Progress report on "Halm til det hele"
Identification of favorable allele(s)/haplotypes for grain protein content in Scandinavian winter wheat population

Deliverables:

List of gene markers:

- Excel-spreadsheet + a short description on background, methods and how to use the list.
- Info on alleles and allele-frequency should be included.

Information about interpretation of results:

The excel file "**selected_SNPs_allele_info.xlsx**" contains the information about the SNPs such as name, physical position, allele, gene_id and the effect on CPC identified from the genome-wide and candidate gene-based association study. For each SNP, three additional files are generated. **File1: heatmap.png:** Contains the results of SNPs/haplotype-based clustering. The number of clusters were determined based on visual inspection.

File2: clusters.txt: The information about clustering of samples determined based on File1.

File3: association.pdf: The boxplot showing the distribution of CPC and GPD among the clusters identified from File1.

Materials and Methods:

Genotypic data:

A total of 255 accessions were used in the current study. The SNP data for accessions genotyped using various marker platforms was obtained from Sejet (109 accessions) and NordicSeeds (146 accessions). As most of the genomic studies in wheat have been carried out using RefSeqV1.0 of Chinese spring genome assembly, the physical positions of SNPs were obtained by mapping the SNP sequences against RefSeqV1.0 of Chinese spring (**ref**). The genotypic data was combined based on the SNP marker information and hapmap file containing chromosome and physical position information was generated. The SNPs were filtered to remove SNPs with >20% missing data and the Principal Component Analysis (PCA) was carried out using SNPRelate package (**ref**) from R statistical environment. The genotypic data was also used to calculate identity-by-state (IBS) based kinship matrix.

Phenotypic data:

The phenotypic data for grain yield (GY) and crude protein content (CPC) recorded under National Trials (NT) and Official Variety Testing (OVT) from years 2000 till 2023 for various locations were downloaded separately from https://sortinfo.dk. The best linear unbiased predictions (BLUPs) were carried out to minimize location and environmental effect. Initially, the mean values of GY and CP were calculated if an accession is phenotyped under both NT and OVT trials for respective year and location. The BLUPs were estimated lmer function from lme4 package (**ref**) in R statistical environment using the formula:

Pheno ~ (1|genotype) + (1|location) + (1|year) + (1|sample: location)

The BULPs were estimated separately for GY and CP and used for genome-wide association study along with above mentioned genotypic data.

Genome-wide association Scan (GWAS):

GWA studies were conducted using genotypic and phenotypic dataset mentioned above using gemma (v0.98.3) software (**ref**). Only SNPs with less than 30% missing data and minor allele frequency (MAF) greater than 0.01 were used for the analysis. The significance of association was visualized by generating Manhattan plot of p-values against the physical position of SNP.

Candidate gene-based association study:

Several genes viz. Glutamine synthetase, nitrate transporter, etc controlling nitrogen uptake and protein biosynthesis have been identified in plants. Further, orthologs of those genes in wheat have also been identified. The SNP/haplotype-based clustering approach was used to study association of candidate genes with CPC. Briefly, SNP(s) with less than 20% missing data and within 500 Kb upstream and downstream region of a candidate gene were extracted and converted to 0,1 and 2 format using vcftools (**ref**). The clustering analysis was carried out using pheatmap package (**ref**) and heatmap was generated. The number of clusters/haplotypes were identified based on visual inspection of heatmap while distribution of GPC from each cluster was visualized to identify cluster with favorable allele/haplotype. The effect of allele/haplotype was estimated using 'lm' function in R statistical environment.

Depth of coverage analysis:

The peculiarity of bread wheat is that it can be crossed with ~ 300 different relatives or species. The region introgressed from diverged wild relatives often inherit as a big block because of suppressed recombination in the introgressed region. Further, the reads from introgressed region do not match to reference causing drop in read coverage. To identify the introgressed region, depth of coverage analysis was carried out as mentioned in **Schulthess et al. 2022.** Briefly, the short reads from each accession were mapped against the reference sequence of Chinse Spring (RefSeq V1.0) and BAM file was generated. The trimmed reads from Chinese Spring were also mapped against RefSeq v.1.0 and a BAM file was generated. From the BAM file for each accession, the number of reads in the 5Mb window were calculated. Reads from each window were first normalized to sequencing coverage and then to the number of reads from the same window for Chinese Spring. The $\log_2$ of normalized count was used to create a genome-wide coverage plot.

Results:

A total of 255 accessions were used in present study, out of which 41, 25 and 23 were genotyped using 7K, 15K and 20K SNP array, respectively while the remaining accessions were genotyped using 25K SNP array (**Table 1**). Out of 24,145 SNPs from 25K SNP array, 70% (16,845) SNPs mapped uniquely against RefSeqV1.0 of cv. chinse spring (**Table 2**). PCA using SNPs with less than 20% missing data showed scattered distribution of accessions from both Sejet and NordicSeeds indicating uniformity in SNP calling (**Figure 1**). The GY and CPC found to be distributed uniformly across years and locations, except for years 2013,2014 and 2015 and therefore BLUP values were estimated to minimize environmental effect (**Figure 2**).

GWAS using 11,219 SNPs and BLUPs of CPC identified significant MTA on long arm of chr2B (**Figure 3**) and SNP based cluster analysis showed the alternate allele “G” for SNP (BS00022717_51) at position 680573507 was found to be associated with high GPC (**Figure 4**) and explained 15% of variation. Depth of coverage analysis identified a drop in coverage for region from 650 to 750 Mb on chr2B (**Figure 5**). This region which is an introgression from spelt wheat has been reported to provide resistance against YR (**ref**). Most of the accessions carrying the intorgression showed reduction in CPC (**Figure 5)** indicating negative correlation of the introgressed region with CPC in winter wheat.

For candidate gene-based analysis, a total of 77 genes involved in nitrogen transport or metabolism were identified from literature survey and 22 genes with some effect on CPC were identified. Interestingly, a haplotype from chr6A carrying nitrate transporter gene “TraesCS6A02G032500” while a SNP from chr4B around flaking region of Glutamine synthetase gene “TraesCS4B02G047400” explained 11% and 7% variation in CPC in the current germplasm (**Table 3**).

Tables and Figures:

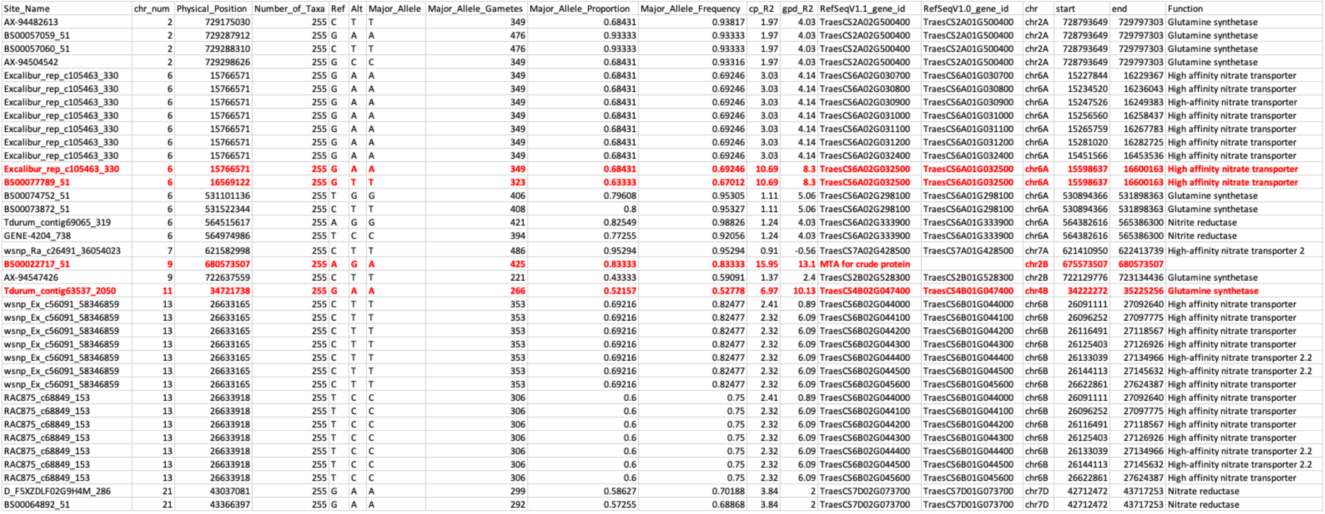
Table 1: Genotyping information about the accessions used in present study.

Array_Name	#SNPs	#Samples	Company
7K	6731	41	Sejet
15K	13006	25	Sejet
20K	17267	23	Sejet
25K	24145	20	Sejet
25K	24145	146	NordicSeeds
Total		**255**	

Table 2: Distribution of SNPs from 25K array on reference assembly of Chinese spring (RefSeqV1.0).

Chromosomes	A	B	D
chr1	954	1214	462
chr2	944	1256	456
chr3	965	1302	266
chr4	638	612	130
chr5	1169	1328	350
chr6	897	1027	305
chr7	1193	901	315
chrUn	161		
Total		**16,845**	

Table 3: List of significant SNP(s)/Haplotypes identified from candidate gene-based association analysis.



Site_Name	chr_num	Physical_Position	Number_of_Taxa	Ref	Alt	Major_Allele	Major_Allele_Gametes	Major_Allele_Proportion	Major_Allele_Frequency	cp_R2	gpd_R2	RefSeqV1.1_gene_id	RefSeqV1.0_gene_id	chr	start	end	Function
AX-94482613	2	729175030	255	C	T	T	349	0.68431	0.93817	1.97	4.03	TraesCS2A02G500400	TraesCS2A01G500400	chr2A	728793649	729797303	Glutamine synthetase
BS00057059_51	2	729287912	255	G	A	A	476	0.93333	0.93333	1.97	4.03	TraesCS2A02G500400	TraesCS2A01G500400	chr2A	728793649	729797303	Glutamine synthetase
BS00057060_51	2	729288310	255	C	T	T	476	0.93333	0.93333	1.97	4.03	TraesCS2A02G500400	TraesCS2A01G500400	chr2A	728793649	729797303	Glutamine synthetase
AX-94504542	2	729298626	255	G	C	C	349	0.68431	0.93316	1.97	4.03	TraesCS2A02G500400	TraesCS2A01G500400	chr2A	728793649	729797303	Glutamine synthetase
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G030700	TraesCS6A01G030700	chr6A	15227844	16229367	High affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G030800	TraesCS6A01G030800	chr6A	15234520	16236043	High affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G030900	TraesCS6A01G030900	chr6A	15247526	16249383	High-affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G031000	TraesCS6A01G031000	chr6A	15256560	16258437	High affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G031100	TraesCS6A01G031100	chr6A	15265759	16267783	High affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G031200	TraesCS6A01G031200	chr6A	15281020	16282725	High affinity nitrate transporter
Excalibur_rep_c105463_330	6	15766571	255	G	A	A	349	0.68431	0.69246	3.03	4.14	TraesCS6A02G032400	TraesCS6A01G032400	chr6A	15451566	16453536	High affinity nitrate transporter
Excalibur_rep_c105463_330	**6**	**15766571**	**255**	**G**	**A**	**A**	**349**	**0.68431**	**0.69246**	**10.69**	**8.3**	**TraesCS6A02G032500**	**TraesCS6A01G032500**	**chr6A**	**15598637**	**16600163**	**High affinity nitrate transporter**
BS00077789_51	**6**	**16569122**	**255**	**G**	**T**	**T**	**323**	**0.63333**	**0.67012**	**10.69**	**8.3**	**TraesCS6A02G032500**	**TraesCS6A01G032500**	**chr6A**	**15598637**	**16600163**	**High affinity nitrate transporter**
BS00074752_51	6	531101136	255	T	G	G	406	0.79608	0.95305	1.11	5.06	TraesCS6A02G298100	TraesCS6A01G298100	chr6A	530894366	531898363	Glutamine synthetase
BS00073872_51	6	531522344	255	C	T	T	408	0.8	0.95327	1.11	5.06	TraesCS6A02G298100	TraesCS6A01G298100	chr6A	530894366	531898363	Glutamine synthetase
Tdurum_contig69065_319	6	564515617	255	A	G	G	421	0.82549	0.98826	1.24	4.03	TraesCS6A02G333900	TraesCS6A01G333900	chr6A	564382616	565386300	Nitrite reductase
GENE-4204_738	6	564974986	255	T	C	C	394	0.77255	0.92056	1.24	4.03	TraesCS6A02G333900	TraesCS6A01G333900	chr6A	564382616	565386300	Nitrite reductase
wsnp_Ra_c26491_36054023	7	621582998	255	C	T	T	486	0.95294	0.95294	0.91	-0.56	TraesCS7A02G428500	TraesCS7A01G428500	chr7A	621410950	622413739	High-affinity nitrate transporter 2
BS00022717_51	**9**	**680573507**	**255**	**A**	**G**	**A**	**425**	**0.83333**	**0.83333**	**15.95**	**13.1**	**MTA for crude protein**		**chr2B**	**675573507**	**680573507**	
AX-94547426	9	722637559	255	C	T	T	221	0.43333	0.59091	1.37	2.4	TraesCS2B02G528300	TraesCS2B01G528300	chr2B	722129776	723134436	Glutamine synthetase
Tdurum_contig63537_2050	**11**	**34721738**	**255**	**G**	**A**	**A**	**266**	**0.52157**	**0.52778**	**6.97**	**10.13**	**TraesCS4B02G047400**	**TraesCS4B01G047400**	**chr4B**	**34222272**	**35225256**	**Glutamine synthetase**
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.41	0.89	TraesCS6B02G044000	TraesCS6B01G044000	chr6B	26091111	27092640	High affinity nitrate transporter
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G044100	TraesCS6B01G044100	chr6B	26096252	27097775	High affinity nitrate transporter
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G044200	TraesCS6B01G044200	chr6B	26116491	27118567	High affinity nitrate transporter
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G044300	TraesCS6B01G044300	chr6B	26125403	27126926	High affinity nitrate transporter
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G044400	TraesCS6B01G044400	chr6B	26133039	27134966	High-affinity nitrate transporter 2.2
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G044500	TraesCS6B01G044500	chr6B	26144113	27145632	High-affinity nitrate transporter 2.2
wsnp_Ex_c56091_58346859	13	26633165	255	C	T	T	353	0.69216	0.82477	2.32	6.09	TraesCS6B02G045600	TraesCS6B01G045600	chr6B	26622861	27624387	High affinity nitrate transporter
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.41	0.89	TraesCS6B02G044000	TraesCS6B01G044000	chr6B	26091111	27092640	High affinity nitrate transporter
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G044100	TraesCS6B01G044100	chr6B	26096252	27097775	High affinity nitrate transporter
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G044200	TraesCS6B01G044200	chr6B	26116491	27118567	High affinity nitrate transporter
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G044300	TraesCS6B01G044300	chr6B	26125403	27126926	High affinity nitrate transporter
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G044400	TraesCS6B01G044400	chr6B	26133039	27134966	High-affinity nitrate transporter 2.2
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G044500	TraesCS6B01G044500	chr6B	26144113	27145632	High-affinity nitrate transporter 2.2
RAC875_c68849_153	13	26633918	255	T	C	C	306	0.6	0.75	2.32	6.09	TraesCS6B02G045600	TraesCS6B01G045600	chr6B	26622861	27624387	High affinity nitrate transporter
D_F5XZDLF02G9H4M_286	21	43037081	255	G	A	A	299	0.58627	0.70188	3.84	2	TraesCS7D02G073700	TraesCS7D01G073700	chr7D	42712472	43717253	Nitrate reductase
BS00064892_51	21	43366397	255	G	A	A	292	0.57255	0.68868	3.84	2	TraesCS7D02G073700	TraesCS7D01G073700	chr7D	42712472	43717253	Nitrate reductase

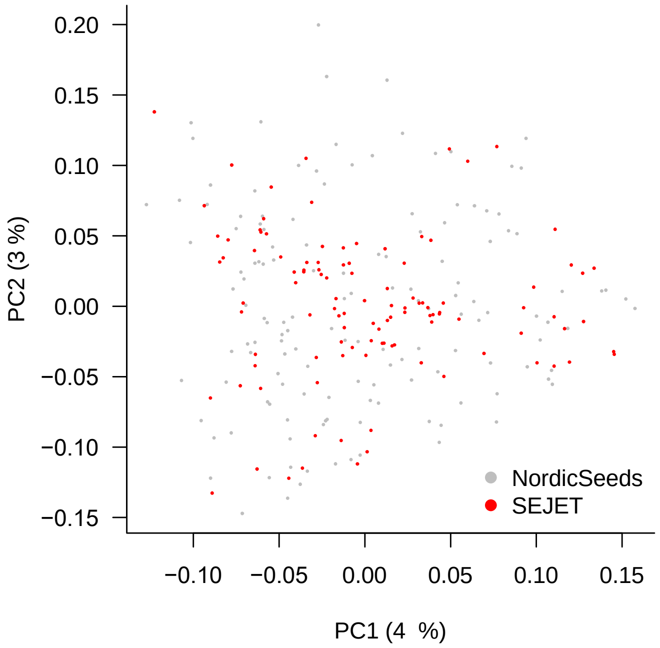


Figure 1: Principal component analysis (PCA) plot showing distribution of 255 accessions from Sejet and NordicSeeds used in current study.

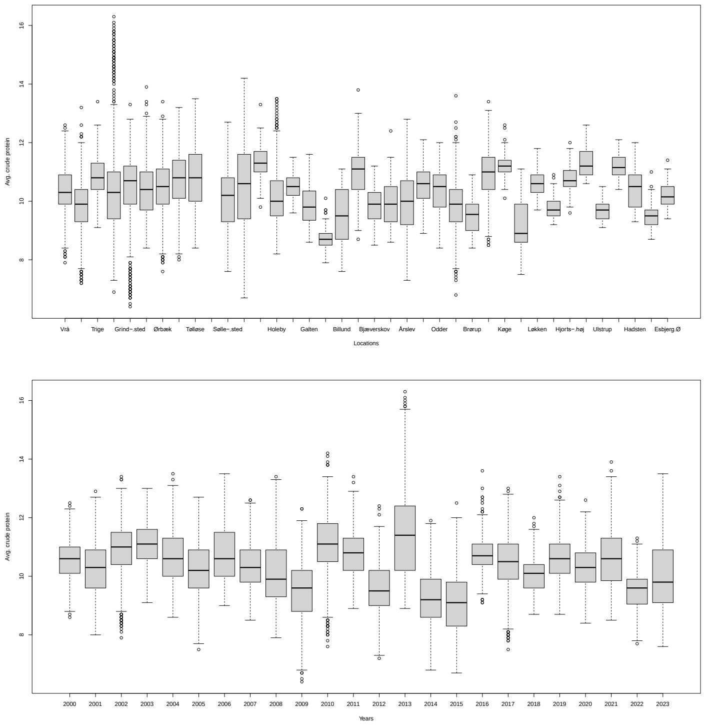


Figure 2: Distribution of CPC of accessions used in present study across multiple years and locations. The CPC values of accessions in years 2013 and 2014 were below the average CPC over multiple years.

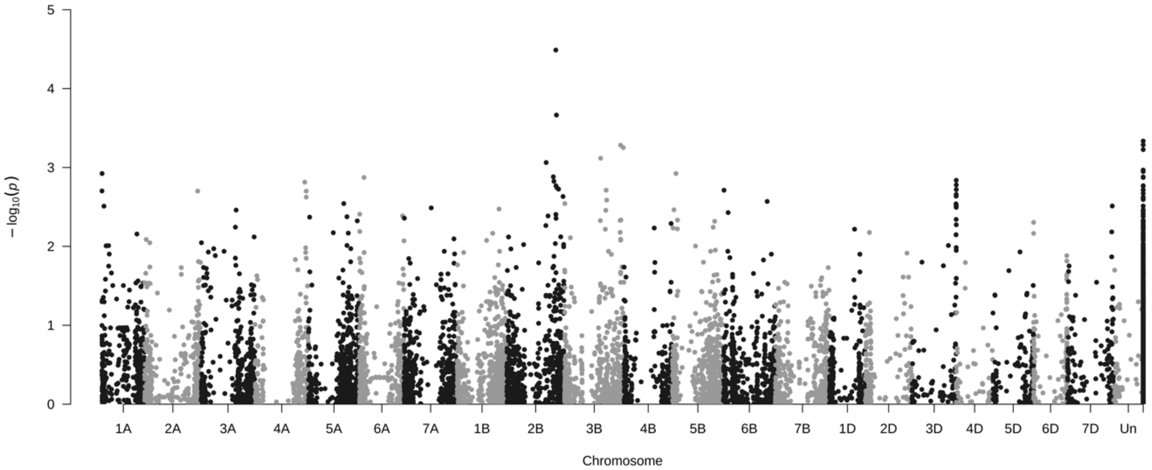


Figure 3: Manhattan plot showing distribution of p-values for CPC against the positions of SNPs on the reference assembly of Chinese spring (RefSeqV1.0). Significant p-values were identified for SNPs on the long arm of chr2B indicating association of the region with CPC in the current population.

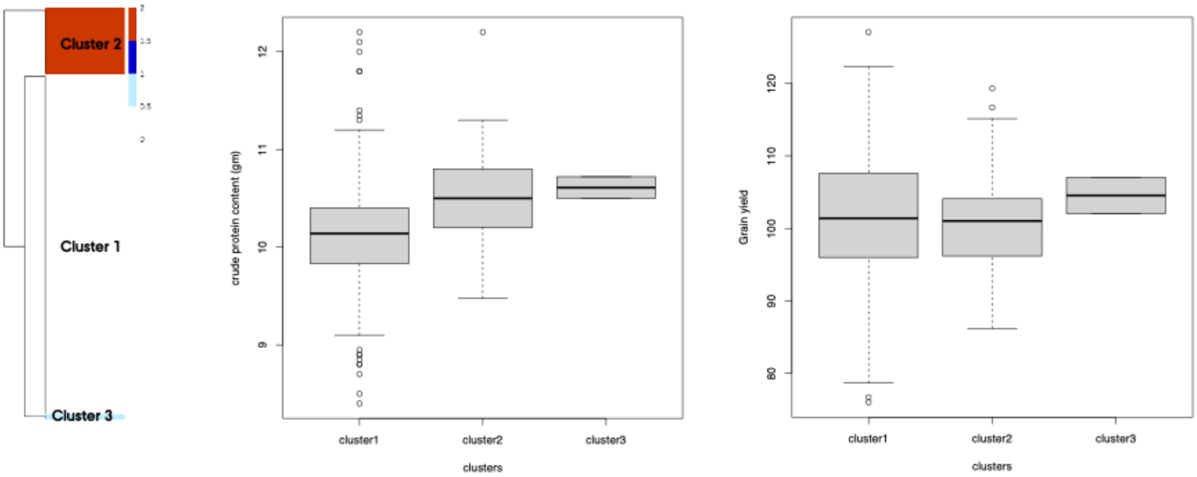
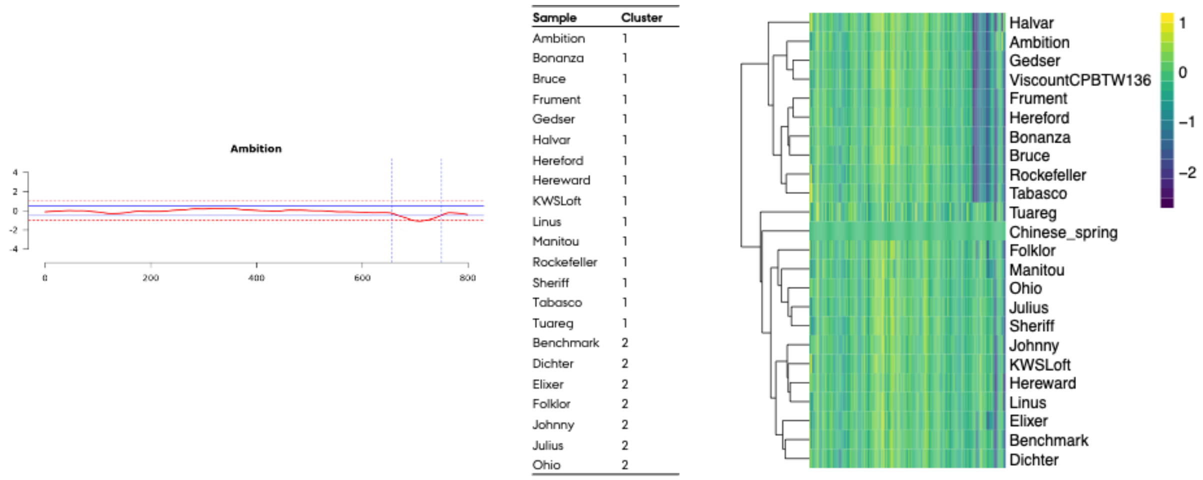


Figure 4: Clustering of accessions using the genotypic information of SNP from chr2B associated with CPC. Accessions from cluster 1 had low CPC than the accessions from cluster 2 and cluster 3. However, no strong association with GY was observed.



Sample	Cluster
Ambition	1
Bonanza	1
Bruce	1
Frument	1
Gedser	1
Halvar	1
Hereford	1
Hereward	1
KWSLoft	1
Linus	1
Manitou	1
Rockefeller	1
Sheriff	1
Tabasco	1
Tuareg	1
Benchmark	2
Dichter	2
Elixer	2
Folklor	2
Johnny	2
Julius	2
Ohio	2

Figure 5: Depth of coverage analysis using GBS data generated under Genebank2.0 project (**ref**) identified drop in coverage in cv. Ambition in the region spanning from 650 Mb to 750 Mb on chr2B in reference assembly of Chinese spring (RefSeqv1.0). The region was present in Ambition and other accessions from cluster 1 while it was absent in accessions from cluster 2 and 3 (**Figure 4**).