) to describe this phenomenon. GPD is an indicator of the relative ability of a cultivar to translocate nitrogen into the developing grain with cultivars exhibiting positive GPD being more efficient.

Several studies in bread (*T. aestivum* ssp. *aestivum*) and durum (*T. turgidum* ssp. *durum*) wheats have shown that GPD is partially under genetic control and therefore amenable to selection (Rapp et al., 2018; Nigro et al., 2019; Mosleth et al.,2020; Geyer et al., 2022; Paina and Gregersen, 2023). We have therefore investigated the genetic architecture of GPD, GPC, GY and thousand grain weight (TGW) in a doubled haploid (DH) population from a cross between the breadmaking wheat cultivars Malacca (negative GPD) and Hereward (positive GPD) grown in field trials for three years.

2. Materials and methods

2.1. Field trials and grain samples

A doubled haploid (DH) population of 104 lines was developed from the cross Malacca x Hereward by RAGT Seeds (UK) as described by Millar et al. (2008). This population was grown in three different environments (combination of year and location): at Rothamsted Research in 2019–2020 (51°48′06″N, 000°23′42″W), abbreviated to RR2020, and at Reading University experimental station at Sonning-on-Thames in 2020–2021 (51°28′47″N, 000°53′59″W) and 2021–2022 (51°28′41″N, 000°54′06″W), abbreviated to RU2021 and RU2022 respectively. The same level of nitrogen fertilisation (150 kg ha^{-1}) was used for all three trials but the application times and other agronomic treatments were those used as standard for the two sites. Large plots were used in order to provide accurate yield data.

The DH population (104 lines) and the two parental lines were grown in three field trials. Most lines were grown in three replicate blocks in all three years but limited availability of grain meant that a small number of lines could only be grown in one replicate (12 lines) or two replicates (8 lines) in year one. Hence, the experimental design in the RR2020 trial consisted of a Balanced Incomplete Block Design (BIBD) with 3 blocks of 100 (4.15m × 1.8m - 7.47 m^2) plots. A Randomised Complete Block Design (RCBD) was used in the two Reading field trials with three blocks of 5m × 1.9m (9.5 m^2) plots of the 106 lines (104 DH and 2 parental lines). The sowing density was 250 seeds.m^{-2} in RR2020 and 350 seeds. m^{-2} in the two Reading trials.

Nitrogen fertilisation was applied at a rate of 150kgN.ha^{-1} in two splits with the RR2020 trial receiving 50kgN.ha^{-1} and 100kgN.ha^{-1} as ammonium sulphate and ammonium nitrate, respectively, and the two trials at Reading receiving 75kgN.ha^{-1} as a mix of ammonium sulphate and ammonium nitrate and 75kgN.ha^{-1} as ammonium nitrate. The ammonium sulphate fertiliser therefore also provided sulphur at 44kgS. ha^{-1} at Rothamsted and 40kgS.ha-1 at Reading.

2.2. Determination of GPC by near infrared spectroscopy (NIRS)

A small metallic plate was filled with cleaned grains and inserted into a FieldSpec 4 Standard-Res spectroradiometer (Malvern Panalytical, UK) which had been calibrated for nitrogen (AACCI Method 46–30) (Approved Methods of Analysis (cerealsgrains.org)). The NIRS spectra were then analysed with the software Indico Pro (Malvern Panalytical, UK) and the module IQ Predict (Alphasoft, Dhaka, Bangladesh) to calculate the grain nitrogen content which was converted to protein by applying a conversion factor of 5.7.

2.3. Determination of GY and TGW

The grain weight (kg) from each plot was measured at harvest by the combine harvester and the grain yield (g.m^{-2}) calculated by dividing the grain fresh weight by the plot area. Fresh grain samples of 70–80g were dried overnight at 105 °C to determine their water content and grain yield at 15% moisture was calculated. The value at 15% moisture was chosen as this is standard for the grain industry and allowed comparison with other studies such as Bogard et al. (2010). Two sub-samples of five hundred dried grains were prepared using an Elmor C1 seed counter (Elmor, Switzerland). These were weighed and the mean values used to calculate TGW on a dry weight basis.

2.4. Calculation of GPD genotypic means

Simple linear regressions between the individual values (including the individual field replicates or blocks) for GPC and GY were calculated for the three separate environments, with 284, 298, and 304 plots for RR2020 and RU2021 and RU2022, respectively, using the statistical software R (v4.1.1; RCore Team 2021) to retrieve the residuals (raw GPD values).

For the RR2020 trial, the Best Linear Unbiased Estimators (BLUES) for GPD were calculated using a mixed model with a fixed structure, "line", a random structure, "block", and a random structure "row-*column" nested into "block" with the R package lme4 (v1.1.30; Bates et al., 2015). A mixed model was selected to account for the imbalance of the line treatment. For the RU2021 and RU2022 trials, the arithmetic means for GPD were calculated with a linear model with a treatment “line” and a structure “block” on the untransformed GPD values for RU2021 and on the log$_{10}$.transformed GPD values for RU2022 to improve the normality and homoscedasticity of the residuals.

2.5. Genotyping

The genotyping procedure and the construction of the genetic linkage map are detailed in Min et al., 2020.

2.6. Calculation of descriptive statistics, correlations and broad sense heritability

All statistics were calculated in the R software suite (v4.1.1; RCore Team 2021). In-built functions mean, median and standard deviation were used to calculate the arithmetic mean, median and the standard deviation, respectively. The correlations between replicates or between measurements in different environments were calculated using function "cor" and method “pearson”.

Broad-sense heritability (H^2) measures the percentage of phenotypic variance that is explained by the genetic variance. A high H^2 value indicates that the trait has a strong genetic basis in the set of environments under study and would be suitable for selection by breeders.

The following fixed effect model was used to calculate the broad-sense heritability (H^2):

$$y_{ikt} = \mu + g_i + e_t + (ge)_{it} + \epsilon_{ikt}$$

where y_{ikt} is the kth observation of the ith genotype at the tth environment, μ is the intercept, g_i is the main effect for the ith genotype, e_t is the main effect for the tth environment, $(ge)_{it}$ is the itth genotype-by-environment interaction effect, and ε_{ikt} is the plot error effect corresponding to y_{ikt}.

The variance components of the model: V_g, V_e, V_{ge},V_ε were calculated using the package VCA (v.1.5.1.) (Schuetzenmeister and Dufey, 2024) and were used to replace the parameters in following equation based on Schmidt et al. (2019) to calculate H^2:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2 + \frac{\sigma_{ge}^2}{n_e} + \frac{\sigma_\varepsilon^2}{n_r n_e}}$$

Where σ_g^2 is the genetic variance, σ_{ge}^2 represents the variance of the interaction between the genotype and the environment, σ_ε^2 is the residual variance, n_e the number of environments and n_r the number of replicates or blocks.

2.7. QTL. Analysis

The QTL analysis was performed in R using the package qtl (v.1.52; Broman et al., 2003) with a custom made script available from https:github.com/wingenl/rqtl_jic/tree/rqtl_jic_vs1.9. The script used the CIM (composite interval mapping) function to scan the genome for QTL locations testing between 2 and 20 co-variates and selecting the model with the highest overall LOD support. The co-variates are introduced in the model to control the influence of QTL outside the genetic interval which is tested. The CIM QTL follow the following statistical model:

$$y_j = \mu + \alpha_q\, x_j + \sum b_k\, x_{jk} + e_j$$

where: q is the qtl being tested; y_j is the trait value for individual j; μ is the overall mean; α_q is the effect of the putative QTL in the marker interval (*i*,i+*1*); x_j is the genetic predictor for individual j (taking value 1 or 0 with probability depending on the genotypes at the markers *i* and *j* and the position tested for this QTL), b_k is the partial regression coefficient of the phenotype *μ* on the *k*th marker, x_{jk} is a known coefficient for the *k*th marker in the *j*th individual taking a value 1 or 0 depending on the marker type and e_j is the random error, all errors assumed to be normally distributed.

The significance of each individual QTL selected in the final model was assessed by backward multiple regression using the R2 criteria. The QTL confidence intervals were defined as the closest markers to the genomic positions of a LOD drop of 1.5 from the QTL peak.

The CIM model uses a permutation test (with 1000 permutations) to derive a genome-wide LOD significance threshold at the 5% level. QTL with LOD scores over this threshold are significant at the 5% level. We also recorded QTL with LOD scores less than 10% below this threshold if the backward regression test showed statistical significance (at the 5% level).

2.8. Identification of putative candidate genes within the two GPD QTL confidence intervals

The positions of markers bordering the QTL confidence interval on the IWGSC RefSeqv1.0 assembly were identified as described in Shorinola et al. (2022). The BioMart tool from the EnsemblPlant software (Release 59, May 2024) was used to search for candidate genes in the Ensembl Plants Genes 59 database within the *Triticum aestivum* genes IWGSC dataset (Harrison et al., 2024) in the confidence interval region. This dataset was screened for protein coding genes only within the confidence interval (CI) of the two GPD QTL.

2.9. Orthology and gene set enrichment analysis

The functional enrichment analysis was performed using g:Profiler (version e111_eg58_p18_f463989d) with g:SCS multiple testing correction method applying a significance threshold of 0.05 (Kolberg et al., 2023). The search for orthologue genes in the model species *Arabidospis thaliana* was carried out with the version (e111_eg58_p18_f463989d).

3. Results

104 DH lines from the cross Malacca (negative GPD) and Hereward (positive GPD) were grown in three field trials (called RR2020, RU2021 and RU2022) with a fertilisation rate of 150kgN.ha^{-1} to investigate the genetic architecture of GPD under sub-optimal nitrogen nutrition (i.e. below the UK national average rate for breadmaking wheat of 200kgN. ha^{-1}).

3.1. Descriptive statistics

Descriptive statistics (arithmetic mean, median, and standard deviation) for GPD, GPC, GY and TGW were calculated for individual field trials and are presented in Table 1.

GPD showed the widest variation of the traits measured, with the coefficient of variation (cv) around the mean ranging from 12.67 to 158, and absolute values ranging from between −1.25 and +2.36 % protein in RR2020 to between −0.90 and +0.90 % protein in RU2021. By contrast, GPC, GY and TGW showed less variation, with cv values from 0.04 (GPC RU2021) to 0.09 (GY RR2020 and RU2022) (Table 1).

3.2. Correlations between trait measurements

The correlation coefficients between the four traits (TGW, GPD, GPC, GY) measured in the different environments (shown Fig. 1d) are generally in good agreement: 0.72, 0.72 and 0.80 for TGW, 0.34, 0.43 and 0.44 for GPD, and 0.39, 0.37 and 0.36 for GY. For GPC, the correlation was greater between RU2021 and RU2022 (0.60) than between these trials and RR2020 (0.43, 0.49).

Correlations between the four traits in each environment were calculated and are shown in Fig. 1 a-c. Within each field trial, GPD was strongly and positively correlated with GPC (0.97, 0.92, 0.95) with weak negative correlations with TGW at RU2021 and RU2022 and with GY at all sites. GPC was negatively correlated with GY at all sites (−0.44, −0.53, −0.51) and with TGW at the RU2021 and RU2022 sites (−0.42, −0.37), but not at RR2020. TGW was positively correlated with GY at all sites, but more strongly at RU2021 and RU2022 (0.55, 0.49) than at RR2020 (0.27).

3.3. Linear regression between GPC and GY

The linear relationships between GPC and GY were analysed separately for the individual environments (Fig. 2). Statistically significant ($p < 0.05$) slightly negative (slope = -0.002) relationships of similar magnitude were found between GPC and GY (Fig. 2), with an increase in GY of 100 g m^{-2} being accompanied by a decrease in GPC of 0.2% dry weight. The values for the two parents, Malacca and Hereward, were situated below and above the regression lines, respectively, in the RR2020 and at RU2021 sample sets (Fig. 2a and b), which is in agreement with previous reports (Millar et al., 2008; Mosleth et al., 2015, 2020). However, the separation was less clear in the RU2022 sample set (Fig. 2c). The regression models for the RU2021 and RU2022 sample sets had a higher coefficient of determination (R2=0.18) than that for the RR2020 sample set (R2=0.08) suggesting a weaker linear relationship between the two variables in the latterenvironment (Fig. 2a–c).

3.4. Broad-sense heritability

The broad sense heritability (H^2) varied from 0.57 for GPD to 0.78 for TGW (Table 2) showing that more than half (0.57) of the observed variation in GPD is due to the genetic differences between cultivars.

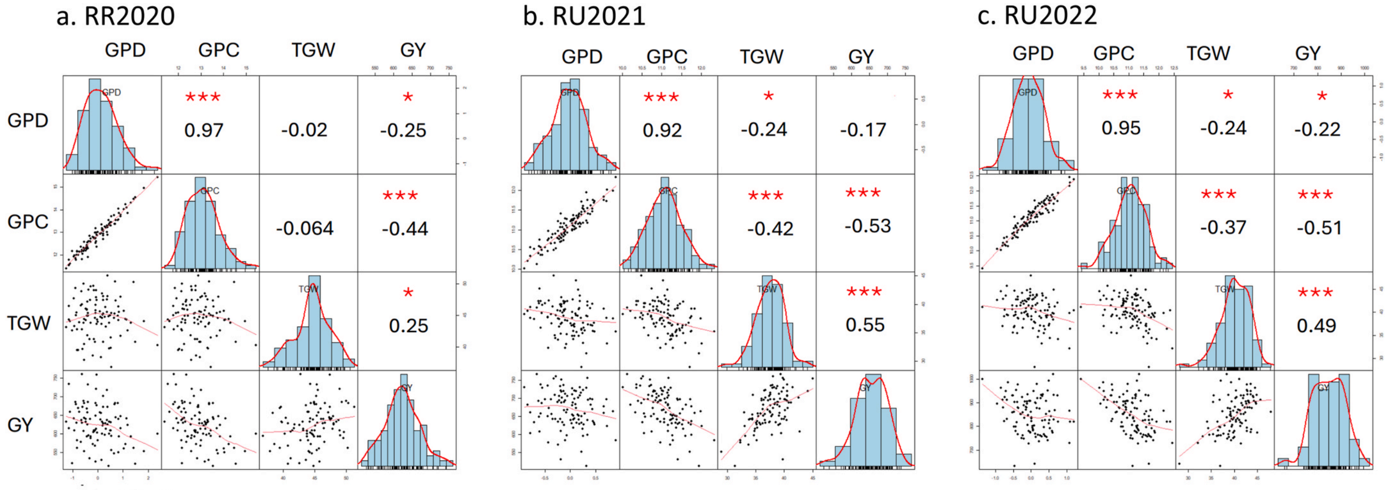
3.5. QTL analysis

The CIM identified nine statistically significant QTLs (with LOD scores above 5) for the four traits. These were located on six chromosomes with the greatest number being five for GPC and the lowest one each for GY and TGW (Table 3). A further five QTL which were just below significance in the CIM model (LOD scores of 0.2–0.4 below the LOD threshold) but significant in a statistical ‘leave-one-out test’ are presented in Supplementary Table S1.

Table 1
Descriptive statistics of GPD, GPC, GY, and TGW in the RR2020, RU2021 and RU2022 sample sets.

Trait	Environment	N	Range	Mean	Median	SD	CV
GPD	RR2020	106	−1.25:+2.36	4.50–0.3	−0.07	0.71	158
GPD	RU2021	105	−0.90; 0.90	−0.03	−0.06	−0.38	12.67
GPD	RU2022	106	−1.36; 1.19	0.01	5.00–03	0.48	48
Across environments				−0.02			
GPC	RR2020	106	11.39; 15.43	12.86	12.83	0.77	0.06
GPC	RU2021	106	10.02; 12.33	11.08	11.09	0.44	0.04
GPC	RU2022	106	9.40; 12.44	11.12	11.13	0.54	0.05
Across environments				11.69			
GY	RR2020	106	514.15; 760.33	626.06	626.68	55.55	0.09
GY	RU2021	106	511.59; 766.54	668.47	670.54	50.73	0.08
GY	RU2022	106	632.26; 1020.34	850.05	852.02	76.56	0.09
Across environments				714.86			
TGW	RR2020	98	36.81; 51.25	44.69	44.86	3.04	0.07
TGW	RU2021	106	29.01; 44.99	37.61	37.72	2.68	0.07
TGW	RU2022	106	28.14; 47.89	40.18	40.27	3.18	0.08
Across environments				40.83			

GPD is expressed as % protein at 15% moisture, GPC as protein % dry weight, GY as g.m^{-2} dry weight, TGW as g dry weight.
The values were rounded up to two decimal places. Sample size (N), standard deviation (SD), and coefficient of variation around the mean (CV).



d. All sites

	GPD RU2021	GPD RU2022		GPC RU2021	GPC RU2022		TGW RU2021	TGW RU2022		GY RU2021	GY RU2022
GPD RR2020	0.34	0.43	GPC RR2020	0.43	0.49	TGW RR2020	0.72	0.72	GY RR2020	0.39	0.37
GPD RU2021		0.44	GPC RU2021		0.60	TGW RU2021		0.80	GY RU2021		0.36

Fig. 1. Correlation coefficients for GPD, GPC, TGW and GY within individual field trials (a-c) and between the three trials (d). Red stars indicate levels of significance at 5% (*), 1% (**), and 0.1% (***) thresholds. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

All of the increasing alleles (ie. alleles associated with higher values for the traits) for GPC and GPD were contributed by Hereward, which was expected as it was chosen as having higher GPC and GPD than Malacca. Malacca contributed the single increasing allele for GY, on chromosome 3B, which was also expected as it has a higher yield potential, being released in 1997, almost a decade after Hereward (1989). Hereward also contributed the single increasing allele for TGW.

Three QTL for GPC (Q.Gpc-3A, Q.Gpc-3B and Q.Gpc-5B) each explained more than 14% of the variance of the trait in the sample sets. The two GPC QTL Q.Gpc-3A and Q.Gpc-5B are mirrored by the GPD QTL Q.Gpd-3A and Q.Gpd-5B, which have similar locations and explain similar proportions of the GPD variance. Similarly, the significant GPC QTL Q.Gpc.3B and Q.Gpc-7A are mirrored by the GPD QTL Q.Gpd.3B and Q.Gpc-7A which were just below significance in the CIM analysis (Supplementary Table S1).

Two co-locations of QTL confidence intervals were noticed: Q.Gpc-5B and Q.Gpd-5B and Q.Gpc-3B and Q.Gy-3B (Table 3). Q.Gpd-5B and Q.Gpc-5B share the same peak marker (AX-95242218) and have fully overlapping QTL confidence intervals (Table 3, highlighted in red). Q. Gpc-3B, which was identified in the RU2022 sample set, co-locates with Q.Gy-3B identified in the RR2020 sample set with the peak marker (AX-94896615) being the same and similar confidence intervals (Table 3, highlighted in blue). However, whereas Hereward exhibited the increasing allele for GPC at RU2022, the increasing allele for GY in RR2020 was from Malacca. This is consistent with the known trade-off between GY and GPC in the two parental cultivars.

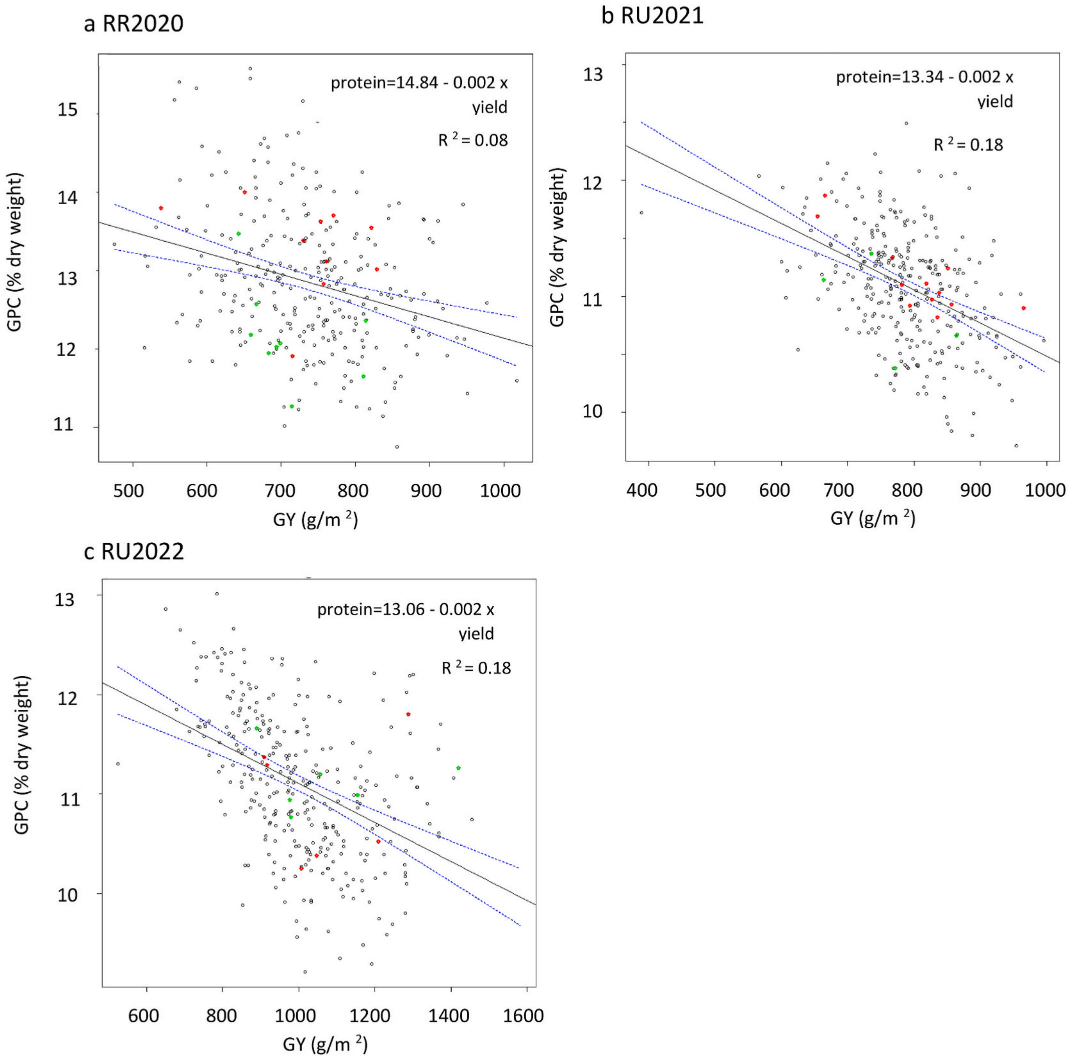


Fig. 2. Linear regressions between GPC and GY at RR2020 (a), RU2021 (b) and RU2022 (c). Blue dotted lines denote the 95% confidence intervals around the regression slopes. Individual observations for the DH parents are color-coded in red for Hereward and green for Malacca. The linear regressions were performed on 284 (a), 298 (b), and 304 (c) plots, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Broad sense heritability (H^2) for the three field trials.

Traits	GPD	GPC	GY	TGW
H^2	0.57	0.74	0.66	0.78

3.6. Gene content of the 3A and 5B GPD QTLs

Q.Gpd-3A and Q.Gpd-5B may correspond to QTLs reported by other authors (as discussed below). A survey of the gene content between the markers bordering the confidence intervals for these QTLs on the IWGCS v1.0 Reference Sequence was therefore carried out. For Q.Gpd-3A, the confidence interval extends between markers AX-94557706 and AX-94535468 (3A:479,474,983 bp – 635,102,746 bp) and for Q.Gpd-5B between AX-94974270 and AX-94892126. However, marker AX-94974270 could not be placed on chromosome 5B as it is absent from the reference sequence. Instead, the adjacent marker on the map, AX-95242218, was used to define the confidence interval (5B:587,128, 030 bp – 602, 244, 888 bp).

136 and 704 protein coding genes were inferred from the reference annotations for Q.Gpd-3A and Q.Gpd-5B, respectively. From these, 66 and 639 genes, respectively, have orthologues in the model species Arabidopsis.

Gene set enrichment analysis on g:Profiler showed a significant over-representation ($p = 0.023$) of genes associated with calmodulin binding (GO:0005516) for Q.Gpd-5B and eight significant over-representations for Q.Gpd-3A; one of them (for GO:0009987-Cellular process) being highly significant ($p = 1.17x10^{-5}$). Two of the identified groups (GO:0042937 and GO:0071916) are associated with peptide trans-membrane transport activity and both contained four genes.

4. Discussion

Improving nitrogen use efficiency (NUE) of wheat is a key sustainability target, in order to reduce the use of nitrogen fertiliser and hence the energy requirement, cost and nitrogen footprint of production. NUE is a complex trait affected by many factors but can be broadly described as the relationship between available nitrogen and crop productivity. It has been described by a range of indices including the relationship between applied nitrogen and nitrogen recovered in the grain (Congreaves et al., 2021). The progressive increases in wheat yields which have been achieved by scientific breeding are associated with decreases in grain protein content due to dilution with starch. Hence, positive GPD is a key

Table 3
QTLs identified for GPC, GY, GPD, and TGW
Two co-locations of QTL confidence intervals are highlighted: Q.Gpc-5B and Q.Gpd-5B (highlighted in red) and Q.Gpc-3B and Q.Gy-3B (highlighted in blue).

QTL name	Linkage group	Trait	Parent with increasing allele	LOD RR2020	LOD RU2021	LOD RU2022	Variance explained (%)	Additive effect	Nearest marker	Position (cM)	CI start (cM)	CI end (cM)
Q.Gpc-3A	3A	GPC	Hereward		7.5		21.7	1.17	AX-95105613	131.74	128.46	136.97
Q.Gpc-3B	3B	GPC	Hereward			5.9	19.2	0.24	AX-94896615	157.1	147.2	159.74
Q.Gpc-5B	5B	GPC	Hereward	5.5			14.4	0.27	AX-95242218	88.49	87.52	89.6
Q.Gpc-5D	5D	GPC	Hereward		6.9		9.3	0.12	AX-95652871	30.41	25.23	36.44
Q.Gpc-7A	7A	GPC	Hereward	5.2			12.6	0.28	AX-94627425	70.58	63.49	77.61
Q.Gpd-3A	3A	GPD	Hereward		5		23.3	0.17	AX-95105613	131.74	128.46	136.97
Q.Gpd-5B	5B	GPD	Hereward	5.4			16.6	0.28	AX-95242218	88.49	87.52	89.6
Q.Gy-3B	3B	GY	Malacca	6.1			23.2	-25	AX-94896615	157.1	154.46	159.74
Q.Tgw-2A	2A	TGW	Hereward	6.1			24	1.32	AX-94512334	114.24	107.94	117.32

sustainability trait as it can be exploited to breed for higher grain protein content without the requirement for additional nitrogen fertilisation (Hawkesford, 2014).

Dissecting the genetic architecture of GPD is challenging because it reflects the relationship between GPC and GY, traits which are strongly influenced by environmental factors (E) and the interactions of these with genotype (G x E). In fact, our previous analyses showed that the genotype contributed only 30% of the variation in GPD, compared with 48% for nitrogen content (a proxy for protein content) and 42% for GY (Mosleth et al., 2020).

The parents of the cross used for this study, Hereward and Malacca, were selected based on previous studies (Mosleth et al., 2015, 2020) which showed that they exhibited either strong positive GPD (Hereward) or negative GPD (Malacca). The simple linear regression for GPC and GY showed slight negative trends in the three environments, which confirmed the inverse relationship between the two traits that has been widely reported (for example, Bogard et al., 2010; Oury et al., 2003).

In this study, the DH lines displayed wide variation for GPD (high CV), which was greater than the variation between the parents. This transgressive segregation suggests the trait is controlled by multiple genes with relatively small effects. The broad-sense heritability (H^2) for GPD in the three environments studied here (0.57) was higher than that reported by Mosleth et al. (2020), who reported a heritability for GPD of 0.44 for a set of genotypes grown in 17 environments. However, in eleven of the environments much higher values for heritability (up to 0.84) were reported than in the combined dataset. The high heritability reported here may, therefore, reflect the low number of environments and greater similarity between them.

QTL analysis showed a total of nine significant QTL for all traits with the percentage of phenotypic variance explained ranging between 9.3% and 23.3%. A further five QTL were just below statistical significance and explained between 6.2 and 17.4 % of phenotypic variance (Supplementary Table S1).

Two major QTL (i.e explaining more than 15% of the phenotypic variance) for GPD were identified, Q.Gpd-3A and Q.Gpd-5B, which accounted for 24% of the phenotypic variance in RU2021 and 16.6% in RR2020, respectively. The high percentages of the variance that were not accounted in these sample sets suggest the presence of other loci with small effects as well as effects of E and G × E interactions.

The two GPD QTL (Q.Gpd-3A and Q.Gpd-5B) had additive effects of 0.17% and 0.28 % protein/g dry weight, respectively (Table 3), with substitution effects (when the decreasing allele is replaced by the increasing allele) of 0.34% and 0.56 % protein/g dry weight.

It is notable that neither of the GPD QTLs was detected in all three sample sets. Differences in the detection of QTLs in sample sets grown in different environments are frequently observed in studies of this type, particularly when the traits are controlled by multiple QTLs with relatively small individual effects. Furthermore, because GPD is a derived trait, calculated from GPC and GY, the analysis will be affected by effects of environment on the two primary traits (GPC and GY).

However, comparisons with published studies show that both GPD QTLs corresponded to previously reported QTLs, with Q.Gpd-3A overlapping with a 478.6–488.7 Mb region reported by Ruan et al. (2021) and Q.Gpd-5B being located between QGpd.mgb-5B.1 (20.8 Mb downstream of the peak marker) and QGpd.mgb-5B.2 (13.3 MB upstream of the peak marker) reported by Nigro et al. (2019) (Supplementary Table S2).

Protein coding genes underlying the confidence intervals of the two QTLs for GPD were predicted and enrichment analysis carried out. This showed that the region around Q.Gpd-5B includes genes that may encode calmodulin-binding proteins, which modulate calcium signalling in a range of biological processes, while the region around Q.Gpd-3A includes genes which may regulate peptide transport across membranes (Supplementary Table S3). It is possible that the latter contribute to greater transport of nitrogen into the developing grain of lines with the Hereward allele, which could be explored by comparing their expression levels in genotypes with the Hereward and Malacca alleles in different tissues and time points between anthesis and harvest (GPD being correlated with to post-anthesis N uptake (Bogard et al., 2010).

In conclusion, we have demonstrated that the Malacca x Hereward DH population is a useful resource to study the genetic architecture of GPD. Our results indicate that the genetic architecture of GPD is complex, involving multiple loci with small effect sizes. Nevertheless, we have identified two major QTLs on chromosomes 3A and 5B which correspond to previously reported QTLs for GPD. These QTL could therefore be used to underpin the development of markers for use in breeding. However, this would require the analyses of further crosses for more precise mapping and the validation of the markers using panels of genotypes grown in field trials. Preliminary analyses of the gene content within these QTL regions also indicate the presence of genes which could contribute to the regulation of protein accumulation in the grain, but further work is required to identify precise candidates and confirm their functions.

CRediT authorship contribution statement

Rohan Richard: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Alison Lovegrove:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Paola Tosi:** Writing – review & editing, Supervision, Conceptualization. **Richard Casebow:** Writing – review & editing, Methodology, Formal analysis. **Mervin Poole:** Conceptualization, Supervision. **Luzie U. Wingen:** Writing – review & editing, Supervision, Methodology, Formal analysis. **Simon Griffiths:** Writing – review & editing, Supervision, Conceptualization. **Peter R. Shewry:** Writing – review & editing, Writing – original draft, Supervision,

Resources, Project administration, Methodology, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work was supported by the Agriculture and Horticulture Development Board (AHDB) Studentship 061 "Exploiting variation in grain protein content and quality to determine effects of environment on processing quality". Rothamsted Research receives strategic funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and this work is part of the Delivering Sustainable Wheat (BB/X011003/1) Institute Strategic Programme.

We thank RAGT Seeds Ltd. (Grange Road, Ickleton, Saffron Walden UK) for providing the Malacca x Hereward DH population used in the study. We thank Caroline Hadley and David Casebow (University of Reading) for assistance with field trials and Malcolm Hawkesford and Andrew Riche (Rothamsted) for discussions and advice on GPD in wheat. We are particularly grateful to Dhan Bhandari (AHDB) for advice and discussion throughout the project.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jcs.2024.104099.

Data availability

Data will be made available on request.

References

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. R package 1 (35), 1. https://cran.r-project.org/web/packages/lme4/.

Bogard, M., Allard, V., Brancourt-Hulmel, M., Heumez, E., Machet, J.-M., Jeuffroy, M.-H., Gate, P., Martre, P., Le Gouis, J., 2010. Deviation from the grain protein concentration–grain yield negative relationship is highly correlated to post-anthesis N uptake in winter wheat. J. Exp. Bot. 61, 4303–4312. https://doi.org/10.1093/jxb/erq238.

Broman, K.W., Wu, H., Sen, Ś., Churchill, G.A., 2003. R/qtl: QTL mapping in experimental crosses. R package 1.66. https://cran.r-project.org/web/packages/qtl/.

Congreves, K.A., Otchere, O., Ferland, S., Williams, S., Arcand, M.M., 2021. Nitrogen use efficiency definitions today and tomorrow. Front Pl Sci 12, 637108. https://doi.org/10.3389/fpls.2021.637108.

Geyer, M., Mohler, V., Hartl, L., 2022. Genetics of the inverse relationship between grain yield and grain protein content in common wheat. Plants 11, 2146. https://doi.org/10.3390/plants11162146.

Harrison, P.W., Amode, M.R., Austine-Orimoloye, O., Azov, A.G., Barba, M., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S.K., Boddu, S., Branco Lins, P.R., Brooks, L., Ramaraju, S.B., Campbell, L.I., Martinez, M.C., Charkhchi, M., Chougule, K., Yates, A.D., 2024. Ensembl 2024. Nucleic Acids Res. 52, D891–D899. https://doi.org/10.1093/nar/gkad1049.

Hawkesford, M.J., 2014. Reducing the reliance on nitrogen fertilisation for wheat production. J. Cereal. Sci. 59, 276–283. https://doi.org/10.1016/j.jcs.2013.12.001.

He, H., Hoseney, R.C., 1992. Factors controlling gas retention in nonheated doughs. Cereal Chem. 69, 1–6.

Kolberg, L., Raudvere, U., Kuzmin, I., Adler, P., Vilo, J., Peterson, H., 2023. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping. Nucleic Acids Res. 51. https://doi.org/10.1093/nar/gkad347.

Millar, S.J., Snape, J., Ward, J., Shewry, P.R., Belton, P., Boniface, K., Summers, R., 2008. Investigating Wheat Functionality through Breeding and End Use (FQS 23) HGCA (No. 429). Project Report.

Min, B., Salt, L., Wilde, P., Kosik, O., Hassall, K., Przewieslik-Allen, A., Burridge, A.J., Poole, M., Snape, J., Wingen, L., Haslam, R., Griffiths, S., Shewry, P.R., 2020. Genetic variation in wheat grain quality is associated with differences in the galactolipid content of flour and the gas bubble properties of dough liquor. Food Chem. 6. https://doi.org/10.1016/j.fochx.2020.100093.

Monaghan, J.M., Snape, J.W., Chojecki, A.J.S., Kettlewell, P.S., 2001. The use of grain protein deviation for identifying wheat cultivars with high grain protein concentration and yield. Euphytica 122, 309–317. https://doi.org/10.1023/A:1012961703208.

Mosleth, E.F., Lillehammer, M., Pellny, T.K., Wood, A.J., Riche, A.B., Hussain, A., Griffiths, S., Hawkesford, M.J., Shewry, P.R., 2020. Genetic variation and heritability of grain protein deviation in European wheat genotypes. Field Crops Res 255, 107896. https://doi.org/10.1016/j.fcr.2020.107896.

Mosleth, E.F., Wan, Y., Lysenko, A., Chope, G.A., Penson, S.P., Shewry, P.R., Hawkesford, M.J., 2015. A novel approach to identify genes that determine grain protein deviation in cereals. Plant Biotechnol. J. 13, 625–635. https://doi.org/10.1111/pbi.12285.

Nigro, D., Gadaleta, A., Mangini, G., Colasuonno, P., Marcotuli, I., Giancaspro, A., Giove, S.L., Simeone, R., Blanco, A., 2019. Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. Planta 249, 1157–1175. https://doi.org/10.1007/s00425-018-03075-1.

Oury, F.-X., Berard, P., Brancourt-Hulmel, M., Depatureaux, C., Doussinault, G., Galic, N., Giraud, A., Heumez, E., Lecomte, C., Pluchard, P., 2003. Yield and grain protein concentration in bread wheat: a review and a study of multi-annual data from a French breeding program. J. Genet. Breed. 57, 59–68. https://www.cabidigitallibrary.org/doi/full/10.5555/20043025770.

Oury, F.-X., Godin, C., 2007. Yield and grain protein concentration in bread wheat: how to use the negative relationship between the two characters to identify favourable genotypes? Euphytica 157, 45–57. https://doi.org/10.1007/s10681-007-9395-5.

Paina, C., Gregersen, P.L., 2023. Recent advances in the genetics underlying wheat grain protein content and grain protein deviation in hexaploid wheat. Plant Biol 25, 661–670. https://doi.org/10.1111/plb.13550.

Payne, P.I., Nightingale, M.A., Krattiger, A.F., Holt, L.M., 1987. The relationship between HMW glutenin subunit composition and the bread-making quality of British-grown wheat varieties. J. Sci. Food Agric. 40, 51–65. https://doi.org/10.1002/jsfa.2740400108.

R Core Team, 2021. R: A Language and Environment for Statistical Computing, 4.3.0. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rapp, M., Lein, V., Lacoudre, F., Lafferty, J., Müller, E., Vida, G., Bozhanova, V., Ibraliu, A., Thorwarth, P., Piepho, H.P., Leiser, W.L., Würschum, T., Longin, C.F.H., 2018. Simultaneous improvement of grain yield and protein content in durum wheat by different phenotypic indices and genomic selection. Theor. Appl. Genet. 131, 1315–1329. https://doi.org/10.1007/s00122-018-3080-z.

Ruan, Y., Yu, B., Knox, R., Zhang, W., Singh, A., Cuthbert, R., Fobert, P., DePauw, R., Berraies, S., Sharpe, A., Fu, B.X., Sangha, J., 2021. Conditional mapping identified quantitative trait loci for grain protein concentration expressing independently of grain yield in Canadian durum wheat. Front. Plant Sci. 12, 642955. https://doi.org/10.3389/fpls.2021.642955.

Schmidt, P., Hartung, J., Bennewitz, J., Piepho, H.-P., 2019. Heritability in plant breeding on a genotype-difference basis. Genetics 212, 991–1008. https://doi.org/10.1534/genetics.119.302134.

Schuetzenmeister, A., Dufey, F., 2024. VCA: variance component analysis. R package 1 (5), 1. https://CRAN.R-project.org/package=VCA.

Shorinola, O., Simmonds, J., Wingen, L.U., Uauy, C., 2022. Trend, population structure, and trait mapping from 15 years of national varietal trials of UK winter wheat. G3 Genes|Genomes|Genetics 12, jkab415. https://doi.org/10.1093/g3journal/jkab415.