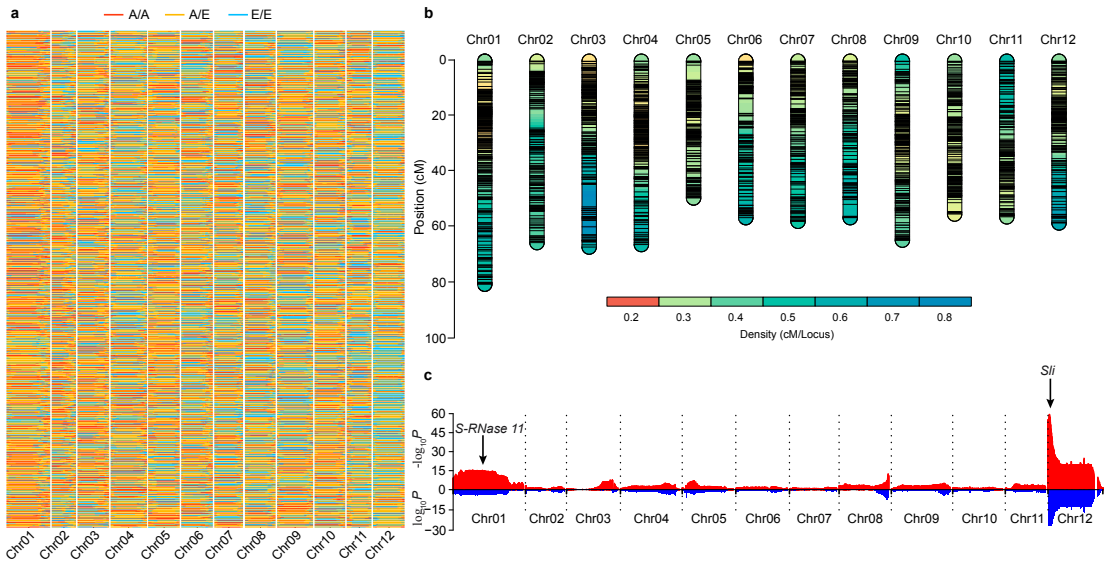
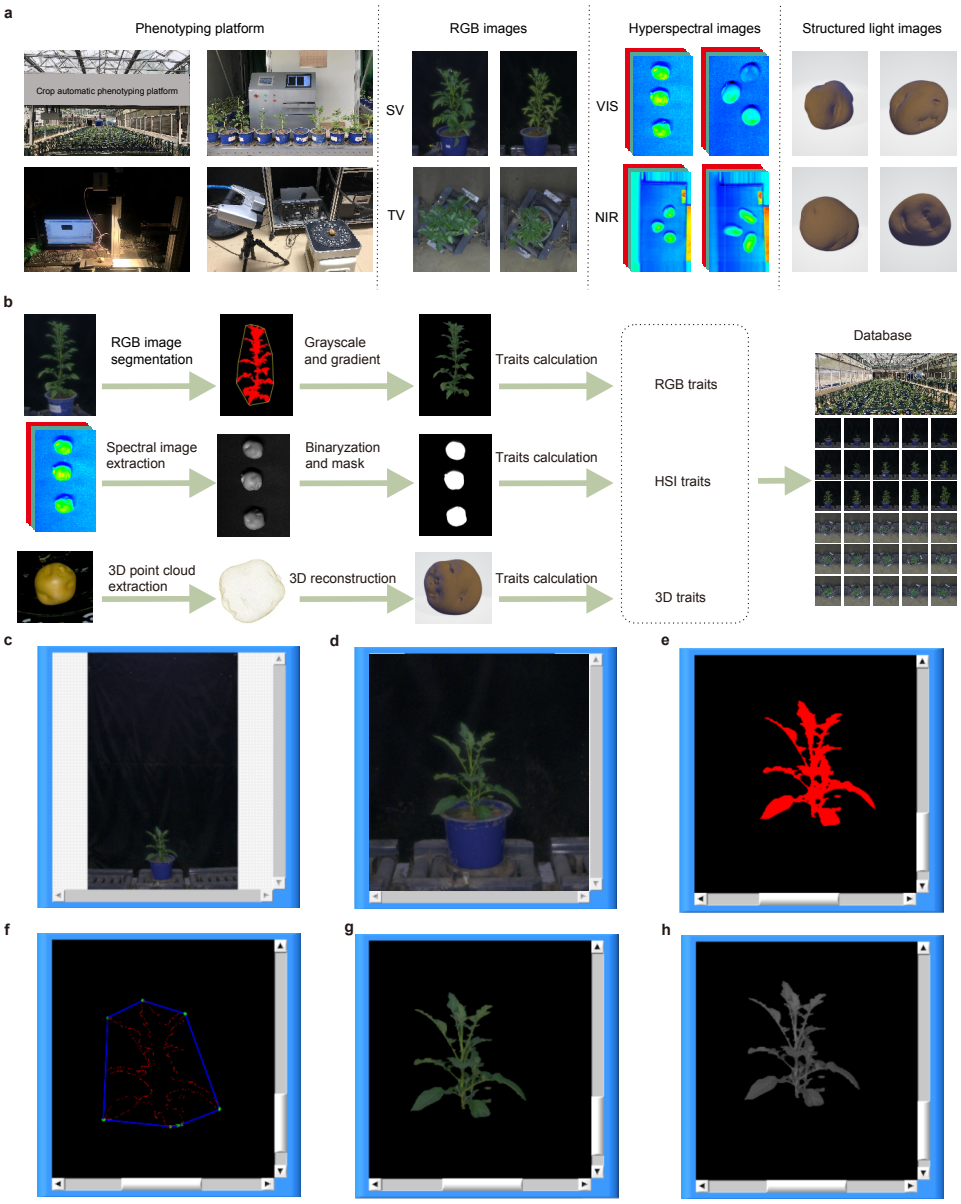
Integrative multi-omics analysis reveals genetic and heterotic contributions to male fertility and yield in potato

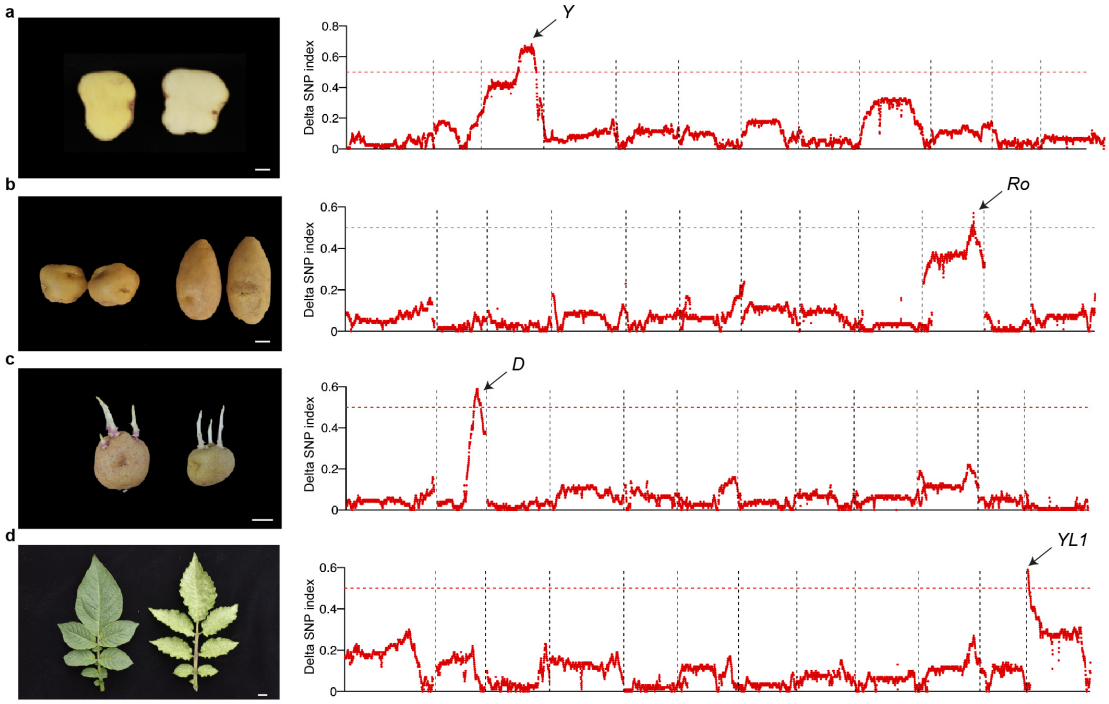
Li *et al.*



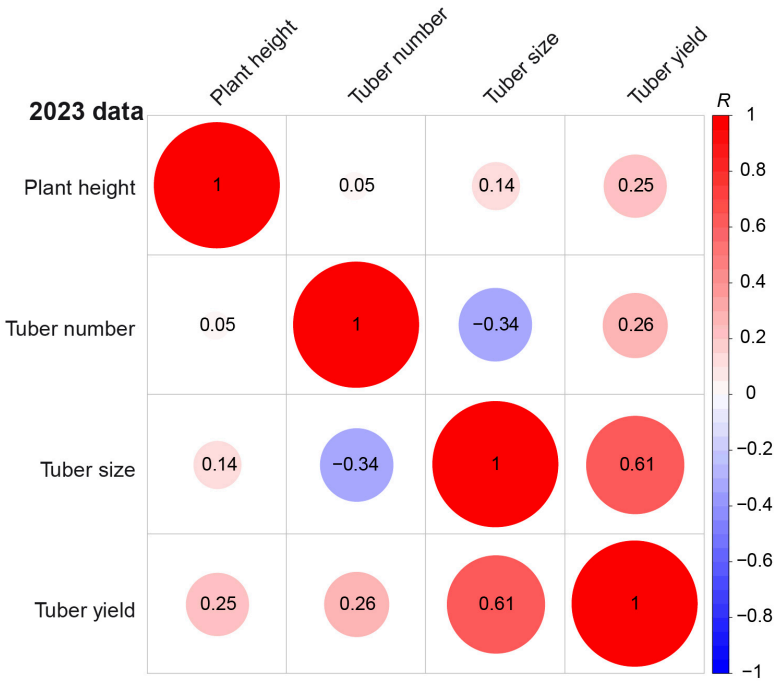
Supplementary Fig. 1. Population analysis by whole-genome sequencing.
a, The genotype map of the F_2 population (n = 1,064). The horizontal axis indicates the 12 chromosomes (columns). The vertical axis presents genotypic data for the 1,064 individuals. A/A, the A6-26 homozygous genotype (A6-26/A6-26); A/E, the heterozygous genotype (A6-26/E4-63); E/E, the E4-63 homozygous genotype (E4-63/E4-63). **b,** The genetic map of the F_2 population; black horizontal lines represent bins. **c,** Segregation distortion analysis of zygotes (red) and gametes (blue). The y axis represents the $-\log_{10}P$ and $\log_{10}P$ for the χ^2 value of each bin. Source data are provided as a Source Data file.



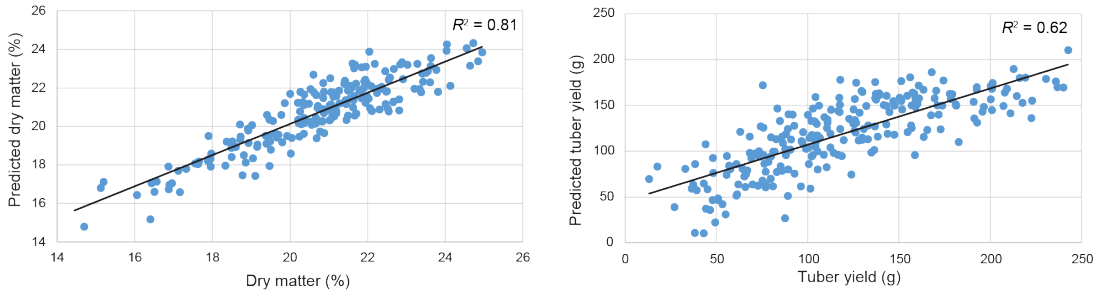
Supplementary Fig. 2. Image processing and image-traits (i-traits) extraction. **a.** Phenomics image acquisition process: Potatoes are grown in a greenhouse and detected using a crop plant phenotyping system, including RGB acquisition system, hyperspectral acquisition system, structured light acquisition system. Side-view (SV) and top-view (TV) images are collected at the 60 days after transplanting. Near-infrared (NIR) and visible light (VIS) hyperspectral data, as well as 3D structured light point cloud data, are collected for the potato tubers. **b.** I-traits extraction process. The image analysis performed for i-trait extraction. It includes image preprocessing in the RGB, HSI, and 3D directions, grayscale threshold segmentation, point cloud extraction, and trait extraction work. **c,** Side-view image, the original acquired image. **d,** Cropped image, removal of background impurities. **e,** Segmented binary image, threshold segmentation is transformed into a binary graph. **f,** Convex hull image, computational morphological characteristics. **g,** Mask image, final segmentation effect. **h,** Gray image (Intensity) for calculating the gray histogram texture feature.



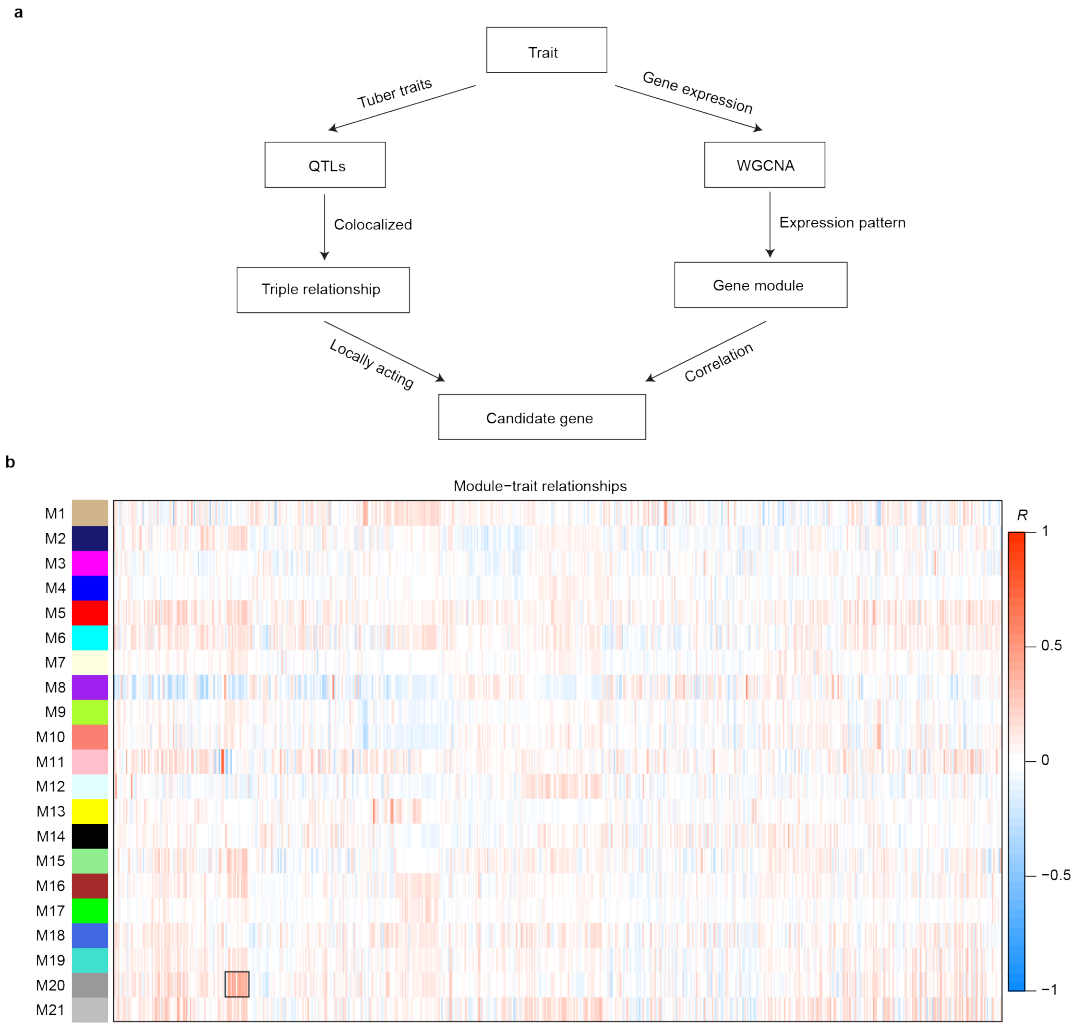
Supplementary Fig. 3. Bulked-segregant analysis of four qualitative traits. **a,** Tuber flesh color. **b,** Tuber shape. **c,** Tuber bud color. **d,** Yellow leaf. The vertical axis indicates the delta SNP index (the difference in allele frequency between the two pools), and the horizontal axis represents the 12 chromosomes (0–12), partitioned by vertical dashed lines. The cutoff line (horizontal dashed line) is set to 0.5. Arrows indicate the positions of candidate genes. Scale bars, 1 cm. Source data are provided as a Source Data file.



Supplementary Fig. 4. Correlation of tuber yield-related traits in 2023.
The first-row concerns tuber size, indicated by an arrow. Tuber size shows a pattern similar to that of total reflectance of near-infrared. Tuber size is positively correlated with total reflectance of near-infrared and negatively correlated with average reflectance of visible light. Red indicates positive correlations (*R*), and blue indicates negative correlations. Source data are provided as a Source Data file.

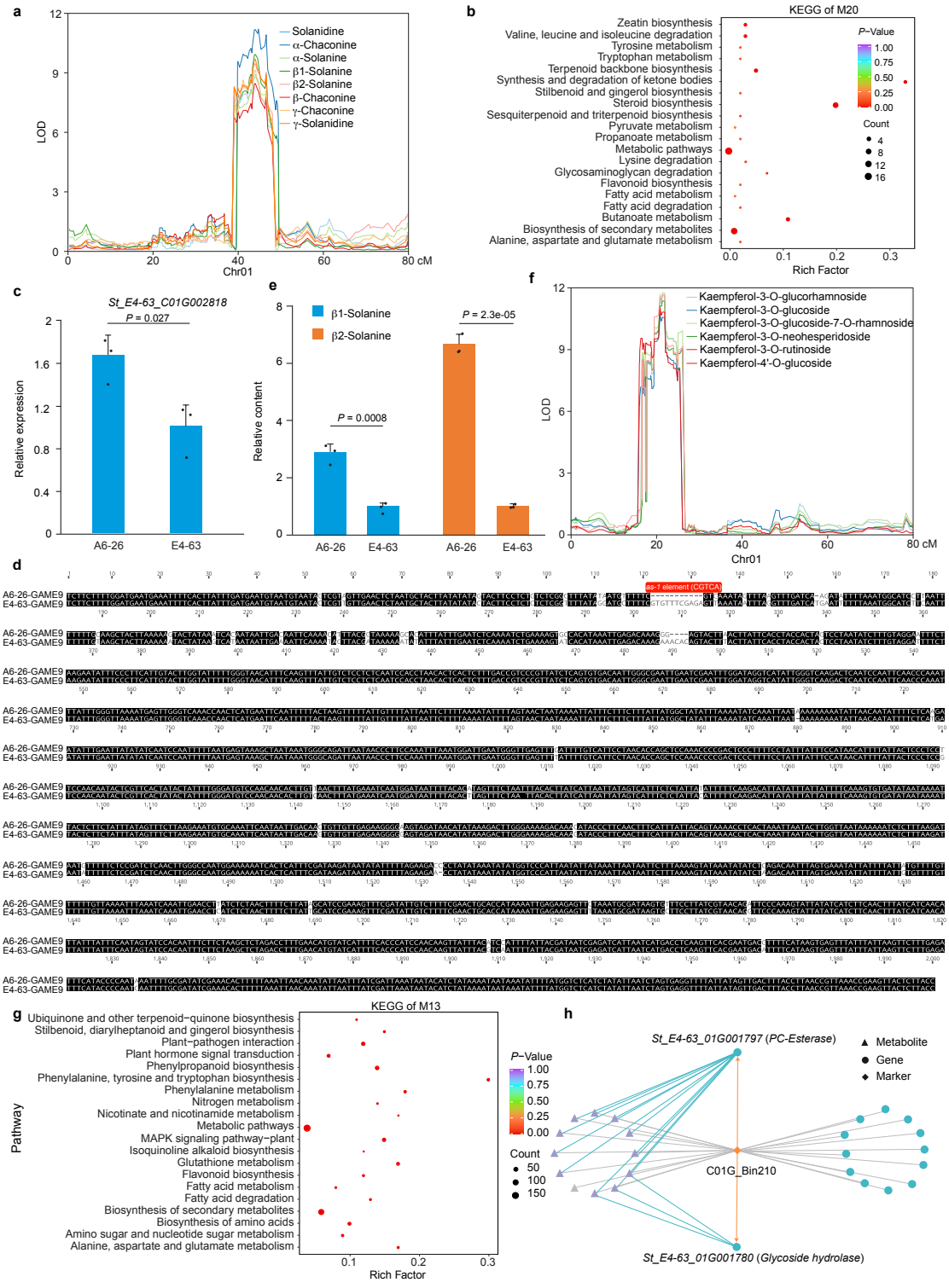


Supplementary Fig. 5. The prediction of tuber dry matter and tuber yield.
The prediction of tuber dry matter and tuber yield using tuber reflectance via stepwise linear regression analysis. Prediction accuracy was evaluated by the square of correlation coefficients (R^2) between the observed values and the predicted values. Source data are provided as a Source Data file.



Supplementary Fig. 6. Flow diagram of potato systems genetics and WGCNA analysis.

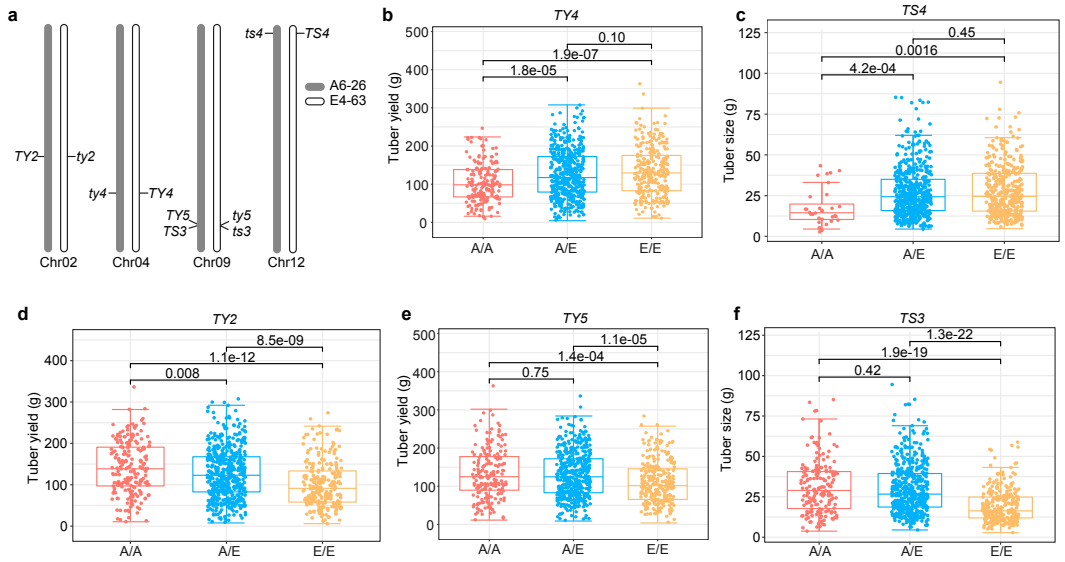
a, Flow diagram of potato systems genetics. **b,** Module–trait relationships generated by WGCNA. The black rectangle (lower left) indicates correlation of M20 and solanines. The horizontal axis presents different traits (metabolites and yield-related traits). Red indicates positive correlations, and blue indicates negative correlations (*R*).



Supplementary Fig. 7. The power of systems genetics in identifying candidate genes.

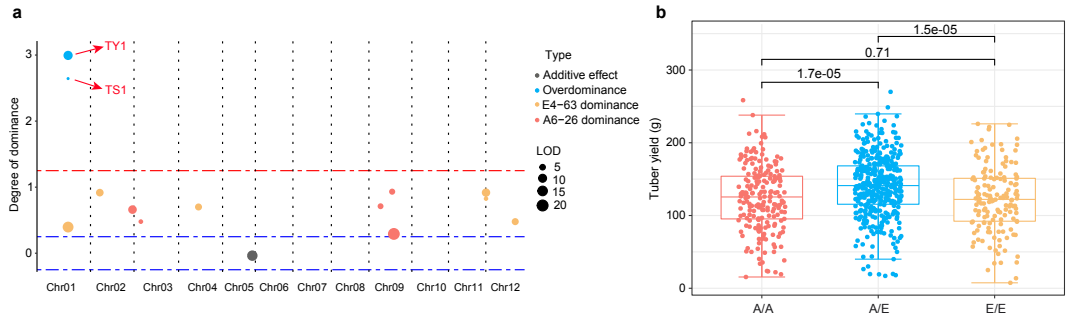
a, QTLs of dimensionality reduction solanines. **b,** KEGG analysis of genes in module M20. The horizontal axis (Rich Factor) quantifies the degree of enrichment. **c,** The expression level of *GAME9* in tubers of the parental lines. *P* values obtained by Student's *t* tests (two-tailed) are indicated. Data are presented as mean values +/– SD. *n* = 3. **d,** DNA sequence alignment of the 2000-bp promoter region of *GAME9* in the

two parental lines. The *as-1* element is indicated as a red box. **e,** The content of solanines in tubers of the parental lines. *P* values obtained by Student's *t* tests (two-tailed) are indicated. Data are presented as mean values +/– SD. $n = 3$. **f,** QTLs of different kinds of flavonoids on chr01. **g,** KEGG analysis of genes in module M13. The horizontal axis (Rich Factor) quantifies the degree of enrichment. **h,** The regulatory network of two candidate genes associated with flavonoid accumulation. The orange arrows indicate local regulation, and the purple triangles indicate flavonoids. Source data are provided as a Source Data file.



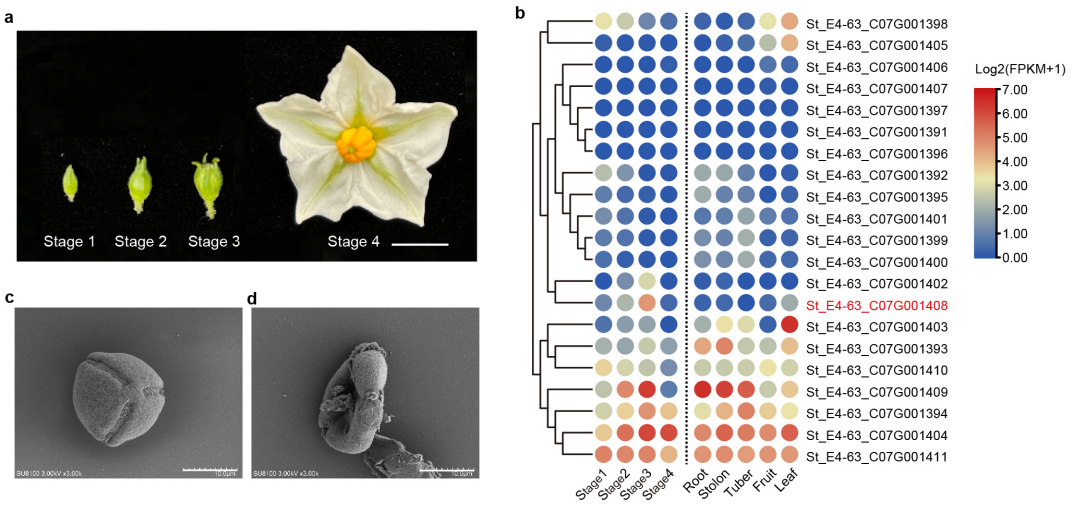
Supplementary Fig. 8. The complementation of tuber yield-related QTLs in F_1 hybrids.

a, The position of five dominant QTLs at the whole-genome scale. **b–f,** The boxplots present the trait values of these five dominant QTLs for tuber yield. *TY4* (b, n = 176, 488 and 274 for A/A, A/E and E/E, respectively.) and *TS4* (c, n = 32, 503 and 309 for A/A, A/E and E/E, respectively.) are contributed by E4-63, while *TY2* (d, n = 207, 495 and 241 for A/A, A/E and E/E, respectively.), *TY5* (e, n = 198, 461 and 269 for A/A, A/E and E/E, respectively.) and *TY3* (f, n = 186, 451 and 267 for A/A, A/E and E/E, respectively.) are contributed by A6-26. *P*-values were obtained by Student's *t*-tests (two-tailed). The upper and lower edges of the boxes denote 75% and 25% quartiles, and the central line indicates the median. Whiskers extend to the lower hinge –1.5× interquartile range and upper hinge +1.5× interquartile range of the data (b–f). Source data are provided as a Source Data file.



Supplementary Fig. 9. The heterotic effects of yield-related QTLs in 2023. **a,** The heterotic effects of yield QTLs (plant height, tuber yield, tuber number and tuber size) in 2023. The y axis indicates *d*/*a* values; *d*/*a* values > 3 are displayed as 3. The red dotted line indicates 1.25, and the dotted blue lines indicate ±0.25. **b,** The yield of different genotypes of *TY1* in 2023. A/A, the A6-26 homozygous genotype (A6-26/A6-26); A/E, the heterozygous genotype (A6-26/E4-63); E/E, the E4-63 homozygous genotype (E4-63/E4-63). n = 197, 381 and 148 for A/A, A/E and E/E, respectively. The upper and lower edges of the boxes denote 75% and 25% quartiles, and the central line indicates the median. Whiskers extend to the lower hinge –1.5× interquartile range and upper hinge +1.5× interquartile range of the data. *P*-values were obtained by Student's *t*-tests (two-tailed). Source data are provided as a Source Data file.

Supplementary Fig. 10. The genetic effects of gene expression and metabolites. **a,** The proportion of different genetic effects of eQTLs. **b,** KEGG analysis of genes with positive eQTLs. **c,** The genetic effects of mQTLs. The vertical axis indicates *d*/*a* values; *d*/*a* values > 3 (and < –3) are displayed as 3 (–3). The red dotted lines indicate ±1.25 and the dotted blue lines indicate ±0.25. **d,** The proportion of different genetic effects of primary mQTLs. **e,** The proportion of different genetic effects of secondary mQTLs. **f,** The proportion of different genetic effects of mQTLs using the quantity of dry matter. In parts a and d–f, numbers in parentheses show QTL numbers and percentages, respectively. Source data are provided as a Source Data file.



Supplementary Fig. 11. The candidate-gene identification of *PV1*.
a, The four developmental stages of anther used for transcriptome sequencing. **b,** The expression level of genes located in the mapping interval in different tissues. **c,** Scanning electron microscopy of WT pollen grains. **d,** Scanning electron microscopy of KO-2 pollen grains. Source data are provided as a Source Data file.

Supplementary Method 1. Potato image processing and image based digital trait (i-trait) extraction

The RGB data processing program

Program running environment: Windows 64 bit, LabVIEW2015 64 bit

In the greenhouse experiment, a randomized block design with three replicates is implemented. All plants are grown under uniform conditions: one plant per pot with standardized soil and fertilizer and irrigation.

1. Image collection process

The collection system includes a conveyor, a rotating platform, a controller, an industrial camera (Basler aca5472) and a computer. Plants are transported to the rotating platform via the conveyor. The controller rotates the platform 360 degrees, capturing six side-view and one top-view photos for each plant. The image resolution is 3244 × 2064 pixels. A total of 17 rounds of photos are taken, resulting in 94,962 images.

2. Trait extraction process

The steps for image processing are as follows: (1) flip and crop the original images to remove background noise. (2) apply EG and HSL segmentation methods to generate a binary image and calculate morphological features. (3) use the OpenCV dynamic link library to obtain the convex hull image and compute traits related to the convex hull. (4) process the binary image with the original image mask to extract color-related parameter features. (5) extract the grayscale image of the I (Intensity) components and calculate the texture features based on the gray histogram.

3. Filtering outliers

Measurements with a coefficient of variation greater than 15% are considered outliers and removed. The coefficient of variation is defined as the ratio of the standard deviation to the mean. A 15% deviation threshold is used to filter gross error.

The hyperspectral data processing program

Program running environment: Windows 64 bit, LabVIEW2015

1. Spectral image acquisition

The hyperspectral imaging system includes a near-infrared hyperspectral camera (1000–1700 nm, Headwall Hyperspec Starter Kit-VNIR, USA), a visible light hyperspectral camera (400–1000 nm, Headwall VNIR A-Series, USA), a halogen light source and a translation stage. The near-infrared hyperspectral camera covers 172 bands with a spectral band interval of 4.65 nm, while the visible light hyperspectral camera covers 314 bands with an interval of 0.635 nm. Tubers are placed on the translation

stage for data collection, which moves at a speed of 2 mm/s in a darkroom, with only the halogen light operating during collection.

2. Spectral image calibration

Before data collection, black-and-white reference calibration is required. For dark current calibration, a frame is captured with the lens covered. For white reference calibration, a standard white board is placed under the lens, and a frame is captured. The corrected grayscale value (R_0) is calculated using the following formula:

$$R_0=(\frac{R-R_d}{R_w-R_d}) \quad (1)$$

where R is the original image, R_d is the dark current correction, and R_w is the white reference correction. Calibration is repeated every hour to ensure data stability.

3. Spectral image extraction

The OTSU algorithm is used for image segmentation, determining an optimal threshold that maximizes inter-class variance. This allows segmentation of the potato tubers, generating a binary image for spectral analysis.

4. Spectral trait extraction

Using the binary image and the masked spectral image, we calculate spectral parameters including total reflectance (T), average reflectance (A), first derivative (dA), second derivative (ddA), and logA.

5. Stepwise linear regression

Stepwise regression is used to identify the most predictive spectral parameters for dry matter, yield, and tuber metabolites. Redundant parameters are removed, and sets of predictive spectral features are identified for each dependent variable.

6. Model evaluation

The model's accuracy and stability are assessed using 10-fold cross-validation. The dataset is randomly divided into 10 groups, with nine used for training and one for testing. The average of these 10 models is taken as the cross-validation result. Metrics such as root mean square error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R^2) are calculated for both the training and testing sets.

The structured light data processing program

Program running environment: Windows 64 bit, CloudCompare V2.1.2beta, Visual studio 2019

The system includes a Reeyee Pro structured light scanner, a rotating platform, a controller and a computer. Potato tubers are placed on the platform, which rotates for three minutes per scan.

1. Scanner calibration

The structured light scanner is calibrated before use. A calibration board is positioned at four angles (0°, 90°, 180°, 270°), with five images captured at each position. The distance between the scanner and the board is adjusted (350 mm–450 mm) for optimal calibration at each angle.

2. Potato image acquisition

Potato tubers are placed on the platform at a fixed working distance of 340 mm. Multiple exposure levels are used to minimize ambient light interference.

3. Point cloud segmentation

The tuber point cloud is separated from the platform using down-sampling, coordinate calibration and filtering techniques. The RANSAC plane fitting method is applied to isolate tubers, after which individual tubers are clustered for phenotypic measurement.

4. Trait extraction

A 3D phenotypic measurement algorithm is designed using the PCL and VTK libraries to extract parameters such as length, width, height, volume, surface area, projected area and perimeter of the tubers.

5. Data correlation comparison

Pearson correlation analysis is conducted between manual measurements of length, width, and height and those derived from structured light scanning. The R-values for length, width, and height are 0.997, 0.976, and 0.953, respectively.

Supplementary Method 2. Definition of the macro-phenomics traits

Near-infrared red

1. T_I1~T_I172: total reflectance under different wavelengths.

$$T\_I_n = \sum R_n(i,j) \qquad (2)$$

where $R_n(i, j)$ is the reflectance of coordinate (i, j) under n wavelength, $(i, j) \in$ plant ROI, n = 1, 2, ……, 172.

2. dT_I1~dT_I172: the first-order derivative of the total reflectance under different wavelengths.

$$dT\_I_n = \frac{1}{2dt}(T\_I_{n+1} - T\_I_{n-1}) \qquad (3)$$

n = 2, 3, ……, 171.

3. ddT_I1~ddT_I172: the second-order derivative of the total reflectance under different wavelengths.

$$ddT\_I_n = \frac{1}{2dt}(dT\_I_{n+1} - dT\_I_{n-1}) \quad \textbf{(4)}$$

n = 2, 3, ……, 171.

4. A_I1~A_I172: average reflectance under different wavelengths.

$$A\_I_n = \frac{\sum R_n(i,j)}{Area} \quad \textbf{(5)}$$

where $R_n(i, j)$ is the reflectance of coordinate (i, j) under n wavelength, $Area$ is the area of plant ROI, $(i, j) \in$ plant ROI, n = 1, 2, ……, 172.

5. dA_I1~dA_I172: the first-order derivative of the average reflectance under different wavelengths.

$$dA\_I_n = \frac{1}{2dt}(A\_I_{n+1} - A\_I_{n-1}) \quad \textbf{(6)}$$

n = 2, 3, ……, 171.

6. ddA_I1~ddA_I172: the second-order derivative of the average reflectance under different wavelengths.

$$ddA\_I_n = \frac{1}{2dt}(dA\_I_{n+1} - dA\_I_{n-1}) \quad \textbf{(7)}$$

n = 2, 3, ……, 171.

7. lgT_I1~lgT_I172: the logarithm of the total reflectance under different wavelengths.

$$lg\, T\_I_n = log_{10}(T\_I_n) \quad \textbf{(8)}$$

n = 1, 2, ……, 172.

8. lgA_I1~lgA_I172: the logarithm of the average reflectance under different wavelengths.

$$lg\, A\_I_n = log_{10}(A\_I_n) \quad \textbf{(9)}$$

n = 1, 2, ……, 172.

Visible light

1. T1~T314: total reflectance of the plant under different wavelengths.

$$T_n = \sum R_n(i,j) \quad \textbf{(10)}$$

where $R_n(i, j)$ is the reflectance of coordinate (i, j) under n wavelength, $(i, j) \in$ plant ROI, n = 1, 2, ……, 314.

2. dT1~dT314: the first-order derivative of the total reflectance under different wavelengths.

$$dT_n = \frac{1}{2dt}(T_{n+1} - T_{n-1}) \quad \textbf{(11)}$$

n = 2, 3, ……, 313.

3. ddT1~ddT314: the second-order derivative of the total reflectance under different wavelengths.

$$ddT_n = \frac{1}{2dt}(dT_{n+1} - dT_{n-1}) \quad \textbf{(12)}$$

n = 2, 3, ……, 313.

4. A1~A314: average reflectance under different wavelengths.

$$A_n = \frac{\sum_{R_n(i,j)}}{Area} \quad \textbf{(13)}$$

where $R_n(i, j)$ is the reflectance of coordinate (i, j) under n wavelength, $Area$ is the area of plant ROI, $(i, j) \in$ plant ROI, n = 1, 2, ……, 314.

5. dA1~dA314: the first-order derivative of the average reflectance under different wavelengths.

$$dA_n = \frac{1}{2dt}(A_{n+1} - A_{n-1}) \quad \textbf{(14)}$$

n = 2, 3, ……, 313.

6. ddA1~ddA314: the second-order derivative of the average reflectance under different wavelengths.

$$ddA_n = \frac{1}{2dt}(dA_{n+1} - dA_{n-1}) \quad \textbf{(15)}$$

n = 2, 3, ……, 313.

7. lgT1~lgT314: the logarithm of the total reflectance under different wavelengths.

$$lg\, T_n = log_{10}(T_n) \quad \textbf{(16)}$$

n = 1, 2, ……, 314.

8. lgA1~lgA314: the logarithm of the average reflectance under different wavelengths.

$$lg\, A_n = log_{10}(A_n) \quad \textbf{(17)}$$

n = 1, 2, ……, 314.

RGB

1. GPA (green projected area): number of foreground pixels attributed by green part.
2. YPA (yellow projected area): number of foreground pixels attributed by yellow part.
3. GPAR (green projected area ratio): GPA/TPA. GPA is the number of foreground pixels attributed by green part of plant.
4. YTR (the ratio of yellow projected area and total projected area): YPA/TPA.
5. PC1–PC6 (plant compactness): the image is divided into several smaller sub-images using a (5 × 5) window. Each sub-image consists of 25 pixels. For each sub-image, calculate the ratio of foreground pixels (representing the plant) to the total number of pixels (25 pixels). This ratio is defined as plant compactness (PCs) for the sub-image and are categorized into six classes based on the percentage of foreground pixels in each sub-image: C1: <10%, C2: 10-20%, C3: 20-40%, C4: 40-60%, C5: 60-80%, C6: 80-100%. Then count the number of PCs that fall into each class, denoted as NDi (i=1,2…6). Leaf compactness of class i (PCi) is computed as the percentage of NDi compared to the sum of NDi.
6. P: plant perimeter in a side view.
7. PAR (perimeter/projected area ratio): calculate the outline length of the plant, then divide the length by projected area.
8. SW: width of the bounding rectangle of the object.
9. SFD: superimpose boxes with box size of δ_k on the interested object, and calculate the number of boxes that are needed to cover the object, denoted as N_{δ_k}. Repeat this process with reducing δ_k until δ_k approaches pixel size. Fractal dimension is calculated using the following equation.

$$FD = \lim_{\delta_k \to 0} \frac{\ln N_{\delta_k}}{-\ln \delta_k} \quad (18)$$

10. TBR: total projected area/bounding rectangle area ratio.
11. CBA: circumscribed box area.
12. TPA (total projected area): number of foreground pixels attributed by plant.
13. HWR (height/width ratio): the six histogram traits, including the mean value (M_I) the standard error (SE_I), the third moment (MU3_I), the uniformity (U_I), the smoothness (S_I) and the entropy (E_I) are calculated using the following equations:

$$M\_I = \sum_{i=0}^{L-1} G_i \times p(G_i) \quad (19)$$

$$SE\_I=\sqrt{\sum_{i=0}^{L-1}(G_i-M\_I)^2 p(G_i)} \quad (20)$$

$$MU3\_I=\sum_{i=0}^{L-1}(G_i-M\_I)^3 p(G_i) \quad (21)$$

$$U\_I=\sum_{i=0}^{L-1}p^2(G_i) \quad (22)$$

$$S\_I=1-\frac{1}{1+SE\_I} \quad (23)$$

$$E\_I=\sum_{i=0}^{L-1}p(G_i)\times log_2 p(G_i) \quad (24)$$

Where G_i is the i-th gray level, and $p(G_i)$ is the probability of G_i. L is the maximum gray level.

Structured light

1. TL (tuber length): X axis point cloud maximum minus minimum value.
2. TW (tuber width): Y axis point cloud maximum minus minimum value.
3. TH (tuber height): Z axis point cloud maximum minus minimum value.
4. TSF (tuber surface): the surface of the grain model, reconstructed in three dimensions, consists of multiple triangular facets. The area of each triangular facet can be calculated using Heron's formula, based on the coordinates of its three vertices. The total surface area of the tuber is then determined by summing the areas of all these triangular facets.
5. TV (tuber volume): the three-dimensional model of the tuber is a closed space formed by triangular facets. Its volume can be computed as the volume integral of the convex polyhedron, projected onto the central plane of each triangular facet.