

Three Methodological Puzzles Inspired by Trees



Milan Lstibůrek



Faculty of Forestry
and Wood Sciences

REALIZED HERITABILITY IN PANMICTIC POPULATIONS

Lstibůrek M., Bittner V., Hodge G.R., Picek J., & Mackay T.F. (2018). Estimating realized heritability in panmictic populations. *Genetics* 208(1): 89-95.





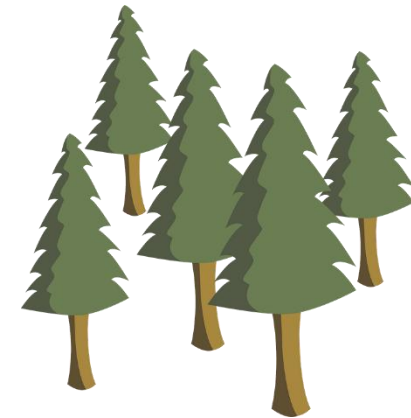
PARENTS gen $n + 1$



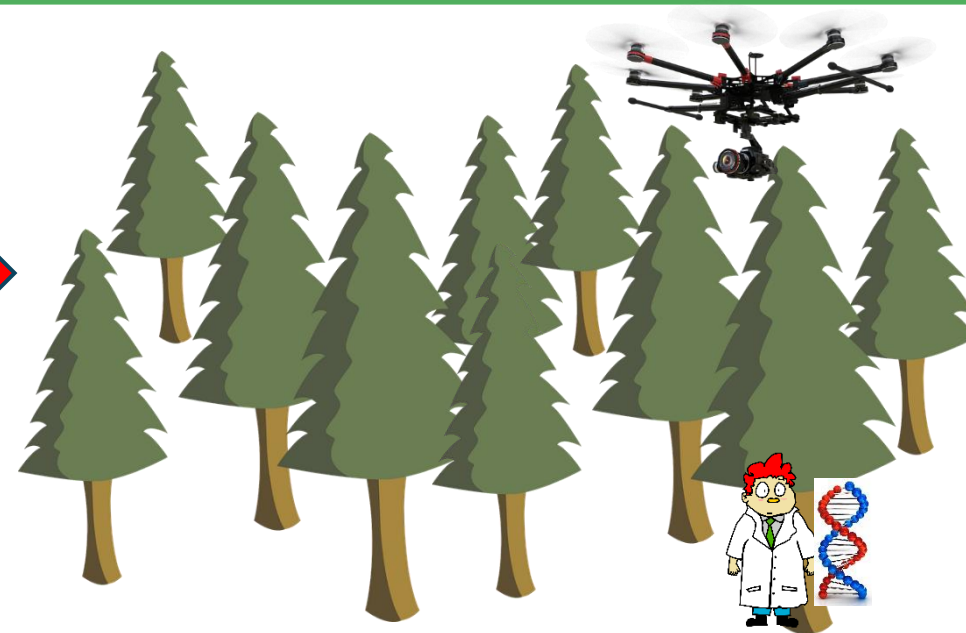
PROGENY TRIAL



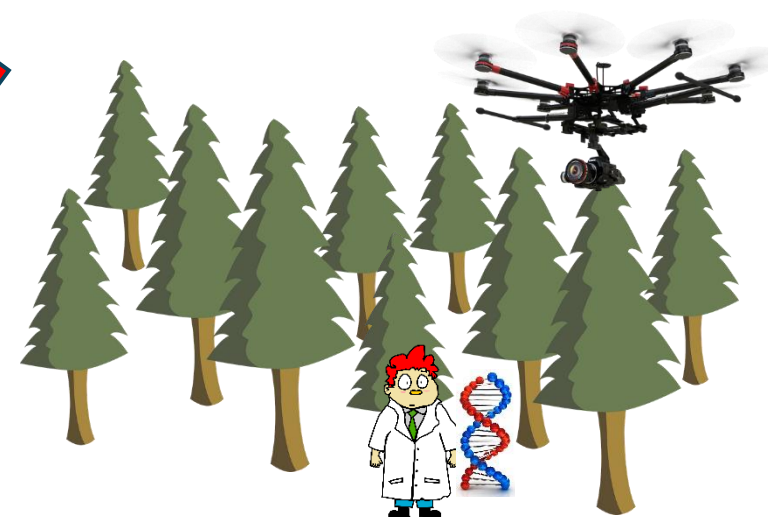
PARENTS gen n



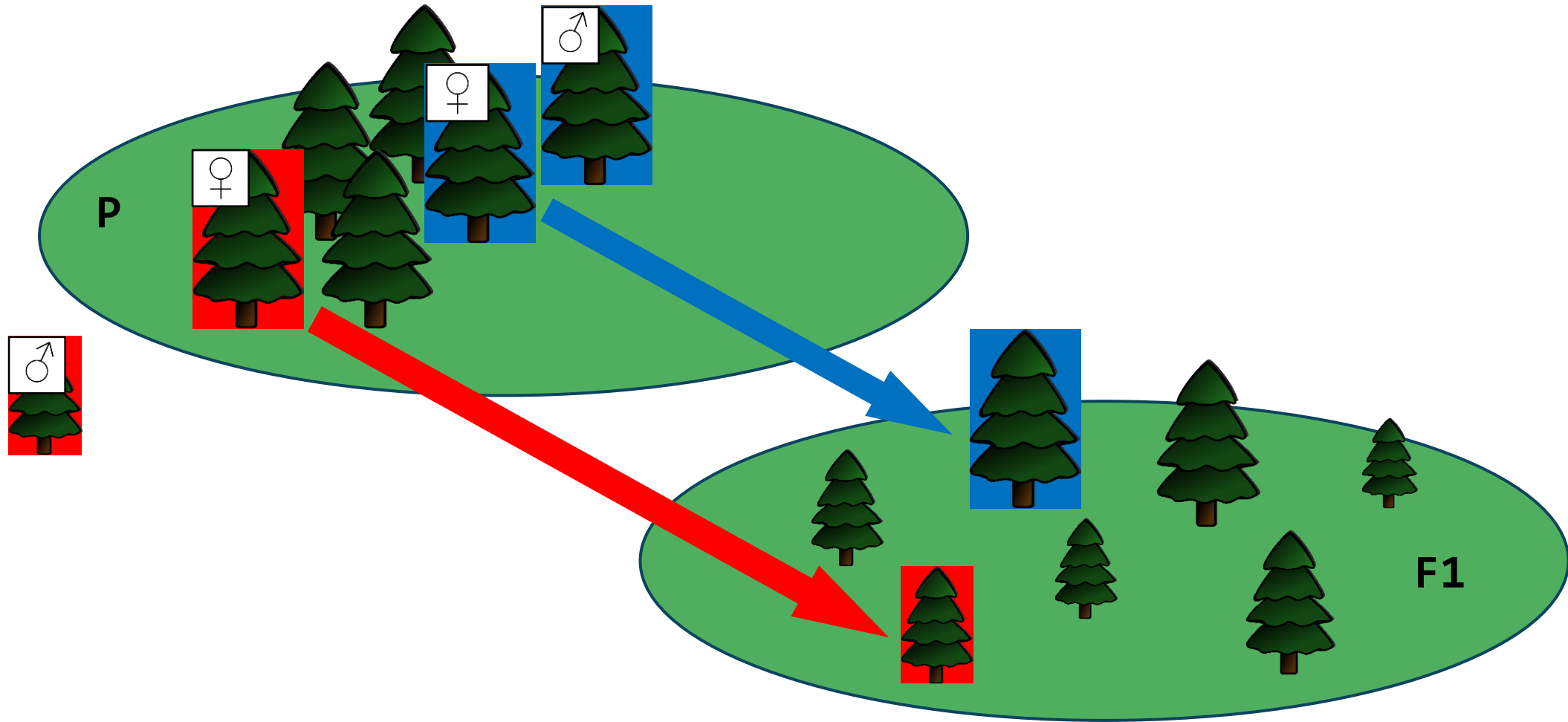
PARENTS gen $n + 1$



PARENTS gen n



El-Kassaby, Y. A., & Lstibůrek, M. (2009).
Breeding without breeding. *Genetics Research*, 91(2): 111-120.



Lstibůrek M. et al. (2012). Breeding without breeding: effect of gene flow on fingerprinting effort. *Tree Genetics & Genomes* 8: 873-877.

Founder population

$$\sigma_P^2 = \sigma_A^2 + \sigma_E^2$$

where

$$P \sim N(\mu, \sigma_P^2)$$

$$A \sim N(\mu, \sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$



Seed orchard

$$P \sim N(S_{SO}, (1 - k)\sigma_P^2)$$

$$A \sim N(S_{SO}, (1 - h^2k)\sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$

$$S_{SO} = i\sigma_P$$

Founder population

$$\sigma_P^2 = \sigma_A^2 + \sigma_E^2$$

where

$$P \sim N(\mu, \sigma_P^2)$$

$$A \sim N(\mu, \sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$

Seed orchard

$$P \sim N(S_{SO}, (1 - k)\sigma_P^2)$$

$$A \sim N(S_{SO}, (1 - h^2k)\sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$

$$S_{SO} = i\sigma_P$$



r



When the offspring is sired by outside pollen, it can be described as follows:

$$P \sim N(h^2(S_{SO} + \mu)/2, (1 - 0.5h^4k')\sigma_P^2)$$

$$A \sim N(h^2(S_{SO} + \mu)/2, (1 - 0.5h^2k')\sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$

where k' is calculated as the unweighted average (different selection intensity on male and female component).

$(1 - r)$



When the offspring is produced from within-orchard's pollen pool, it can be described as follows:

$$P \sim N(h^2 S_{SO}, (1 - 0.5h^4k)\sigma_P^2)$$

$$A \sim N(h^2 S_{SO}, (1 - 0.5h^2k)\sigma_A^2)$$

$$E \sim N(0, \sigma_E^2)$$

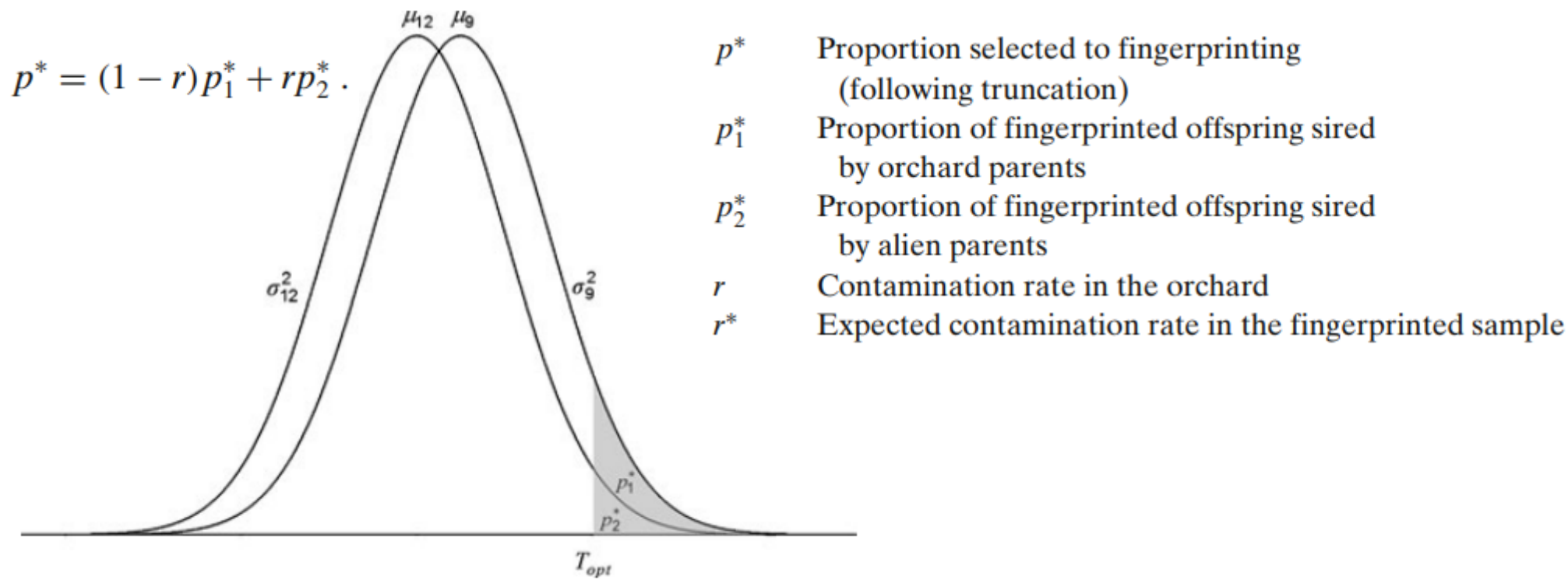


Fig. 1 Conceptual representation of two offspring populations produced from within the orchard's pollen pool (*right*, μ_9 and σ_9^2 corresponding to Eq. 9) and those sired by outside pollen (*left*, μ_{12} and σ_{12}^2 corresponding to Eq. 12). The area beyond the truncation point denotes the probabilities p_1^* (Eq. 9) and p_2^* (Eq. 12), representing within-orchard and outside pollination, respectively. The actual parameters (i.e., means and variances) depend on the input. T_{opt} is the truncation point produced by Solver optimization

Lstibůrek M. et al. (2012). Breeding without breeding: effect of gene flow on fingerprinting effort. Tree Genetics & Genomes 8: 873-877.

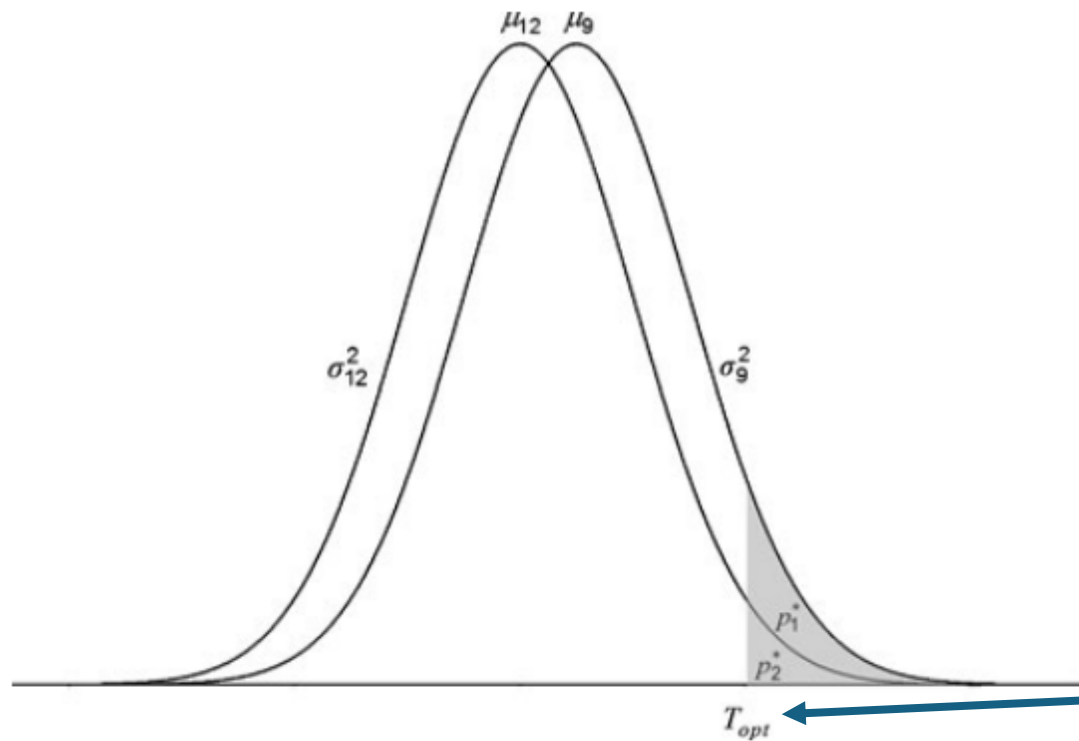


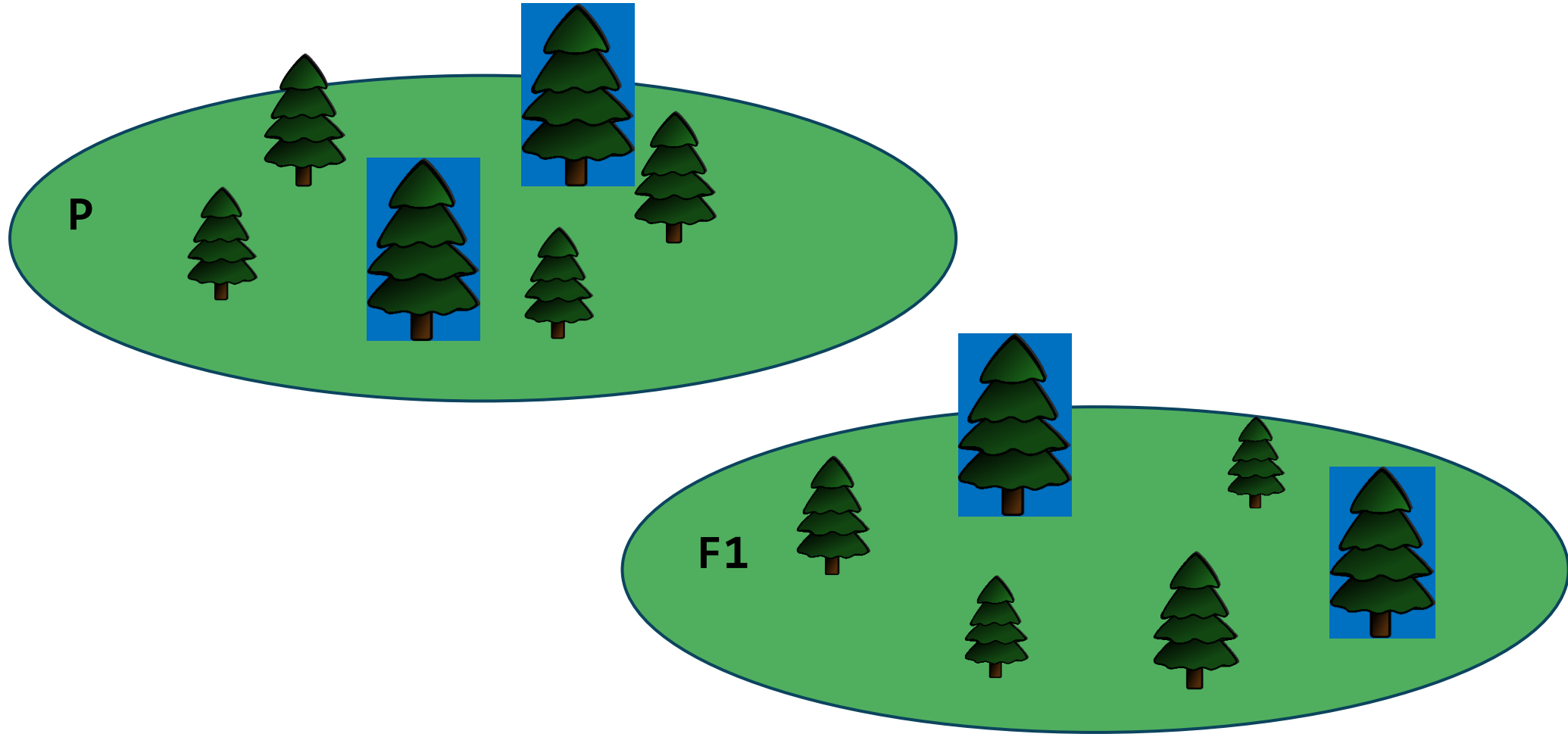
Fig. 1 Conceptual representation of two offspring populations produced from within the orchard's pollen pool (*right*, μ_9 and σ_9^2 corresponding to Eq. 9) and those sired by outside pollen (*left*, μ_{12} and σ_{12}^2 corresponding to Eq. 12). The area beyond the truncation point denotes the probabilities p_1^* (Eq. 9) and p_2^* (Eq. 12), representing within-orchard and outside pollination, respectively. The actual parameters (i.e., means and variances) depend on the input. T_{opt} is the truncation point produced by Solver optimization

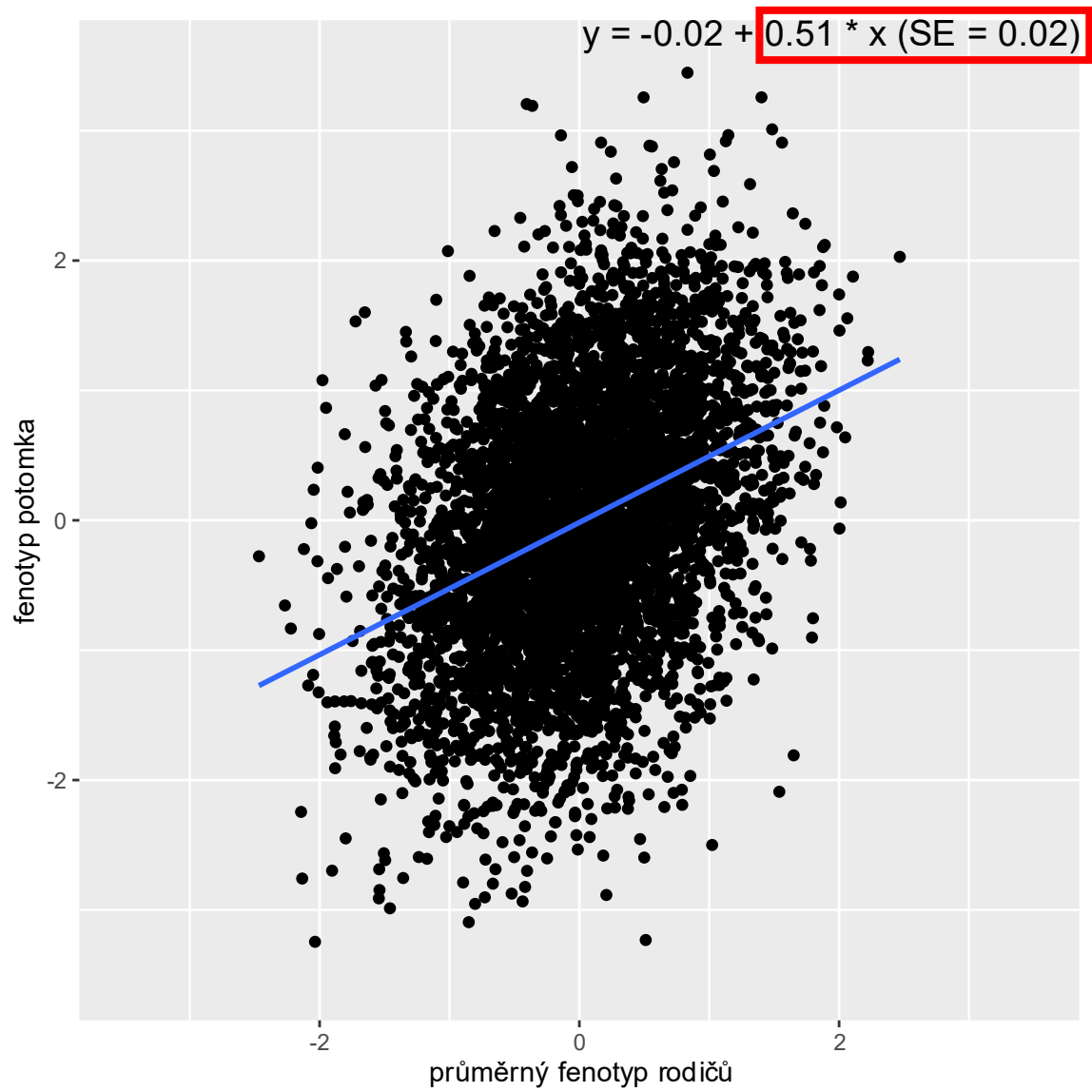
$$p^* = (1 - r)p_1^* + rp_2^* .$$

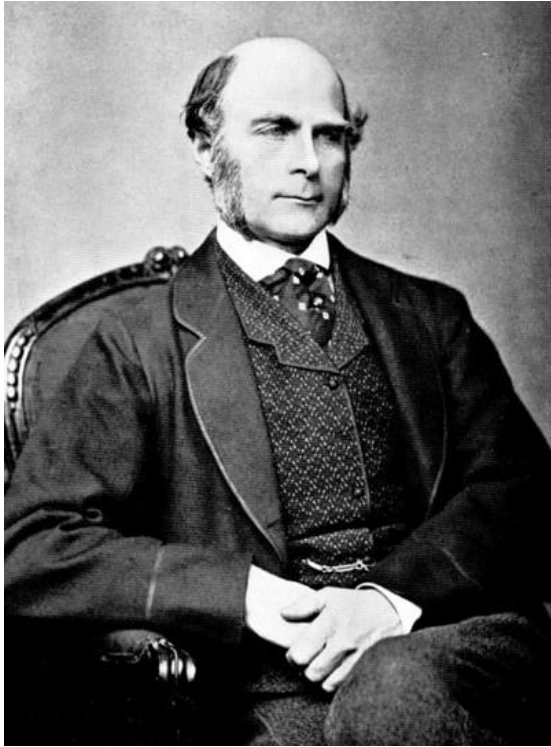
The value of r is set as an input in Eq. 15, and p_1^* and p_2^* are calculated using the normal distribution function in Excel (NORMDIST). This function returns the probability that a number falls at or below a given value of a normal distribution (for the specified number, mean, and standard deviation). Each distribution's mean and variance are known; thus, the only unknown variable is the truncation point (common to both distributions, Fig. 1). Next, with the use of the "Solver" optimization tool in Excel, the truncation point is determined, so that p^* is meeting the requested proportion of progenies that are fingerprinted (set as input). The Microsoft Excel Solver is an optimization tool, which is included in the Microsoft Office.

Lstibůrek M. et al. (2012). Breeding without breeding: effect of gene flow on fingerprinting effort. *Tree Genetics & Genomes* 8: 873-877.

Using the same reasoning, but what if the narrow-sense heritability is unknown, i.e.







For demonstration purposes, we used F. Galton's human height data. The dataset contains 205 families with the adult heights of parents and offspring (934 offspring, with family sizes ranging from 1 to 15).

Galton F. (1886). Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* 15: 246–263.

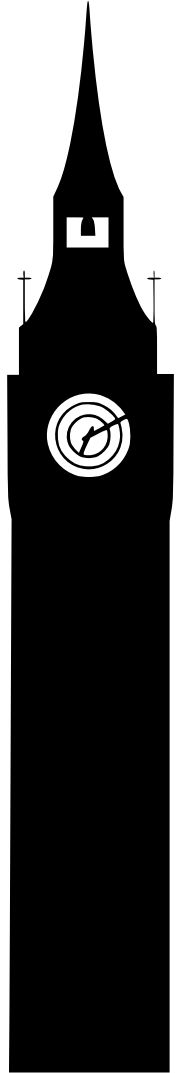


TABLE I.
NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1.08).

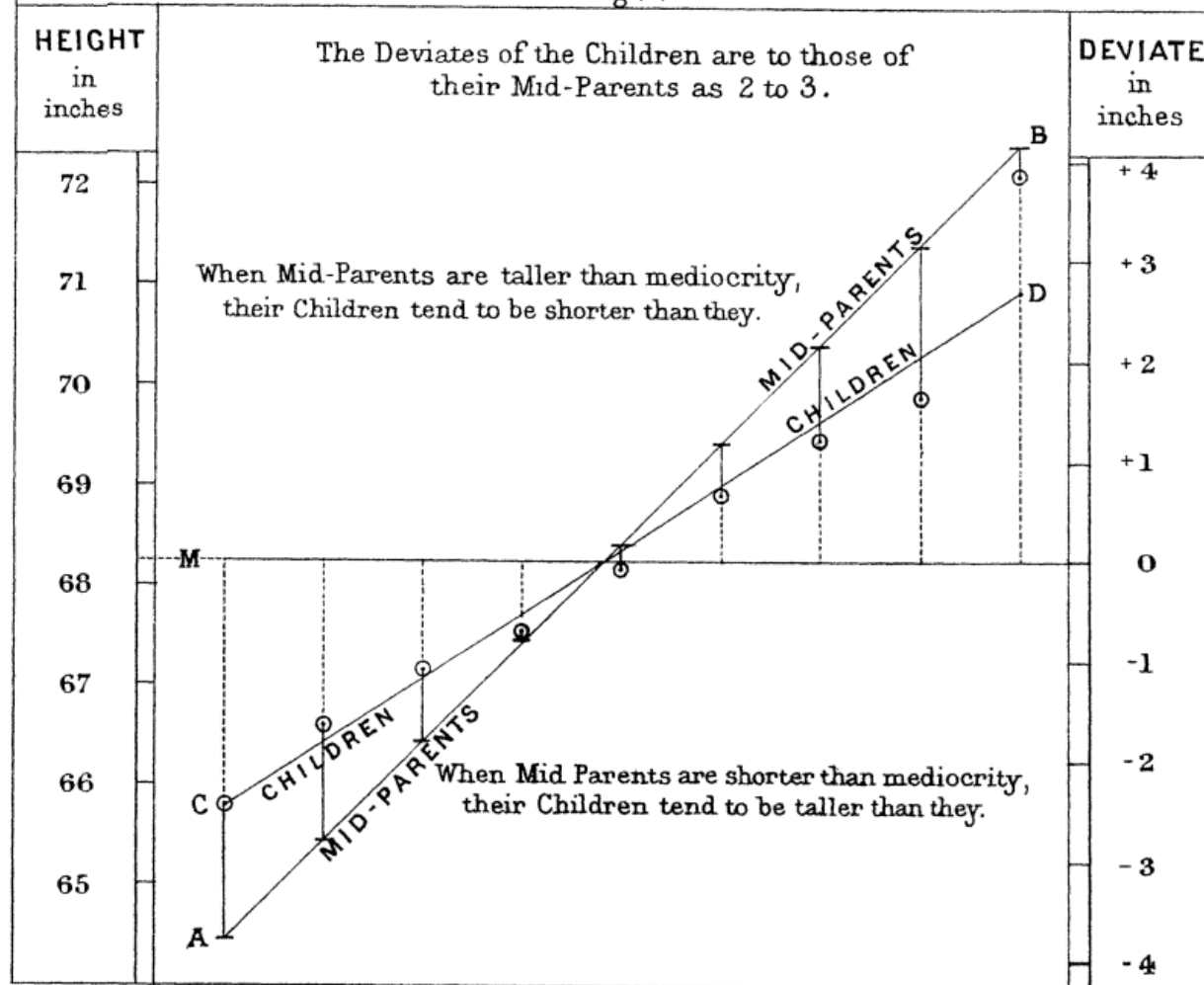
Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72.5 ..	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72.2
71.5 ..	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69.9
70.5 ..	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5 ..	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5 ..	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68.2
67.5 ..	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67.6
66.5 ..	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67.2
65.5 ..	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66.7
64.5 ..	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65.8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

Galton F. (1886). Regression towards mediocrity in hereditary stature.
J. Anthropol. Inst. G. B. Irel. 15: 246–263.

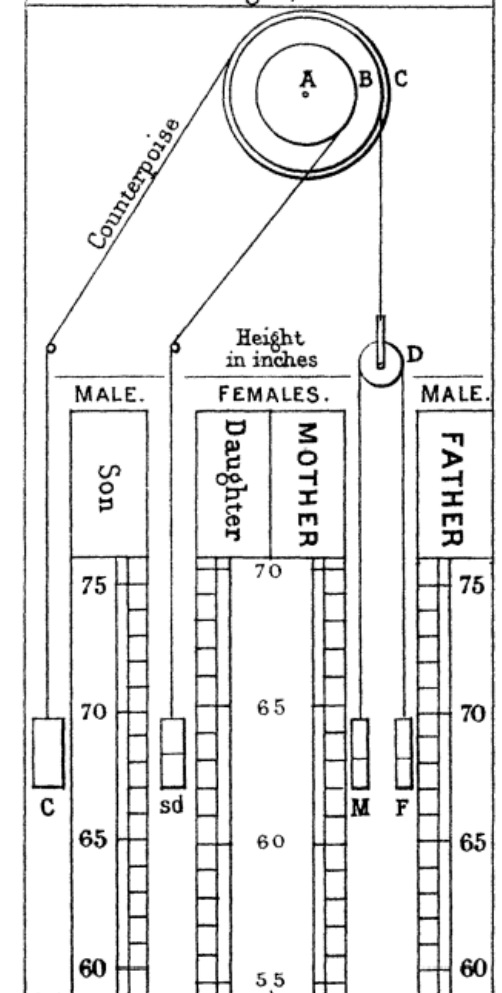
RATE OF REGRESSION IN HEREDITARY STATURE.

Fig. (a)



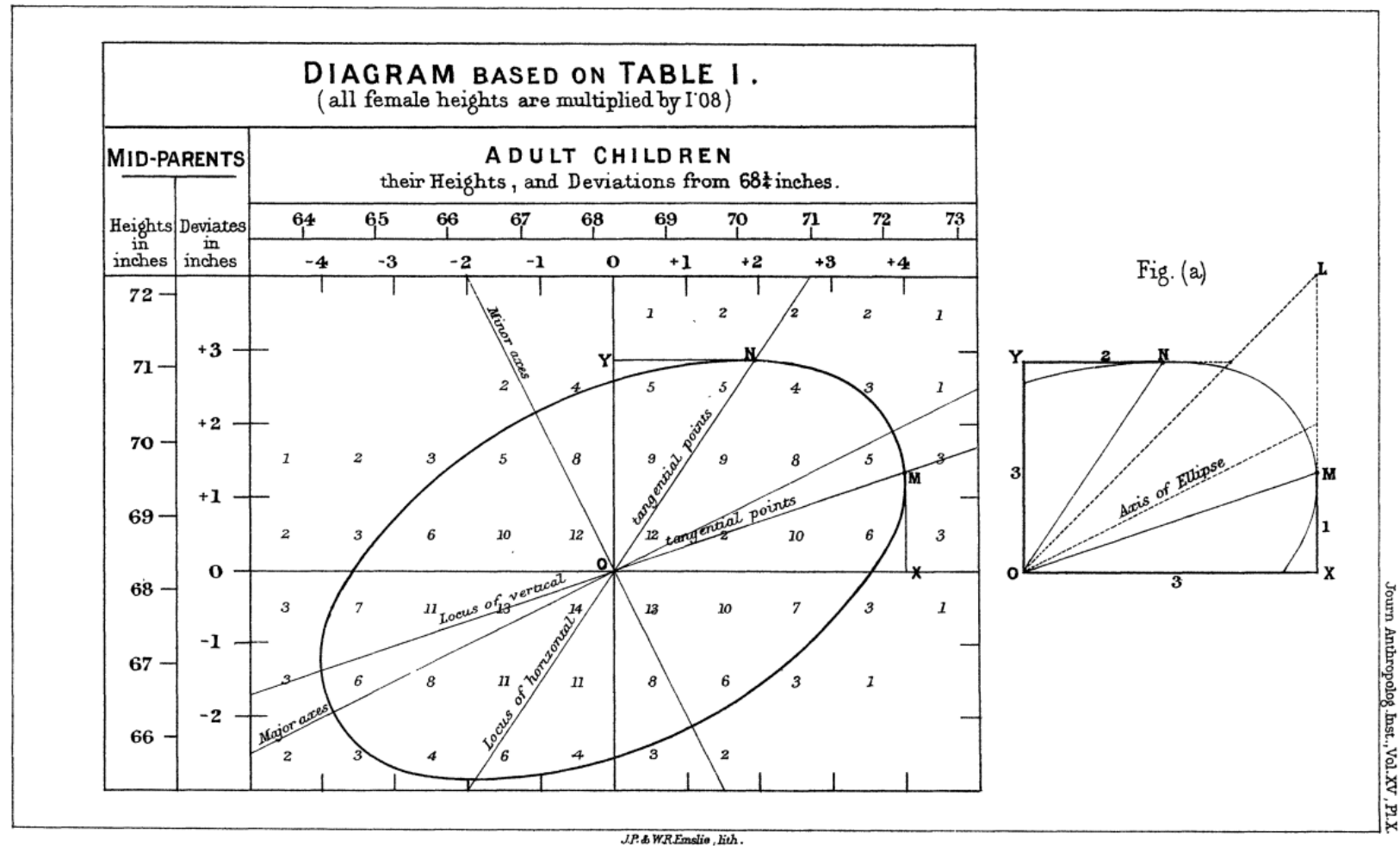
FORECASTER OF STATURE

Fig. (b)



J.P. & W.R. Emslie, lith.

Galton F. (1886). Regression towards mediocrity in hereditary stature.
 J. Anthropol. Inst. G. B. Irel. 15: 246–263.

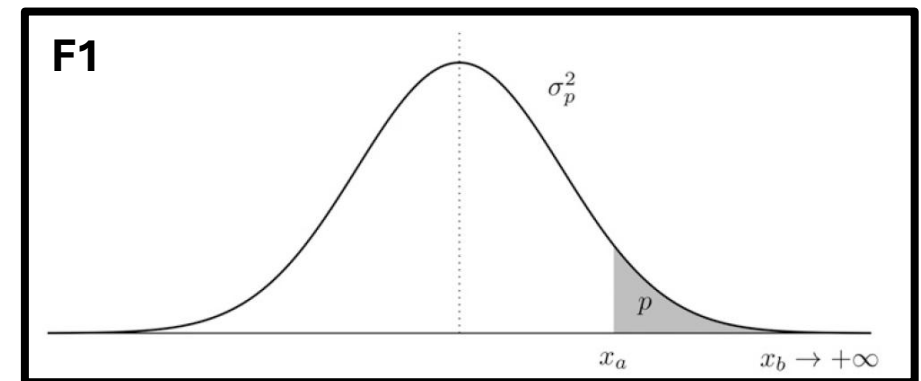
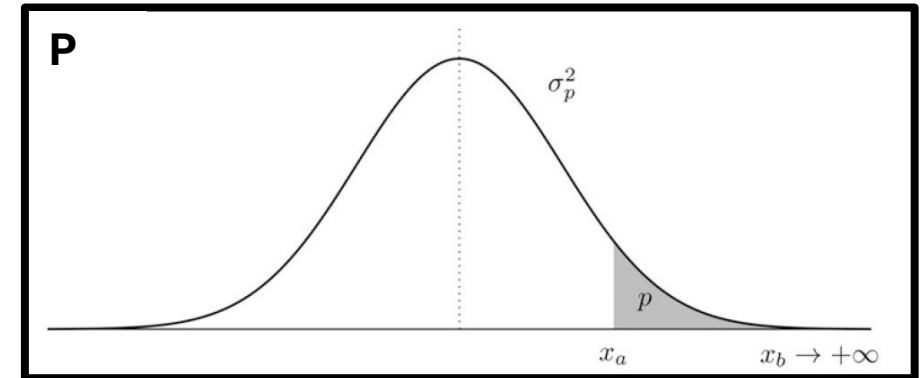


Galton F. (1886). Regression towards mediocrity in hereditary stature.
 J. Anthropol. Inst. G. B. Irel. 15: 246–263.

1. Estimates of the phenotypic mean and variance in P and F1.
2. Random sample of offspring in truncated area of F1 (>0.5 SE), 278 out of 934.
3. Determination, if their parent(s) originate in the equivalent truncated area of P.

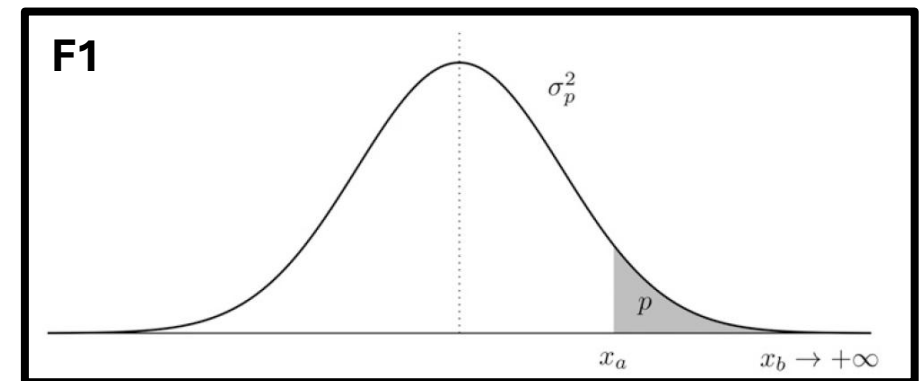
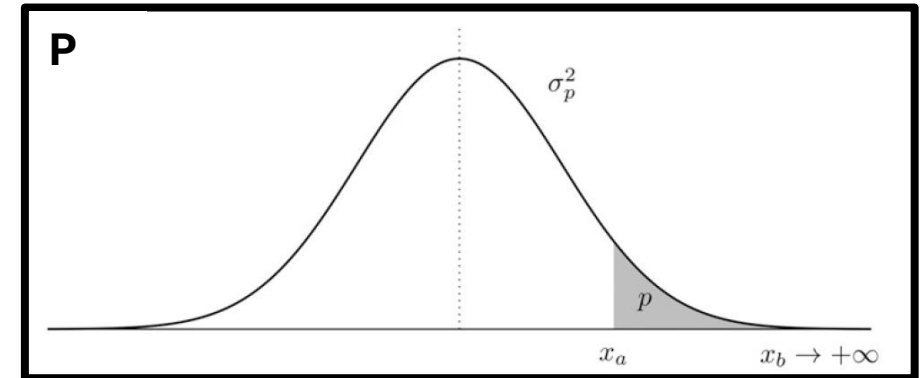
$$h^2 = \frac{2\alpha Bi \pm [4\alpha^2 B^2 i^2 - 2(2B^2 i^2 + BQ'^2 k)(\alpha^2 - Q'^2)]^{\frac{1}{2}}}{(2B^2 i^2 + BQ'^2 k)}$$

$$SE\{h^2\} = \frac{(1 - 0.5h^4 kB)^{\frac{3}{2}}}{|-\phi(a)(-iB + 0.5h^2 \alpha kB)|} \frac{p}{P(Y)} \sqrt{r(1-r)} \frac{1}{\sqrt{n}}$$

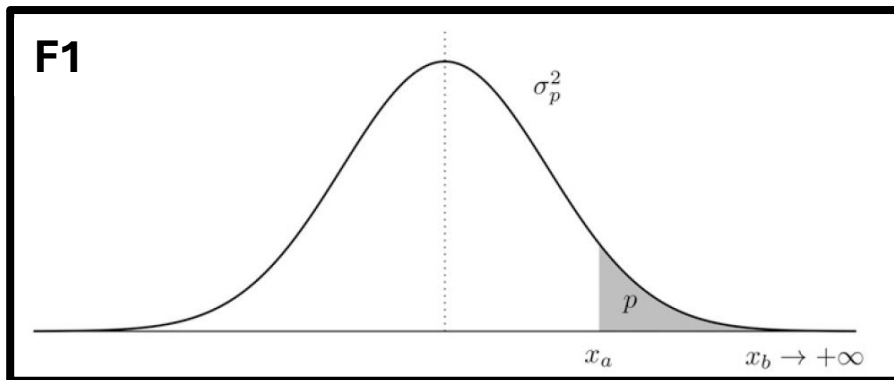
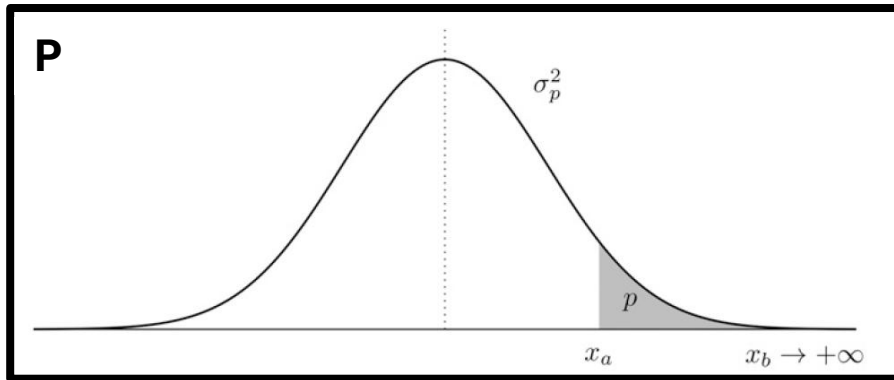


Lstibůrek M., Bittner V., Hodge G.R., Picek J., & Mackay T.F. (2018). Estimating realized heritability in panmictic populations. *Genetics* 208(1): 89-95.

Estimated $h^2 = 0.71 - 0.73$, statistically equivalent to linear regression.

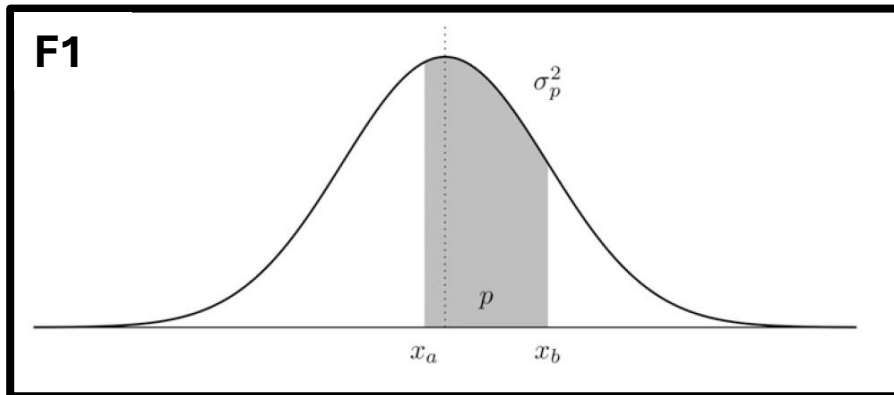
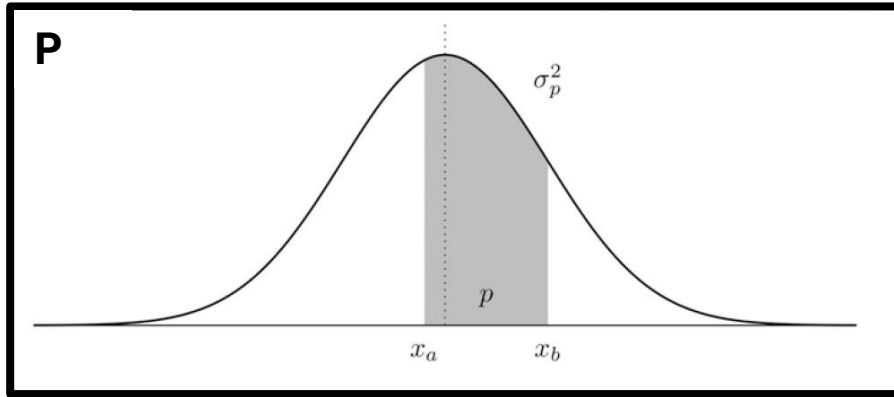


Lstibůrek M., Bittner V., Hodge G.R., Picek J., & Mackay T.F. (2018). Estimating realized heritability in panmictic populations. *Genetics* 208(1): 89-95.



„Under the assumptions of Hardy-Weinberg equilibrium, the h^2 of a quantitative trait in a given population is directly related to the likelihood of a phenotypic subset of parents passing their respective alleles onto the corresponding phenotypic subset of offspring.“

Lstibůrek M., Bittner V., Hodge G.R., Picek J., & Mackay T.F. (2018). Estimating realized heritability in panmictic populations. *Genetics* 208(1): 89-95.



„Under the assumptions of Hardy-Weinberg equilibrium, the h^2 of a quantitative trait in a given population is directly related to the likelihood of a phenotypic subset of parents passing their respective alleles onto the corresponding phenotypic subset of offspring.“

Lstibůrek M., Bittner V., Hodge G.R., Picek J., & Mackay T.F. (2018). Estimating realized heritability in panmictic populations. *Genetics* 208(1): 89-95.

IMPACT OF ERRONEOUS MARKERS ON HERITABILITY

Sagariya C., Bittner V., Pook T., Roy A., Lstibůrek M. (2026). Impact of erroneous marker data on the accuracy of narrow-sense heritability. *Genetics* 232(1): iyaf230.

$$\mathbf{M} = (m_{ij} \in \mathbb{R}^{n \times m}) : m_{ij} \in \{-1; 0; 1\}$$

$$\mathbf{G} = (g_{ij} \in \mathbb{R}^{n \times n})$$

$$\mathbf{G}^{\text{method1}} = \frac{(\mathbf{M} - \mathbf{1}_n \mathbf{p})(\mathbf{M} - \mathbf{1}_n \mathbf{p})^T}{2 \sum p_i (1 - p_i)}$$

VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>.

Error Scenario	Genotype	$\rightarrow A_1A_1$	$\rightarrow A_1A_2$	$\rightarrow A_2A_2$
1: Directional genotype perturbation	A_1A_1	0	$2/3$	$1/3$
	A_1A_2	$1/3$	$1/3$	$1/3$
	A_2A_2	$1/3$	$2/3$	0
2: HWE redraw	$\forall$	$(1 - p_i)^2$	$2p_i(1 - p_i)$	p_i^2
3: Uniform redraw	$\forall$	$1/3$	$1/3$	$1/3$

The table presents the probabilities of substituting specific elements of **M** under each of the three error scenarios (column 1). In Scenario 1, the probabilities depend on the existing genotypes, as listed in column 2. The error corresponds to redrawing the genotype uniformly at random from the full set $\{A_1A_1, A_1A_2, A_2A_2\}$. This definition allows the possibility that the resampled genotype coincides with the original one (e.g. $A_1A_2 \rightarrow A_1A_2$). In Scenario 2, substitution probabilities are based on the allelic frequency p_i in the i th locus. These probabilities are independent of the original genotype in **M**. In Scenario 3, substitution probabilities are uniform across all elements of **M**.

Population simulation

- **Simulation Rationale:** Used MoBPS software to create a known, error-free reference to systematically isolate and quantify the impact of genotyping errors.
- **Genome Structure:** Simulated 4 linkage groups (100 cM each) across three evenly spaced SNP marker densities (2,500, 5,000, and 20,000 markers).
- **Trait Architecture:** Modeled a single quantitative trait driven by 300 purely additive QTLs with uniformly distributed initial allele frequencies.
- **Genetic Variance:** Additive QTL effects were sampled from a normal distribution, with trait heritabilities set at 0.2 and 0.6.
- **LD Generation:** 5,000 unrelated founder individuals underwent 100 generations of random mating to establish realistic linkage disequilibrium.
- **Final Population:** 100 parents were randomly selected from the final generation and mated to produce a cohort of 1,000 offspring.

Genetic evaluation

To estimate the h^2 , we utilized the **G** matrix within a univariate mixed model implemented in the rrBLUP R package (Endelman2011).

The mixed model fitted by the restricted maximum likelihood (REML)

$$\mathbf{y} = \mu \mathbf{1} + \mathbf{a} + \mathbf{e}$$

$$\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \sigma_a^2 \mathbf{G}), \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{I})$$

Error Scenario	Genotype	$\rightarrow A_1A_1$	$\rightarrow A_1A_2$	$\rightarrow A_2A_2$
1: Directional genotype perturbation	A_1A_1	0	2/3	1/3
	A_1A_2	1/3	1/3	1/3
	A_2A_2	1/3	2/3	0

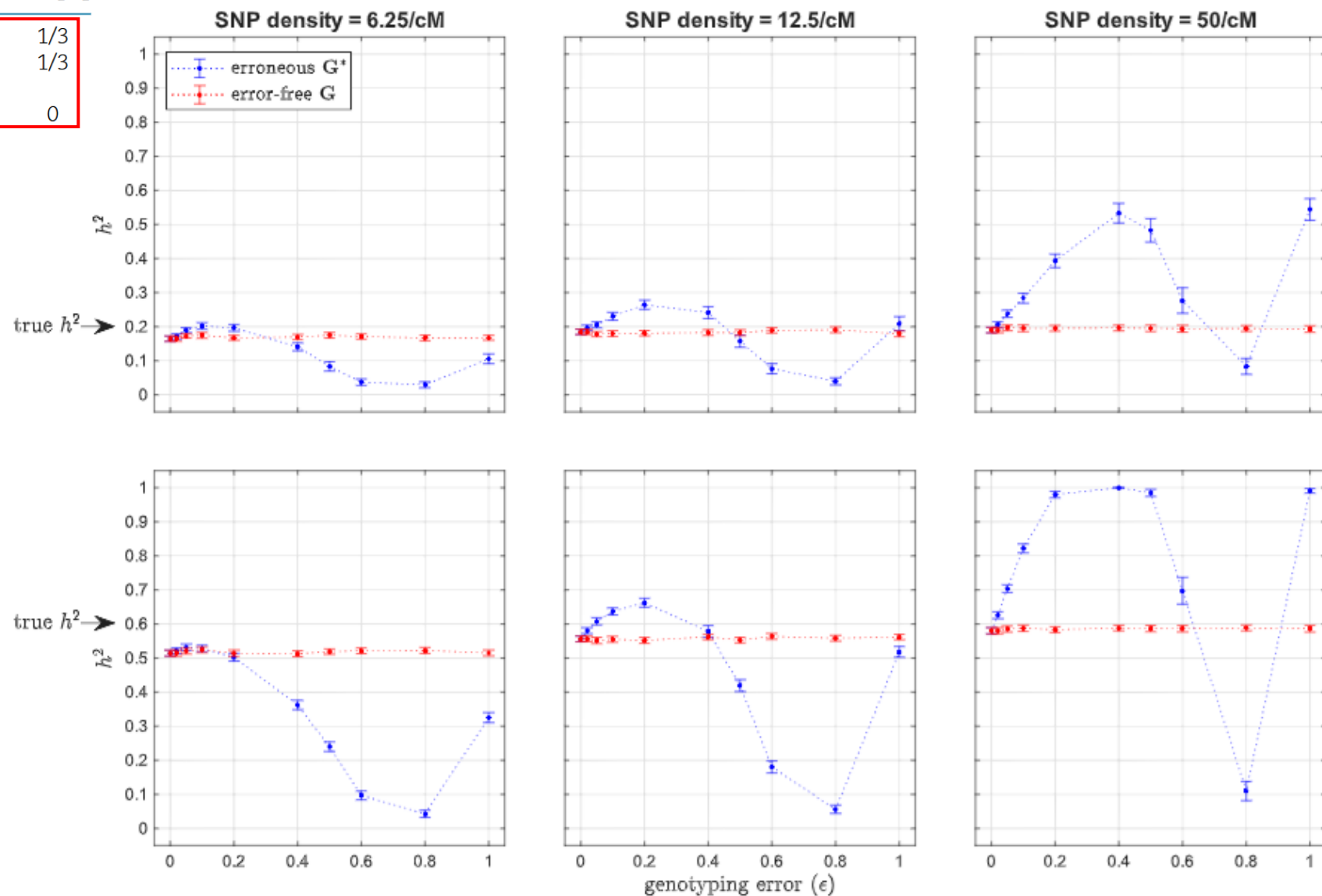


Fig. 1. Scenario 1 (substitution based on existing genotypes). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ϵ); the Y-axis shows h^2 , with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free G are represented by the lines parallel to the X-axis, while estimates based on erroneous G^* are represented by the lines exhibiting a polynomial bias pattern.

Error Scenario	Genotype	$\rightarrow A_1A_1$	$\rightarrow A_1A_2$	$\rightarrow A_2A_2$
1: Directional genotype perturbation	A_1A_1	0	$2/3$	$1/3$
	A_1A_2	$1/3$	$1/3$	$1/3$
	A_2A_2	$1/3$	$2/3$	0
2: HWE redraw	$\forall$	$(1 - p_i)^2$	$2p_i(1 - p_i)$	p_i^2

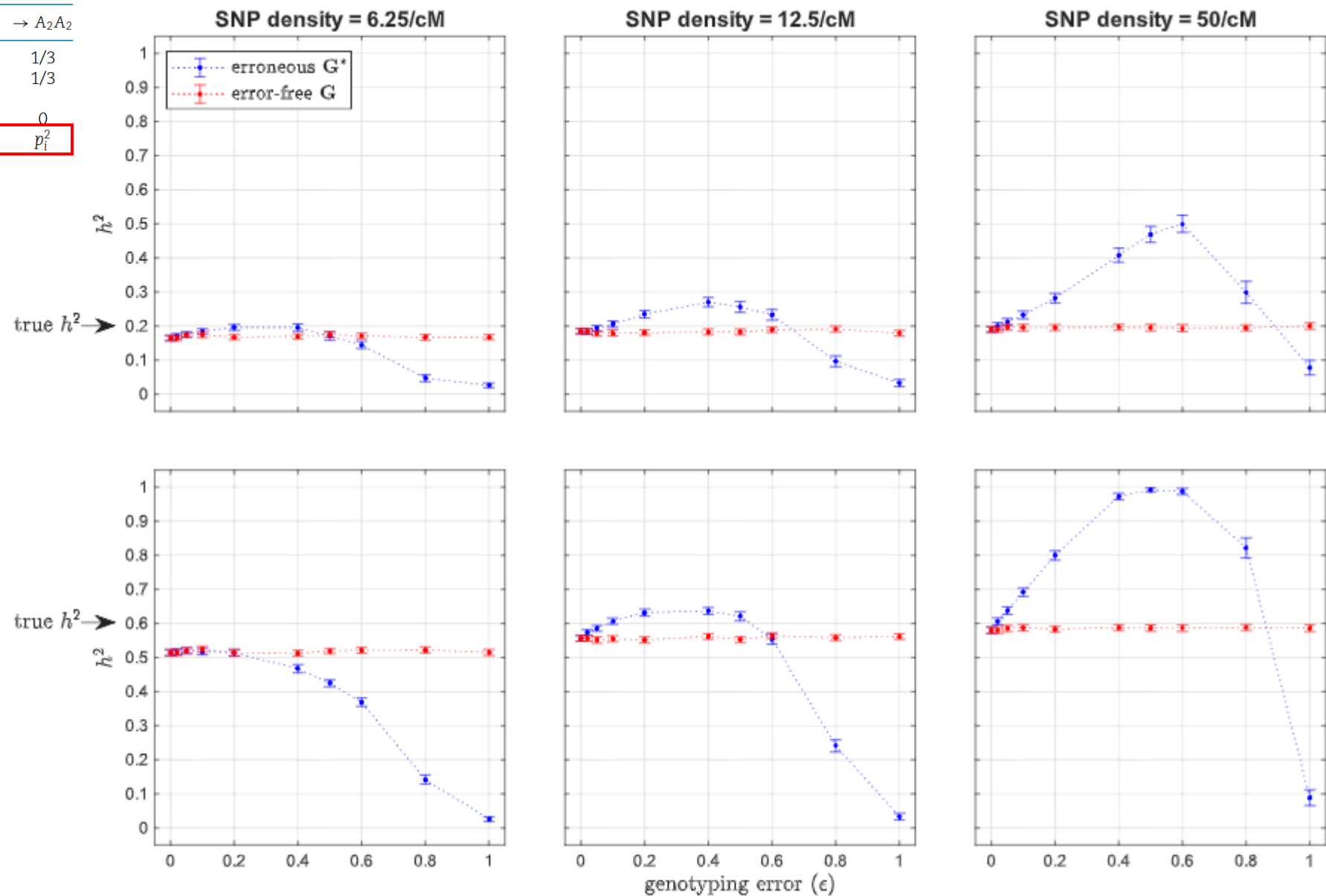


Fig. 2. Scenario 2 (substitution based on H-W frequencies). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ϵ); the Y-axis shows h^2 , with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free G are represented by the lines parallel to the X-axis, while estimates based on erroneous G^* are represented by the lines exhibiting a polynomial bias pattern.

Error Scenario	Genotype	$\rightarrow A_1A_1$	$\rightarrow A_1A_2$	$\rightarrow A_2A_2$
1: Directional genotype perturbation	A_1A_1	0	2/3	1/3
	A_1A_2	1/3	1/3	1/3
2: HWE redraw	A_2A_2	1/3	2/3	0
3: Uniform redraw	$\forall$	$(1 - p_i)^2$	$2p_i(1 - p_i)$	p_i^2
	$\forall$	1/3	1/3	1/3

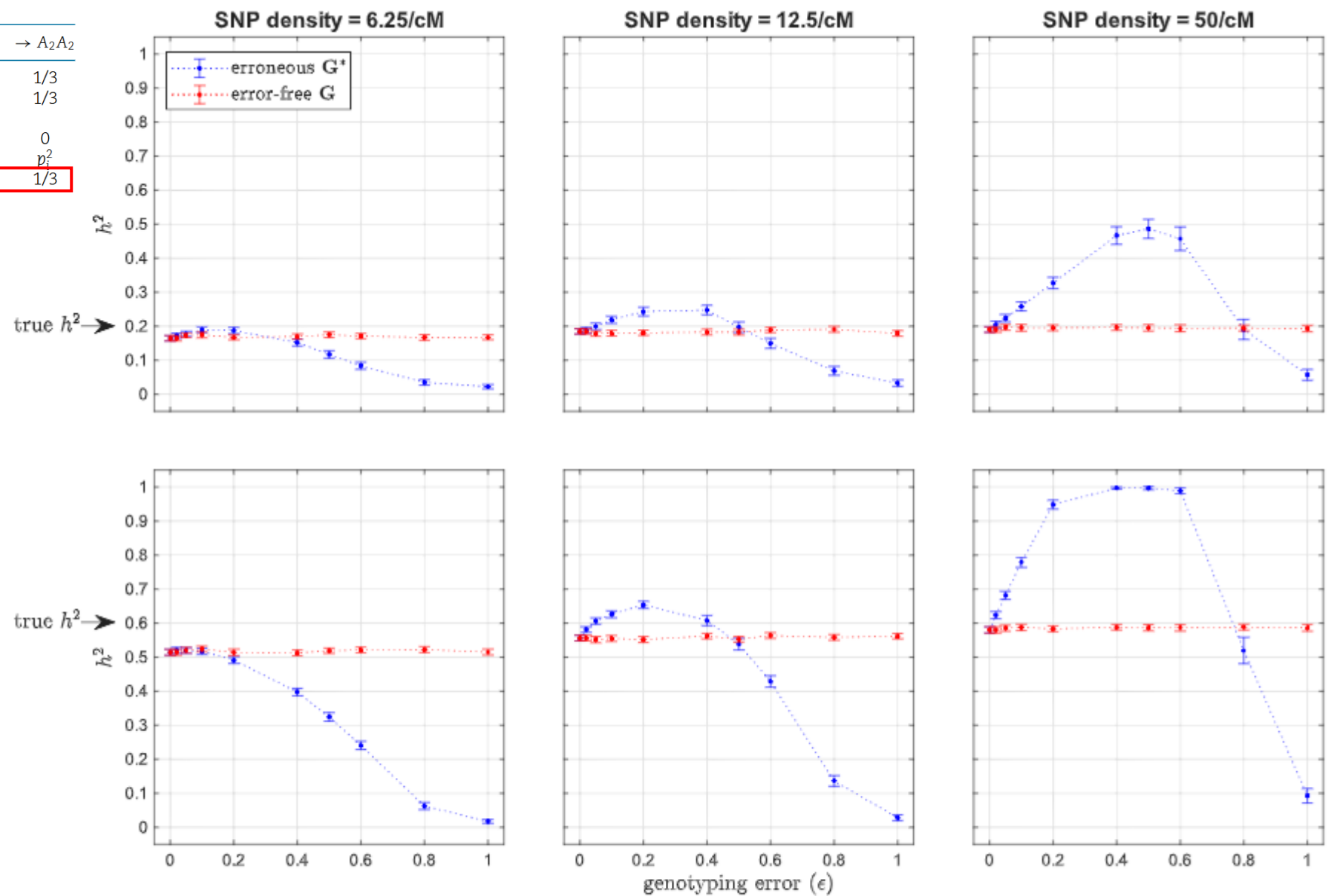


Fig. 3. Scenario 3 (uniform substitution probabilities). The top row represents true $h^2 = 0.2$, and the bottom row true $h^2 = 0.6$. From left to right, graphs correspond to SNP densities of 6.25, 12.5, and 50 SNPs per cM. The X-axis indicates the percentage of modified genotypes (ϵ); the Y-axis shows h^2 , with mean values across 100 replicates connected by dotted lines and 95% confidence intervals. Estimates assuming error-free G are represented by the lines parallel to the X-axis, while estimates based on erroneous G^* are represented by the lines exhibiting a polynomial bias pattern.

Impact of genotyping error

Matrix Distortion: Genotyping errors alter the expected value, variance, and covariance of the genomic relationship matrix ($\mathbf{G}^*$), which distorts its centering, diagonal elements, and off-diagonal elements.

Covariance Shrinkage: The presence of genotyping error artificially shrinks the mean off-diagonal elements of the matrix (representing relatedness) by a factor of approximately $(1 - \epsilon)^2$

The Ratio Effect: The ultimate distortion in heritability estimates is driven by how the error changes the ratio between the mean diagonal elements and the mean off-diagonal elements of $\mathbf{G}^*$.

Heritability Inflation: To compensate for the lost covariance, REML overestimates the additive genetic variance, which leads to a systematic inflation of the heritability estimate.

Eventual Collapse: Heritability estimates peak by a 50% error rate at the latest; beyond this point, the relatedness information vanishes, and the REML method is forced to push h^2 estimate toward zero to prevent a mathematically inadmissible negative environmental variance.

Impact of marker density

Sensitivity Scales with Density: Using a denser marker coverage significantly increases the sensitivity of the heritability estimates to genotyping errors.

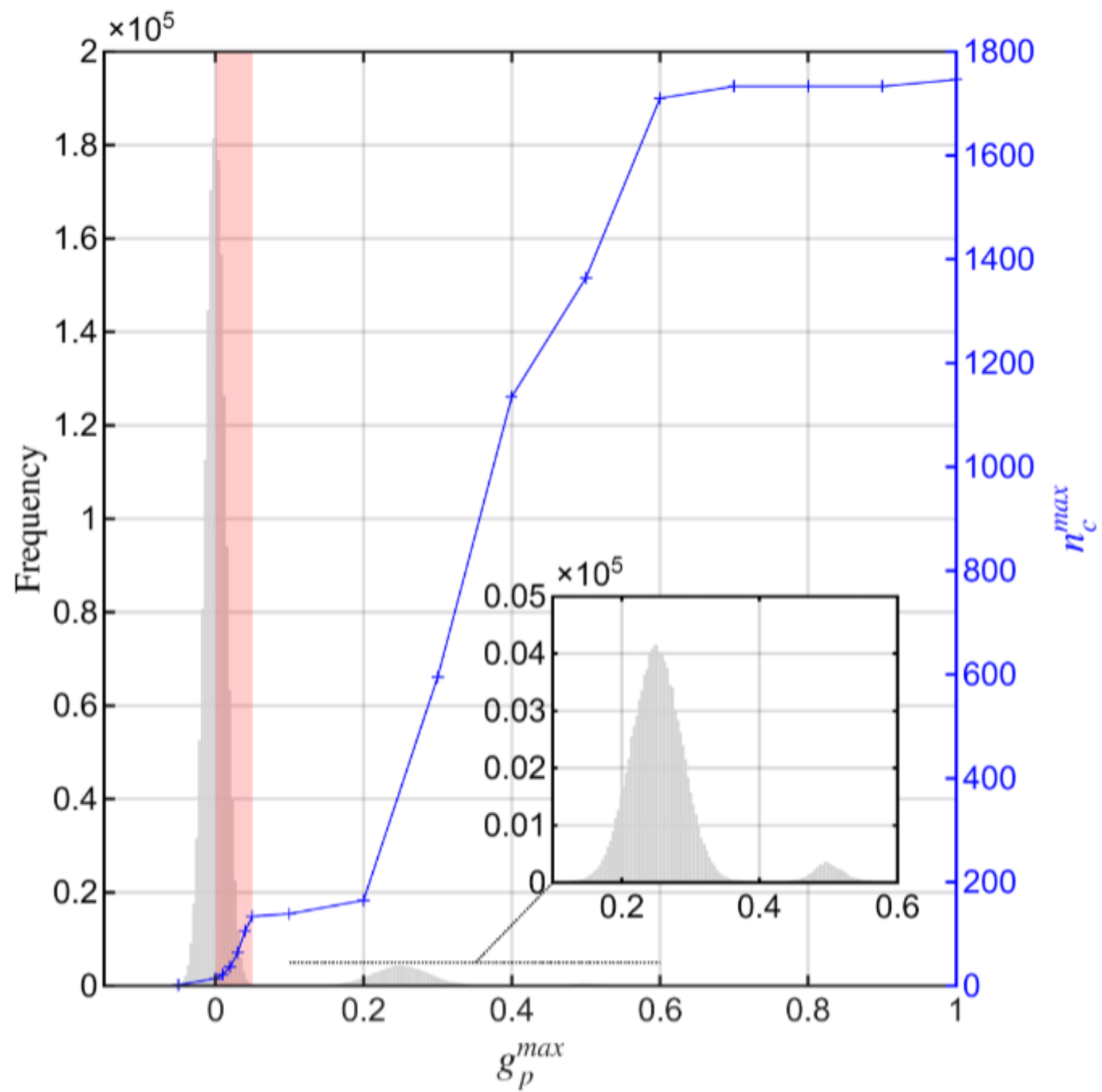
Marker Efficiency: The accuracy of the genetic relationship matrix is tied to a "marker efficiency" parameter $\lambda(m)$, which quantifies how well a specific quantity of markers captures true genetic relationships, entirely independent of actual errors.

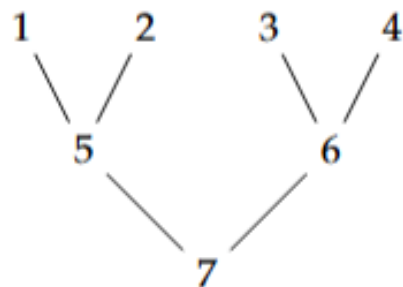
Sparse Panels Underestimate: Even under perfectly error-free conditions, low-density marker panels will systematically underestimate true heritability.

The Ultimate Trade-off: Low-density panels are robust to errors up to roughly 10% but constantly underestimate heritability; conversely, high-density panels are perfectly accurate with zero error but begin to overestimate heritability at error rates as low as 1%.

GENOMIC RELATIONSHIP-BASED SELECTION OF UNRELATED INDIVIDUALS

Lstibůrek M., Bittner V., Solvin T., Korecký J., & Steffenrem A. (2026). Genomic relationship-based selection of unrelated individuals. *Under review*.

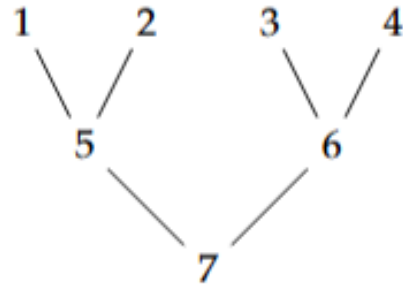




$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 1 & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \quad \tilde{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{b}^T = [10.1 \quad 7.8 \quad 6.5 \quad 8.1 \quad 8.3 \quad 9.4 \quad 12.6]$$

$\mathbf{r} = (r_i) \in \mathbb{R}^{n_t} : r_i \in \{0; 1\}$ is a decision variable (column vector) with 0 for not selected and 1 for selected individuals .



$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 0 & 1 & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} \quad \tilde{\mathbf{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{b}^T = [10.1 \quad 7.8 \quad 6.5 \quad 8.1 \quad 8.3 \quad 9.4 \quad 12.6]$$

$$\max \rightarrow r_1 + r_2 + r_3 + r_4 + r_5 + r_6 + r_7$$

or

$$\begin{aligned} \max &\rightarrow 10.1r_1 + 7.8r_2 + 6.5r_3 + 8.1r_4 + 8.3r_5 + 9.4r_6 + 12.6r_7 \\ &r_1 + r_2 + r_3 + r_4 + r_5 + r_6 + r_7 = n_c \end{aligned}$$

$$r_1 + r_5 \leq 1$$

$$r_3 + r_6 \leq 1$$

$$r_5 + r_7 \leq 1$$

$$r_1 + r_7 \leq 1$$

$$r_3 + r_7 \leq 1$$

$$r_6 + r_7 \leq 1$$

$$r_2 + r_5 \leq 1$$

$$r_4 + r_6 \leq 1$$

$$r_2 + r_7 \leq 1$$

$$r_4 + r_7 \leq 1$$

$$(\mathbf{r})_i + (\mathbf{r})_j \leq 1 \quad \forall i < j : (\tilde{\mathbf{G}})_{ij} = 1$$

