GENÉTICA QUANTITATIVA E MELHORAMENTO DE PLANTAS

2025/2026

Introdução aos estudos de associação genómica (GWAS) e seleção genómica

Elsa Gonçalves, dezembro 2025

Marcador genético

- Um marcador genético é um gene ou sequência de DNA com uma localização conhecida num cromossoma que pode ser utilizado para identificar indivíduos ou espécies. Pode ser descrito como uma variação que pode ser observada (que pode surgir devido à mutação ou alteração dos *loci* genómicos).

Alguns tipos de marcadores genéticos:

SNP (Single nucleotide polymorphism);

RFLP (Restriction fragment length polymorphism); SSLP (Simple sequence length polymorphism); AFLP (Amplified fragment length polymorphism); RAPD (Random amplification of polymorphic DNA); VNTR (Variable number tandem repeat); SSR Microsatellite polymorphism (Simple sequence repeat); STR (Short tandem repeat); SFP (Single feature polymorphism); DArT (Diversity Arrays Technology), RAD markers (Restriction site associated DNA markers).

3

Um marcador genético pode ser uma sequência curta de DNA, **tal como uma sequência em torno de uma única alteração do par de bases (polimorfismo de nucleótidos simples, SNP).**

Referência	GCAACGTTAGA
Ind 1	GCAACGTTAGA
Ind 2	GCAGCGTTAGA
Ind 3	GCATCGTTAGA

⇓
SNP

SNP - substituição de um único nucleótido que ocorre numa posição específica no genoma, onde cada variação está presente numa determinada percentagem na população (por exemplo, >1%).
Por exemplo, numa posição específica no genoma, o nucleótido C pode aparecer na maioria dos indivíduos, mas numa minoria de indivíduos a posição é ocupada por um A. Isto significa que existe um SNP nesta posição específica, e as duas possíveis variações de nucleótidos - C ou A - são ditas como sendo os alelos para esta posição específica.

4

APLICAÇÕES

- **Estudos de associação genómica** (**GWAS**, genome-wide association), é um estudo baseado num conjunto de variantes genómicas em diferentes indivíduos para ver se alguma variante está associada a uma característica. Os estudos GWAS centram-se normalmente em associações entre polimorfismos de um único nucleótido (SNP) e características importantes.

- Estimar a quantidade de genoma partilhado por indivíduos numa população para estudar a semelhança entre indivíduos. Juntando esta informação com os dados fenotípicos é abordado o tema da **seleção genómica.**

5

Os dados de marcadores moleculares não estão geralmente prontos para análise. Quando os dados genotípicos são obtidos, a primeira tarefa é organizar, sumarizar e reformatar os dados para subsequente análise. Existem muitos packages no R para fazer esta preparação (por exemplo: vcf.R, snpReady, Synbreed)

1. Uma situação comum é receber resultados dos SNPs como pares de bases. Vejamos um exemplo para uma amostra de indivíduos diploides:

Marker ID	IND1	IND2	IND3	IND4	IND5	IND6	…
01-256	GG	GG	GG	GG	GG	GG	GG
01-71	AC	AA	AA	CC	AA	AA	AA
01-559	CC	CC	CC	CC	CC	CC	CC
01-431	GG	AG	GG	0	GG	AG	AG
…	AA	GG	GG	0	GA	GG	GG

- Por exemplo, o IND1 é homozigótico (GG) no *locus* 01-256, mas heterozigótico (AC) no *locus* 01-71.
- 0, ou também NA ou espaço em branco, são dados genotípicos em falta.
- Todos os indivíduos da população estudada são homozigóticos CC no *locus* 01-559 (monomórfico). Os *loci* monomórficos são retirados antes da análise.

6

2. O passo seguinte na preparação destes dados é a identificação do alelo secundário (minor allele) e do alelo principal (major allele) em cada locus.
Por definição, o alelo secundário tem uma frequência menor que 0.5.

Transpondo as linhas e colunas da matriz anterior é fácil identificar **o alelo secundário (minor allele, MA)**

	MA=A	MA=C	MA=.	MA=G	MA=A
MARKER ID	**01-256**	**01-71**	**01-559**	**01-431**	**…**
IND1	GG	AC	CC	GG	AA
IND2	GG	AA	CC	AG	GG
IND3	GG	AA	CC	GG	GG
IND4	GG	CC	CC	NA	NA
IND5	GG	AA	CC	GG	GA
IND6	GG	AA	CC	AG	GG
…	GG	AA	CC	AG	GG

- Não há alelo secundário para o *locus* 01-559. Embora não vejamos o alelo A para o *locus* 01-256 na pequena amostra de 6 indivíduos que estamos a visualizar no quadro, o alelo menor para este *locus*, A, encontra-se em alguns indivíduos não incluídos no quadro.

7

3. Seguidamente os elementos da matriz são escritos numericamente em função do alelo secundário:

- Um individuo que é homozigótico para o alelo secundário é codificado como 2;
- Um individuo heterozigótico, tem uma cópia do alelo secundário, é codificado como 1;
- Um individuo homozigótico para o alelo principal tem zero cópias para o alelo secundário e é codificado como 0.

A matriz anterior fica:

	MA=A	MA=C	MA=.	MA=G	MA=A
MARKER ID	01-256	01-71	01-559	01-431	…
IND1	GG	AC	CC	GG	AA
IND2	GG	AA	CC	AG	GG
IND3	GG	AA	CC	GG	GG
IND4	GG	CC	CC	NA	NA
IND5	GG	AA	CC	GG	GA
IND6	GG	AA	CC	AG	GG
…	GG	AA	CC	AG	GG

MARKER ID	**01-256**	**01-71**	**01-559**	**01-431**	**…**
IND1	0	1	0	2	2
IND2	0	0	0	1	0
IND3	0	0	0	2	0
IND4	0	2	0	NA	NA
IND5	0	0	0	2	1
IND6	0	0	0	1	0
…	0	0	0	1	0

Esta matriz é frequentemente designada por matriz M

°

Exemplificando outra forma para a matriz M, com um pequeno exemplo:

	Locus 1	Locus 2	Locus 3	Locus 4
Ind1	0	1	0	2
Ind2	2	1	1	1
Ind3	2	0	0	0

Vamos escrever a **matriz M** com uma nova notação, subtraindo 1 a todos os elementos, ficando-se com **-1, 0, 1**

	Locus 1	Locus 2	Locus 3	Locus 4
Ind1	-1	0	-1	1
Ind2	1	0	0	0
Ind3	1	-1	-1	-1

Este tipo de matriz também é muito usada nos estudos de associação genómica e é o ponto de partida para a construção da matriz de relações entre indivíduos ($\mathbf{G_A}$) usada nos modelos para seleção genómica (e em alguns de GWAS).

9

Modelo para estudo de associação genómica (GWAS)

A análise de associação a nível do genoma é baseada no modelo linear misto (em notação matricial):

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zg} + \boldsymbol{M\tau} + \boldsymbol{e}$$

$\boldsymbol{Y}_{(n\times 1)}$ é o vector das observações (vector dos valores fenotípicos)

$\boldsymbol{X}_{(n\times p)}$ é a matriz de delineamento dos efeitos fixos

$\boldsymbol{\beta}_{(p\times 1)}$ é o vector de efeitos fixos (por exemplo, vector dos efeitos do local, de efeitos associados aos delineamento experimental)

$\boldsymbol{Z}_{(n\times q)}$ é a matriz de delineamento dos efeitos aleatórios

$\boldsymbol{u}_{(q\times 1)}$ é o vector de efeitos aleatórios (vector dos efeitos genotípicos/efeitos genéticos aditivos, de efeitos associados ao delineamento experimental, etc.)

$\boldsymbol{M}$ $(n \times m)$ é uma matriz de marcadores numéricos.

$\boldsymbol{\tau}$ $(m \times 1)$ é o vector dos efeitos dos SNPs (efeitos fixos).

$\boldsymbol{e}_{(n\times 1)}$ é o vector de erros aleatórios

10

- **Exemplo:**

Estudo de associação em Sobreiro; dados fenotípicos da % de suberina na cortiça

Vários packages podem ser usados, por exemplo, *R:*

- *Package sommer*
- *Package ASRgwas*

Locus	Chrom	Position	suberina beta	p-value
locus_824692_55	824692	55	-0.628176898	0.002239
locus_775409_79	775409	79	-0.495213993	0.007894
locus_863500_52	863500	52	-0.558417685	0.009591
locus_713513_61	713513	61	-0.55612366	0.010868
locus_535704_2	535704	2	-0.588855873	0.011051
locus_728960_21	728960	21	-0.432428827	0.012303
locus_848371_26	848371	26	-0.601878818	0.012374
locus_642246_74	642246	74	-0.452600034	0.012808
locus_23702_47	23702	47	-0.435878976	0.015473
locus_824692_57	824692	57	-0.525046439	0.015685
locus_809218_43	809218	43	0.428140818	0.016117
locus_755528_11	755528	11	-0.441133408	0.01647
locus_1066716_31	1066716	31	-0.422888902	0.017849
locus_757108_75	757108	75	0.509398602	0.017919
locus_628260_64	628260	64	-0.481892416	0.025854
locus_590986_67	590986	67	0.444848078	0.027683
locus_923340_11	923340	11	0.430509268	0.028209
locus_535704_33	535704	33	-0.55653351	0.028644
locus_713518_71	713518	71	-0.413019201	0.029462
locus_179052_74	179052	74	0.424631342	0.030052
locus_671731_12	671731	12	-0.483981125	0.032355
locus_1006236_64	1006236	64	0.338746874	0.033167
locus_847026_63	847026	63	-0.368370775	0.037485
locus_503663_21	503663	21	0.412646504	0.038518
locus_770764_67	770764	67	-0.412240207	0.042735
locus_822294_54	822294	54	-0.407880878	0.043189
locus_825747_45	825747	45	0.401194201	0.044692
locus_353499_25	353499	25	-0.351495203	0.04644
locus_800184_26	800184	26	0.35096151	0.04682
locus_976133_63	976133	63	-0.354794191	0.047018
locus_993441_73	993441	73	0.331044395	0.048661
locus_773658_53	773658	53	0.386053505	0.048999
locus_168944_13	168944	13	0.402227359	0.051144
locus_144690_46	144690	46	0.399427172	0.052529
locus_998793_81	998793	81	-0.530724991	0.054277

...

Nota: Como há vários testes de hipóteses e pretende-se controlar o nível de significância global, é muito comum aplicar a correção de Bonferroni ($\alpha*=\alpha/r$, sendo r o número de marcadores). No entanto, essa correção tem a desvantagem de ser excessivamente conservadora quando r é muito grande.

11

***Manhattan Plot*, é normalmente utilizado em estudos de associação genómica para mostrar os SNPs significativos.**

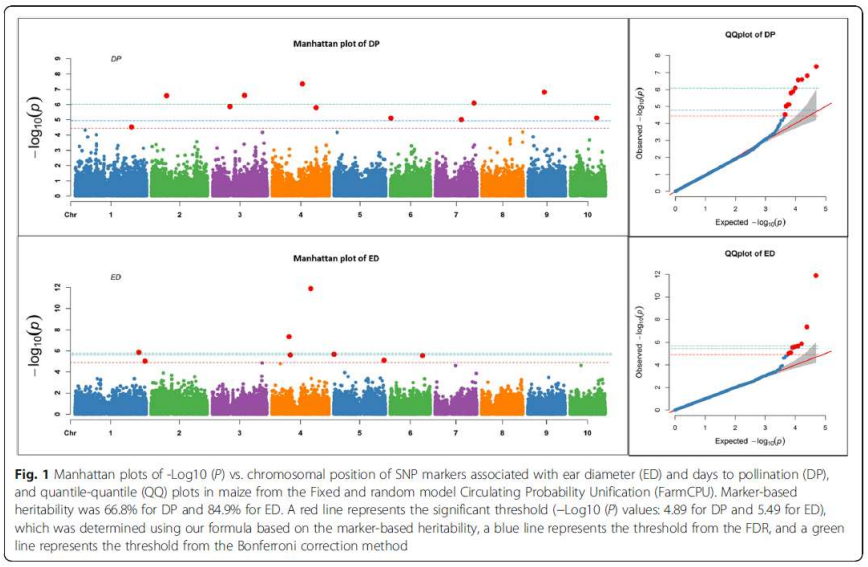


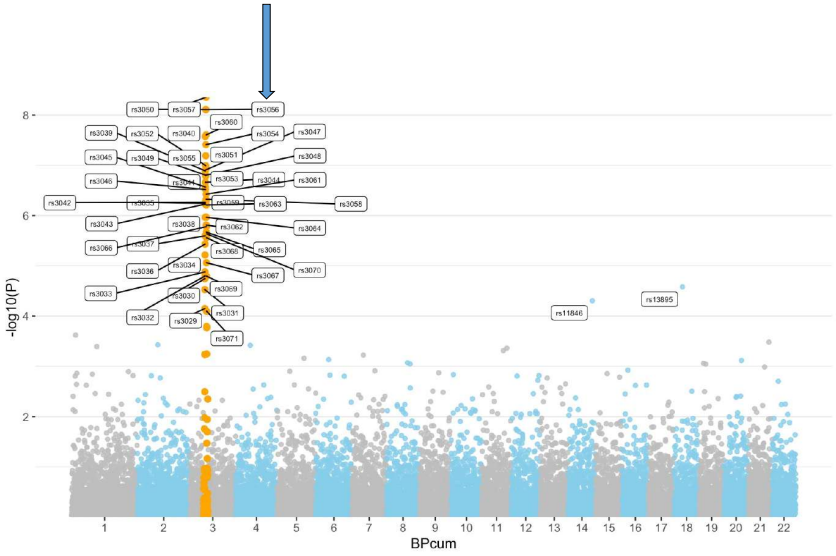
Fig. 1 Manhattan plots of -Log10 (*P*) vs. chromosomal position of SNP markers associated with ear diameter (ED) and days to pollination (DP), and quantile-quantile (QQ) plots in maize from the Fixed and random model Circulating Probability Unification (FarmCPU). Marker-based heritability was 66.8% for DP and 84.9% for ED. A red line represents the significant threshold (–Log10 (*P*) values: 4.89 for DP and 5.49 for ED), which was determined using our formula based on the marker-based heritability, a blue line represents the threshold from the FDR, and a green line represents the threshold from the Bonferroni correction method

Kaler and Purcell *BMC Genomics* (2019) 20:618
https://doi.org/10.1186/s12864-019-5992-7

12

Exemplo de estudo de associação genómica (GWAS)

Manhattan Plot, com identificação dos SNPs significativos



13

Seleção Genómica (GS)

Tradicionalmente as análises genéticas combinam os dados fenótipos e o grau de semelhança entre parentes para prever o mérito genético dos indivíduos. Mais recentemente, a informação de marcadores moleculares tem sido usada para medir a semelhança genética entre indivíduos.

- **As relações genómicas derivadas de marcadores moleculares para predição do mérito genético dos indivíduos tem ganho muita atenção nos últimos anos no melhoramento animal e vegetal**

A seleção genómica requer uma matriz de relações genómicas (semelhança entre indivíduos)

A matriz das relações genómicas estimadas a partir de informação de marcadores moleculares é designada por matriz $\boldsymbol{G_A}$. O cálculo desta matriz requer os métodos de preparação dos dados referidos anteriormente.

14

Seleção Genómica (GS)

Construir modelos de previsão utilizando os dados fenotípicos da população de referência e marcadores moleculares que captam a maior parte da variação quantitativa.

Para fazer seleção genómica é necessário:

- **Um ensaio** com muitos genótipos/famílias da espécie/variedade a selecionar (uma amostra representativa da variabilidade genética)

População de referência
(Training Population)
Obtêm-se dados fenotípicos de todos os indivíduos e os dados genómicos para todos os indivíduos

- **Novos genótipos em que apenas se dispõe de dados genómicos (prediction population).** Estes são incluídos numa análise em que é incorporada a informação da população de referência. Para cada novo genótipo faz-se uma previsão do seu valor genético para determinada característica da qual existem dados na população de referência. Para estes genótipos obtêm-se os *Genomic BLUPS* (GBLUPS).

15

Existem muitos métodos para fazer seleção genómica

Alguns exemplos:

- BLUP-Based: G-BLUP, RR-BLUP
- Bayes-Based: Bayes A, Bayes B, Bayes Cπ, Bayes RR
- LASSO-Based: Bayesian Lasso, Improved Lasso
- Semi-Parametric Regression: RKHS
- Non-Parametrics: Suport Vector Machine, Neural-Networks

Vários Packages no R, por exemplo: ASRgenomics

16

Agora, em notação matricial, o modelo linear misto pode ser escrito genericamente da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

$\boldsymbol{Y}_{(n\times 1)}$ é o vector das observações (vector dos valores fenotípicos)

$\boldsymbol{X}_{(n\times p)}$ é a matriz de delineamento dos efeitos fixos

$\boldsymbol{\beta}_{(p\times 1)}$ é o vector de efeitos fixos (por exemplo, vector dos efeitos do local, do ano, de efeitos associados aos delineamento experimental,…)

$\boldsymbol{Z}_{n\times q}$ é a matriz de delineamento dos efeitos aleatórios (efeitos genéticos aditivos)

$\boldsymbol{u}_{(q\times 1)}$ é o vector de efeitos aleatórios (vector dos efeitos genéticos aditivos (*breeding values*) dos individuos, $\mathcal{N}_q(\mathbf{0}, \sigma_a^2 \boldsymbol{G_A})$

- A matriz $\boldsymbol{G_A}$ é agora derivada de marcadores moleculares

$\boldsymbol{e}_{(n\times 1)}$ é o vector de erros aleatórios, $\mathcal{N}_n(\mathbf{0}, \sigma_e^2 \boldsymbol{I_n})$

17

A matriz $\boldsymbol{G_A}$ é agora derivada de marcadores moleculares

$\boldsymbol{G_A}$ é a matriz de relações observadas entre indivíduos derivada de marcadores moleculares (também conhecida por matriz genómica).

Exemplo:

$$\mathbf{G}_A = \begin{bmatrix} 0.98 & 0.42 & 0.23 & -0.02 \\ 0.42 & 0.99 & 0.26 & 0.01 \\ 0.23 & 0.26 & 1.03 & 0.20 \\ -0.02 & 0.01 & 0.20 & 0.99 \end{bmatrix}$$

O vector dos melhores preditores empíricos lineares não enviesados dos efeitos reprodutivos genómicos (GBLUP) vai ser dado por:

$$\tilde{\boldsymbol{u}} = \boldsymbol{G_A Z^T V^{-1}}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right).$$

18