A Method for Identifying Environmental Stimuli and Genes Responsible for Genotype-by-Environment Interactions From a Large-Scale Multi-Environment Data Set

Akio Onogi[1]*[†], Daisuke Sekine[2,3], Akito Kaga[3], Satoshi Nakano[4], Tetsuya Yamada[3], Jianming Yu[5] and Seishi Ninomiya[6]

[1]Department of Plant Life Science, Faculty of Agriculture, Ryukoku University, Otsu, Japan, [2]Institute of Vegetable and Floriculture Science, National Agriculture and Food Research Organization, Tsukuba, Japan, [3]Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Japan, [4]Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Tsukuba, Japan, [5]Department of Agronomy, Iowa State University, Ames, IA, United States, [6]Graduate School of Agricultural and Life Science, The University of Tokyo, Nishitokyo, Japan

It has not been fully understood in real fields what environment stimuli cause the genotype-by-environment (G × E) interactions, when they occur, and what genes react to them. Large-scale multi-environment data sets are attractive data sources for these purposes because they potentially experienced various environmental conditions. Here we developed a data-driven approach termed Environmental Covariate Search Affecting Genetic Correlations (ECGC) to identify environmental stimuli and genes responsible for the G × E interactions from large-scale multi-environment data sets. ECGC was applied to a soybean (*Glycine max*) data set that consisted of 25,158 records collected at 52 environments. ECGC illustrated what meteorological factors shaped the G × E interactions in six traits including yield, flowering time, and protein content and when these factors were involved in the interactions. For example, it illustrated the relevance of precipitation around sowing dates and hours of sunshine just before maturity to the interactions observed for yield. Moreover, genome-wide association mapping on the sensitivities to the identified stimuli discovered candidate and known genes responsible for the G × E interactions. Our results demonstrate the capability of data-driven approaches to bring novel insights on the G × E interactions observed in fields.

Keywords: genotype-by-environment interactions, genetic correlation, genome-wide association, historical data, multi-environmental trial, environmental covariate

INTRODUCTION

Genotype-by-environment (G × E) interactions have been one of main interests in plant research for decades (Mather and Jones, 1958; van Eeuwijk et al., 2005; Des Marais et al., 2013). However, understanding the interactions in fields is not an easy task because a number of players including environmental stimuli and genes can be involved in at various growth stages. Thus, it has been a great

challenge to depict comprehensive landscapes on how genes and environment stimuli cause the G × E interactions together in fields. So far, studies have successfully used environmental stimuli in statistical models to map quantitative trait loci (QTLs) and/or predict crop phenotypes (Malosetti et al., 2013; Jarquin et al., 2014; Li et al., 2018; Millet et al., 2019; Guo et al., 2020). These studies show the usefulness of a reaction norm approach where phenotypes/genotypic values are regressed on quantitative indices of environment stimuli to model the sensitivity of genotypes (or phenotype plasticity). An important point of this approach is, however, that the sensitivity of genotypes (or phenotype plasticity) is not necessary associated with the observed G × E interactions (see Model description in Materials and Methods and Interpretation of ECGC in Results and Discussion). Methods to identify environmental stimuli and genes directly related to the G × E interactions have been lacked.

Here, we propose a novel method termed Environmental Covariate Search Affecting Genetic Correlations (ECGC) to reveal quantitative environmental stimuli (referred to as environmental covariates) and genetic architecture underpinning the G × E interactions using large-scale multi-environment data sets. ECGC searches environmental covariates whose similarity matrices between environments are significantly correlated with genetic correlation matrices between environments which can be regarded as indicators of the G × E interactions (Hayes et al., 2016). This proposed method is able to identify the environmental stimuli and genes directly related to the observed G × E interactions. Moreover, because genetic correlations between environments can be estimated using mixed models, ECGC is robust to missing records and unbalanced data structure that often characterize multi-environment data. Here we applied the proposed method to a large-scale multi-environment data of soybean (*Glycine max*). The traits were days to flowering (DTF), days to maturity (DTM), stem length (SL, cm), protein content of seeds (PR, %), yield (YI, kg/a) and seed weight (SW, g/100 seeds).

MATERIALS AND METHODS

Model Description
As described above, ECGC searches environmental covariates whose similarity matrices between environments are correlated with genetic correlation matrices between environments. Here it will be illustrated how the similarity of environmental covariates is associated with genetic correlation between environments.

For the jth environment, the phenotype adjusted for variations due to years and management conditions, $\tilde{y}_j$, is decomposed as

$$\tilde{y}_j = \mathbf{u}_j + \mathbf{e}_j$$

where $\mathbf{u}_j$ and $\mathbf{e}_j$ are the additive genetic effect and residual, respectively. Note that $\tilde{y}_j$ and $\mathbf{u}_j$ include all adjusted phenotypes and additive genetic effects of genotypes evaluated at the environment. $\mathbf{u}_j$ and $\mathbf{e}_j$ are assumed to follow multivariate normal distributions (MVN) as $\mathbf{u}_j \sim \mathrm{MVN}(0, \mathbf{G}\sigma_{u_j}^2)$ and

$\mathbf{e}_j \sim \mathrm{MVN}(0, \mathbf{I}\sigma_{e_j}^2)$, respectively, where $\mathbf{G}$ is the genomic relationship matrix and $\mathbf{I}$ is the identity matrix. After scaling with the genetic standard deviation (σ_{u_j}), the additive genetic effect is assumed to be further decomposed into two components, one affected by an environmental covariate, x_j, and the other is free of it, as

$$\mathbf{u}_j = \sigma_{u_j}\big(\boldsymbol{\beta}x_j + \tilde{\mathbf{u}}_j\big)$$

where $\boldsymbol{\beta}$ is the slopes of varieties that represent the sensitivity to the environmental covariate, and $\tilde{\mathbf{u}}_j$ is the residual genetic effects. Here $\boldsymbol{\beta}$ is assumed to be $\boldsymbol{\beta} \sim \mathrm{MVN}(0, \mathbf{G}\sigma_{\beta}^2)$ which means $\tilde{u}_j \sim \mathrm{MVN}[0, \mathbf{G}\sigma_{\tilde{u}_j}^2]$ where $\sigma_{\tilde{u}_j}^2 = \sigma_{u_j}^2 - x_j^2\sigma_{\beta}^2$. Then the genetic covariance of the additive genetic effects between environments j and k can be represented as

$$
\begin{aligned}
\mathrm{cov}(\mathbf{u}_j, \mathbf{u}_k) &= \mathrm{cov}\big[\sigma_{u_j}\big(\boldsymbol{\beta}x_j + \tilde{\mathbf{u}}_j\big), \sigma_{u_k}\big(\boldsymbol{\beta}x_k + \tilde{\mathbf{u}}_k\big)\big] \\
&= \sigma_{u_j}\sigma_{u_j}\big[x_j x_k \mathbf{G}\sigma_{\beta}^2 + x_j\mathrm{cov}(\boldsymbol{\beta}, \tilde{\mathbf{u}}_k) \\
&\quad + x_k\mathrm{cov}(\tilde{\mathbf{u}}_j, \boldsymbol{\beta}) + \mathrm{cov}(\tilde{\mathbf{u}}_j, \tilde{\mathbf{u}}_k)\big] \\
&= \mathbf{G}\sigma_{u_j}\sigma_{u_j}\big[x_j x_k \sigma_{\beta}^2 + x_j\sigma_{k,\beta} + x_k\sigma_{j,\beta} + \sigma_{j,k}\big]
\end{aligned}
$$

by letting $\mathrm{cov}(\boldsymbol{\beta}, \tilde{\mathbf{u}}_k) = \mathbf{G}\sigma_{k,\beta}$ and $\mathrm{cov}(\tilde{\mathbf{u}}_j, \tilde{\mathbf{u}}_k) = \mathbf{G}\sigma_{j,k}$. Thus, the genetic covariance between environments j and k, $\sigma_{u_{jk}}$, is

$$\sigma_{u_{jk}} = \sigma_{u_j}\sigma_{u_j}\big[x_j x_k \sigma_{\beta}^2 + x_j\sigma_{k,\beta} + x_k\sigma_{j,\beta} + \sigma_{j,k}\big]$$

and the genetic correlation between these environments, $\mathrm{P}_{j,k}$, is

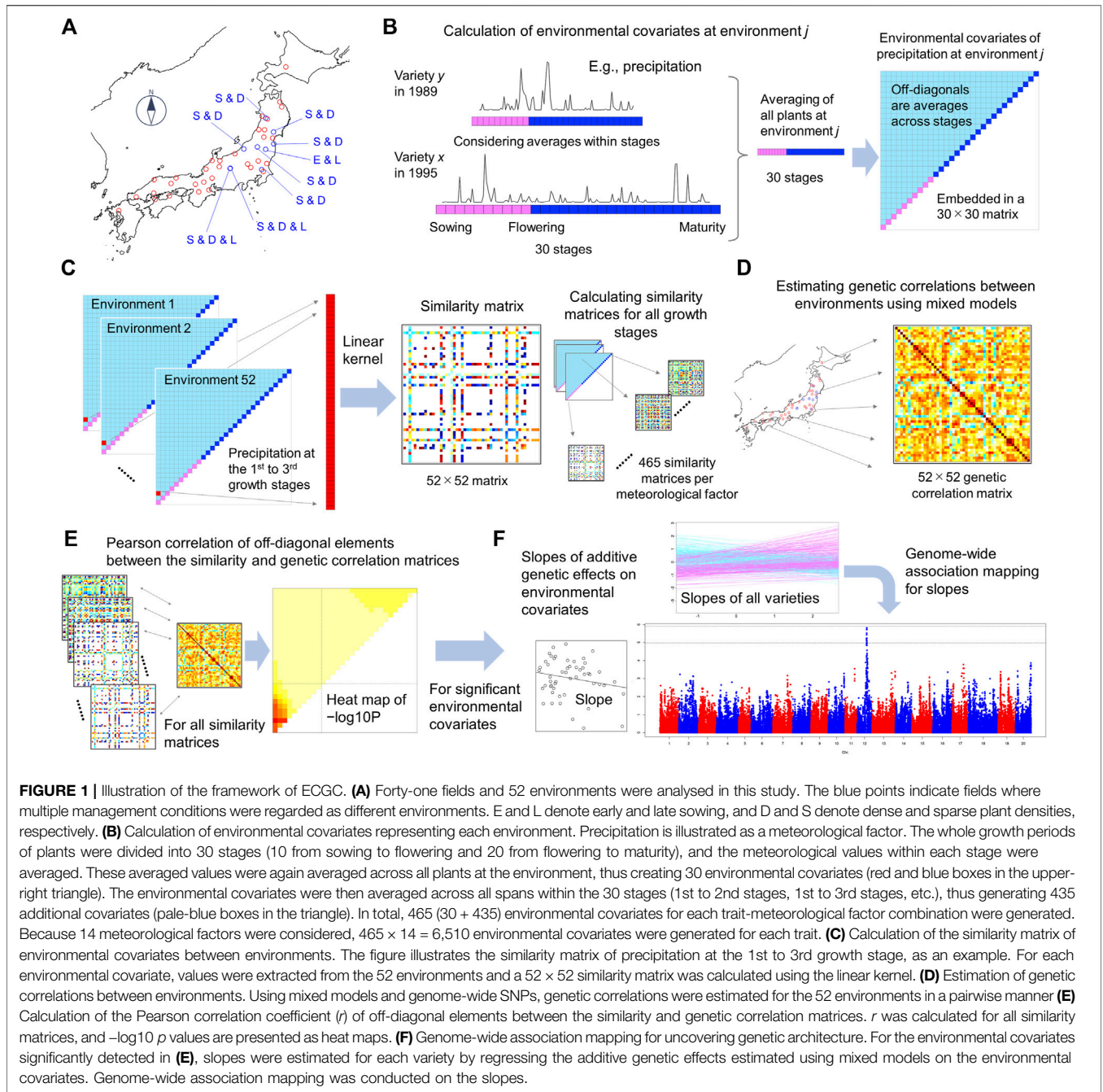$$P_{j,k} = x_j x_k \sigma_{\beta}^2 + x_j\sigma_{k,\beta} + x_k\sigma_{j,\beta} + \sigma_{j,k} \tag{1}$$

Equation 1 illustrates how the similarity of environmental covariates ($x_j x_k$) is related to genetic correlation between environments ($P_{j,k}$). Genetic correlations between environments poses information of relative magnitudes of genotypic values among varieties that change across environments, i.e., G × E interactions. Thus, ECGC searches the environment covariates that are associated with the G × E interactions by scanning $\mathrm{cor}_{j,k}(P_{j,k}, x_j x_k)$ where $\mathrm{cor}_{j,k}$ means taking correlations across all combinations of environments. The squared $\mathrm{cor}_{j,k}(P_{j,k}, x_j x_k)$ (i.e., r^2) represents the proportion of the variance of $x_j x_k \sigma_{\beta}^2$ in the variance of $P_{j,k}$. The significantly detected environmental covariates will be able to be used for various purposes. Here we focus on revealing genes underlying the G × E interactions. To this end, GWA mapping was conducted on the slopes ($\boldsymbol{\beta}$) associated with the detected environmental covariates as the final step of ECGC.

Data Analysis Overview
To apply ECGC to real data, five steps are required.

(1). Calculation of environmental covariates (x_j).
(2). Calculation of similarities of environmental covariates between environments ($x_j x_k$).
(3). Estimation of genetic correlation between environments ($P_{j,k}$).
(4). Scanning environmental covariates associated with genetic correlation based on $\mathrm{cor}_{j,k}(P_{j,k}, x_j x_k)$.

FIGURE 1 | Illustration of the framework of ECGC. **(A)** Forty-one fields and 52 environments were analysed in this study. The blue points indicate fields where multiple management conditions were regarded as different environments. E and L denote early and late sowing, and D and S denote dense and sparse plant densities, respectively. **(B)** Calculation of environmental covariates representing each environment. Precipitation is illustrated as a meteorological factor. The whole growth periods of plants were divided into 30 stages (10 from sowing to flowering and 20 from flowering to maturity), and the meteorological values within each stage were averaged. These averaged values were again averaged across all plants at the environment, thus creating 30 environmental covariates (red and blue boxes in the upper-right triangle). The environmental covariates were then averaged across all spans within the 30 stages (1st to 2nd stages, 1st to 3rd stages, etc.), thus generating 435 additional covariates (pale-blue boxes in the triangle). In total, 465 (30 + 435) environmental covariates for each trait-meteorological factor combination were generated. Because 14 meteorological factors were considered, $465 \times 14 = 6,510$ environmental covariates were generated for each trait. **(C)** Calculation of the similarity matrix of environmental covariates between environments. The figure illustrates the similarity matrix of precipitation at the 1st to 3rd growth stage, as an example. For each environmental covariate, values were extracted from the 52 environments and a 52×52 similarity matrix was calculated using the linear kernel. **(D)** Estimation of genetic correlations between environments. Using mixed models and genome-wide SNPs, genetic correlations were estimated for the 52 environments in a pairwise manner **(E)** Calculation of the Pearson correlation coefficient (r) of off-diagonal elements between the similarity and genetic correlation matrices. r was calculated for all similarity matrices, and $-\log 10$ p values are presented as heat maps. **(F)** Genome-wide association mapping for uncovering genetic architecture. For the environmental covariates significantly detected in **(E)**, slopes were estimated for each variety by regressing the additive genetic effects estimated using mixed models on the environmental covariates. Genome-wide association mapping was conducted on the slopes.

(5). Estimation of slopes (**β**) which representing the sensitivity of varieties to the environmental covariates detected in Step 4 and conduct GWA mapping on the slopes.

These steps correspond with **Figures 1B–F**. In the following sections whose titles start as "ECGC step", we explain how these steps were conducted in our analyses using a large-scale multi-environment data of soybean. Preceding to these steps, we also conducted model development and fitting for each environment-trait combination to calculate the phenotypes adjusted with fixed effects ($\tilde{y}_j$). This step may not be always necessary for ECGC, but required for our data. This step is referred to as Step 0.

Multi-Environment Data of Soybean

The data set used in this study consisted of 25,158 records of 624 varieties evaluated at 41 fields (**Supplementary Data S1**). This data set was extracted as described below from a historical data of soybean including 72,829 records of 6,106 varieties evaluated at 440 fields over 55 years in Japan (from 1961 to 2015). The overview of this historical data is provided in **Supplementary Methods S1**. In nine fields, multiple management conditions

(e.g., early or late sowing dates) were consistently conducted in multiple years. For these fields, different management conditions were defined as different environments as described later. As a result, a total of 52 environments were defined from 41 fields (**Supplementary Table S2**, **Figure 1A**). Thus, in this study, "environments" denote the combinations of fields (locations) and management conditions.

DNA Extraction and SNP Genotyping

Seeds of varieties included in the historical data were collected from the breeding centres and NARO genebank (as many seeds as possible). The seeds were sown in pots and grown until the first trifoliate leaves emerged in greenhouses. Total genomic DNA was extracted from the first trifoliate leaves of one plant using a method based on guanidine hydrochloride and proteinase K (Khosla et al., 1999), with modifications. Among the ~2,000 varieties with extracted DNA, 573 varieties were genotyped using the Axiom SNP custom array (ThermoFisher, MA, United States), which was designed for Japanese soybean varieties. These 573 varieties consisted mainly of breeding lines that had been evaluated until later generations (typically F_8 or later) as promising lines. In addition, from the ~2,000 varieties, 149 varieties that were registered as major varieties by the Ministry of Agriculture, Forestry and Fisheries and 187 varieties included in the soybean core collection (Kaga et al., 2012) were also genotyped using the array. The 573 varieties were selected to avoid overlap with the 336 (149 + 187) varieties. Among the SNPs genotyped, SNPs that showed high genotyping quality (i.e., SNPs termed "PolyHighResolution" in the Axiom genotyping system) and could be mapped to the Williams 82 genome assembly version 2.0 (Wm82.a2.v1/Glyma 2.0) were extracted (138,555 SNPs) (**Supplementary Table S1**). The average call rate of SNPs was 0.998 (SD ± 0.003). Missing genotypes were imputed using Beagle 4.1 (08Jun17.d8b) (Browning and Browning, 2016).

Extraction of Phenotypic Records

The varieties that had SNP genotypes from the Axiom custom array were subjected to subsequent analyses. Because the names of varieties usually change with the advancement of generations, the variety names were integrated using the names at the last generations. Among the 440 fields included in the historical data, 41 fields were selected for analyses because of the number of records and balanced geological locations across Japan (**Figure 1A**). For each field, phenotypic values that were out of the mean ± 3SD (standard deviation) range were treated as missing values.

The following six traits were selected for analyses: DTF (days), DTM (days), SL (cm), PR (%), YI (kg/a), and SW (g/100 seeds). These traits were selected because of their importance for breeding and the number of phenotyped records. In Japanese soybean breeding, to unify the measuring methods of traits, regulations on trait measurement were established in 1954 by a committee where the Ministry of Agriculture and Forestry at the time took the leading role. Although these regulations seem to have undergone several minor updates to date, trait definitions were largely common across the study period. Briefly, DTF was

defined as the date when 40–50% of the buds of the strains reached flowering. DTM was the date when 80–90% of the pods of the strains showed variety-unique colours at maturity. SL was the length of the main culm between the ground surface and the growth point. PR was mainly measured using near-infrared spectroscopy, but there seemed to be some variations in the measuring methods and used machines. YI and SW were defined as the weight with 15% moisture, although there also seemed to be some variations in the moisture percentage among stations and years.

Records that had at least one observation for these six traits were extracted for subsequent analyses. As a result, the extracted records consisted of 25,158 records of 624 varieties evaluated at 41 fields (**Supplementary Data S1**). The summaries of the extracted phenotypic data and used varieties are presented in **Supplementary Tables S2, S3**, respectively.

Meteorological Factors

Meteorological information was taken from MeteoCropDB ver. 1.0 (Kuwagata et al., 2011). The database was designed to help the utilisation of crop growth models and provides daily values at each weather station of the Japan Meteorological Agency (JMA). Meteorological information for each environment was taken from the nearest JMA station. In this study, 13 meteorological factors, i.e., mean temperature (℃), maximum temperature (℃), minimum temperature (℃), precipitation (mm), vapour pressure (hPa), vapour pressure deficit (hPa), relative humidity (%), minimum relative humidity (%), wind speed (m/s), maximum wind speed (m/s), hours of sunshine (h), solar radiation (W/m^2) and potential evapotranspiration (mm), were used. Among these factors, wind speed and maximum wind speed were adjusted as observed at 2.5 m of altitude by the developer of the database from the original values of the JMA. Solar radiation was estimated by the database developer from the observations of the JMA regarding hours of sunshine. Potential evapotranspiration was also estimated by the developer based on mean temperature, vapour pressure, wind speed, solar radiation and air pressure. The values of the other factors were according to the observations of the JMA. In addition to these factors, photoperiod (h) was calculated for each environment. For simplicity, daily length was also considered as a meteorological factor. The values of these meteorological factors are included in **Supplementary Data S1**, together with the phenotype records.

ECGC Step 0: Developing Mixed Models at Each Environment

Single-trait mixed models were fitted to the records at each environment. This process had two purposes; the first was to eliminate variations attributable to years and management conditions from phenotypic values, that is, to create $\tilde{y}_j$. The adjusted phenotypes $\tilde{y}_j$ were then used for estimating genetic correlations between environments (ECGC Step 3). The second was to estimate the additive genetic effects of varieties at each environment. The additive genetic effects were then used to estimate the slopes of genetic effects on environmental covariates (ECGC Step 5).

The year effects were added as fixed effects in the mixed models. The conditions of three management methods (plant density, fertiliser and sowing date) were modelled in various ways. In nine fields (F04, F07, F08, F09, F22, F24, F25, F27, and F34), multiple management conditions were consistently conducted in multiple years. For these fields, different conditions were defined as different environments resulting in a total of 52 environments (**Supplementary Table S2**, **Figure 1A**).

In most fields, however, the management conditions often varied across years. When the effects of management conditions were indistinguishable from those of year effects, the effects were absorbed to the year effects and not modelled. Otherwise, the effects were added in the mixed models. The modelling schemes varied according to the environments and management methods. Although the conditions of plant density, fertilizer, and sowing date are intrinsically continuous variables, when the number of conditions at an environment was few, these conditions were regarded as categorical variables and the effects were modelled as fixed effects. Otherwise, the conditions were regarded as continuous and the effects were model using basis functions (e.g., B-splines, Hastie et al., 2009). In each case, the interactions between genotypes and management were also considered. Variance components were estimated using REML implemented in airemlf90 (Misztal et al., 2002), and model selection was conducted using the AICs provided by the programme. The covariance structure of additive genetic effects among varieties (i.e., genomic relationship matrix) was defined using the genome-wide SNPs (VanRaden, 2008). The A.mat function provided by the R package, rrBLUP (ver. 4.6.1) (Endelman, 2011; R Core Team, 2020), was used for calculation (**Supplementary Data S2**). Thus, the mixed models applied here were the so-called genomic or genome-enabled BLUP (GBLUP) (de los Campos et al., 2013). The details of this procedure are described for each environment in the **Supplementary Methods S2**. Narrow-sense SNP heritability estimated by the selected models are presented in **Supplementary Table S4**.

ECGC Step 1: Calculation of Environmental Covariates

Environmental covariates were calculated for each environment, i.e., combination of fields and management conditions. Because trials had been conducted for multiple years at each field (**Supplementary Table S1**), in principle, environmental covariates were calculated by averaging meteorological values across years. The details are explained below.

Even on the same calendar day, the growth stages of plants can vary depending on the sowing date and growth speed, which depend on the environmental conditions (e.g., temperature) and genotypes. Thus, a meteorological event (e.g., high or low temperature) on a calendar day can have different effects on plants with different growth stages. To consider the difference in growth stage, the growth period between the sowing dates and days of flowering was divided into 10 equal-sized stages (on

average 5.5 ± 1.2 days per stage), and the period between the days of flowering and days of maturity was divided into 20 equal-sized stages (3.8 ± 0.5 days per stage). For each variety at each environment and year, the daily meteorological values within stages were averaged (**Figure 1B**).

Meteorological values representing each environment were then obtained by averaging the above-mentioned meteorological values of all plants included in the environment, yielding 30 values for the 30 growth stages for each environment (**Figure 1B**, **Supplementary Tables S5–S10**). In addition, meteorological values were averaged across stages, e.g., averages across the 1st to 2nd stages, across the 1st to 3rd stages, etc. This procedure was inspired by the joint genomic regression analysis (Li et al., 2018). As a result, 435 (30 × 29/2) additional representative values were generated, resulting in 465 values for each environment (**Figure 1B**). That is, 465 × 14 (meteorological factors) = 6,510 environmental covariates were generated.

Three issues were notable in this procedure. First, the number of stages will depend on the prior knowledge or assumptions on growth stages. For soybean, typically five and eight growth stages are assumed for the vegetative and reproductive phases (Fehr and Caviness, 1977). The number of stages used here (30) were determined to be able to discriminate these known growth stages. If a crop experiences more growth stages, setting a larger number will facilitate interpretation. Second, although the growth stages of plants can be represented using cumulative temperature and/or photoperiod, we avoided using these indices, to simplify the procedures as much as possible. Lastly, although equal-sized stages were used here, unequal-sized stages also can be used. In particular, because environments during anthesis are important for plant growth, finer stages during anthesis may bring more precise information on G × E interactions.

ECGC Step 2: Calculation of Similarities of Environmental Covariates Between Environments

For each environmental covariate (i.e., combination of growth stage and meteorological factor), 52 representative values corresponding to 52 environments were used to calculate the similarity between environments (**Figure 1C**). The similarity was defined using a linear kernel. That is, the similarity between the jth and kth environments were calculated as $x_j x_k$ where x_j and x_k indicate the deviations (i.e., differences from the mean of the environmental covariate) of the environmental covariate values at environments j and k, respectively. As a result, a 52×52 similarity matrix was obtained for each environmental covariate. Total 465 growth stages × 14 meteorological factors = 6,510 similarity matrices were obtained for each trait.

ECGC Step 3: Estimation of Genetic Correlation Between Environments

Genetic correlations between environments were estimated bivariate mixed models in a pairwise manner. The model can be written as

$$\begin{bmatrix} \tilde{\boldsymbol{y}}_j \\ \tilde{\boldsymbol{y}}_k \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_k \end{bmatrix} \begin{bmatrix} \mathbf{u}_j \\ \mathbf{u}_k \end{bmatrix} + \begin{bmatrix} \mathbf{e}_j \\ \mathbf{e}_k \end{bmatrix} \qquad (2)$$

where $\mathbf{Z}_j$ and $\mathbf{Z}_k$ are the design matrices. Here it is assumed that

$$\begin{bmatrix} \mathbf{u}_j \\ \mathbf{u}_k \end{bmatrix} \sim \mathrm{MVN}\left(0, \quad \boldsymbol{\Sigma}_{u_{jk}}^2 \otimes \mathbf{G} \right) \qquad \text{and}$$

$$\begin{bmatrix} \mathbf{e}_j \\ \mathbf{e}_k \end{bmatrix} \sim \mathrm{MVN}\left(0, \quad \boldsymbol{\Sigma}_{e_{jk}}^2 \otimes \mathbf{I} \right) \quad \text{where} \quad \boldsymbol{\Sigma}_{u_{jk}}^2 \text{ and } \boldsymbol{\Sigma}_{e_{jk}}^2 \text{ are the}$$

genetic and residual covariance matrices, respectively and $\otimes$ denotes the Kronecker product. To fit this model, first, the records at each environment (i.e., records in $\tilde{\boldsymbol{y}}_j$ and $\tilde{\boldsymbol{y}}_k$) were matched to each other. Records were matched according to the varieties and evaluation years. That is, same varieties evaluated at the same year were matched to each other. When multiple records could be matched (i.e., when a variety was evaluated at an environment multiple times in a single year), matching was determined randomly. As a result of matching, the phenotypic data consisting of 25,158 records was arranged to a matrix of 7,887 by 52 (**Supplementary Data S3, S4**). Note that most records did not match to any other records. Thus, the resulting 7,887 by 52 matrix was sparse: the non-missing proportion was at most $25,158/(7,887 \times 52) = 0.061$. Because PR was absent in 22 environments, the matrix for PR becomes 7,887 by 30 (**Supplementary Data S3, S4**).

Subsequently, for each trait, genetic correlations between environments were estimated in a pairwise manner. First, the bivariate mixed model (**Eq. 2**) was fitted to all pairs of environments ($52 \times 51/2 = 1,326$ pairs), to estimate the genetic covariances between environments ($\boldsymbol{\Sigma}_{u_{jk}}^2$). The diagonal elements of the covariance matrix (i.e., genetic variances) were estimated by averaging the estimates of these bivariate model analyses, which were duplicated 51 times. This pairwise manner of estimation of covariance matrices can be interpreted as a pseudo-likelihood-based approach (Fieuws and Verbeke, 2006). The resulting genetic covariance matrix was then set as a positive definite using the nearPD function of the Matrix package (ver. 1.2-18) of R, and a genetic correlation matrix was calculated by standardising the covariance matrix (**Figure 1D, Supplementary Tables S11–S16**).

It is notable that genetic correlations between environment can be estimated using methods other than the pairwise estimation described here (e.g., standard multivariate models or factor analytic models). But the pairwise estimation will be one of the easiest methods to conduct in particular when the number of environments is great.

ECGC Step 4: Scanning Environmental Covariates Associated With the Genetic Correlation

Now we have the 6,510 (465 growth stages × 14 meteorological factors) similarity matrices of environmental covariates and six (number of traits) genetic correlation matrices. The sizes of both kind matrices were 52 (number of environments) × 52 for the traits except for PR where the sizes were 30 × 30. The Pearson's product moment correlation coefficients between the upper (or

lower) triangle off-diagonals of the similarity matrices and genetic correlation matrices were then calculated (**Figure 1E**) using the cor.test function of the R package stats (ver. 4.0.3). p values of the coefficients were also calculated using the function. The significance of the coefficients was judged after Bonferroni correction of the p values. Considering the number of environmental covariates (6,510) and traits (6), the threshold of significance was set at $0.05/(6,510 \times 6) = 1.28e{-}6$.

Even after Bonferroni correction, 10,397 environmental covariate/trait combinations remained significant. This was attributed to redundancy in the environmental covariates. For example, a value at the 10th to 12th growth stages was often highly correlated with a value at the 9th to 13th growth stages. To eliminate this redundancy, environmental covariates were clustered using the hierarchical clustering implemented in the hclust function of R. Similarities were defined based on Euclidean distance, and complete-linkage clustering was used. The number of clusters was determined using the gap statistic (Tibshirani et al., 2001). The clustering results are presented in **Supplementary Figure S1**. When multiple environmental covariates were significant at a cluster, the combination with the highest r^2 was selected, resulting in 555 significant environmental covariate/trait combinations (**Supplementary Table S17**).

ECGC Step 5: Estimation of Slopes and GWA Mapping on the Slopes

For each significant environmental covariate/trait combination, the additive genetic effects of the 624 varieties at each environment were regressed on the environmental covariates, and the slopes ($\boldsymbol{\beta}$) were calculated for each variety (**Figure 1F, Supplementary Methods S3, Supplementary Table S17**). The additive genetic effects were estimated using single-trait mixed models, as described in "ECGC Step 0: Developing mixed models at each environment". The additive genetic effects were scaled with the additive genetic variance (σ_{u_j}), and the meteorological values were standardised before this regression analysis. Considering these slopes as phenotypic values, GWA mapping was conducted (**Figure 1F**) using the GWAS function of the rrBLUP package (ver. 4.6.1). The genomic relationship matrix was calculated using the A.mat function of the package. No principal components were included in the models because p values did not inflate (**Supplementary Figure S2**). SNPs with minor allele frequencies >0.05 were subjected to the statistical tests. This conservative threshold (0.05) was adopted to prevent false positives. This GWA mapping involved more than 6.40e+7 statistical tests (on average, 115,312 SNPs × 555 environmental covariate/trait combinations); thus, the application of strict multiple testing correction, such as Bonferroni correction, is unrealistic. Instead, we applied a method that controlled the false discovery rate (FDR) (Storey and Tibshirani, 2003). The threshold of significance was calculated using an R script implemented in the GWAS function of rrBLUP, with modifications. The threshold for FDR <0.05 was $9.086278e{-}06$ (5.041614 in the $-\log10p$ expression). The distribution of the p values is shown in **Supplementary Figure S2**. By grouping

significant SNPs that were less than 100 k bp apart from each other into the same regions, 1,486 regions were detected for 270 environmental covariate/trait combinations with overlaps (**Supplementary Table S19**).

To verify the validity of these associations, subsampling analyses were conducted for these 270 combinations. Specifically, 500 (80%) varieties were randomly selected from the 624 varieties and subjected to GWA mapping. Random sampling was adopted because the aim of this subsampling is to perturbate the genetic structure which might produce false positives. This procedure was repeated 10 times. Then replications where the regions detected using the full data also showed significant associations were counted (threshold −log10p = 5.041614; **Supplementary Table S19**). Regions with counts >4 were regarded as reliable results. As a result, 948 regions were remained for 179 environmental covariate/trait combinations. These regions were found to be constituted with 39 regions by removing overlaps between combinations (**Supplementary Table S19**).

Orthologs of *Arabidopsis thaliana* mapped in the detected regions were extracted from JBrowse provided by Phytozome ver.12.1 (https://phytozome.jgi.doe.gov/jbrowse/index.html?data=genomes/Gmax). Gene ontology of these orthologs were surveyed with overrepresentation analyses provided by PANTHER ver.16.0 (http://pantherdb.org/) using *Arabidopsis* genes as a reference.

Simulation Analyses

Simulation analyses were conducted to assess the performance to ECGC to detect environment covariates. As illustrated in **Eq. 1**, the detection power of ECGC depends on the proportion of the variance of $x_j x_k \sigma_\beta^2$ to that of $P_{j,k}$ (i.e., r^2) and the number of environments denoted as M. The detection power was expected to increase as r^2 and M increase. For the sake of simplification, $\sigma_{j,\beta} = 0$ and $\sigma_{u_j}^2 = 1$ for any environment j, and $\sigma_\beta^2 = 1$ throughout the simulations. Consequently, $P_{j,k} = x_j x_k + \sigma_{j,k}$. The genetic correlation matrices between environments were simulated as follows.

(1). Assign values for r^2 and M from grids 0.01, 0.018, 0.025, 0.035, 0.053, 0.075, 0.1, and 0.15, and 5, 10, 15, 20, 30, 40, 50, and 60, respectively.
(2). Generate x_j from the standard normal distribution for all j ($1 \le j \le M$), and calculate the variance of $x_j x_k$ ($1 \le j < k \le M$).
(3). Determine the variance of $\sigma_{j,k}$ according to r^2 and the variance of $x_j x_k$.
(4). Generate $\sigma_{j,k}$ from the LKJ distribution using the R package rethinking (ver. 2.13) (McElreath, 2020). Parameter η of the distribution was arbitrary set to four. Scale $\sigma_{j,k}$ according to the variance of $\sigma_{j,k}$.
(5). Generate a symmetric matrix by adding $x_j x_k$ to $\sigma_{j,k}$. Here $x_j x_j$ (i.e., diagonal elements) is set to the average of $x_j x_j$ ($1 \le j \le M$).
(6). Convert the symmetric matrix of (5) to the correlation matrix.

For each combination of r^2 and M, the genetic correlation matrix was simulated 2000 times. For each simulated matrix, 99 additional environment covariates that were not associated with the correlation matrix were also simulated as true negatives. Then the Pearson correlation between $P_{j,k}$ and $x_j x_k$ was tested for these simulated environmental covariates. The performance of ECGC was assessed using the ROC curves drawn by the R package ROCR (ver. 1.0-11) (Sing et al., 2005).

RESULTS AND DISCUSSION

Interpretation of ECGC

Eq. 1 shows that environmental covariates ($x_j x_k$) are associated with G × E interactions ($P_{j,k}$) when σ_β^2 (variance of genotype sensitivity or phenotype plasticity) is not zero and the term $x_j \sigma_{k,\beta} + x_k \sigma_{j,\beta} + \sigma_{j,k}$ does not conceal $x_j x_k \sigma_\beta^2$. This fact implies that σ_β^2 itself is not an evidence that the environmental covariate is associated with G × E interactions; even when σ_β^2 significantly deviates from zero, the effect of σ_β^2 on the genetic correlation can be concealed by the term $x_j \sigma_{k,\beta} + x_k \sigma_{j,\beta} + \sigma_{j,k}$. That is, non-zero σ_β^2 is a necessity condition for the association rather than a sufficient condition. On the other hand, significant correlations between $P_{j,k}$ and $x_j x_k$ can be the direct evidence of association between the environmental covariate and G × E interactions because, when the term $x_j \sigma_{k,\beta} + x_k \sigma_{j,\beta} + \sigma_{j,k}$ conceals $x_j x_k \sigma_\beta^2$, $x_j x_k$ is no longer correlated with $P_{j,k}$. This fact also implies that statistical testing on σ_β^2 is not an alternative of ECGC. Both approaches (ECGC and testing on σ_β^2) have different roles (detecting environment covariates associated with G × E interactions and genotype sensitivity/phenotype plasticity, respectively).

Detection of Environmental Covariates

The multi-environment data set used here consisted of 25,158 records of 624 varieties that were evaluated from 1961 to 2015 at 41 fields. The included varieties consisted of cultivars and breeding lines developed in Japan for the sake of food production, such as tofu (**Supplementary Table S3**). Among the 41 fields, in nine fields, multiple management conditions (early/late sowing dates and sparse/dense plant densities) were consistently conducted in multiple years. For these fields, different management conditions were defined as different environments as described in Materials and Methods. As a result, a total of 52 environments were defined from 41 fields (**Figure 1A** and **Supplementary Table S2**).

The steps of ECGC are shown in **Figures 1B–F**. In the first step, environmental covariates representing each environment were calculated (**Figure 1B**). Here, we considered 14 meteorological factors as the environmental stimuli (Materials and Methods). We divided the whole growth period of a plant, from sowing to maturity, into 30 stages; this was achieved by dividing the growth period (from sowing to flowering) into 10 stages, and the period from flowering to maturity into 20 stages (**Figure 1B**). For each meteorological factor (e.g., daily mean temperature), the meteorological values within each stage were averaged, yielding 30 values for each plant. These values were then averaged across plants evaluated at the environment for all the years, thus yielding 30 values for each environment (**Supplementary Tables S5–S10**). Subsequently, the values
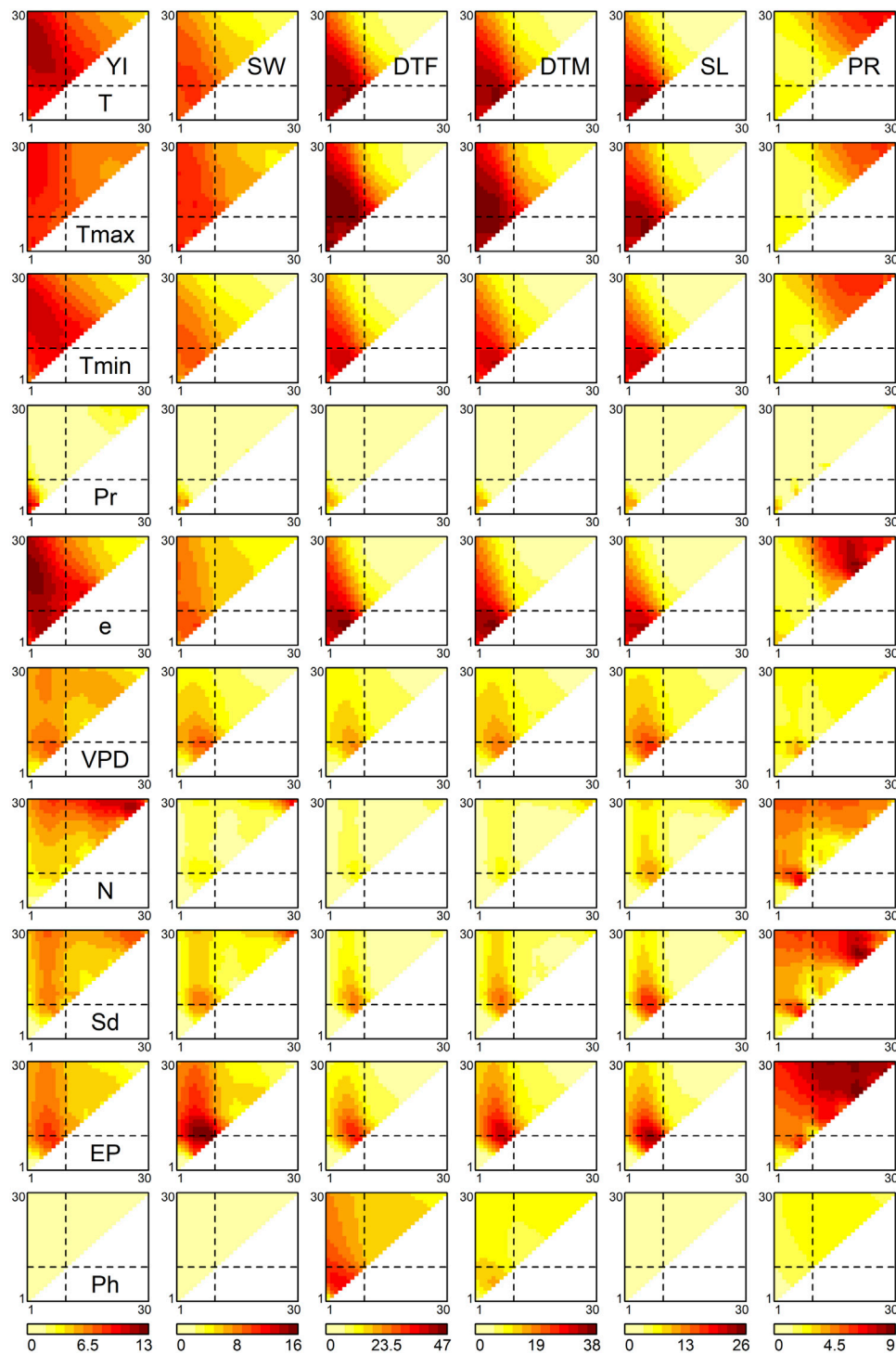
FIGURE 2 | Associations of environmental covariates with genetic correlations between environments. The heat maps represent the −log10 *p* values of correlation coefficients of off-diagonal elements between the similarity matrix of each environmental covariate and the genetic correlation matrix. The diagonal elements of the triangles correspond with the 1st to 30th growth stages, from the lower left to the upper right. The off-diagonal elements correspond to the growth periods that span multiple stages, where the *x* and *y* axes denote the start and end of the periods, respectively. The broken lines indicate flowering time. Abbreviations: YI, yield; SW, seed weight; DTF, days to flowering; DTM, days to maturity; SL, stem length; PR, protein content; T, mean temperature; Tmax, maximum temperature; Tmin, minimum temperature; Pr, precipitation; e, vapour pressure; VPD, vapour pressure deficit; N, hours of sunshine; Sd, solar radiation; EP, potential evapotranspiration; Ph, photoperiod.

were further averaged across the 30 stages, i.e. across the 1st to 2nd stages, the 1st to 3rd stages, etc. This procedure yielded 465 ($30 + 30 \times 29/2$) environmental covariates for each trait–meteorological factor combination. Thus, total 465×14 (meteorological factors) = 6,510 environmental covariates were generated.

In the second step, for each environmental covariate (e.g., precipitation at the 1st to 3rd stages), values were extracted from the 52 environments, and the linear kernel (i.e., the similarity matrix) between environments was calculated (**Figure 1C**). In the third step, genetic correlations of additive genetic effects between environments were estimated using mixed models for each trait (**Figure 1D**, **Supplementary Tables S11–S16**, **Supplementary Figure S3**). Estimated genetic correlations suggest strong G × E interactions particularly for traits other than SW (**Supplementary Figure S3**). In the fourth step, the genetic correlation matrices were compared with the linear kernels (i.e., similarity matrices) of the environmental covariates (**Figure 1E**). Stronger associations between these matrices indicated that the environmental covariate affected the G × E interactions with greater magnitude. These associations were measured using the Pearson's moment product correlation coefficient (r) between the off-diagonal elements of the matrices. **Figure 2** shows the distributions of the $-\log10 p$ values of r calculated for 10 meteorological factors (the complete results are presented in **Supplementary Figure S4**). Generally, YI and SW, and DTT, DTM and SL shared similar p values patterns, whereas PR exhibited unique patterns. A total of 555 environmental covariates were significantly detected for the six traits ($p < 0.05$ after Bonferroni correction, **Supplementary Table S17**).
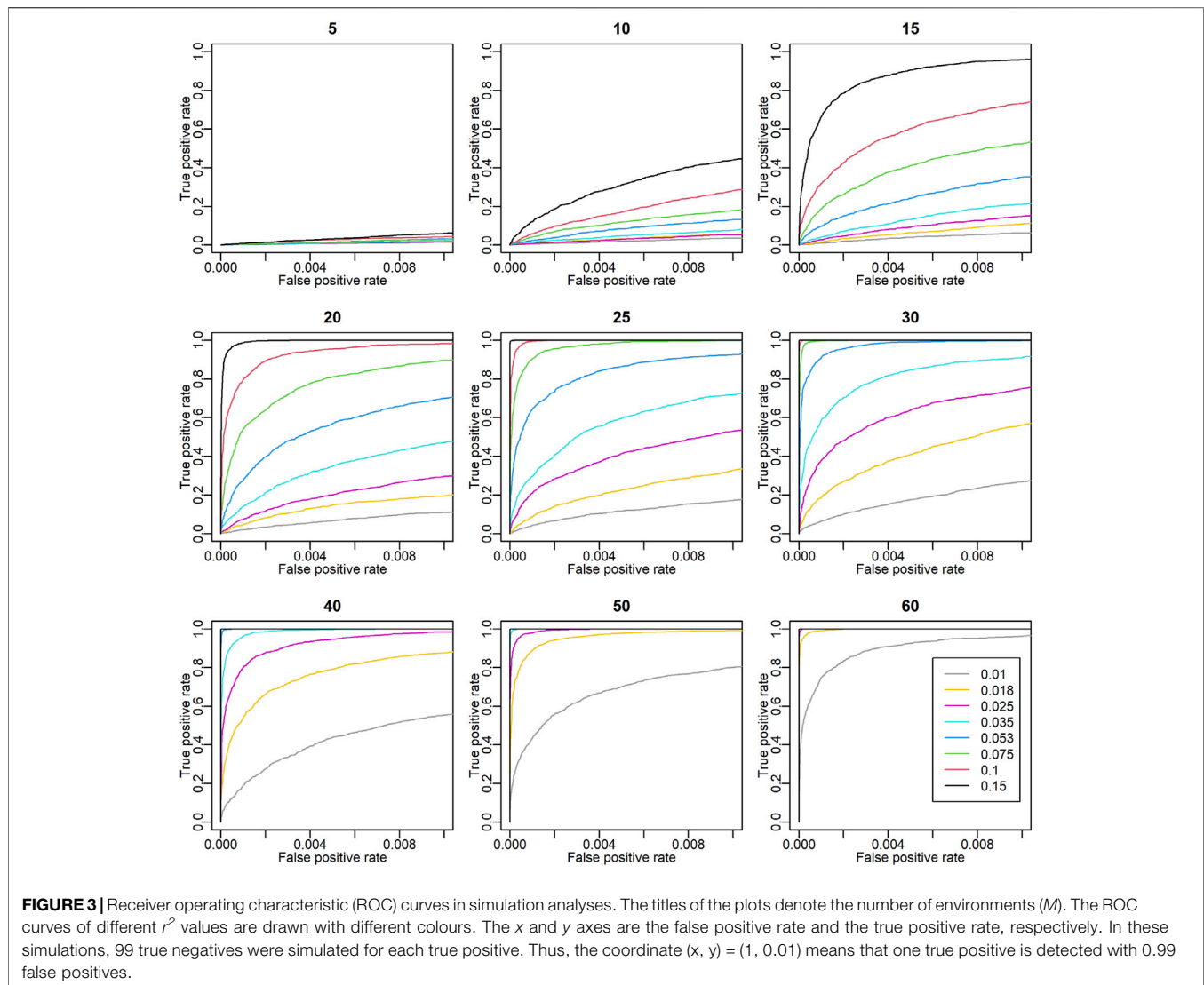
For YI, three regions in **Figure 2** showed high $-\log10 p$ values. The first region was detected around the 2nd to 22nd growth stages (upper-left part of the triangles in **Figure 2**) of the temperature-related environmental covariates, such as mean temperature, minimum temperature and vapour pressure. The highest $-\log10 p$ values of these meteorological factors were observed at the 2nd to 22nd growth stages ($r^2 = 0.036$, $-\log10P = 11.554$), the 4th to 17th stages ($r^2 = 0.033$, $-\log10P = 10.677$) and the 2nd to 19th stages ($r^2 = 0.039$, $-\log10P = 12.460$), respectively. The 22nd growth stage corresponded on average to 30.4–34.2 days before maturity. According to the well-known system used to stage soybean development (Fehr and Caviness, 1977), this period largely corresponds to the R5 soybean growth stage (Egli, 2010). Because the R5 stage is defined as the beginning of seed filling (Fehr and Caviness, 1977), it is likely that the temperature-related covariates affect the G × E interactions on YI by modifying the upper limit of the number of seeds. The second region was a period around sowing in relation to precipitation. The highest $-\log10 p$ value was 10.888 at the 1st to 3rd stages ($r^2 = 0.034$). This stage corresponds to a period from sowing to, on average, 16.5 days after sowing. Soybeans are vulnerable to waterlogging, and, in particular, waterlogging around germination has a severe impact on YI (Kokubun, 2013). In addition, genetic variations exist in root development in flood conditions (Sakazono et al., 2014). Thus, our results are reasonable and clearly suggest the

importance of precipitation around germination for the G × E interactions regarding YI. The third region was a period just before maturity in relation to hours of sunshine and solar radiation. The highest $-\log10 p$ value was 11.682 observed for hours of sunshine at the 26th to 29th stages ($r^2 = 0.037$). During this period, these meteorological factors also showed a high $-\log10 p$ value for SW; the highest value for SW around this period was 11.101 for hours of sunshine at the 29th to 30th stage ($r^2 = 0.035$).

The highest $-\log10 p$ value for SW was observed in relation to potential evapotranspiration at the 8th to 10th stages ($r^2 = 0.050$, $-\log10P = 15.715$). Potential evapotranspiration (or potential evaporation) can be regarded as the upper limit of evapotranspiration from a crop field (Kuwagata et al., 2011), which can reflect photosynthesis activity. The other meteorological factors detected, i.e., mean temperature and maximum temperature (highest $r^2 = 0.032$ and 0.036 at the 5th stage; $-\log10P = 10.199$ and 11.570, respectively) and vapour pressure deficit ($r^2 = 0.029$ at the 9th stage, $-\log10P = 9.261619$), also suggest the relevance of photosynthesis. The influence of these factors started before flowering, when seed filling has not started. Thus, it is likely that these factors affected SW indirectly via the modification of the number of seeds, because soybean shows compensation effects between SW and seed number.

For DTF, the significant environmental covariates included mean temperature at the 6th stage ($r^2 = 0.140$, $-\log10P = 44.930$), minimum temperature at the 6th stage ($r^2 = 0.127$, $-\log10P = 40.286$), maximum temperature at the 5th to 14th stages ($r^2 = 0.146$, $-\log10P = 46.602$) and vapour pressure at the 6th stage ($r^2 = 0.145$, $-\log10P = 46.353$). Photoperiod before flowering was also a significant covariate ($r^2 = 0.106$ at the 4th to 5th stages, $-\log10P = 33.286$). These results are reasonable considering that temperature and photoperiod are the main determinants of soybean flowering time (Sinclair et al., 1991). The results obtained for DTM and SL were close to those obtained for DTF, suggesting that the influence of meteorological factors on the G × E interactions for DTM and SL occurred via those of DTF. In other words, if DTF does not show a G × E interaction, DTM and SL will also not show. Finally, the results obtained for SL are reasonable, because the varieties commonly cultivated in Japan exhibit the determinate stem type, in which the prolongation of the stem ends shortly after flowering begins (Bernard, 1972).

The p value patterns of PR were different from those of the other traits. Meteorological factors generally exert their effects after flowering, around the 21st to 27th stages, as observed for mean temperature (highest $-\log P = 6.854$ at the 27th stage, $r^2 = 0.062$), maximum temperature ($-\log P = 5.544$ at the 27th stage, $r^2 = 0.049$), minimum temperature ($-\log P = 6.207$ at the 27th stage, $r^2 = 0.056$), vapor pressure (at the 21st stage, $-\log10P = 8.330$, $r^2 = 0.076$), solar radiation ($-\log P = 8.605$ at the 21st to 24th stages, $r^2 = 0.079$) and potential evapotranspiration ($-\log P = 8.658$ at the 21st to 22nd stages, $r^2 = 0.079$). The 21st to 27th stages occurred on average 38–64.6 days after flowering, when the accumulation of protein is past its rapidest growth period (20–40 days after flowering), but is still ongoing (Gayler and Sykes, 1981). The meteorological

FIGURE 3 | Receiver operating characteristic (ROC) curves in simulation analyses. The titles of the plots denote the number of environments (*M*). The ROC curves of different r^2 values are drawn with different colours. The *x* and *y* axes are the false positive rate and the true positive rate, respectively. In these simulations, 99 true negatives were simulated for each true positive. Thus, the coordinate (x, y) = (1, 0.01) means that one true positive is detected with 0.99 false positives.

factors detected potentially affected the G × E interactions of PR via the photosynthetic activity at this stage, which is essential for nitrogen fixation.
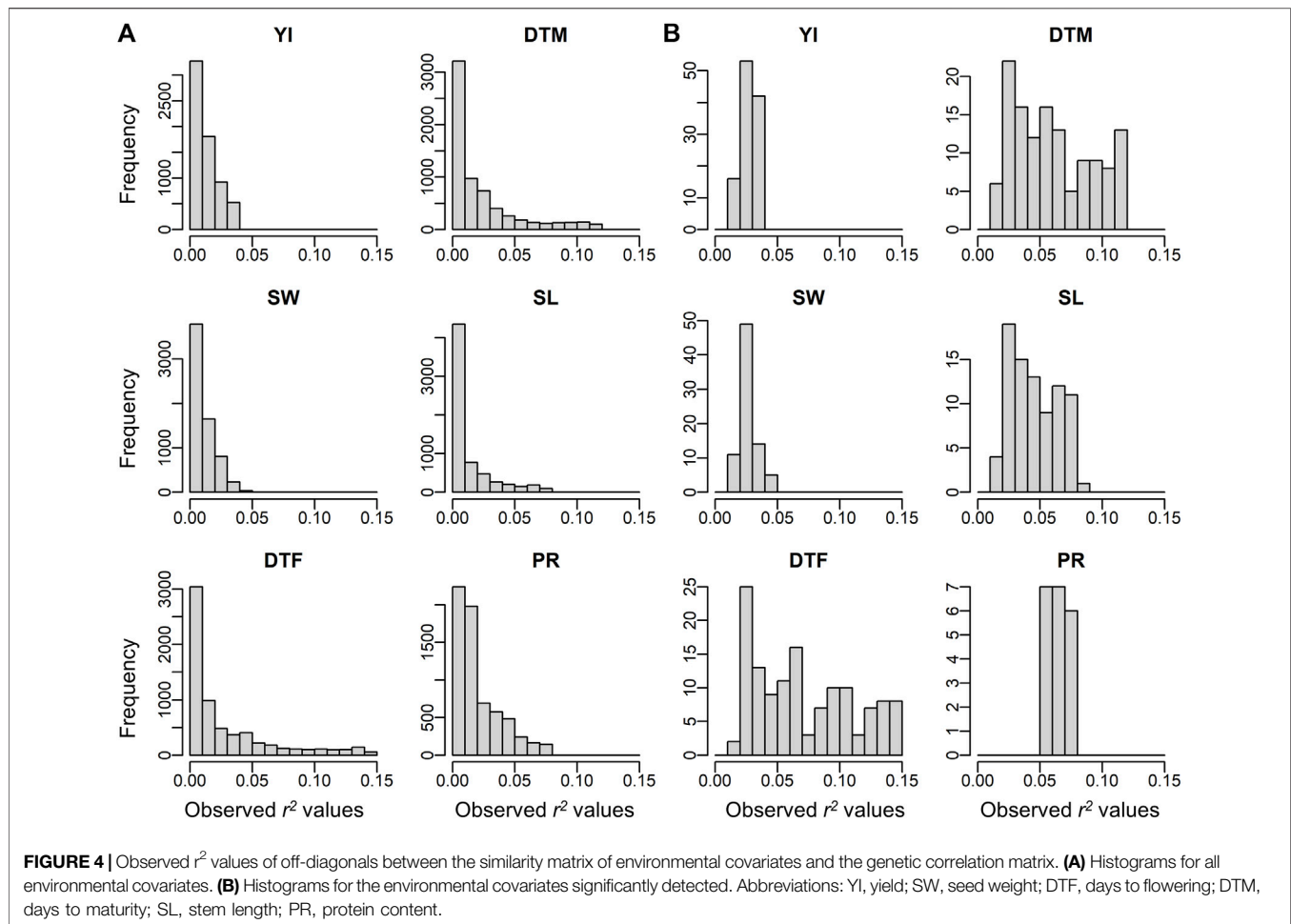
Simulation Results

To assess the validity of these detected environmental covariates, the detective power of ECGC was verified with simulations. As illustrated in the Methods, the performance of ECGC is affected by r^2 and the number of environment *M*. The receiver operating characteristic (ROC) curves obtained under different combinations of r^2 and *M* are shown in **Figure 3**. As expected, the performance of ECGC gained as r^2 and *M* increased. In **Figures 4A,B**, observed r^2 values for each trait are shown for comparison with **Figure 3**. The r^2 values of the detected environmental covariates were greater than 0.053 for PR (*M* = 30) and greater than 0.018 for the other traits (*M* = 52) (**Figure 4B**). The ROC curves under the corresponding combinations of r^2 and *M* in simulation results (0.053 and 30, and 0.018 and 50, respectively) suggest that ECGC detected the

environmental covariates with reasonable accuracy under these r^2 and *M* values.

Association Mapping on Genotype Sensitivity

In the final step of ECGC, the genetic architecture underpinning the sensitiveness of the varieties to the detected environmental covariates was examined using association mapping (**Figure 1F**). For the environmental covariates detected in the fourth step, slopes were estimated for each variety by regressing the additive genetic effects at each environment on the environmental covariates (**Supplementary Table S18**). The slope represented the sensitivity of the variety to the changes in the environmental covariates, and genome-wide association (GWA) mapping was conducted on the slopes. By pruning with false discovery rate (FDR) < 0.05 and subsampling analyses, 39 chromosomal regions were significantly detected for traits except for SW (**Supplementary Figures S5–S10** and **Supplementary Table S19**). DTF shared one and two regions

FIGURE 4 | Observed r² values of off-diagonals between the similarity matrix of environmental covariates and the genetic correlation matrix. **(A)** Histograms for all environmental covariates. **(B)** Histograms for the environmental covariates significantly detected. Abbreviations: YI, yield; SW, seed weight; DTF, days to flowering; DTM, days to maturity; SL, stem length; PR, protein content.
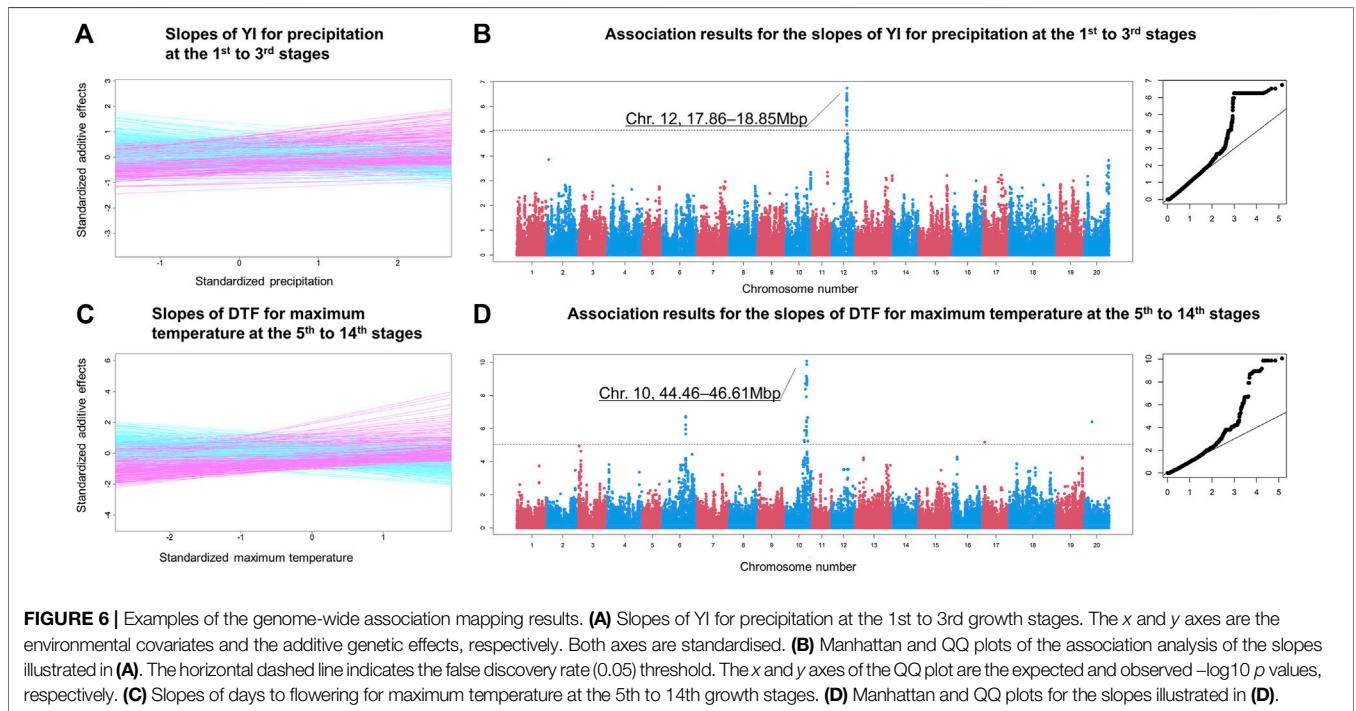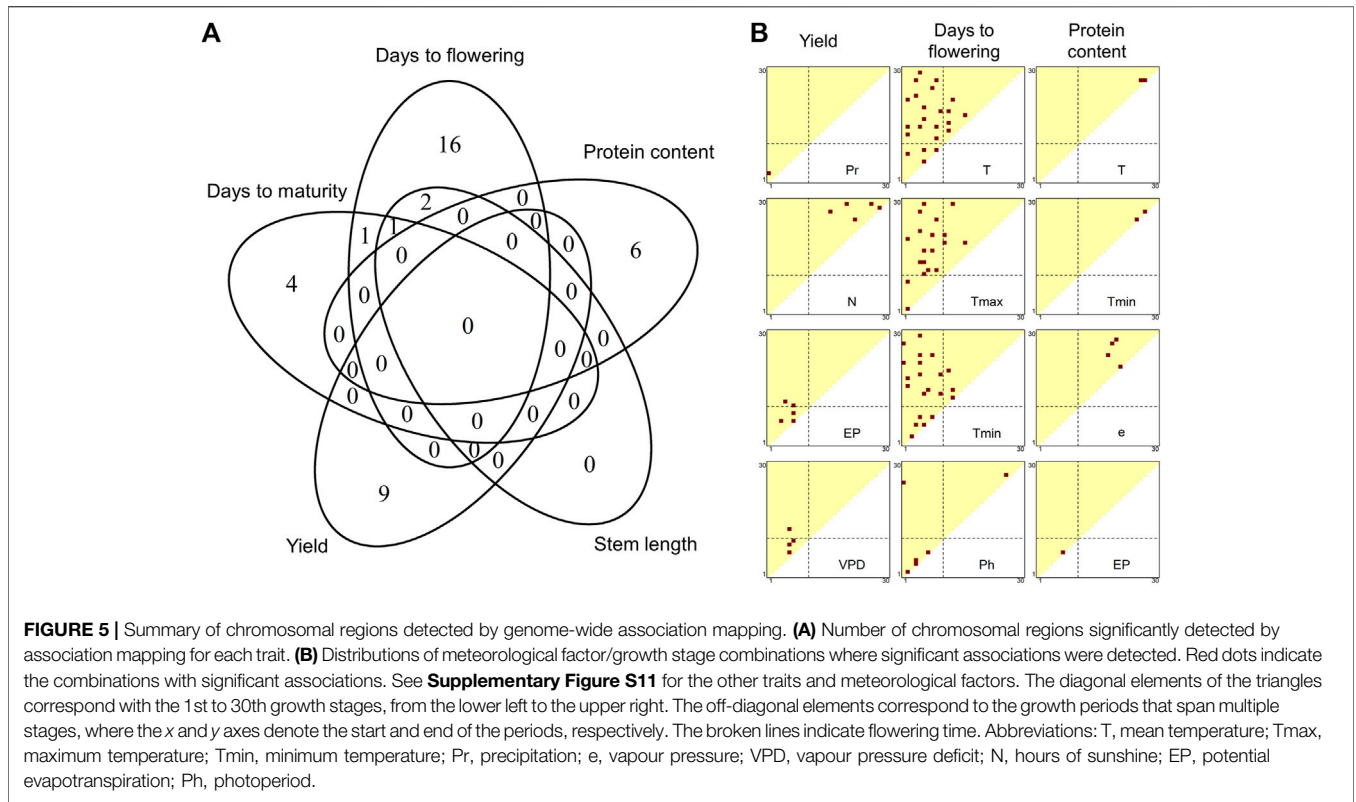
with DTM and SL, respectively, and these three traits shared one region (**Figure 5A**) which is located at near the known flowering gene (*E2*) (Watanabe et al., 2011). On the other hand, YI and PR shared no regions with each other, and nor with DTF, DTM, and SL (**Figure 5A**). These results suggest that major genes responsible for the G × E interactions triggered by meteorological factors are generally not pleiotropic. This suggestion can be confirmed by the distributions of the meteorological factor/growth stage combinations where significant associations were detected (**Figure 5B** and **Supplementary Figure S11**). For DTF, DTM, and SL, significant associations were generally detected for the temperature-related factors and photoperiod before flowering. For YI, associations were significantly detected for precipitation around sowing, hours of sunshine during maturity, and potential evapotranspiration and vapor pressure deficit (i.e., factors related to photosynthesis) around flowering. For PR, significant associations were mainly detected for temperature-related factors during maturity. That is, the major genes responsible for these traits affect the sensitivities to different meteorological stimuli occurred at different growth stages.

GWA mapping of ECGC could narrow down a genomic region that was suggested to be responsible for the G × E interactions by QTL mapping. Three close regions on chromosome 12 spanning 17.86–18.85 Mbp were detected for

the slopes of YI for precipitation at the 1st to 3rd stages (**Figures 6A,B**, **Supplementary Table S19**). These regions include the QTL (14.39–35.1 Mbp) for hypoxia tolerance of Japanese soybean breeds (Van Nguyen et al., 2017), which is related to root extension under flood conditions. Gene ontology analyses revealed that two orthologs of *Glyma.12G142900* (18.503–18.505 Mbp) in *Arabidopsis thaliana* (*AT4G27280* and *AT5G54490*) are involved in cell response to hypoxia. In addition, the gene (*Glyma.12G142900*) was reported to show higher expression on root tissues (Phytozome 12, accessed on March 30, 2021), suggesting the gene is a plausible candidate.
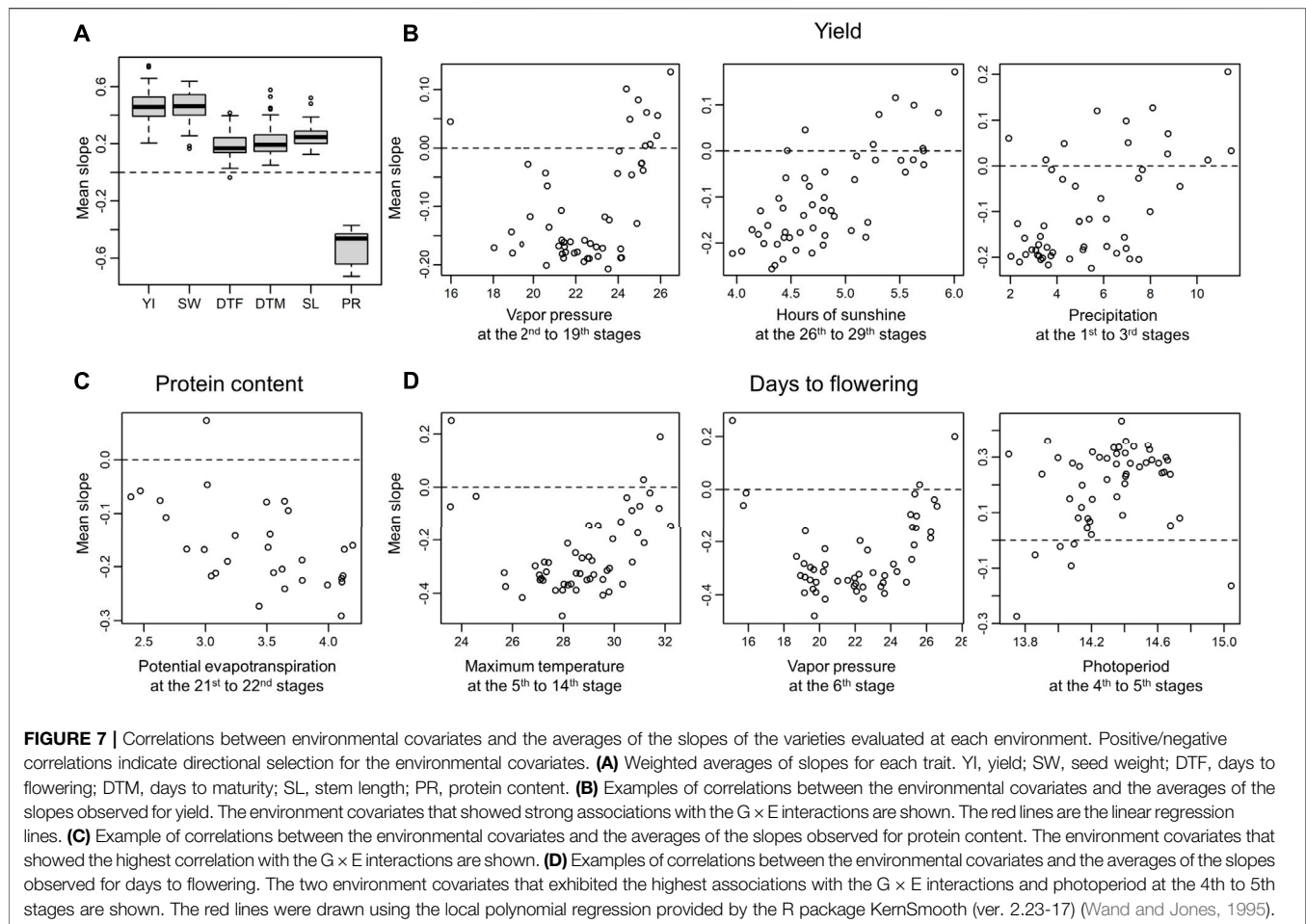
Strong associations were observed on multiple adjacent regions on chromosome 10 spanning 44.46–46.61 Mbp for various meteorological factors, including temperature-related covariates and photoperiod (**Figures 6C,D**, **Supplementary Table S19**, and **Supplementary Figure S9**). These regions were also detected for DTM and SL, reflecting the similar tendencies of the G × E interactions (**Supplementary Figures S7, S10**). These regions included *E2* (45.29–45.32 Mbp), which is an analogous gene of *GIGANTEA* of *Arabidopsis* and a major gene for flowering time in soybean (Watanabe et al., 2011). These regions also included two orthologs (*Glyma.10G180600* and *Glyma.10G209600*) of *Arabidopsis* flowering genes (*CRY2* and

FIGURE 5 | Summary of chromosomal regions detected by genome-wide association mapping. **(A)** Number of chromosomal regions significantly detected by association mapping for each trait. **(B)** Distributions of meteorological factor/growth stage combinations where significant associations were detected. Red dots indicate the combinations with significant associations. See **Supplementary Figure S11** for the other traits and meteorological factors. The diagonal elements of the triangles correspond with the 1st to 30th growth stages, from the lower left to the upper right. The off-diagonal elements correspond to the growth periods that span multiple stages, where the x and y axes denote the start and end of the periods, respectively. The broken lines indicate flowering time. Abbreviations: T, mean temperature; Tmax, maximum temperature; Tmin, minimum temperature; Pr, precipitation; e, vapour pressure; VPD, vapour pressure deficit; N, hours of sunshine; EP, potential evapotranspiration; Ph, photoperiod.

FIGURE 6 | Examples of the genome-wide association mapping results. **(A)** Slopes of YI for precipitation at the 1st to 3rd growth stages. The x and y axes are the environmental covariates and the additive genetic effects, respectively. Both axes are standardised. **(B)** Manhattan and QQ plots of the association analysis of the slopes illustrated in **(A)**. The horizontal dashed line indicates the false discovery rate (0.05) threshold. The x and y axes of the QQ plot are the expected and observed $-\log10$ p values, respectively. **(C)** Slopes of days to flowering for maximum temperature at the 5th to 14th growth stages. **(D)** Manhattan and QQ plots for the slopes illustrated in **(D)**.

ELF6, respectively). Known flowering genes of soybean responsible for photosensitivity, such as *E1* (Xia et al., 2012), *E3* (Watanabe et al., 2009), and *E4* (Liu et al., 2008), were not detected for any of the environmental covariates. It is notable that association mapping using slopes can detect genes with effects that vary according to environmental covariates.

FIGURE 7 | Correlations between environmental covariates and the averages of the slopes of the varieties evaluated at each environment. Positive/negative correlations indicate directional selection for the environmental covariates. **(A)** Weighted averages of slopes for each trait. YI, yield; SW, seed weight; DTF, days to flowering; DTM, days to maturity; SL, stem length; PR, protein content. **(B)** Examples of correlations between the environmental covariates and the averages of the slopes observed for yield. The environment covariates that showed strong associations with the G × E interactions are shown. The red lines are the linear regression lines. **(C)** Example of correlations between the environmental covariates and the averages of the slopes observed for protein content. The environment covariates that showed the highest correlation with the G × E interactions are shown. **(D)** Examples of correlations between the environmental covariates and the averages of the slopes observed for days to flowering. The two environment covariates that exhibited the highest associations with the G × E interactions and photoperiod at the 4th to 5th stages are shown. The red lines were drawn using the local polynomial regression provided by the R package KernSmooth (ver. 2.23-17) (Wand and Jones, 1995).

Because the loss-of-function alleles of *E1* (Tsubokura et al., 2014) were strictly used at higher latitudes (**Supplementary Methods S4** and **Supplementary Table S20**), the effects at lower latitudes could not be estimated. Thus, these effects would not be reflected in the slopes. Conversely, *E3* and *E4* were probably not involved in the G × E interactions because of constant gene effects across environments. Additional analyses that examined the allele effects of these flowering genes using sommer (Covarrubias-Pazaran, 2016) showed that the *E2* gene effect was the most variable across environments (**Supplementary Methods S5** and **Supplementary Figure S12**).

Insights on Selection History

The slopes provide insights on how the G × E interactions were involved in the selection of modern varieties. For example, for YI, varieties with positive slopes for an environmental covariate are expected to be more preferred at environments with greater environmental covariates. In other words, if a trait is directionally preferred, it is expected that the environmental covariates will be correlated with the averages of slopes of the varieties evaluated at each environment. In fact, the correlations were positive for YI and SW (**Figures 7A–C**) and often significant ($p < 0.05$ after Bonferroni correction, **Supplementary Table S21**).

The negative correlations found for PR are attributable to the trade-off between YI and PR (i.e., a higher yield tends to lower PR). These results clearly suggest that varieties that exhibit high YI under the detected environmental covariates have been selected by breeders, intendedly or unintendedly. Conversely, for DTF, DTM, and SL, the correlations were more moderate, and no significant correlation was found. This result is reasonable because these traits were not directionally preferred. Rather, a characteristic U-shaped trend was found for temperature-related environmental covariates and photoperiod (**Figure 7D**). For DTF, this tendency indicates that a longer DTF is acceptable in lower latitudes (more temperate climates), whereas a shorter DTF is preferred in higher latitudes (less temperate climates). These results are coherent because early flowering is required in cold climates to avoid frost damage.

Advantages of ECGC

The base model of ECGC is a conventional model for reaction norm where the genotypic value is divided into a component representing sensitivity to environmental covariates and a component free from them (van Eeuwijk et al., 2005; Hayes et al., 2016). The novelty of ECGC is to associate similarity matrices of environmental covariates with the genetic

correlation matrix, and the idea of ECGC can be derived by extending the base model as described in Materials and Methods. This approach brings the following advantages. First, the environmental covariate search by ECGC is directly related to the G × E interactions whereas the search based on the variance of slopes (sensitivities of genotypes) is not necessarily related to the G × E interactions. Second, because of the properties of mixed models used for estimating genetic correlations between environments, ECGC is applicable to data with missing values and/or unbalanced structure. Genetic correlations between environments can be estimated using mixed models and genomic relationship matrices even when varieties are not overlapped between environments. Estimation of genetic correlation without overlapped records can be found in, for example, estimation of between-sex genetic correlation (Crews and Kemp, 2001). Lastly, genetic covariances/correlations between multiple environments can be estimated in pairwise manner (Fieuws and Verbeke, 2006) and parallelized, which enables ECGC to be scalable with the number of environments. The last two advantages are particularly useful in analysis of large-scale multi-environment data sets and/or historical data sets.

CONCLUSION

Owing to these advantages, our proposed ECGC was applicable to the large-scale multi-environment data set of soybean, and able to depict comprehensive landscapes on how environment stimuli are involved in the G × E interactions for agronomic traits evaluated in real fields. Moreover, candidate QTLs/genes responsible for the interactions were detected. ECGC also provided interesting insights on how the G × E interactions are related to the selection of modern Japanese varieties. Thus, it can be concluded that ECGC will be a promising approach to understand the G × E interactions and to reveal the gene-by-environment stimuli interactions.

DATA AVAILABILITY STATEMENT

The data used in this study is subject to the following licenses/restrictions: the SNP data cannot be freely disclosed. Requests to access the dataset should be directed to onogiakio@gmail.com. The other data (phenotype and meteorological values, Supplementary Data 1–4) is freely available at Dryad (https://doi.org/10.5061/dryad.rr4xgxd6r).

REFERENCES

AUTHOR CONTRIBUTIONS

AO collected and organised historical data, conceived the ECGC framework, conducted all statistical analyses, drafted the manuscript and collaborated on DNA extraction and genotyping. DS extracted DNA and collaborated in drafting the manuscript. AK designed the SNP array, extracted DNA and collaborated on genotyping and drafting the manuscript. SN, TY, and JY collaborated on drafting the manuscript. SN suggested the importance of historical data and collaborated on drafting the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material (Supplementary Methods S1-5 and Figures S1-12) for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.803636/full#supplementary-material. The Supplementary Tables S1-21 and an example R script to execute ECGC are available at Dryad (https://doi.org/10.5061/dryad.rr4xgxd6r)

Bernard, R. L. (1972). Two Genes Affecting Stem Termination in Soybeans 1. *Crop Sci.* 12, 235–239. doi:10.2135/cropsci1972.0011183X001200020028x

Browning, B. L., and Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *Am. J. Hum. Genet.* 98, 116–126. doi:10.1016/j.ajhg.2015.11.020

Covarrubias-Pazaran, G. (2016). Genome-assisted Prediction of Quantitative Traits Using the R Package Sommer. *PLoS One* 11, e0156744. doi:10.1371/journal.pone.0156744

Crews, D. H., Jr., and Kemp, R. A. (2001). Genetic Parameters for Ultrasound and Carcass Measures of Yield and Quality Among Replacement and slaughter Beef Cattle. *J. Anim. Sci.* 79, 3008–3020. doi:10.2527/2001.79123008x

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L. (2013). Whole-genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193, 327–345. doi:10.1534/genetics.112.143313

Des Marais, D. L., Hernandez, K. M., and Juenger, T. E. (2013). Genotype-by-environment Interaction and Plasticity: Exploring Genomic Responses of Plants to the Abiotic Environment. *Annu. Rev. Ecol. Evol. Syst.* 44, 5–29. doi:10.1146/annurev-ecolsys-110512-135806

Egli, D. B. (2010). "Soybean Yield Physiology: Principles and Processes of Yield Production," in *The Soybean Botany, Production and Uses*. Editor G. Singh (Wallingford, Oxfordshire: CAB International). doi:10.1079/9781845936440.0113

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome* 4, 250–255. doi:10.3835/plantgenome2011.08.0024

Fehr, W. R., and Caviness, C. E. (1977). *Stages of Soybean Development*. Special Report 87. Ames, Iowa: Iowa State University, Agricultural and Home Economics Experiment Station. Available at: https://lib.dr.iastate.edu/specialreports/87.

Fieuws, S., and Verbeke, G. (2006). Pairwise Fitting of Mixed Models for the Joint Modeling of Multivariate Longitudinal Profiles. *Biometrics* 62, 424–431. doi:10.1111/j.1541-0420.2006.00507.x

Gayler, K. R., and Sykes, G. E. (1981). β-Conglycinins in Developing Soybean Seeds. *Plant Physiol.* 67, 958–961. doi:10.1104/pp.67.5.958

Guo, T., Mu, Q., Wang, J., Vanous, A. E., Onogi, A., Iwata, H., et al. (2020). Dynamic Effects of Interacting Genes Underlying rice Flowering-Time Phenotypic Plasticity and Global Adaptation. *Genome Res.* 30, 673–683. doi:10.1101/gr.255703.119

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.

Hayes, B. J., Daetwyler, H. D., and Goddard, M. E. (2016). Models for Genome × Environment Interaction: Examples in Livestock. *Crop Sci.* 56, 2251–2259. doi:10.2135/cropsci2015.07.0451

Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., et al. (2014). A Reaction Norm Model for Genomic Selection Using High-Dimensional Genomic and Environmental Data. *Theor. Appl. Genet.* 127, 595–607. doi:10.1007/s00122-013-2243-1

Kaga, A., Shimizu, T., Watanabe, S., Tsubokura, Y., Katayose, Y., Harada, K., et al. (2012). Evaluation of Soybean Germplasm Conserved in NIAS Genebank and Development of Mini Core Collections. *Breed. Sci.* 61, 566–592. doi:10.1270/jsbbs.61.566

Khosla, S., Augustus, M., and Brahmachari, V. (1999). Sex-specific Organisation of Middle Repetitive DNA Sequences in the Mealybug *Planococcus Lilacinus*. *Nucleic Acids Res.* 27, 3745–3751. doi:10.1093/nar/27.18.3745

Kokubun, M. (2013). Genetic and Cultural Improvement of Soybean for Waterlogged Conditions in Asia. *Field Crops Res.* 152, 3–7. doi:10.1016/j.fcr.2012.09.022

Kuwagata, T., Yoshimoto, M., Ishigooka, Y., Hasegawa, T., Utsumi, M., Nishimori, M., et al. (2011). MeteoCrop DB: an Agro-Meteorological Database Coupled with Crop Models for Studying Climate Change Impacts on rice in Japan. *J. Agric. Meteorol.* 67, 297–306. doi:10.2480/agrmet.67.4.9

Li, X., Guo, T., Mu, Q., Li, X., and Yu, J. (2018). Genomic and Environmental Determinants and Their Interplay Underlying Phenotypic Plasticity. *Proc. Natl. Acad. Sci. USA* 115, 6679–6684. doi:10.1073/pnas.1718326115

Liu, B., Kanazawa, A., Matsumura, H., Takahashi, R., Harada, K., and Abe, J. (2008). Genetic Redundancy in Soybean Photoresponses Associated with Duplication of the Phytochrome A Gene. *Genetics* 180, 995–1007. doi:10.1534/genetics.108.092742

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The Statistical Analysis of Multi-Environment Data: Modeling Genotype-By-Environment Interaction and its Genetic Basis. *Front. Physiol.* 4, 44. doi:10.3389/fphys.2013.00044

Mather, K., and Jones, R. M. (1958). Interaction of Genotype and Environment in Continuous Variation: I. Description. *Biometrics* 14, 343–359. doi:10.2307/2527879

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Second edition. Boca Raton, FL: CRC Press.

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic Prediction of maize Yield across European Environmental Conditions. *Nat. Genet.* 51, 952–956. doi:10.1038/s41588-019-0414-y

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., and Lee, D. H. (2002). "BLUPF90 and Related Programs," in Proceedings of the 7th World Congress On Genetics Applied To Livestock Production, Montpellier, France.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.

Sakazono, S., Nagata, T., Matsuo, R., Kajihara, S., Watanabe, M., Ishimoto, M., et al. (2014). Variation in Root Development Response to Flooding Among 92 Soybean Lines during Early Growth Stages. *Plant Prod. Sci.* 17, 228–236. doi:10.1626/pps.17.228

Sinclair, T. R., Kitani, S., Hinson, K., Bruniard, J., and Horie, T. (1991). Soybean Flowering Date: Linear and Logistic Models Based on Temperature and Photoperiod. *Crop Sci.* 31, 786–790. doi:10.2135/cropsci1991.0011183X003100030049x

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing Classifier Performance in R. *Bioinformatics* 21, 3940–3941. doi:10.1093/bioinformatics/bti623

Storey, J. D., and Tibshirani, R. (2003). Statistical Significance for Genomewide Studies. *Proc. Natl. Acad. Sci.* 100, 9440–9445. doi:10.1073/pnas.1530509100

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the Number of Clusters in a Data Set via the gap Statistic. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 63, 411–423. doi:10.1111/1467-9868.00293

Tsubokura, Y., Watanabe, S., Xia, Z., Kanamori, H., Yamagata, H., Kaga, A., et al. (2014). Natural Variation in the Genes Responsible for Maturity Loci *E1, E2, E3* and *E4* in Soybean. *Ann. Bot.* 113, 429–441. doi:10.1093/aob/mct269

van Eeuwijk, F. A., Malosetti, M., Yin, X., Struik, P. C., and Stam, P. (2005). Statistical Models for Genotype by Environment Data: from Conventional ANOVA Models to Eco-Physiological QTL Models. *Aust. J. Agric. Res.* 56, 883–894. doi:10.1071/AR05153

Van Nguyen, L., Takahashi, R., Githiri, S. M., Rodriguez, T. O., Tsutsumi, N., Kajihara, S., et al. (2017). Mapping Quantitative Trait Loci for Root Development under Hypoxia Conditions in Soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet.* 130, 743–755. doi:10.1007/s00122-016-2847-3

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *J. Dairy Sci.* 91, 4414–4423. doi:10.3168/jds.2007-0980

Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing. Volume 60 of Monographs on Statistics and Applied Probability*. Boca Raton, FL: Chapman and Hall/CRC. .

Watanabe, S., Hideshima, R., Xia, Z., Tsubokura, Y., Sato, S., Nakamoto, Y., et al. (2009). Map-based Cloning of the Gene Associated with the Soybean Maturity Locus *E3*. *Genetics* 182, 1251–1262. doi:10.1534/genetics.108.098772

Watanabe, S., Xia, Z., Hideshima, R., Tsubokura, Y., Sato, S., Yamanaka, N., et al. (2011). A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that the *GIGANTEA* Gene Is Involved in Soybean Maturity and Flowering. *Genetics* 188, 395–407. doi:10.1534/genetics.110.125062

Xia, Z., Watanabe, S., Yamada, T., Tsubokura, Y., Nakashima, H., Zhai, H., et al. (2012). Positional Cloning and Characterization Reveal the Molecular Basis for Soybean Maturity Locus *E1* that Regulates Photoperiodic Flowering. *Proc. Natl. Acad. Sci.* 109, E2155–E2164. doi:10.1073/pnas.1117982109