Plant Biotechnology Journal

Plant Biotechnology Journal (2015) **13**, pp. 625–635 doi: 10.1111/pbi.12285

A novel approach to identify genes that determine grain protein deviation in cereals

Ellen F. Mosleth[1,2], Yongfang Wan[2], Artem Lysenko[2], Gemma A. Chope[3], Simon P. Penson[3], Peter R. Shewry[2] and Malcolm J. Hawkesford[2]*

[1]*Nofima AS, Ås, Norway*
[2]*Rothamsted Research, Harpenden, Hertfordshire, UK*
[3]*Cereals and Ingredients Processing, Campden BRI, Chipping Campden, Gloucestershire, UK*

Received 26 June 2014;
revised 17 September 2014;
accepted 24 September 2014.

**Correspondence* (Tel +44 (0)1582 763133; fax +44 (0)1582 763010; email malcolm.hawkesford@rothamsted.ac.uk)

Keywords: wheat, grain protein content, transcriptome, grain protein deviation, ANOVA–PCA, one-block means of scores regression.

Summary

Grain yield and protein content were determined for six wheat cultivars grown over 3 years at multiple sites and at multiple nitrogen (N) fertilizer inputs. Although grain protein content was negatively correlated with yield, some grain samples had higher protein contents than expected based on their yields, a trait referred to as grain protein deviation (GPD). We used novel statistical approaches to identify gene transcripts significantly related to GPD across environments. The yield and protein content were initially adjusted for nitrogen fertilizer inputs and then adjusted for yield (to remove the negative correlation with protein content), resulting in a parameter termed corrected GPD. Significant genetic variation in corrected GPD was observed for six cultivars grown over a range of environmental conditions (a total of 584 samples). Gene transcript profiles were determined in a subset of 161 samples of developing grain to identify transcripts contributing to GPD. Principal component analysis (PCA), analysis of variance (ANOVA) and means of scores regression (MSR) were used to identify individual principal components (PCs) correlating with GPD alone. Scores of the selected PCs, which were significantly related to GPD and protein content but not to the yield and significantly affected by cultivar, were identified as reflecting a multivariate pattern of gene expression related to genetic variation in GPD. Transcripts with consistent variation along the selected PCs were identified by an approach hereby called one-block means of scores regression (one-block MSR).

Introduction

Wheat is the most important food crop in temperate zones, with 713 million tonnes being produced globally in 2013 (http://faostat3.fao.org/faostat-gateway/go/to/home/E). It is also the most important crop in the UK, with up to 15 million tonnes being harvested annually and about 6 million tonnes milled for making bread and other food products. However, the yields of major crops, including wheat, are highly dependent on inputs, particularly of nitrogen fertilizer which is required for canopy development and carbon capture. Wheat production is particularly dependent on nitrogen availability as the quality for bread making is largely determined by the amount and composition of the grain storage proteins (see Shewry, 2007), and it may be necessary to apply additional nitrogen (i.e. above the optimum required for grain yield) in order to achieve an adequate content of grain protein for processing. Nitrogen is currently the major production cost for wheat farmers in the UK and Europe and may also have a significant environmental footprint when applied at high levels. Increases in cereal production must therefore be viewed against this economic and environmental background (Hawkesford, 2014).

Plant breeders have been highly successful in increasing wheat yields, by an average of about 1% a year in the UK (Mackay *et al.*, 2011). However, increased yield is associated with lower protein concentration in grain (Barraclough *et al.*, 2010) and the high protein content required for bread making (a minimum of 13% dry weight in the UK) means that modern bread-making cultivars require about 35 kg N/ha more than older cultivars. For example, Dampney *et al.* (2006) reported that six of 16 modern cultivars required >280 kg N/ha to achieve 13% dry weight protein, while four of 16 required >300 kg N/ha. The sustainability of such farming practices is now being questioned, in terms of economic returns, diffuse pollution and water framework compliance.

There is a well-established negative relation between grain yield and protein concentration (see i.e. Frey, 1951; Krapp *et al.*, 2005; Lam *et al.*, 1996; Simmonds, 1995) which reflects the inter-relationships between these traits. One hypothesis is that the negative correlation between grain yield and grain protein concentration results from the dilution of protein by carbohydrates (Acreche and Slafer, 2009). Another hypothesis is competition between carbon and nitrogen for energy (Munier-Jolain and Salon, 2005).

The negative relationship between yield and grain protein content is similar for most bread-making wheat cultivars when grown under the same conditions of nitrogen availability. However, some cultivars show reproducible deviation from this relationship, with high yield being combined with high grain protein content. This relationship has been called GPD (Monaghan *et al.*, 2001) calculated as the residual from a regression analysis of grain yield on protein content. In some studies, GPD was calculated within each growth environment and compared across environments (i.e. Bogard *et al.*, 2010;

Oury and Godin, 2007), whereas in other studies, environmental factors were incorporated into the regression (i.e. Monaghan *et al.*, 2001).

It has been reported that GPD is under genetic control (Bogard *et al.*, 2010; Oury *et al.*, 2003). However, the analysis of GPD is complicated by the fact that both grain protein and grain yield have strong genotype–environmental interactions (Oury and Godin, 2007). Bogard *et al.* (2010) compared wheat grown under different conditions and showed that in most situations, GPD was correlated with postanthesis nitrogen uptake, but not with nitrogen remobilization, or with remobilization efficiency, although there was some variation between the different growth conditions. Uauy *et al.* (2006) also showed that *Gpc-B1*, a QTL associated with high contents of protein and minerals in wheat grain, encodes a transcription factor that controls nutrient remobilization from the leaves to the grain during senescence.

As much of the final grain nitrogen is accumulated in the plant before flowering and later mobilized to the grain (Barneix, 2007; Triboi and Triboi-Blondel, 2002), we hypothesize that genetic differences in GPD could directly or indirectly be reflected in differential expression of genes in the developing grain. We have therefore compared the expression patterns of gene transcripts in developing grain of six UK wheat cultivars grown in the field over three seasons at two different sites. This required the development of a novel statistical approach to dissociate differences in grain protein content and yield from the direct effects of nitrogen supply and from indirect effects related to yield and growth environment, in order to identify gene transcripts associated with GPD alone. We also suggest that this approach may have wider applicability in dissecting the transcriptional control of other complex phenotypic traits.

Results

Calculation of corrected grain protein deviation

Six UK cultivars were selected on the basis of differences in grain protein content: five high protein bread-making cultivars (Hereward, Marksman, Cordiale, Malacca and Xi19) and Istabraq which is a feed wheat cultivar known to have lower protein content. These cultivars were grown over three seasons (2008–2009, 2009–2010 and 2010–2011) at Rothamsted Research (Harpenden, UK) and at four other sites in the south-east of the UK, and at three N levels: 100 kg/ha as a 'low input' level, 200 kg/ha to reflect modern practice for bread-making wheats in the UK and 350 kg/ha as an extreme high input to achieve high grain protein (see Barraclough *et al.*, 2010; Chope *et al.*, 2014). The total number of samples was 594. Transcriptome data were determined for the experiments grown at Rothamsted Research and RAGT at three N levels in 2009 and 2010 and for one N level in 2011 giving a total of 161 samples (with one missing value). The trials grown at Rothamsted Research in 2009 and 2010 were used for feature extraction, while the remaining field trials were used to study the consistency of the expression of the selected genes across growth environments.

The yields at Rothamsted in 2009 ranged between 8.2 and 12.7 t/ha (at 85% dry matter), with grain %N ranging from 1.4 to 2.4. Istabraq had the highest yields and lowest %N which is consistent with the fact that it was the only feed cultivar. The yields in 2010 were substantially lower than in 2009, from 7.3 to 10.2 t/ha, with grain %N varying from 1.4 to 2.8. Both yield and grain %N were very responsive to N inputs in 2009, but yield was much less responsive in 2010, while %N remained very responsive. Consequently, grain %N was highest at high N inputs in 2010. This may relate to the fact that 2010 had below average rainfall, with the exception of August which was very wet. In 2011, March to May also had exceptionally low rainfall, but this was followed by a relatively wet summer (summaries of temperature and rainfall for the three growth years are provided in Table S1). The yields of the samples where gene expression data are available ranged from 7.6 to 11.5 t/ha and grain %N from 1.6 to 3.2 (Table S2).

A negative relationship between grain %N and grain yield was observed within each year and at each N level, as shown in Figure 1 for the experiments at Rothamsted and RAGT where transcriptome data were available. The different cultivars are represented by different colours, and lines are drawn for the linear relationships between yield and grain %N at the different N inputs.

In order to quantify the extent of GPD, and to identify associated transcripts, novel statistical approaches were developed to dissociate effects on grain protein content from the direct effect of nitrogen and the indirect effect of yield, and to thereby relate transcript expression profiles to this trait alone.

The yields and grain %N contents of the samples grown in 2009 (Figure 2) and 2010 (Figure S1) were initially adjusted for the direct effects of N fertilization, with Figure 2a,b (and Figure S1a,b) showing the uncorrected data and Figure 2c,d (and Figure S1c,d) the data corrected for the impact of the applied N fertilizer. A second correction was then applied to remove the inverse relationship between grain %N content and yield, providing a measure of GPD called corrected GPD (Figure 3). Figure 3(a,b) therefore show grain %N content vs. grain yield for 2009 and 2010, respectively, where both the grain %N content and the yield have been corrected for the direct effect of N level (as illustrated in Figure 2 and Figure S1). Similarly, Figure 3(c,d) show grain %N contents for the same years after correction for yield (i.e. corrected GPD).

Whereas Figure 3(a,b) show the well-established negative correlation between grain %N content and yield, this is replaced by straight lines in Figure 3(c,d) with samples showing positive and negative GPD falling above and below these lines, respectively.

Analysis of variance (ANOVA) was performed to determine the effects of the design parameters on grain %N contents, grain yield and the corrected values (Table 1a). This showed significant effects of the cultivars on GPD as well as on the uncorrected and corrected values for grain %N content and yield. Whereas nitrogen level was significant for the uncorrected values, it was not for the corrected values, showing that the effect of N fertilization had been successfully removed. There were no significant interactions between cultivar (CV) and nitrogen fertilization for any of the parameters (results not shown).

The mean values for the cultivars within each site and year (Table 2) show that Hereward generally had high GPD, whereas Istabraq had low GPD, with Istabraq generally having higher yields than Hereward. This is also seen in Figure 3(c) where Hereward is generally is located to the left in the figure in the low-yield area and Istabraq to the right. To determine whether significant genetic variation in GPD existed in the absence of a relationship to genetic differences in yield, we also performed ANOVA without these two cultivars. This again showed a significant effect of cultivar for GPD (Table 1b). The mean values for yield and grain %N, both corrected for N (Figure 4), for the remaining four cultivars show significant differences in GPD that are not related to variation in grain yield. Malacca had lower

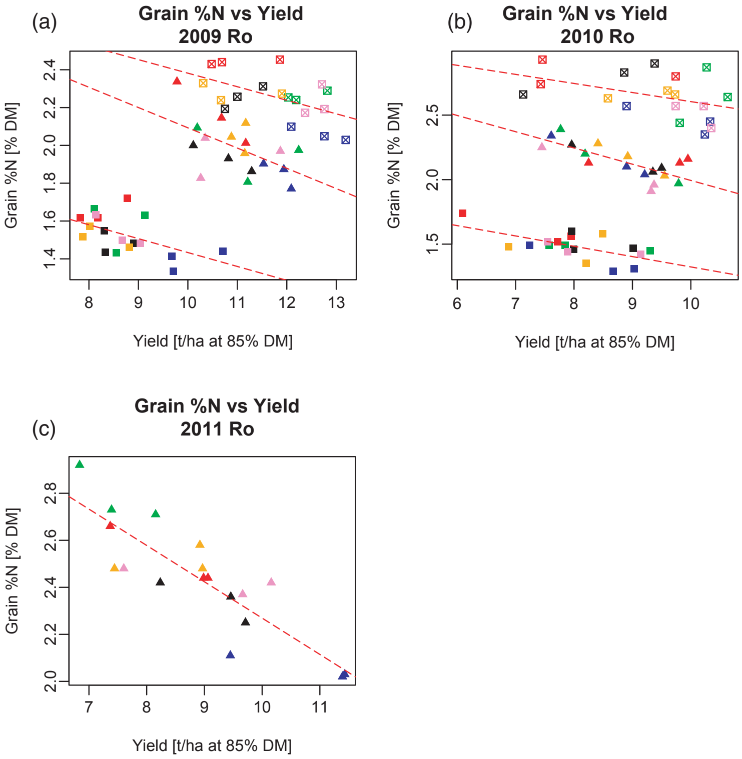

Figure 1 Raw-data plots of grain %N as a function of yield for (a) 2009, (b) 2010 and (c) 2011 at Rothamsted (Ro). In 2009 and 2010, there were three N levels: 100 kg/ha (filled squares), 200 kg/ha (triangles) and 250 kg/ha (open squares). Cultivars are colour-coded: Cordiale (green), Hereward (red), Istabraq (blue), Malacca (black), Marksman (yellow) and Xi19 (purple).

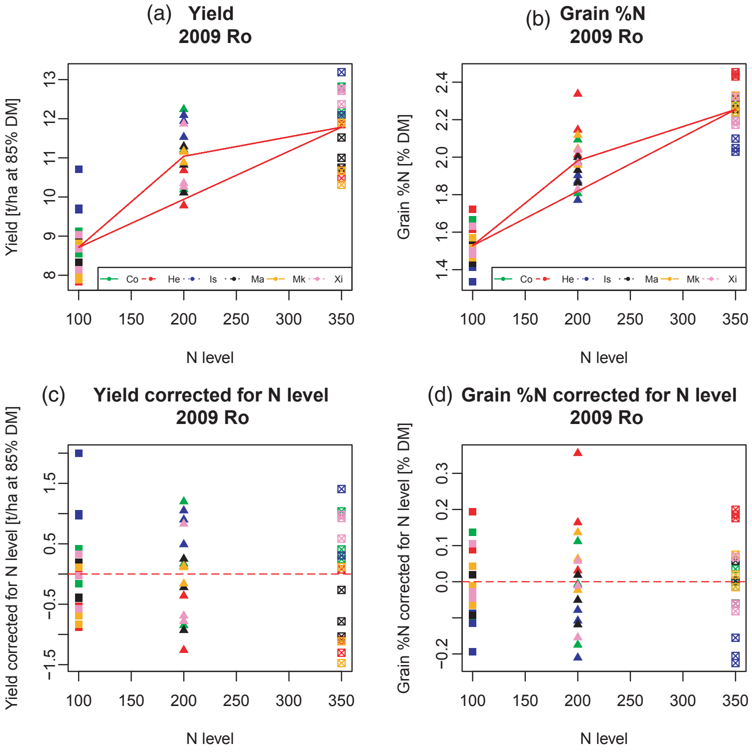

Figure 2 Correction of yield and grain %N content for their relation to N fertilisation (Yield ~N + N^2) and (Grain %N ~N + N^2) for wheat grown at Rothamsted in 2009. The x-axes of all plots are the N levels and the y-axes are as follows: (a) Yield and (b) Grain %N where the red lines are the linear and the quadratic effects. The deviation from the linear regression with N and N^2 is presented as: (c) Yield corrected for N level, and (d) Grain %N corrected for N level. N levels: 100 kg/ha (filled squares), 200 kg/ha (triangle) and 250 kg/ha (open squares). Cultivars are colour-coded: Cordiale (green), Hereward (red), Istabraq (blue), Malacca (black), Marksman (yellow) and Xi19 (purple).

values for both yield and grain %N, (corrected for N level), as well as for GPD compared with the other three cultivars, whereas Cordiale has the highest values for grain %N corrected for N level and GPD and higher Yield corrected for N level than Malacca.

The calculated GPDs (expressed as grain %N dry weight) for the six cultivars are summarized for samples grown at Rothamsted in 2009 in Figure 5(a) and in 2010 in Figure 5(b), which both show each cultivar at all N levels, and in Figure 5(c) which shows data for all sites and years at 200 kg N/ha. Figure 5(d) summarizes

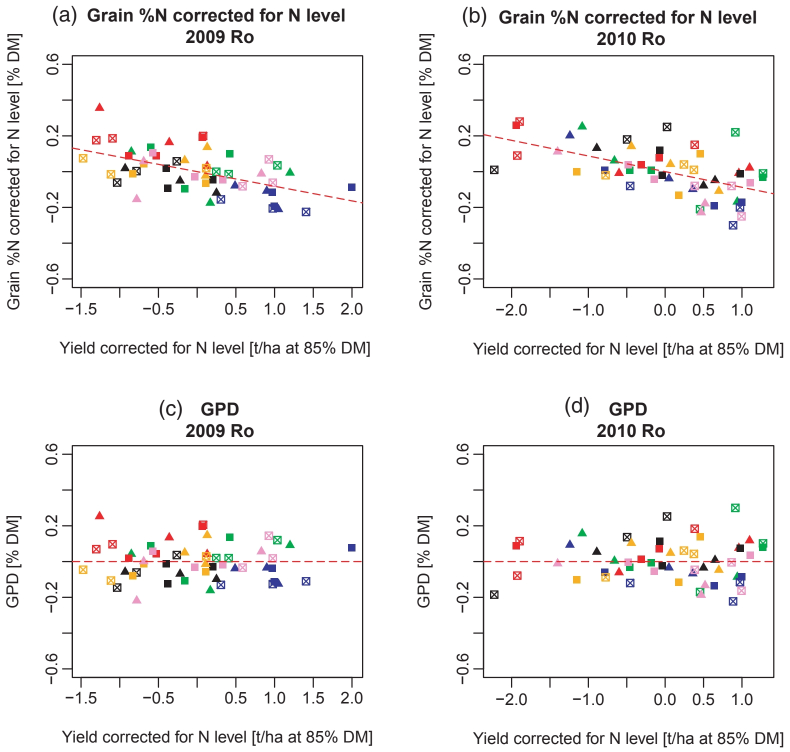

Figure 3 Correction of grain %N for yield for the experiment at Rothamsted (Ro). (a, b) Grain %N content vs. yield corrected for N level for 2009 and 2010, respectively. Cultivars are colour-coded: Cordiale (green), Hereward (red), Istabraq (blue), Malacca (black), Marksman (yellow) and Xi19 (purple). N levels: 100 kg/ha (filled squares), 200 kg/ha (triangle) and 350 kg/ha (open squares). (c, d) grain protein deviation (GPD) vs. yield corrected for N level for 2009 and 2010, with the data corrected for N level.

Table 1 *P*-values from ANOVA on the effect of the cultivar (CV) and N fertilization (N level) on the phenotypic characteristics; Yield and Protein both corrected for the effect of N level, and the double correction of protein to give GPD. The analyses were performed across years and sites

		Yield	Grain %N	Yield corrected for N level	Grain %N corrected for N level	GPD
(a) All data, 584 samples	N level	0.000	0.000	0.993	0.993	0.993
11 sites over 3 years	N level^2	0.020	0.000	0.937	0.937	0.937
3 N levels, 6 CV,	CV	0.000	0.000	0.000	0.000	0.000
3 biological replicates						
10 missing values						
(b) A subset of 391 samples	N level	0.001	0.000	0.993	0.993	0.993
11 sites over 3 years	N level^2	0.073	0.000	0.963	0.895	0.895
3 N levels, 4 CV (no Is or He),	CV	0.022	0.009	0.011	0.009	0.003
3 biological replicates						
5 missing values						

the results of the overall means for all of the trials at Rothamsted and RAGT over 3 years where the transcriptome data are available. Table 2 and Figure 4 shows results for all years and sites: Hereward shows positive GPD and Istabraq negative GPD in all years (2009–11), and Malacca being consistently lower than Hereward and Cordiale and higher than Istabraq in all data sets.

Identification of transcripts correlated with GPD

Principal component analysis (PCA) was used as the first step to identify gene transcripts related to GPD. The analysis was performed separately on the gene transcript profiles from the material grown at Rothamsted in 2009 and 2010. Consequently, there is no relationship between the principal components (PCs) identified for the 2 years.

The PCs were related to the design parameters (Table 3) and to phenotypic characters using the means of scores for the latter (Table 4). The same means of scores of the gene transcripts are used here both towards the phenotypic characters, and as will be seen below, as an internal validation towards the transcriptional data used to generate the scores. The two data blocks of measured variables are thereby connected by their design represented by means over the biological replicates of a multivariate expression, whereas the biological variation within each block is kept for validation. This approach is hereby called means of scores regression (MSR).

The PCs which are significantly related to the cultivar (Table 3) and to GPD without affecting the grain yield (Table 4) reflect the expression of gene transcripts that underlie the genetic variation in GPD.

For 2009, both PC2 and PC7 are related to GDP (Table 4), with PC2 explaining 10.1% and PC7 1.4 % of the total variation (Table 3). PC7 has the strongest relationship to GDP ($P < 0.001$) with no relationship to grain yield ($P = 0.85$), whereas PC2 has *P*-values of $P = 0.065$ for grain yield and of $P = 0.053$ for yield

Table 2 Mean values of the six cultivars across all experiments in all years (in total 11 experiments over 3 years, 584 samples). The sites were Kw (KWS), Limagrain, Ra (RAGT), Ro (Rothamsted), Sy (Syngenta). The yield and grain %N in this tables were mean centred and scaled to unit variance prior the calculation of the corrected values. Corresponding data without standardization is provided in Table S2

	Co	He	Is	Ma	Mk	Xi
Yield corrected for N after mean centring and standardising to unit variance						
2009 Ro	0.26	−0.73	1.28	−0.50	−0.54	0.22
2010 Kw	0.57	−0.80	0.91	−0.54	−0.27	0.20
2010 Li	0.58	−1.15	0.95	−0.91	−0.16	0.69
2010 Ra	0.35	−1.37	0.82	−0.68	0.77	0.11
2010 Ro	0.31	−0.53	0.18	−0.19	−0.04	0.29
2010 Sy	−0.20	−1.65	1.41	0.03	−0.04	0.27
2011 Kw	−0.19	−1.25	1.60	−0.46	0.05	0.25
2011 Li	−0.70	−0.98	1.55	0.05	−0.20	0.29
2011 Ra	−0.54	−0.97	0.93	0.03	−0.48	1.02
2011 Ro	−1.22	−0.35	1.15	0.52	−0.35	0.25
2011 Sy	−0.35	−1.63	0.91	0.11	0.44	0.52
Grain %N corrected for N after mean centring and standardising to unit variance						
2009 Ro	0.09	1.38	−1.28	−0.25	0.21	−0.14
2010 Kw	0.23	1.00	−1.87	0.33	0.40	−0.13
2010 Li	0.51	0.97	−1.66	0.14	0.01	0.03
2010 Ra	0.33	1.32	−1.66	0.08	−0.09	0.01
2010 Ro	0.11	0.73	−0.70	0.43	0.06	−0.62
2010 Sy	0.53	1.49	−1.49	−0.09	−0.05	−0.24
2011 Kw	0.19	1.54	−1.43	−0.32	0.00	0.02
2011 Li	0.67	1.31	−1.26	−0.63	0.06	−0.16
2011 Ra	0.65	1.60	−1.22	−0.48	−0.08	−0.47
2011 Ro	1.53	0.39	−1.34	−0.63	0.10	−0.05
2011 Sy	0.16	1.81	−0.29	−0.79	−1.00	0.12
GPD calculated after mean centring and standardising to unit variance						
2009 Ro	0.27	1.17	−0.70	−0.62	−0.10	−0.03
2010 Kw	0.51	0.73	−1.64	0.11	0.32	−0.05
2010 Li	0.72	0.66	−1.44	−0.14	−0.04	0.24
2010 Ra	0.62	0.70	−1.45	−0.33	0.38	0.08
2010 Ro	0.34	0.52	−0.73	0.39	0.04	−0.56
2010 Sy	0.66	0.21	−0.57	−0.13	−0.16	−0.02
2011 Kw	0.07	0.92	−0.35	−0.99	0.05	0.30
2011 Li	0.22	0.87	−0.14	−0.91	−0.14	0.10
2011 Ra	0.38	1.29	−0.79	−0.64	−0.57	0.34
2011 Ro	0.95	0.19	−0.69	−0.36	−0.40	0.32
2011 Sy	−0.07	1.02	0.35	−0.92	−0.93	0.56

corrected for N level (Table 4). Therefore, PC7 is the most relevant to GPD for 2009. For 2010, PC2 and PC3 are of most interest as they are both related to all protein parameters without any relationships to yield. These PCs accounted for 11.5% and 7.9%, respectively, of the variation in the transcriptome data (Table 4). The PCs above PC10 did not capture any information relevant to GPD. All of these PCs are significantly affected by the cultivar (Table 3).

The means of the selected scores for each cultivar for 2009 and 2010 are shown in Figure 6. As the directions of the scores and loading plots are arbitrary in PCA, the directions have been selected to facilitate the comparison with the GDP plots. These two figures show remarkable similarity when comparing PC7 with GPD for 2009 and PC2 and 3 with GPD for 2010, with Figure 5 showing the GDP values as means of the cultivars and Figure 6 showing the means of PCs from the PCA of the gene expression data selected to represent GDP. Figure 7 shows the relationship between GPD and PC7 as means of cultivars and N levels. There is a close relationship between PC7 and GPD for five of the six cultivars (omitting Xi19) with a correlation coefficient of $r = 0.86$, whereas Xi19 deviates from this. This is consistent with the genes underlying the selected PCs being responsible for, or correlated with, the variation in GPD in the cultivars Cordiale, Hereward, Istabraq, Malacca and Marksman, but not in Xi19.

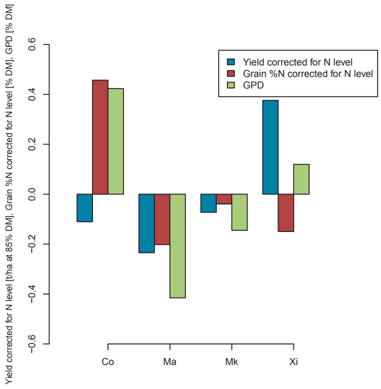

Figure 4 Yield and grain %N, both corrected for N, and corrected GPD, all mean centred and standardised to unit variance, for all data and four of the cultivars: Cordiale, Malacca, Marksman and Xi19. In total, 391 samples.

The selected PCs which are shown in Table 4 to be related to GPD therefore reflect a multivariate pattern of gene expression. However, not all of these genes may be relevant to the traits that are reflected by the PCs. To identify a smaller number of candidate genes to be studied in more detail, we applied ANOVA to each of the selected PCs, using the means of the biological replicates of the scores as inputs and the gene expression values as the responses. Thus, we used the means of the scores obtained on the same data that were used to generate the PCA. We here call this approach one-block MSR.

For 2009, PC7 was of particular relevance as it was strongly related to GPD without any relation to yield or yield corrected for N level ($P < 0.001$) (Table 4). To limit the number of genes, we therefore focused on genes that were significant for PC7 in 2009, and at the same time significant for either PC2 or PC3 in 2010. This gave 959 transcripts as the best candidates for determining the corrected GPD.

The genes selected as significant by one-block MSR as having stable values for the selected PCs for the three different biological replicates were generally those with high or low loadings of the selected PCs (see Figure S2). To further reduce the number of potential candidate genes, we performed partial least-squares regression (PLSR) with Jackknife (see Figure S3). As shown in Figure 7, the cultivar Xi19 is clearly separated from the other cultivars in terms of the relationship between the transcriptome profiles reflected by the selected PCs and the corrected GPD.

The most interesting transcripts in terms of GPD were found to be 136 transcripts positively related to GPD by the PLS regression analysis (shown as pink-filled triangles in the Figure S3). A raw

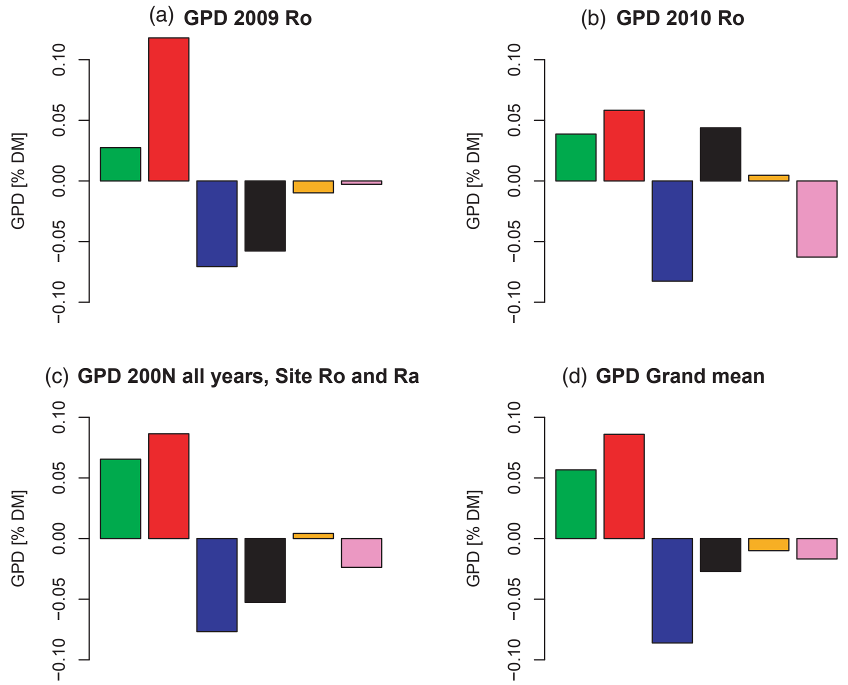

Figure 5 Mean value of GDP where protein content are corrected for N levels, and its relation with yield for (a) 2009 and (b) 2010, (c) All three growth years 2009, 2010 and 2011 at N level 200 kg/ha at Rothamsted and RAGT, corrected for yield and the effect of year and (d) all 161 data points summarised. Cultivars are colour-coded: Cordiale (green), Hereward (red), Istabraq (blue), Malacca (black), Marksman (yellow) and Xi19 (purple).

Table 3 Results of ANOVA (FDR-adjusted *P*-values) showing the effect of the design parameters (CV, linear and quadratic effects of N, and the interaction between N and CV) (input of the model) on the scores of PCA of the gene expression data (output of the model) for (a) 2009 [A total 53 samples: 1 site (Rothamsted), 3 N levels, 6 CV, 3 biological replicates, 1 missing value] and (b) 2010 [a total of 54 samples: 1 site (Rothamsted), 3 N levels, 6 CV, 3 biological replicates]

	ExplVar	CV	N	N^2	CV*N
(a)					
PC1	33.5	0.003	0.414	0.451	0.943
PC2	10.1	0.002	0.603	0.832	0.859
PC3	8.9	0.005	0.808	0.712	0.943
PC4	3.4	0.001	0.838	0.712	0.645
PC5	2.3	0.000	0.273	0.712	0.271
PC6	1.3	0.001	0.414	0.656	0.672
PC7	1.4	0.000	0.669	0.739	0.271
PC8	1.1	0.004	0.371	0.522	0.705
PC9	1.2	0.001	0.808	0.522	0.943
PC10	1.4	0.000	0.002	0.246	0.432
(b)					
PC1	24.2	0.177	0.018	0.547	0.919
PC2	11.5	0.001	0.018	0.691	0.831
PC3	7.9	0.004	0.432	0.115	0.754
PC4	2.0	0.000	0.037	0.122	0.738
PC5	4.7	0.000	0.168	0.115	0.938
PC6	1.9	0.000	0.000	0.115	0.938
PC7	1.7	0.000	0.002	0.115	0.938
PC8	4.0	0.000	0.033	0.370	0.738
PC9	5.0	0.000	0.082	0.115	0.863
PC10	0.8	0.959	0.918	0.961	0.738

Table 4 Results of ANOVA (FDR-adjusted *P*-values) showing the effect of the scores of PCA of the gene expression data (input of the model) on the phenotypic characteristics (output of the model) for (a) 2009 [a total 53 samples: 1 site (Rothamsted), 3 N levels, 6 CV, 3 biological replicates, 1 missing value, and (b) 2010 (a total of 54 samples: 1 site (Rothamsted), 3 N levels, 6 CV, 3 biological replicates]

	Yield	Protein	Yield corrected for N level	Grain%N corrected for N level	GPD
(a)					
Mean PC1	0.078	0.031	0.740	0.519	0.581
Mean PC2	0.065	0.991	0.053	0.005	0.043
Mean PC3	0.152	0.46	0.008	0.006	0.129
Mean PC4	0.736	0.682	0.139	0.911	0.421
Mean PC5	0.362	0.037	0.000	0.015	0.643
Mean PC6	0.357	0.195	0.823	0.572	0.415
Mean PC7	0.853	0.063	0.199	0.000	0.000
Mean PC8	0.071	0.034	0.431	0.636	0.285
Mean PC9	0.314	0.446	0.080	0.067	0.297
Mean PC10	0.000	0.000	0.447	0.088	0.010
(b)					
Mean PC1	0.001	0.000	0.293	0.392	0.756
Mean PC2	0.603	0.000	0.170	0.012	0.038
Mean PC3	0.942	0.021	0.285	0.030	0.062
Mean PC4	0.966	0.085	0.449	0.081	0.111
Mean PC5	0.143	0.255	0.369	0.837	0.708
Mean PC6	0.013	0.000	0.086	0.244	0.120
Mean PC7	0.313	0.005	0.641	0.491	0.608
Mean PC8	0.283	0.295	0.740	0.084	0.061
Mean PC9	0.026	0.245	0.136	0.087	0.301
Mean PC10	0.836	0.948	0.704	0.561	0.329

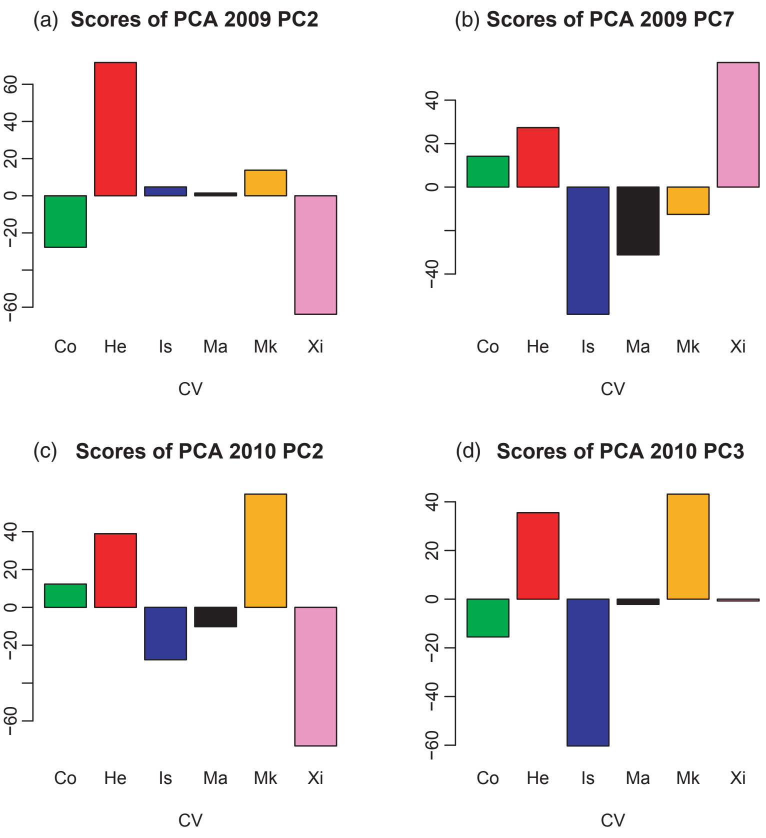

Figure 6 Mean values of the CVs for scores of selected PCs from PCA of the gene expression in 2009 (a,b) and 2010 (c,d). The PCs are selected as reflecting variation in GDP with no relation to the yield. (a,b) 2009: PC2 and PC7, respectively. (c,d) 2010: PC2 and PC3, respectively.

data plot of all these transcripts is given in Figure S4 where the gene transcripts are sorted according to the regression coefficient of the PLS regression.

The molecular functions of the 959 transcripts selected as significant for PC7 in 2009 and PC2 or PC3 in 2010 were predicted (Table S3) and assigned to functional groups (Table S4) by gene ontogeny (GO) analysis, based on sequence similarities with characterized genes from other plant species.

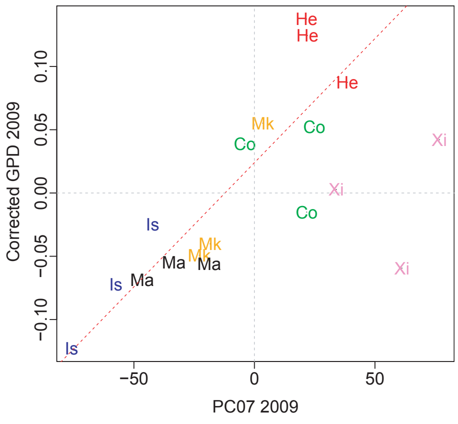

Figure 7 Plot of corrected GPD vs. PC7 for 2009 as means of the CVs and N level: N levels: 100 kg/ha (filled squares), 200 kg/ha (triangle) and 350 kg/ha (open squares). Cultivars are colour-coded: Co = Cordiale (green), He = Hereward (red), Is = Istabraq (blue), Ma = Malacca (black), Mk = Marksman (yellow) and Xi = Xi19 (purple).

To determine how the expression of selected genes was affected by genetic and environmental factors, ANOVA was applied for all samples where transcriptome data were available using the design factors as input and the expression profiles of the selected gene transcripts as output. Most transcripts displayed differences related to the genotype, with the year and N level also affecting a number of the transcripts. However, the relative importance of these factors differed. Four of the 136 gene transcripts identified as candidate genes positively related to GPD are displayed in Figure 8 for the five cultivars Cordiale, Hereward, Istabraq, Malacca and Marksman, for all 161 samples where transcriptome data were available. All four transcripts were significantly related to both the cultivar differences and to the year of growth, but the relative importance of these two factors differed. For transcripts Ta.8367.2.S1_a_at and Ta.14543.2.A1_at, the cultivar differences dominated with environmental factors having little impact, whereas the year of growth had a relatively larger impact on transcripts Ta.6968.1.S1_at and Ta.10471.1.S1_x_at. For Ta.8367.2.S1_a_at, the cultivar difference primarily resulted from lower expression of Istabraq vs. the remaining cultivars, whereas for Ta14543.2.A1_at showed more gradual variation in expression among the five cultivars. Similar expression profile for all 136 gene transcripts across all growth environments are shown in Figure S4.

Discussion

The identification of gene transcripts whose expression is significantly related to traits is a challenge in functional genomics, as the number of features can be high. For wheat, the Affymetrix arrays comprise approximately 60 000 features (gene probe sets

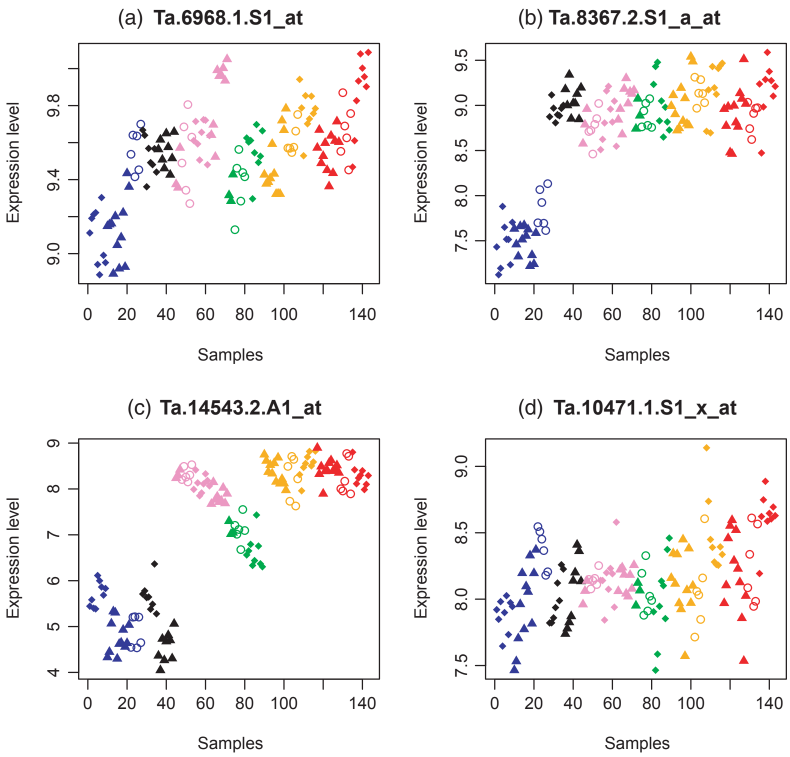

Figure 8 Examples of four genes (a-d) located in quadrant 4 of the PCA plot in Figure S3 and by PLS regression as being positively related to GPD. The colour codes indicate cultivars: Co=Cordiale (green), He=Hereward (red), Is=Istabraq (blue), Ma=Malacca (black), and Mk=Marksman (yellow), sorted by cultivar by their impact on GPD as found in regression analysis. The symbols indicate the year of growth: 2009 (filled squares): year 2010 (triangles) and 2011 (open circles), which to different degrees significantly affected the expression of these genes. N-level did not have significant impact on the expression of any of these four genes (not shown).

corresponding to slightly fewer unique transcripts), which may show separate or coordinated patterns of expression. The multivariate nature of functional genomics data must therefore be taken into account (Faergestad *et al.*, 2009). One family of multivariate methods is based on the projection of the original variables onto new variables defined as linear combinations of the original variables [i.e. PCA (Hotelling, 1933a,b) and PLS regression (Wold *et al.*, 1983)]. These methods present results as multivariate patterns, also called latent factors, which reflect the underlying phenomena that gives rise to the variation observed in the data. However, there is also a need for feature selection at the level of the observed variables (Lazar *et al.*, 2012; Saeys *et al.*, 2007). For the present study, yet another level of complexity had to be taken into account as we did not search for gene transcripts related to single phenotypic traits, but for those that resulted in the optimal deviation of two traits that are negatively related. Namely, genes that were associated with the highest protein content at a given yield (i.e. GPD). We therefore developed a novel methodology, to resolve the phenotypic characters and combine this with multivariate projection and feature extraction. After projection of the main information in the data onto latent factors (PCs), feature extraction was first performed at the level of the latent variation to identify genetically determined patterns of variation relating to the property of interest (here GPD alone). Feature extraction was then repeated within the selected latent factors to identify gene transcripts which consistently showed significant variation along the selected PCs. This is obtained by relating means of scores both to the corrected data for the phenotypic characteristics and to the gene transcripts used to generate the scores. Data from one site over two growth years was used to calculate GPD and identify genes related to the trait, and data for the second site and growth year were used in the validation of the consistency of the selected genes across different growth environments.

We suggest that this approach may have wider applicability in dissecting the transcriptional control of other complex traits. One of the strengths of the present approach is that we use methodologies well known to most biologists in the functional genomics area (PCA and ANOVA), combined in a novel way to solve complex problems.

Conclusions

We developed novel statistical approaches to identify transcripts whose expression in developing wheat caryopses is correlated with GPD at grain maturity. The transcripts that we identified probably represent both genes that control the trait and genes whose expression is affected as a consequence. In total, 136 gene transcripts were identified, and their behaviour was observed across different growth environments. Further work is required to identify the biological functions of these genes and identify those that can be exploited in for crop improvement. The availability of new rapid approaches for transcript analysis, such as next generation sequencing should facilitate the profiling of selected genes in large numbers of samples. It is necessary to identify and compare further material showing positive deviation from the negative relation between grain %N and yield, including larger collections of cultivars for association genetics or crosses between varieties differing in GPD for classical Mendelian analysis. These approaches should lead to the development of markers for breeders as well as providing information on the mechanisms controlling the trait. Although the cost of expression analysis has often limited the application of molecular approaches to crop improvement in the past, this is not likely to be the case in the future. Instead, the practical limitation is more likely to be the production of appropriate plant material and the availability of approaches to handle the complex data sets that are generated. As wheat yields can only be accurately determined in replicated field trials carried out in least three environments (years and/or sites), this not only requires appropriate facilities but also a long-term commitment.

Experimental procedures

Wheat material

Six UK cultivars (Istabraq, Hereward, Marksman, Condiale, Malacca and Xi 19) were grown over three seasons (2008–2009, 2009–2010 and 2010–2011) at Rothamsted Research and at four other sites in the south-east of the UK (RAGT, Ickleton, Cambridge; Limagrain, Woolpit, Suffolk; Syngenta, Whittlesford, Cambridge; KWS-UK, Thriplow, Hertfordshire) in 2009–2010 and 2010–2011 only. Three replicate plots were grown at three N levels: 100 kg/ha (N100), 200 kg/ha (N200) and 350 kg/ha (N350) (see Barraclough *et al.*, 2010; Chope *et al.*, 2014).

Developing heads (10 per plot) were tagged, and caryopses were harvested from the Rothamsted (2009, 2010 and 2011) and RAGT (2010 and 2011) sites at 21 days postanthesis (dpa), which represents the middle of grain filling when gene expression is at its highest (Wan *et al.*, 2008). Gene expression was measured using Affymetrix wheat microarrays giving a total of 161 samples.

Yield and grain protein determination

Trials were performed as previously described (Barraclough *et al.*, 2010; Chope *et al.*, 2014). Yields are standardized to 85% dry matter, after determining moisture content of individual samples. DM total nitrogen was determined in mature grain using the Dumas combustion method (Dumas, 1831), using a CNS (carbon, nitrogen, sulphur) Combustion Analyser (Leco Corp., St. Paul, MN). Nitrogen is presented as % of 100% dry matter content (Grain %N).

Affymetrix Genechip® hybridization

Microarrays were used to profile transcriptome. A time point of 21 dpa was chosen as a key developmental stage (mid-grain filling) in which grain storage proteins were being synthesized. Ten years per plot (three replicate plots per treatment/variety) were tagged at anthesis, and around 100 caryopses per sample were taken from the mid-third of the year and were harvested 21 days later. Gene expression was determined by profiling RNA extracted from this material against a gene chip representing 61 313 probe sets equating to 55 052 transcripts. Data from the profiling are semiquantitative, giving a good indication of the relative levels of expression of all RNAs in the sample simultaneously. Data were collected for 3 years at Rothamsted and for 2010 and 2011 at the RAGT site, for the three N levels in 2009 and 2010, and for the 200 kg N/ha treatment in 2011. One sample from 2009 was omitted (Malacca in 2009 grown at 2003, replicate three) as this sample was not analysed.

Microarray data are submitted to Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/).

Data analysis

Measurements of the deviation from the negative relation between grain %N and yield were obtained by first adjusting yield and protein content for the effect of N level (N) and the second order effect of N (N^2). The corrected value of protein (Grain%N_corrN) was further corrected for its negative relation to the corrected value of yield (Yield_corrN), giving corrected GPD as the residual. ANOVA was performed to investigate how the design factors: cultivars (CV), N level and the interaction between cultivar and N level, affected the phenotypic characters, and their corrected values using *P*-values were adjusted for multiple comparisons by false discovery rate using rotation test (Langsrud, 2005).

Principal component analysis was performed on the gene expression data. By PCA, the original data are decomposed into PCs where the PCs give multivariate patterns that describe in decreasing order the main variation in the data. All variables were centred and scaled to unit variance to allow gene transcript with small variation have the same impact on the model as gene transcript with large variation.

The mean values of each score were calculated across the three biological replicates to leave the variation in the phenotypic character across the biological replicates for validation. By this approach, we could identify multivariate pattern that might cause a positive deviation between grain %N and yield without negatively influencing the yield. The approach of using means of the scores for the regression is here called MSR.

To identify the individual gene transcripts that varied consistently along the relevant latent factors, a regression analysis was performed to relate means of the scores of the selected PCs to the gene expression data. This is an internal analysis performed within one block of data for validation of the consistency of each variable along relevant PCs. As above, we use MSR for this analysis, and we call this one-block MSR. The test was performed by rotation test (Langsrud, 2005) for correction of multiple comparisons by false discovery rate using rotation test.

To visualize positive vs. negative direction of the effects for the selected gene transcripts on GPD, and to further narrow down the number of candidate genes, PLSR was performed to relate the selected gene transcripts to GPD. The model was validated by Jackknife adapted to bilinear models (Martens and Martens, 2000). Gene transcripts significant for the selected PCs which showed a consistent pattern of variation related to the response for the different cross-validation segments and a positive relation to GPD were thereby selected.

For the significant genes positively related to GPD, a gene expression profile was made where the selected gene expression data are plotted for all the available data, also those not used for the selection of the gene transcripts. Whereas only growth year 2009 and 2010 at site Rothamsted were used to identify gene transcripts significantly related to GPD, all three growth years at both sites (in total 161 samples) were investigated for their behaviour over environmental conditions.

GO function analysis

GO functions for significantly over-expressed transcripts are found from the annotation offered by the B2G-FAR resource (Götz *et al.*, 2011). B2G-FAR GO annotation is a broad-specificity data set derived from homology and protein domain annotation; it therefore included some erroneous annotations for functions that are not found in plants. To address this issue, the original annotation set was filtered and only plant-relevant terms where retained. The filtering was carried out based on a high-quality reference set of all plant-relevant terms that was created by taking a nonredundant union of all terms and their ancestor terms from manually annotated rice and Arabidopsis GO annotation sets. In total, about 11% of nonplant annotations were removed.

Acknowledgements

Rothamsted Research is funded by the Biotechnology and Biological Sciences Research Council (BBSRC). The work reported here was supported by the BBSRC Industrial Partnership Award BB/G022437 with support from the Home Grown Cereals Authority grant RD-2007-3409 'Sustainability of UK-grown wheat for breadmaking'. EFM was supported by grants from the Fund for Research Levy on Agricultural Products in Norway. Field trials at Rothamsted Research were supported by the Defra-funded WGIN project. Trials at KWS, Limagrain, RAGT and Syngenta were supported by the respective breeding companies. We are grateful to the reviewers of this paper for their constructive comments.

References

Acreche, M.M. and Slafer, G.A. (2009) Variation of grain nitrogen content in relation with grain yield in old and modern Spanish wheats grown under a wide range of agronomic conditions in a Mediterranean region. *J. Agric. Sci.* **147**, 657–667.

Barneix, A.J. (2007) Physiology and biochemistry of source-regulated protein accumulation in the wheat grain. *J. Plant Physiol.* **164**, 581–590.

Barraclough, P.B., Howarth, J.R., Jones, J., Lopez-Bellido, R., Parmar, S., Shepherd, C.E. and Hawkesford, M.J. (2010) Nitrogen efficiency of wheat: genotypic and environmental variation and prospects for improvement. *Eur. J. Agron.* **33**, 1–11.

Bogard, M., Allard, V., Brancourt-Hulmel, M., Huemez, E., Machet, J.-M., Jeuffroy, M.-H., Gate, P., Martre, P. and 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.

Chope, G.A., Wan, Y., Penson, S.P., Bhandari, D.G., Powers, S.J., Shewry, P.R. and Hawkesford, M.J. (2014) Effects of genotype, season, and nitrogen nutrition on gene expression and protein accumulation in wheat grain. *J. Agric. Food Chem.* **62**, 4399–4407.

Dampney, P.M.R., Edwards, A. and Dyer, C.J. (2006) Managing nitrogen applications to new Group 1 and 2 wheat varieties. *HGCA Proj Rep* No. 400.

Dumas, J.B.A. (1831) Procedes de l'analyse organique. *Ann. Chim. Phys.* **2**, 198–213.

Faergestad, E.M., Langsrud, Ø., Høy, M., Hollung, K., Sæbø, S., Liland, K.H., Kohler, A., Gidskehaug, L., Almergren, J., Anderssen, E. and Martens, H. (2009) Analysis of megavariate data in functional genomics. In *Comprehensive Chemometrics*, vol. **4** (Brown, S., Tauler, R. and Walczak, R., eds), pp. 221–278. Oxford: Elsevier.

Götz, S., Arnold, R., Sebastián-León, P., Martín-Rodríguez, S., Tischler, P., Jehl, M.-A., Dopazo, J., Rattei, T. and Conesa, A. (2011) B2G-FAR, a species-centered GO annotation repository. *Bioinformatics*, **27**, 919.

Hawkesford, M.J. (2014) Reducing the reliance on nitrogen fertiliser for wheat production. *J. Cereal Sci.* **59**, 276–283.

Hotelling, H. (1933a) Analysis of a complex of statistical variables into principal components. *J. Edu. Psychol.* **24**, 417–441.

Hotelling, H. (1933b) Analysis of a complex of statistical variables into principal components. *J. Edu. Psychol.* **24**, 498–520.

Krapp, A., Saliba-Colombani, V. and Daniel-Vedele, F. (2005) Analysis of C and N metabolisms and of C/N interactions using quantitative genetics. *Photosynth. Res.* **83**, 251–263.

Lam, H.-M., Coschigano, K.T., Oliviera, I.C., Melo-Oliviera, R. and Coruzzi, G. (1996) The molecular-genetics of nitrogen assimiliation into amino acids in higher plants. *Annu. Rev. Plant Phys.* **47**, 569–593.

Langsrud, Ø. (2005) Rotation test. *Stat. Comput.* **15**, 53–60.

Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H. and Nowé, A. (2012) A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM. Trans. Comput. Biol. Bioinform.* **9**, 1106–1119.

Mackay, I., Horwell, A., Garner, J., White, J. and Philpott, H. (2011) Reanalyses of the historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *Theor. Appl. Genet.* **122**, 225–238.

Martens, H. and Martens, M. (2000) Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Qual. Prefer.* **11**, 5–16.

Monaghan, J.M., Snape, J.W., Chojecki, J.S. and 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.

Munier-Jolain, N.G. and Salon, C. (2005) Are the carbon costs of seed production related to the quantitative and qualitative performance? An appraisal of legumes and other crops. *Plant, Cell Environ.* **28**, 1388.

Oury, F.X. and 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.

Oury, F.X., Be'rard, P. and Brancourt-Hulmel, M. (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.

Saeys, Y., Inza, I. and Larrañaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Shewry, P.R. (2007) Improving the protein content and composition of cereal grain. *J. Cereal Sci.* **46**, 239–250.

Simmonds, N.W. (1995) The relation between yield and protein in cereal grain. *J. Sci. Food Agric.* **67**, 309.

Triboi, E. and Triboi-Blondel, A.-M. (2002) Productivity and grain or seed composition: a new approach to an old problem. *Eur. J. Agron.* **16**, 163–186.

Uauy, C., Distelfeld, A. and Fahima, T. (2006) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. *Science* **314**, 1298–1301.

Wan, Y., Poole, R.L., Huttly, A.K., Toscano-Underwood, C., Feeney, K., Welham, S., Gooding, M.J., Mills, E.N.C., Edwards, K.J., Shewry, P.R. and Mitchell, R.A.C. (2008) Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, **9**, 121.

Wold, S., Martens, H. and Wold, H. (1983) The multivariate calibration problem in chemistry solved by the PLS method. In: *Proc. Conf. Matrix Pencils. Lecture Notes in Mathematics*. (Ruhe, A. and Kågstrom, B., eds), pp. 286–293. Springer: Heidelberg.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Correction of yield and grain% N content for their relation to N fertilization (yield ~N + N^2) and (grain%N ~N + N^2) for wheat grown at Rothamsted in 2010.

Figure S2 PCA of the gene expression (a–b) 2009 and (c–d) 2010 (a and c) shows the scores of the samples (Is = Istabraq, He = Hereward, Mk = Marksman, Co = Cordiale, Ma = Malacca and Xi = Xi19).

Figure S3 PCA plot of genes transcript selected as being significantly related to GDP but not to grain yield, as they are selected as the significant genes for PC7 in 2009 and PC2 and PC3 in 2010 (see Table 3) (a) Score plot named by the cultivar (**b**) the same score plot named by the year of growth. A regression line is marked when omitting Xi19. The cultivar symbols are Is=Istabraq, He=Hereward, Mk=Marksman, Co=Cordiale Ma=Malacca Quadrant 1: cyan, 2: purple, 3: gray and 4: pink. The gene transcripts in quadrant 2 and 4 are marked according to the results of PLS regression with Jackknife for the genes in these quadrants; gene transcripts with filled symbols in quadrant 4 are significantly positively related to GPD, and gene transcripts in quadrant 2 with filled symbols are significantly negatively related to GPD.

Figure S4 All genes located in the fourth quadrant of the PCA plot in Figure 8 sorted by their cultivar according to the regression coefficient in the prediction of GPD as presented in the score plot in Figure 8.

Table S1 Weather data at Rothamsted for 2009, 2010 and 2011.

Table S2 Mean values of the six cultivars across all experiments in all years (in total 11 experiments over 3 years, 584 samples). The sites were Kw (KWS), Limagrain, Ra (RAGT), Ro (Rothamsted), Sy (Syngenta).

Table S3 Gene ontology enrichment analysis for the four loadings quadrants identified in the subset of 939 best candidate transcripts for determining GPD.

Table S4 ANOVA on the effect of the design variables on the GO terms with groups of genes significant for E-GPD from quadrant 4 in the PCA displayed in Figure 8.