Теория селекции

Лекция 2. Путевые диаграммы Райта.

В.М.Ефимов, д.б.н.

Институт цитологии и генетики СО РАН

Путевые диаграммы Райта

Анализ путей. Одним из современных методов моделирования является анализ путей. Сам метод был предложен еще в 30-х гг. XX в. С. Райтом. Метод основан на графическом представлении причинных и корреляционных связей, или путей, между переменными, включенными в описание модели. Как правило, на диаграмме путей квадратами и кружками с прописными буквенными символами внутри обозначают наблюдаемые переменные (т. е. доступные непосредственному измерению), например фенотипические значения изучаемого признака, и так называемые латентные переменные (недоступные измерению): генотипические значения, параметры общей и различающейся среды и т. п.

Путевые диаграммы Райта

Кружки и квадраты соединяются между собой стрелками, которые обозначают предполагаемые связи между переменными. Если связь причинная, то стрелка имеет направление в одну сторону (от причины к следствию), если корреляционная - то в обе стороны, поскольку при корреляционных зависимостях не предполагается наличие причинно-следственных отношений, а лишь однонаправленность отклонений переменной от среднего. Рядом со стрелками, обозначающими пути от причины к следствию, располагаются путевые коэффициенты, а рядом со стрелками, предполагающими наличие корреляций - коэффициенты корреляции. Эти коэффициенты обозначаются соответствующими величинами (если они известны) или строчными буквами. В диаграмме путей зависимыми (эндогенными) переменными являются те, которые подлежат объяснению (например, фенотипические характеристики), а независимыми (экзогенными) – те, действием которых объясняются зависимые переменные и их связи (чаще всего генетические и средовые параметры).

Связь регрессии, дисперсии и путевых диаграмм Райта

Одномерная линейная регрессия

Ф.Гальтон

TABLE I.

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.

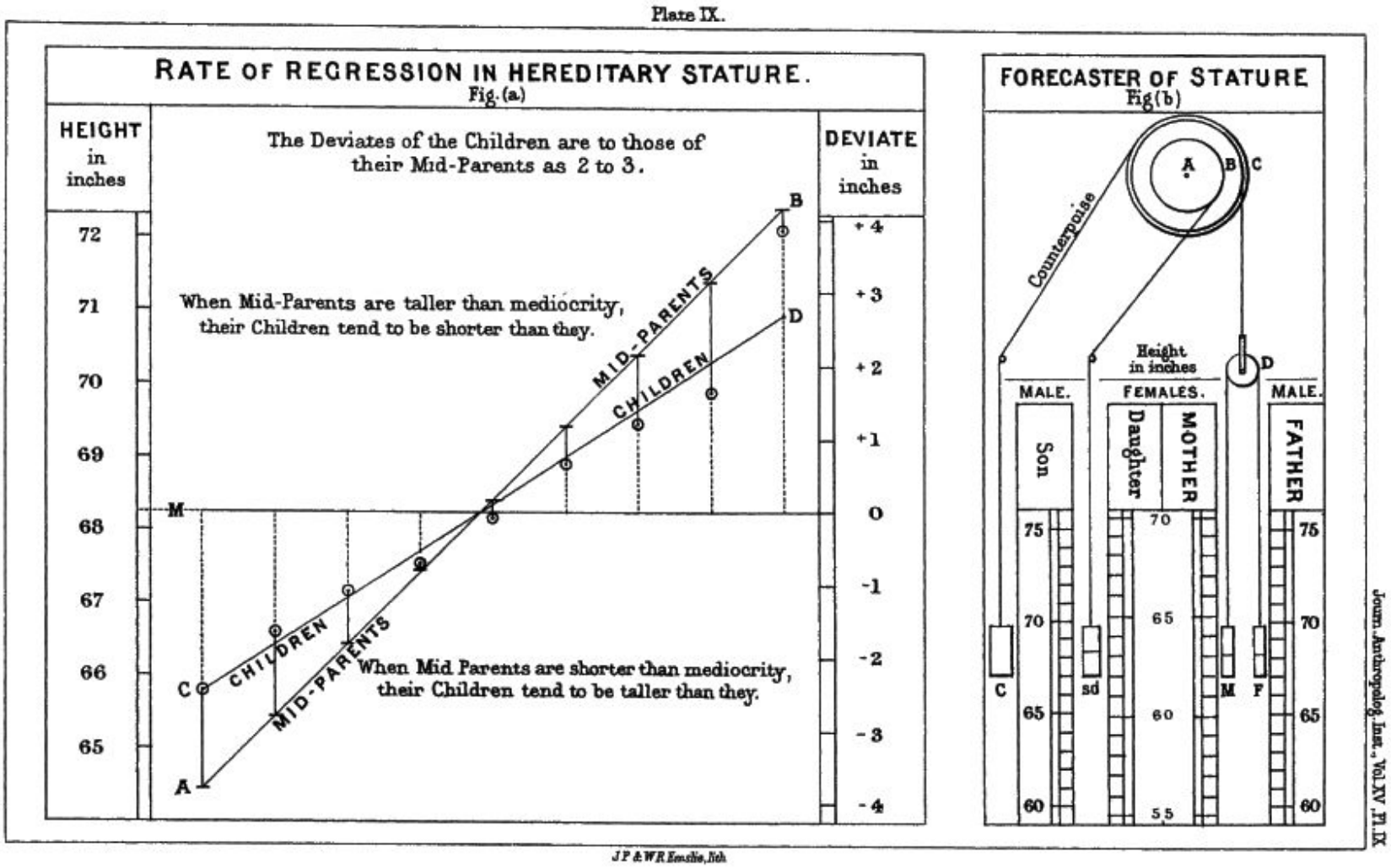
(All Female heights have been multiplied by 1·08).

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.

Одномерная линейная регрессия

Ф.Гальтон



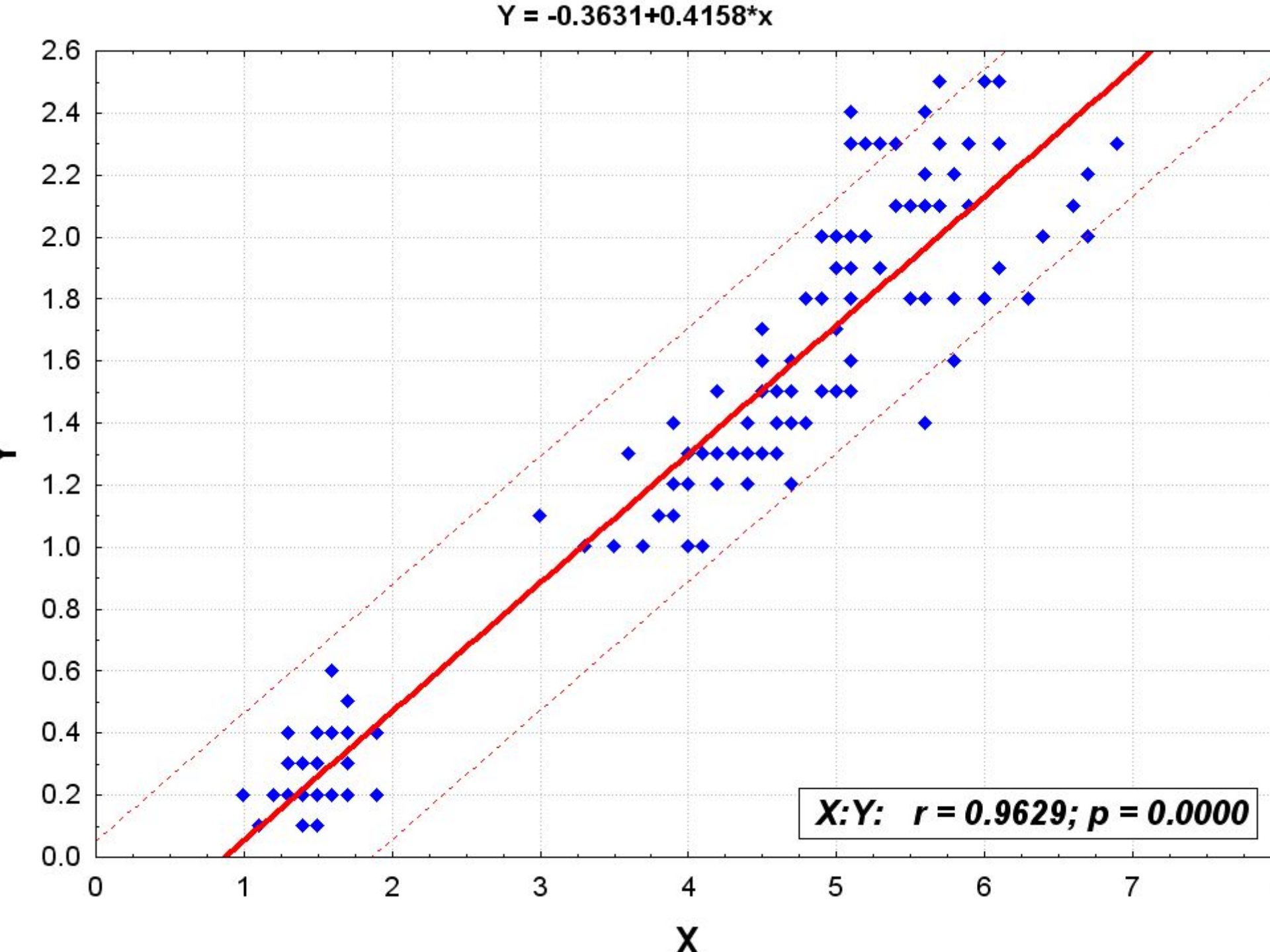
Уравнение линейной регресии.
Метод наименьших квадратов

$$y = ax + b$$

$$y = r\frac{s_y}{s_x}x + \bar{y} - r\frac{s_y}{s_x}\bar{x}$$

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

$$y' = rx'$$


Y = -0.3631+0.4158*x
2.6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Y
0
1
2
3
4
5
6
7
X
X:Y: r = 0.9629; p = 0.0000

Будем считать, что обе переменные центрированы и нормированы.

Уравнение $\tilde{y}=ax+b$ – это оценка зависимости y от x, это наше знание об y, если мы знаем только x. Обозначим ее через $\tilde{y}$. Тогда $\tilde{y}= rx$.

Сколько дисперсии «снимает» такая зависимость? Сколько знания она приносит?

Связь регрессии и дисперсии. Путь X --- Y

$$\sum y_i^2/N = \sigma_y^2 = 1; \qquad \sum x_i^2/N = \sigma_x^2 = 1;$$

$$\sum y_i x_i/N = r;$$

$$\tilde{y} = rx; \quad \tilde{y}_i = rx_i;$$

$$\sigma_{\tilde{y}}^2 = \sum \tilde{y}_i^2/N = \sum (rx_i)^2/N = r^2 \sum x_i^2/N = r^2\sigma_x^2 = r^2$$

Таким образом, доля дисперсии ***X*** в дисперсии ***Y*** равна r^2.

В единицах стандартного отклонения $\sigma_{\tilde{y}} = \sqrt{(r^2)} = r.$

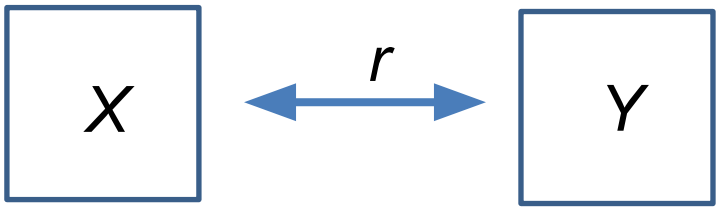
Связь регрессии и дисперсии. Путь X --- Y

Но можно рассмотреть регрессию X на Y.

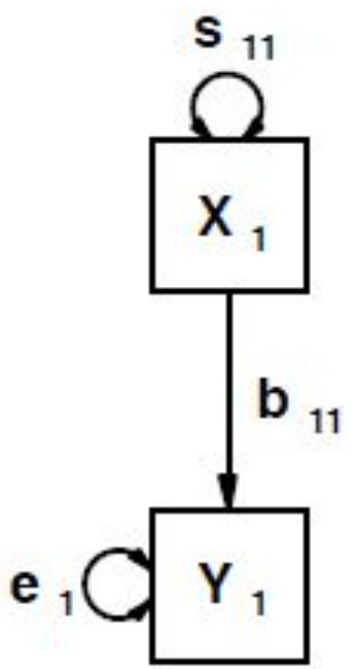
Получим, что доля дисперсии Y в дисперсии X тоже равна r^2.

Таким образом, ситуация симметрична.

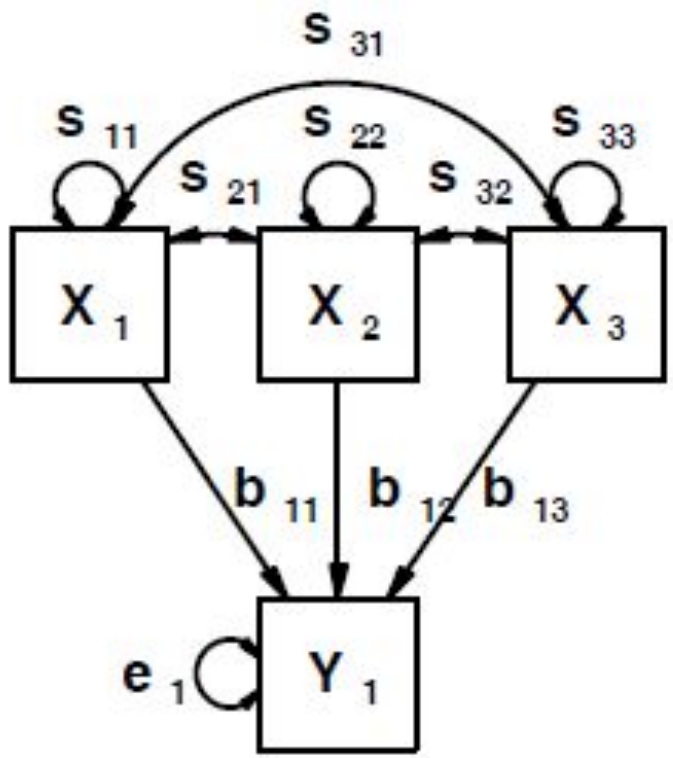
Графически это изображается следующим образом:



Связь регрессии и дисперсии. Путь X --- Y



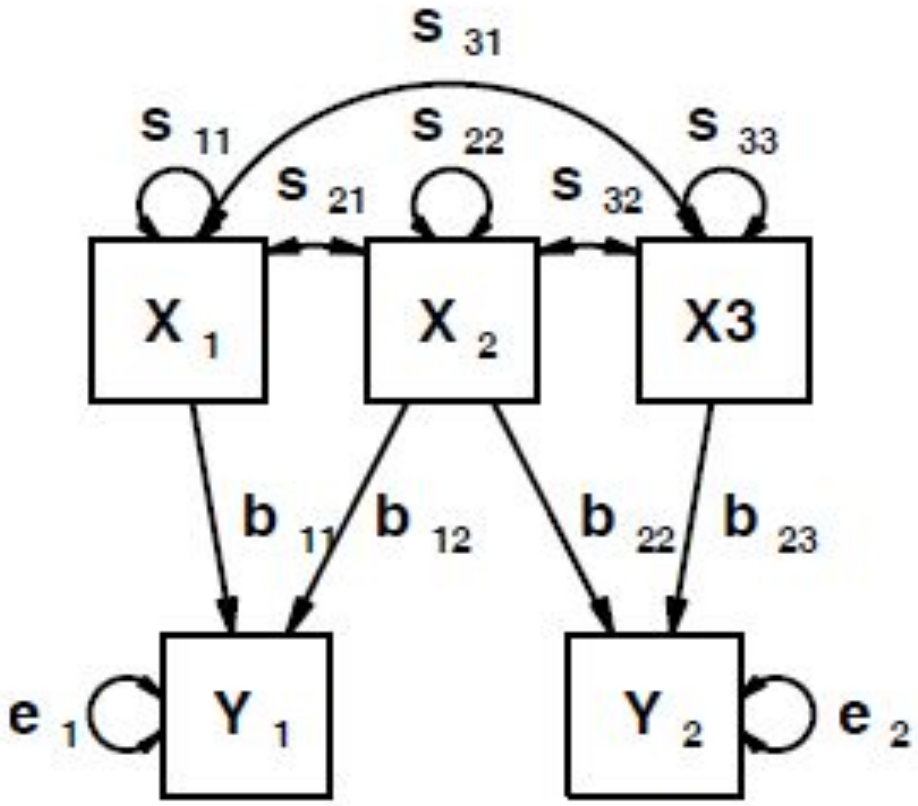
Одномерная регрессия



Множественная регрессия

Neale, M., & Cardon, L. (2013). *Methodology for genetic studies of twins and families* (Vol. 67). Springer Science & Business Media.

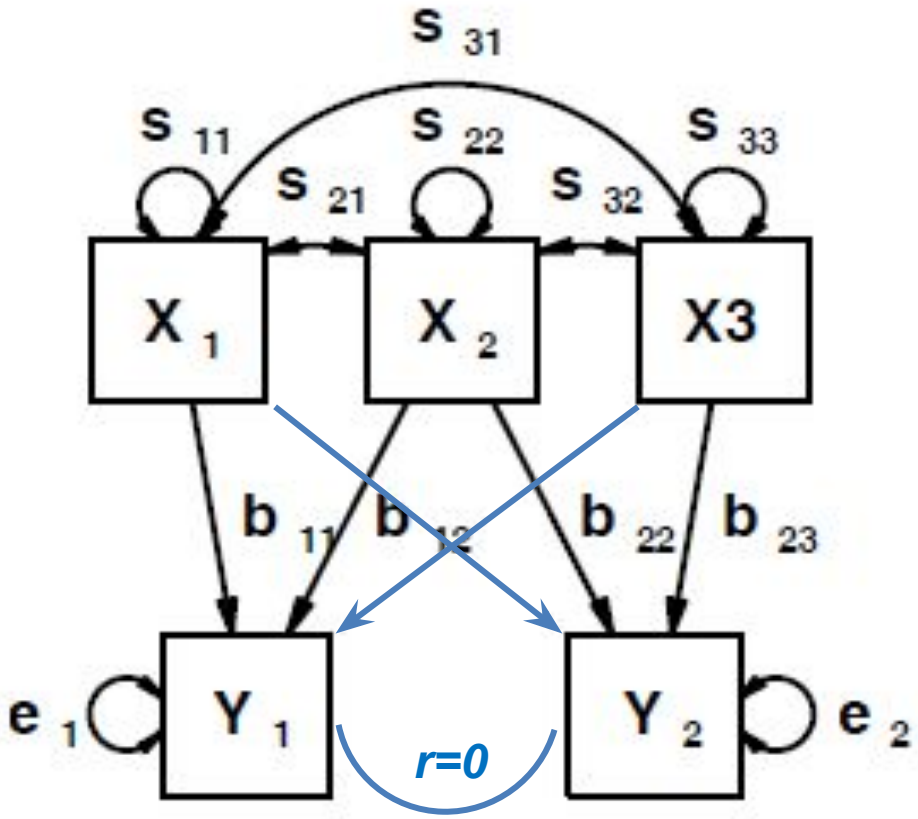
Связь регрессии и дисперсии. Путь X --- Y



Многомерная регрессия

Neale, M., & Cardon, L. (2013). *Methodology for genetic studies of twins and families* (Vol. 67). Springer Science & Business Media.

Связь регрессии и дисперсии. Путь X --- Y



Главные компоненты

Коэффициент наследуемости (гетерогенные родители и потомки)

$$h^2 = \frac{\sigma_P^2 - \sigma_E^2}{\sigma_P^2} = \frac{\sigma_G^2}{\sigma_P^2}$$

где σ_G^2 – генотипическая дисперсия,

σ_P^2 – фенотипическая дисперсия.

Lush. I. L. 1937. Animal Breeding Plans. Collegiate Press, Inc., Ames, IA.

Коэффициент наследуемости (гетерогенные родители и потомки)

Генотипическое значение признака включает в себя три компоненты:

σ_A^2 -- аддитивный эффект всех генов;

σ_D^2-- доминантный эффект всех генов;

σ_I^2 **-- суммарный** эффект, вызванный эпистатическим взаимодействием между генами

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

Нас интересует прежде всего аддитивный эффект. Именно он отвечает за успешность отбора.

Коэффициент наследуемости (гетерогенные родители и потомки)

в широком смысле:

$$h^2 = \frac{\sigma_F^2 - \sigma_E^2}{\sigma_F^2} = \frac{\sigma_G^2}{\sigma_F^2}$$

в узком смысле:

$$h^2 = \frac{\sigma_A^2}{\sigma_F^2}$$

То есть, коэффициент наследуемости – это доля дисперсии **генотипа** в дисперсии **фенотипа**

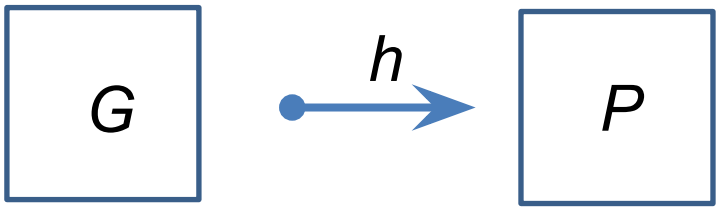
Коэффициент наследуемости (гетерогенные родители и потомки)

В единицах стандартного отклонения: $$\frac{\sigma_G}{\sigma_P} = h$$

или : $$\sigma_G = h\sigma_P$$

Здесь ситуация не симметрична.

Графически это изображается следующим образом:



Коэффициент наследуемости (гетерогенные родители и потомки)

Lush. I. L. 1937. Animal Breeding Plans. Collegiate Press, Inc., Ames, IA.

$$h^2 = \frac{\sigma_F^2 - \sigma_E^2}{\sigma_F^2} = \frac{\sigma_G^2}{\sigma_F^2}$$

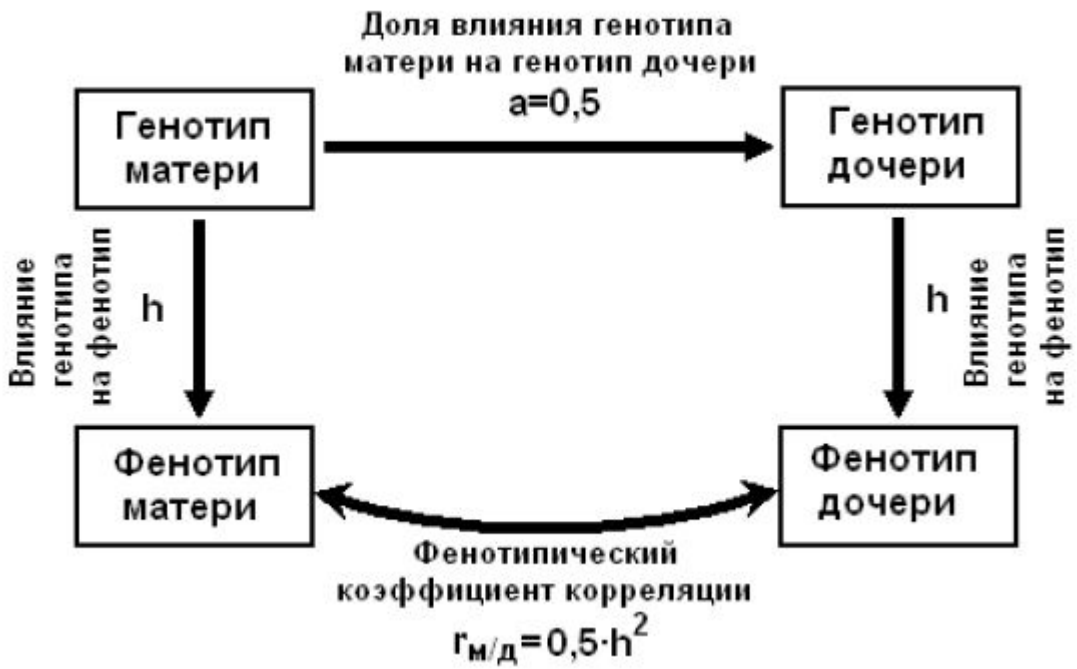
через корреляции “родитель-потомок”:

$$h^2 = 2r$$

один родитель - потомок

Коэффициент наследуемости

через корреляцию фенотипов мать—дочь



Согласно теореме цепных корреляций, корреляция между концами цепи равна произведению корреляции звеньев, их связывающих. Тогда

$$r = 0.5 \cdot h \cdot h = 0.5 \cdot h^2 \qquad h^2 = 2r$$

Коэффициент наследуемости отец-дочь

через корреляцию фенотипов дочерей -- полусибсов по отцу

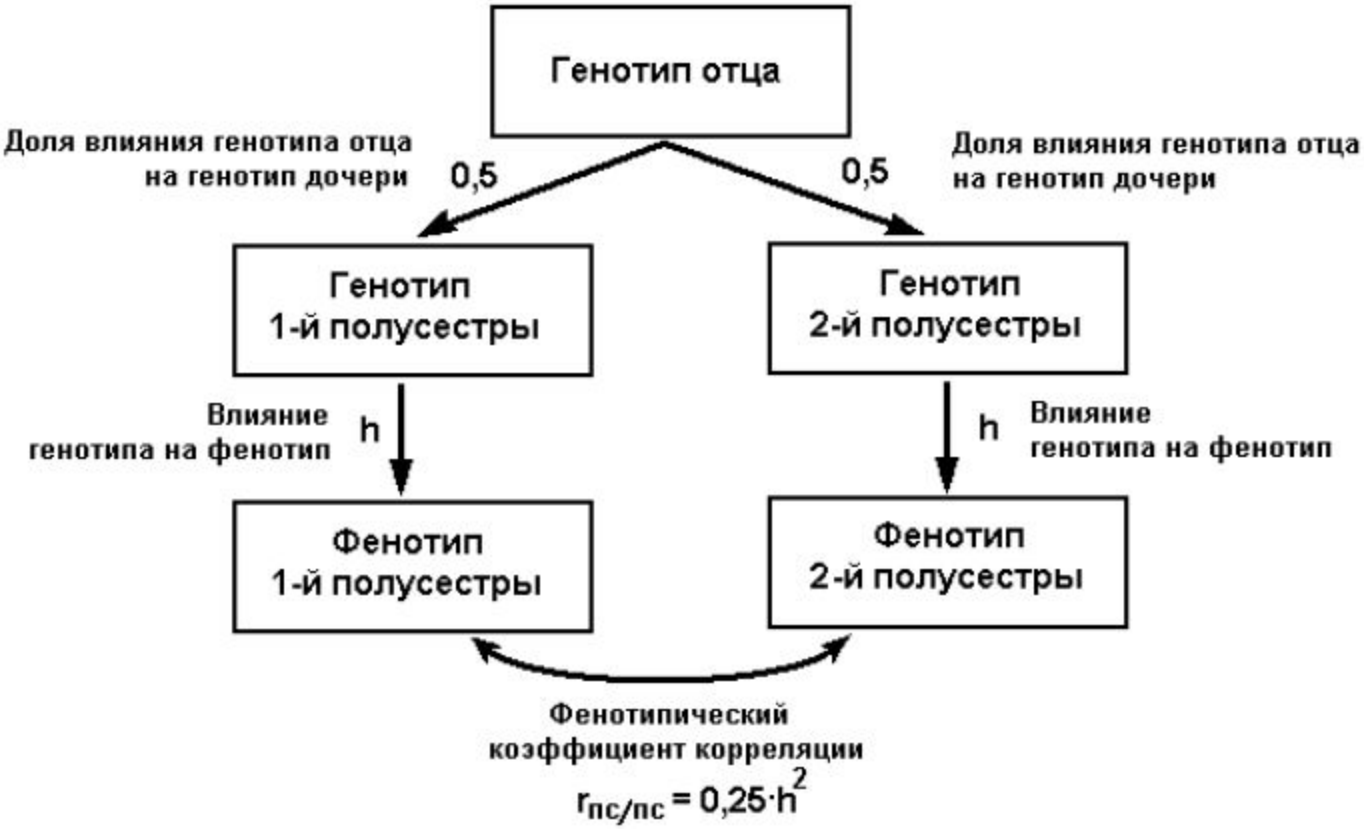
Сходство родственников, принадлежащих к разным поколениям (предки — потомки), обычно оценивается коэффициентом корреляции Пирсона, который называют также межклассовым коэффициентом корреляции. В случае близнецов и сиблингов применяется коэффициент внутриклассовой корреляции, подсчитываемый на основе дисперсионного анализа:

$$r = \frac{2\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \frac{(x_i' - \bar{x})}{+\sum(x_i' - \bar{x})^2}$$

где x_i' и x_i — значения одного и того же признака у членов одной пары.

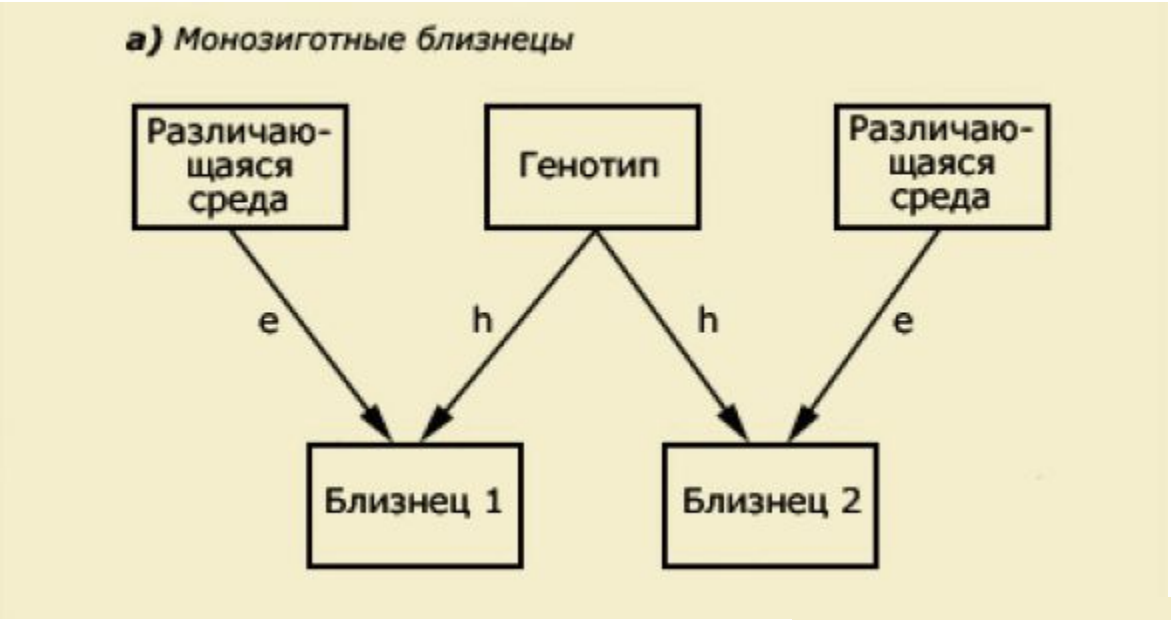
Коэффициент наследуемости отец-дочь

через корреляцию фенотипов дочерей -- полусибсов по отцу



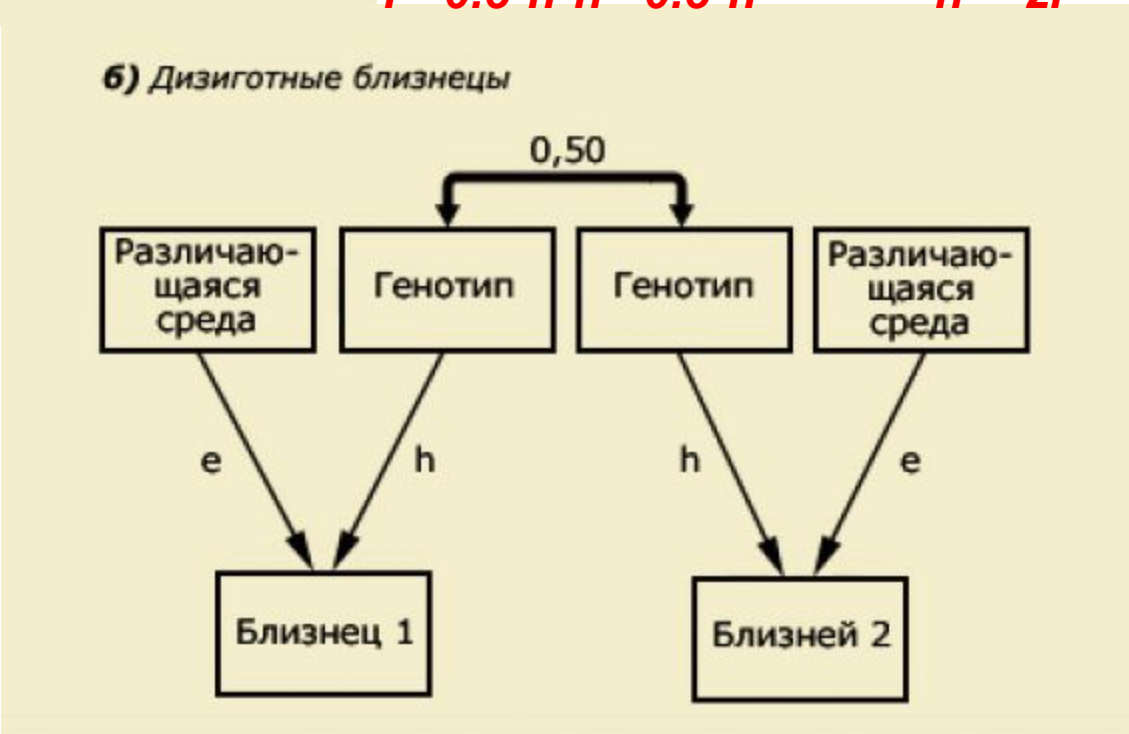
$$r = 0.5 \cdot h \cdot 0.5 \cdot h = 0.25 \cdot h^2 \qquad h^2 = 4r$$

Путевые диаграммы Райта для близнецов



$r = 0.5 \cdot h \cdot h = 0.5 \cdot h^2$ $h^2 = 2r$

$r = h \cdot h = h^2$ $h^2 = r$



Путевые диаграммы Райта для близнецов

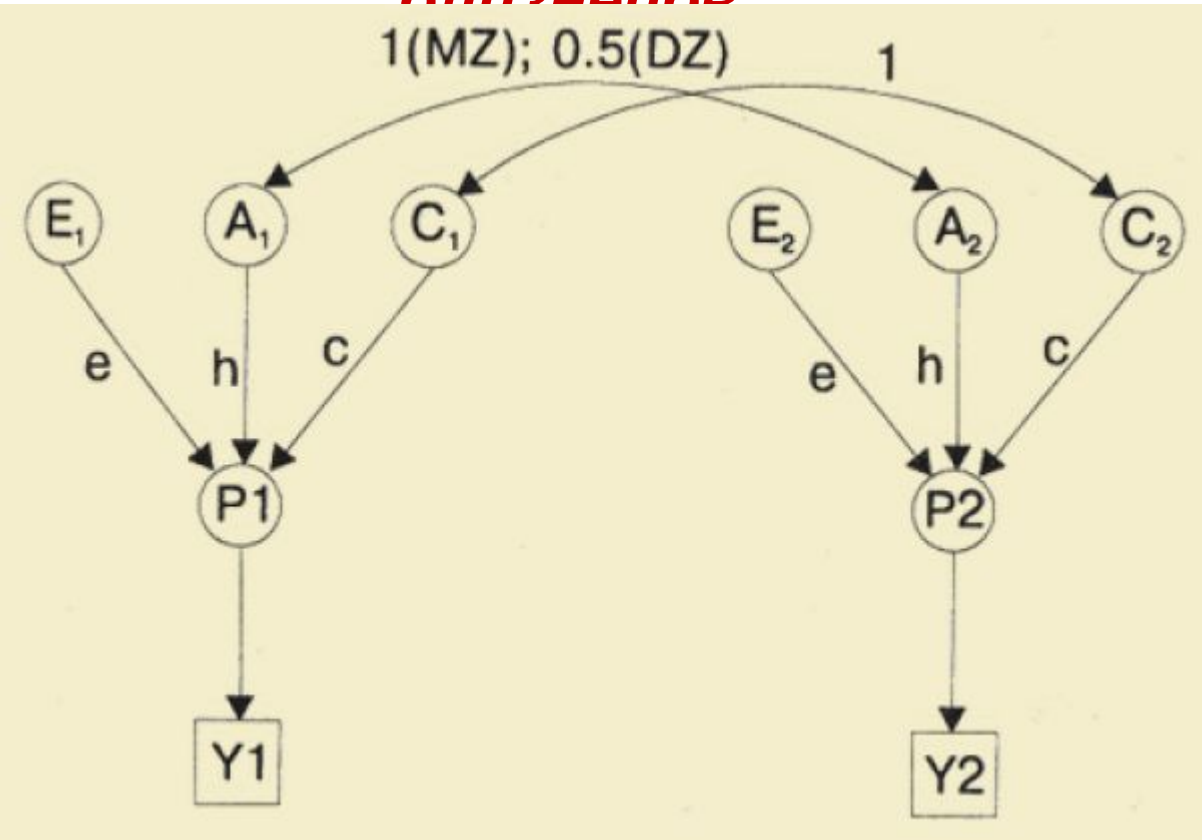
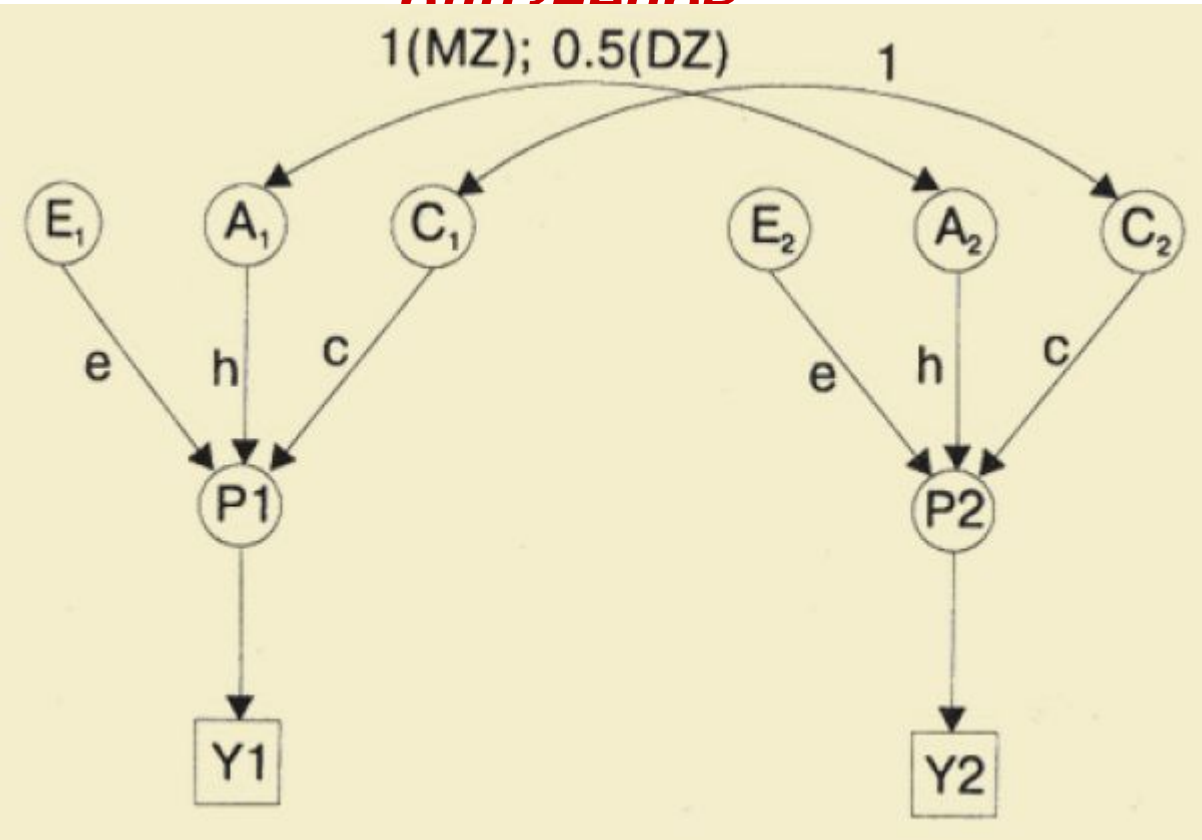


Диаграмма путей для монозиготных и дизиготных близнецов. P1, E1, C1, A1 - латентные переменные для первого близнеца (фенотип, различающаяся среда, общая среда, аддитивное действие генов); P2, E2, C2, A2 - латентные переменные для второго близнеца; Y1, Y2 - наблюдаемые переменные для первого и второго близнеца

Путевые диаграммы Райта для близнецов

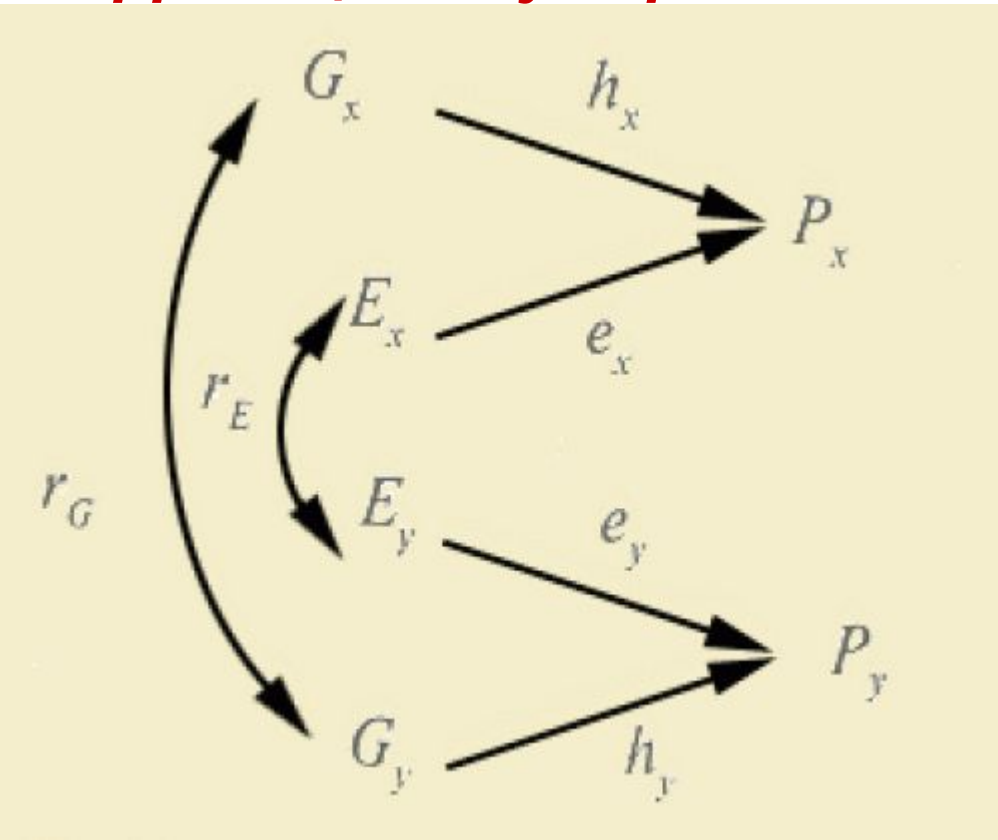


Последовательные пути перемножаются, параллельные складываются

$$r_{MZ} = 1 \bullet h \bullet h + c \bullet c = h^2 + c^2 \qquad r_{DZ} = 0.5 \bullet h \bullet h + c \bullet c = 0.5h^2 + c^2$$

$$h^2 = 2(r_{MZ} - r_{DZ})$$

Диаграмма путей фенотипической корреляции двух признаков



P_x и P_y - коррелирующие фенотипические признаки;
r_G - генетическая корреляция; r_E - средовая корреляция;
G – гены; E – среда; h – наследуемость; e – влияние среды.

$$r(P_x, P_y) = h_x h_y r_G + e_x e_y r_E$$

Структурное моделирование

Структурное моделирование применяется в эконометрике и представляет собой один из наиболее сложных современных методов. По существу, это разновидность множественной регрессии, осложненная возможностью обратного влияния выходных переменных на входные. Применение этого метода требует соответствующей квалификации исследователя и наличия компьютерных программ, специально разработанных для этих целей (LISREL, EQS). Метод используется для анализа большого количества зависимых и независимых переменных, включенных в различные гипотезы исследования. Оценка и тестирование моделей при этом требует наличия больших выборок и современного компьютерного обеспечения. С.Райт является родоначальником этого научного направления.

Структурное моделирование

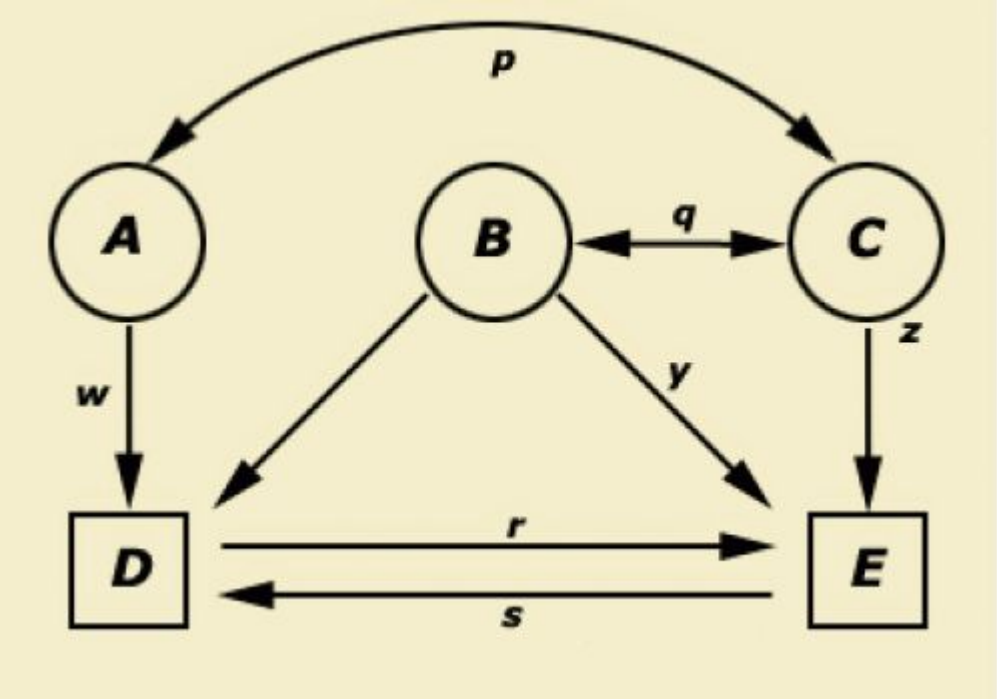


Диаграмма путей, объединяющая три латентных (A, B, C) и две наблюдаемых (D и E) переменных, p и q - корреляции; r, s, w, x, y, z - путевые коэффициенты.

Спасибо за внимание!

© ИЦиГ СО РАН

Многомерный аналог коэффициента наследуемости

Lande R (1979). Quantitative genetic analysis of multivariate evolution applied to brain:body size allometry. Evolution 33:402–416.

$$H = GP^{-1}$$

G **– матрица коэффициентов корреляции между родителями и потомками**

P **– фенотипическая матрица корреляций между признаками**

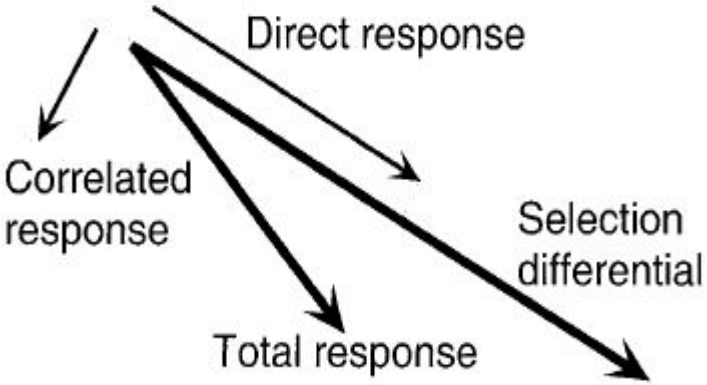
Уравнение селекционера

Smith, H.F. 1936. A discriminant function for plant selection. Ann. Eugen. 7: 240-250.

$$\Delta\mu = GP^{-1}s = Hs$$

s **– селекционный дифференциал**

$\Delta\mu$ **– ответ на отбор**



Родители (X)

X_{11}	X_{12}	…	$X_{1,M-1}$	$X_{1,M}$
X_{21}	X_{22}	…	$X_{2,M-1}$	$X_{2,M}$
X_{31}	X_{32}	…	$X_{3,M-1}$	$X_{3,M}$
…	…	…	…	…
…	…	…	…	…
$X_{N-2,1}$	$X_{N-2,2}$	…	$X_{N-2,M-1}$	$X_{N-2,M}$
$X_{N-1,1}$	$X_{N-1,2}$	…	$X_{N-1,M-1}$	$X_{N-1,M}$
$X_{N,1}$	$X_{N,2}$	…	$X_{N,M-1}$	$X_{N,M}$

$$P = R_{X/X}$$

Потомки (Y)

Y_{11}	Y_{12}	…	$Y_{1,M-1}$	$Y_{1,M}$
Y_{21}	Y_{22}	…	$Y_{2,M-1}$	$Y_{2,M}$
Y_{31}	Y_{32}	…	$Y_{3,M-1}$	$Y_{3,M}$
…	…	…	…	…
…	…	…	…	…
$Y_{N-2,1}$	$Y_{N-2,2}$	…	$Y_{N-2,M-1}$	$Y_{N-2,M}$
$Y_{N-1,1}$	$Y_{N-1,2}$	…	$Y_{N-1,M-1}$	$Y_{N-1,M}$
$Y_{N,1}$	$Y_{N,2}$	…	$Y_{N,M-1}$	$Y_{N,M}$

$$G = R_{X/Y}$$

Поиск осей с максимальной наследуемостью
(в узком смысле)

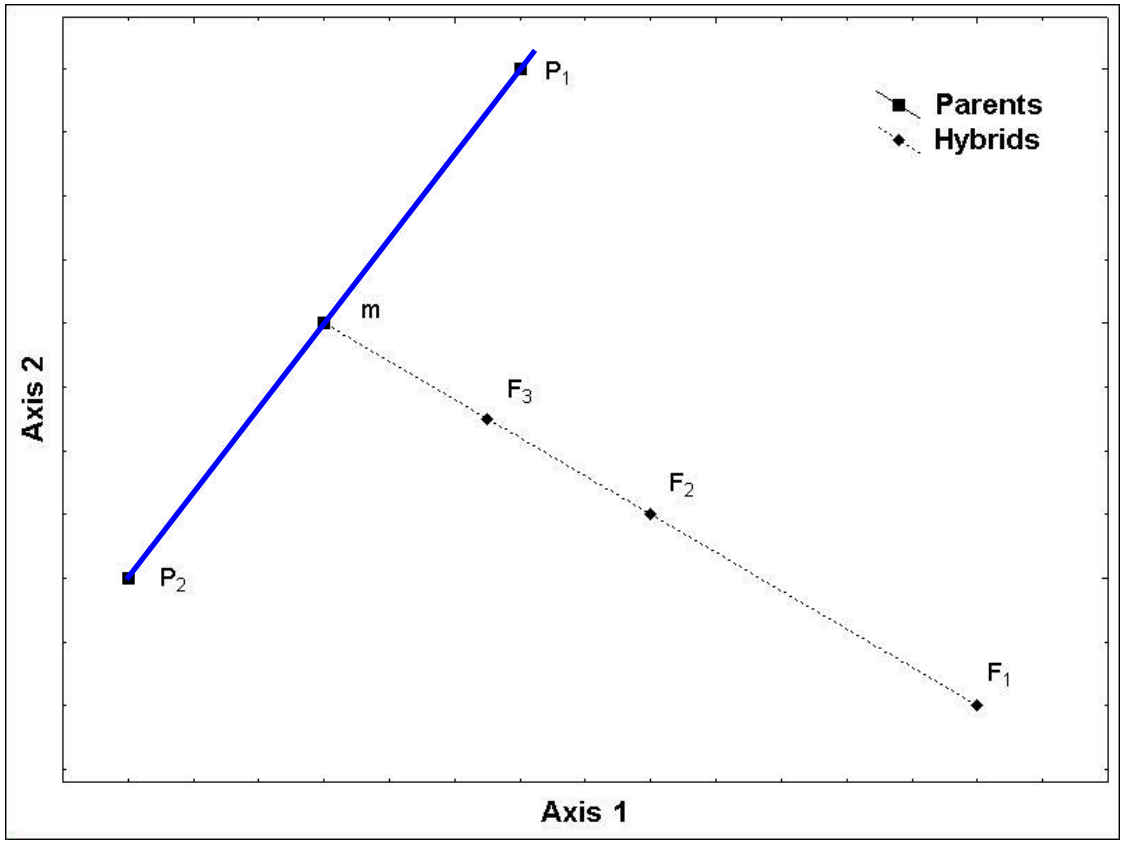
Ott J, Rabinowitz D (1999). A principal-components approach based on heritability for combining phenotype information. Hum Hered 49: 106–111.

Klingenberg CP, Leamy L. 2001. Quantitative genetics of geometric shape in the mouse mandible. Evolution 55(11): 2342–2352.

$$\Delta\mu = GP^{-1}s = Hs = \lambda s$$

Многомерная наследуемость

Расположение центроидов родительских и гибридных выборок в многомерном пространстве признаков при аддитивно-доминантной модели наследования

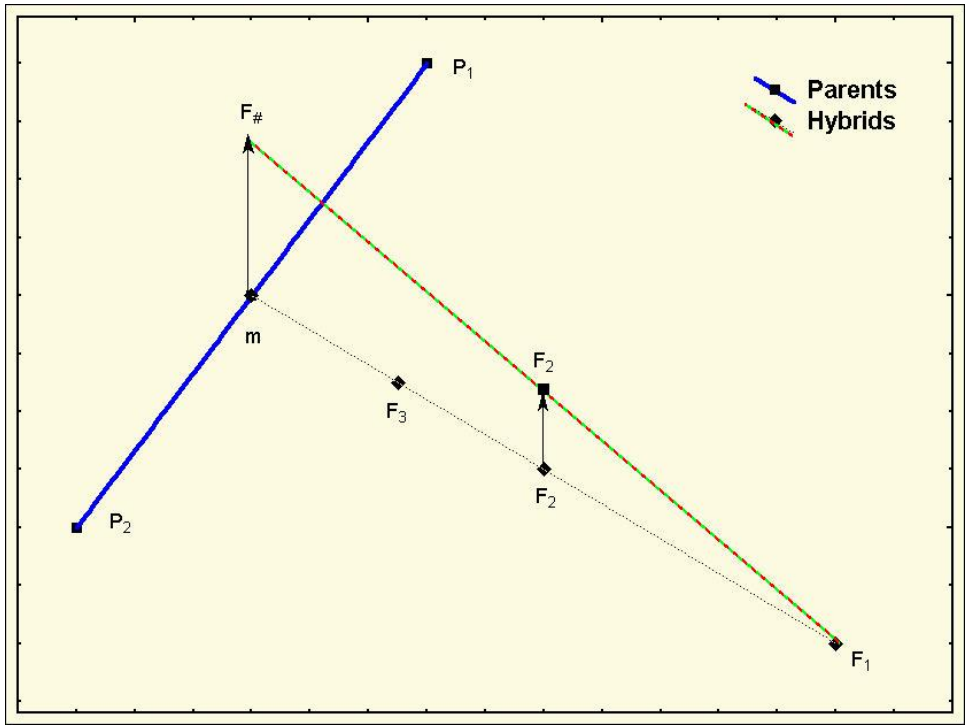


F1 - m – ось гетерозиготности;
P1 - P2 – ось аддитивности

Heredity, 2005. V. 94. P. 101-107.

Многомерная наследуемость

Расположение центроидов родительских и гибридных выборок в многомерном пространстве признаков в общем случае (*HIA*-модель)



F1 - F# – ось гетерозиготности ***H***
P1 - P2 – ось аддитивности ***A***
m - F# – ось эпистаза ***I***
(F1 - m – ось гетерозиготности в аддитивно-доминантной модели)

Материал

Am. J. Hum. Genet. 75:1094–1103, 2004

Genetic Inheritance of Gene Expression in Human Cell Lines

S. A. Monks,[1,2,4] A. Leonardson,[4] H. Zhu,[4] P. Cundiff,[3] P. Pietrusiak,[5] S. Edwards,[4] J. W. Phillips,[6] A. Sachs,[4] and E. E. Schadt[4]

[1]Department of Statistics, Oklahoma State University, Stillwater, OK; Departments of [2]Biostatistics and [3]Pharmacology, University of Washington, and [5]Rosetta Inpharmatics LLC, Seattle; [6]Department of Epidemiology, Johns Hopkins University, Baltimore; and [4]Merck Research Laboratories, Merck & Co., Rahway, NJ

Combining genetic inheritance information, for both molecular profiles and complex traits, is a promising strategy not only for detecting quantitative trait loci (QTLs) for complex traits but for understanding which genes, pathways, and biological processes are also under the influence of a given QTL. As a primary step in determining the feasibility of such an approach in humans, we present the largest survey to date, to our knowledge, of the heritability of gene-expression traits in segregating human populations. In particular, we measured expression for 23,499 genes in lymphoblastoid cell lines for members of 15 Centre d'Etude du Polymorphisme Humain (CEPH) families. Of the total set of genes, 2,340 were found to be expressed, of which 31% had significant heritability when a false-discovery rate of 0.05 was used. QTLs were detected for 33 genes on the basis of at least one P value <.000005. Of these, 13 genes possessed a QTL within 5 Mb of their physical location. Hierarchical clustering was performed on the basis of both Pearson correlation of gene expression and genetic correlation. Both reflected biologically relevant activity taking place in the lymphoblastoid cell lines, with greater coherency represented in Kyoto Encyclopedia of Genes and Genomes database (KEGG) pathways than in Gene Ontology database pathways. However, more pathway coherence was observed in KEGG pathways when clustering was based on genetic correlation than when clustering was based on Pearson correlation. As more expression data in segregating populations are generated, viewing clusters or networks based on genetic correlation measures and shared QTLs will offer potentially novel insights into the relationship among genes that may underlie complex traits.

Introduction

In 1980, Botstein et al. proposed that sequence differences be treated as markers, in order to map genes involved in inherited traits. Since that time, the number of genes mapped to positions in the human genome has grown exponentially. Mapping these genes for inherited traits has been extremely successful for simple Mendelian diseases; however, finding such genes for diseases—and their associated risk traits—that are of large public health interest has proven difficult. Reasons for this difficulty include disease heterogeneity (disease subtypes with some or no overlapping genetic causes), misclassification (from using discrete classifications of disease from thresholds and combinations of thresholds), and unaccounted-for environmental influences. With the advent of technology to measure changes in molecular profiles—for example, changes in mRNA transcript abundance, protein levels, and metabolite levels—it should be possible to unravel some of the complexity of these complex diseases. In particular, gene expression can be viewed as a more refined phenotype, since it is a measure of phenotypic variation at the molecular level. In addition, each gene-expression phenotype provides annotation, pathway, and genome location data. Combining these data with genetic-inheritance information, for both molecular profiles and complex traits, is a promising strategy not only for detecting QTLs for complex traits but for understanding which genes, pathways, and biological processes are also under the influence of a given QTL.

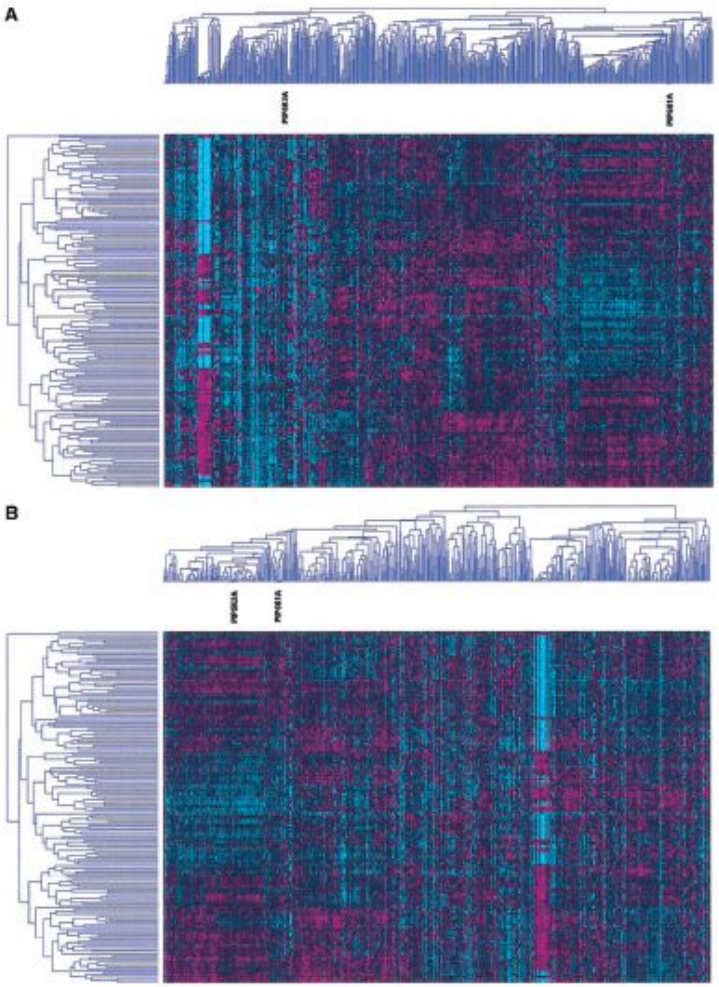
Jansen and Nap (2001) were among the first to suggest the use of expression profiles in segregating populations. They discussed the power of using well-developed methods and designs available for dissecting quantitative traits along with the rapidly expanding collection of methods for large-scale sets of phenotypes. They provided an illustration that combined linkage data from a set of genes with known genomic locations, to construct a putative pathway. Jin et al. (2001) studied the contributions of sex, genotype, and age on transcription in *Drosophila melanogaster* through a study of two inbred lines of *Drosophila*. They observed a large sex effect on expression and less of an effect due to

Received July 9, 2004; accepted for publication October 1, 2004; electronically published October 21, 2004.

Address for correspondence and reprints: Dr. Stephanie A. Monks, Department of Statistics, 301G Mathematics, Statistics, and Computer Science Building, Oklahoma State University, Stillwater, OK 74078-1056. E-mail: stephanie.monks@okstate.edu

© 2004 by The American Society of Human Genetics. All rights reserved.

0002-9297/2004/7506-0014$15.00



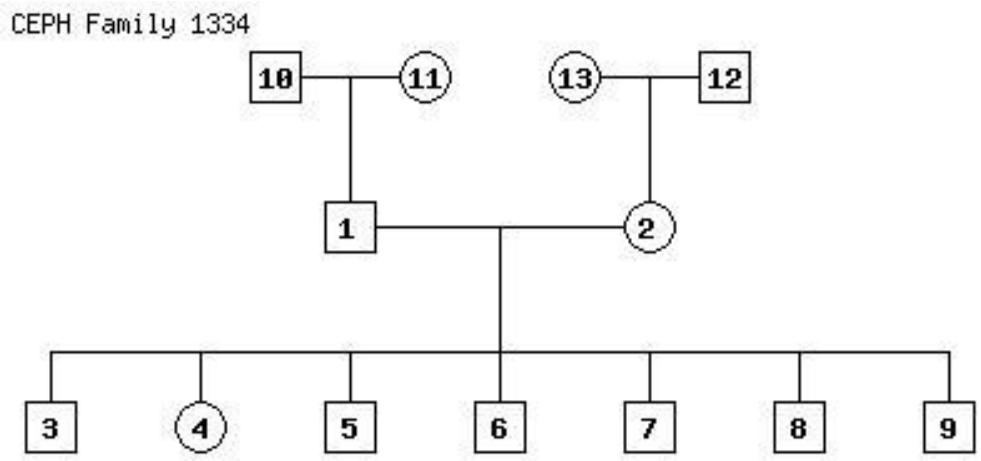
Family 1424 Members

Records Return: (14)

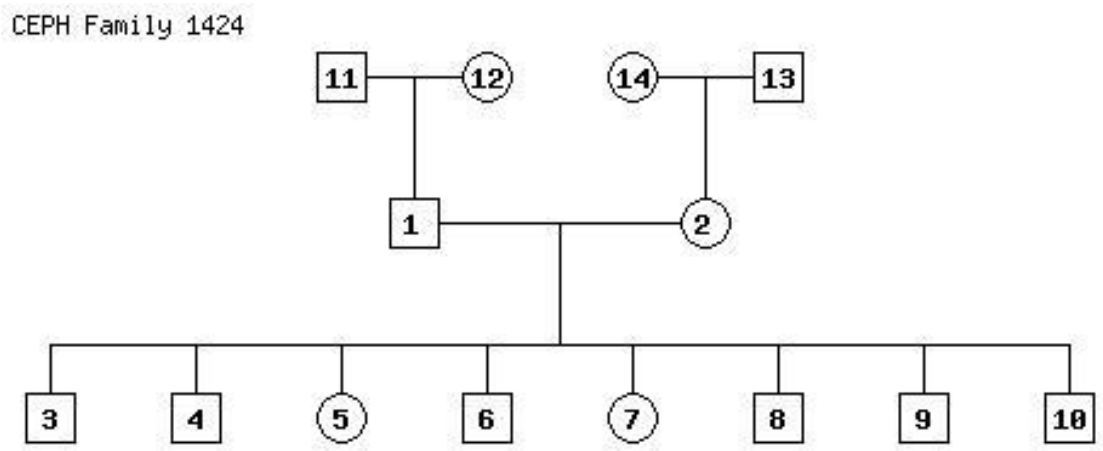
Show 20

Catalog ID	Description	Sex	Age	Race	Affected	Family Member	Relationship to Proband	Cell Type
GM10844	CEPH/UTAH PEDIGREE 1424	Female	50 YR	Caucasian		2	mother	B-Lymphocyte
GM10845	CEPH/UTAH PEDIGREE 1424	Male	49 YR	Caucasian		1	father	B-Lymphocyte
GM11922	CEPH/UTAH PEDIGREE 1424	Male	30 YR	Caucasian		3	son	B-Lymphocyte
GM11923	CEPH/UTAH PEDIGREE 1424	Male	29 YR	Caucasian		4	son	B-Lymphocyte
GM11924	CEPH/UTAH PEDIGREE 1424	Female	27 YR	Caucasian		5	daughter	B-Lymphocyte
GM11925	CEPH/UTAH PEDIGREE 1424	Male	25 YR	Caucasian		6	son	B-Lymphocyte
GM11926	CEPH/UTAH PEDIGREE 1424	Female	21 YR	Caucasian		7	daughter	B-Lymphocyte
GM11927	CEPH/UTAH PEDIGREE 1424	Male	19 YR	Caucasian		8	son	B-Lymphocyte
GM11928	CEPH/UTAH PEDIGREE 1424	Male	15 YR	Caucasian		9	son	B-Lymphocyte
GM11929	CEPH/UTAH PEDIGREE 1424	Male	13 YR	Caucasian		10	son	B-Lymphocyte
GM11930	CEPH/UTAH PEDIGREE 1424	Male		Caucasian		11	paternal grandfather	B-Lymphocyte
GM11931	CEPH/UTAH PEDIGREE 1424	Female	78 YR	Caucasian		12	paternal grandmother	B-Lymphocyte
GM11932	CEPH/UTAH PEDIGREE 1424	Male	76 YR	Caucasian		13	maternal grandfather	B-Lymphocyte
GM11933	CEPH/UTAH PEDIGREE 1424	Female	74 YR	Caucasian		14	maternal grandmother	B-Lymphocyte

Родственные связи



Родственные связи



Коэффициенты корреляции родителей с потомками по первым пяти компонентам с максимальной аддитивной наследуемостью

(выделены достоверные при $p<0.05$; N=196)

Parents	Descedants				
	Factor1	Factor2	Factor3	Factor4	Factor5
Factor1	***0.66***	0.05	0.06	-0.04	0.02
Factor2	-0.06	***0.35***	***0.15***	-0.09	0.06
Factor3	-0.09	-0.12	***0.32***	0.08	0.02
Factor4	0.00	0.06	-0.06	***0.14***	0.09
Factor5	-0.06	-0.06	-0.01	-0.10	0.03

Коэффициент наследуемости

через корреляции “родитель-потомок”:

$$h^2 = 2r$$

один родитель - потомок

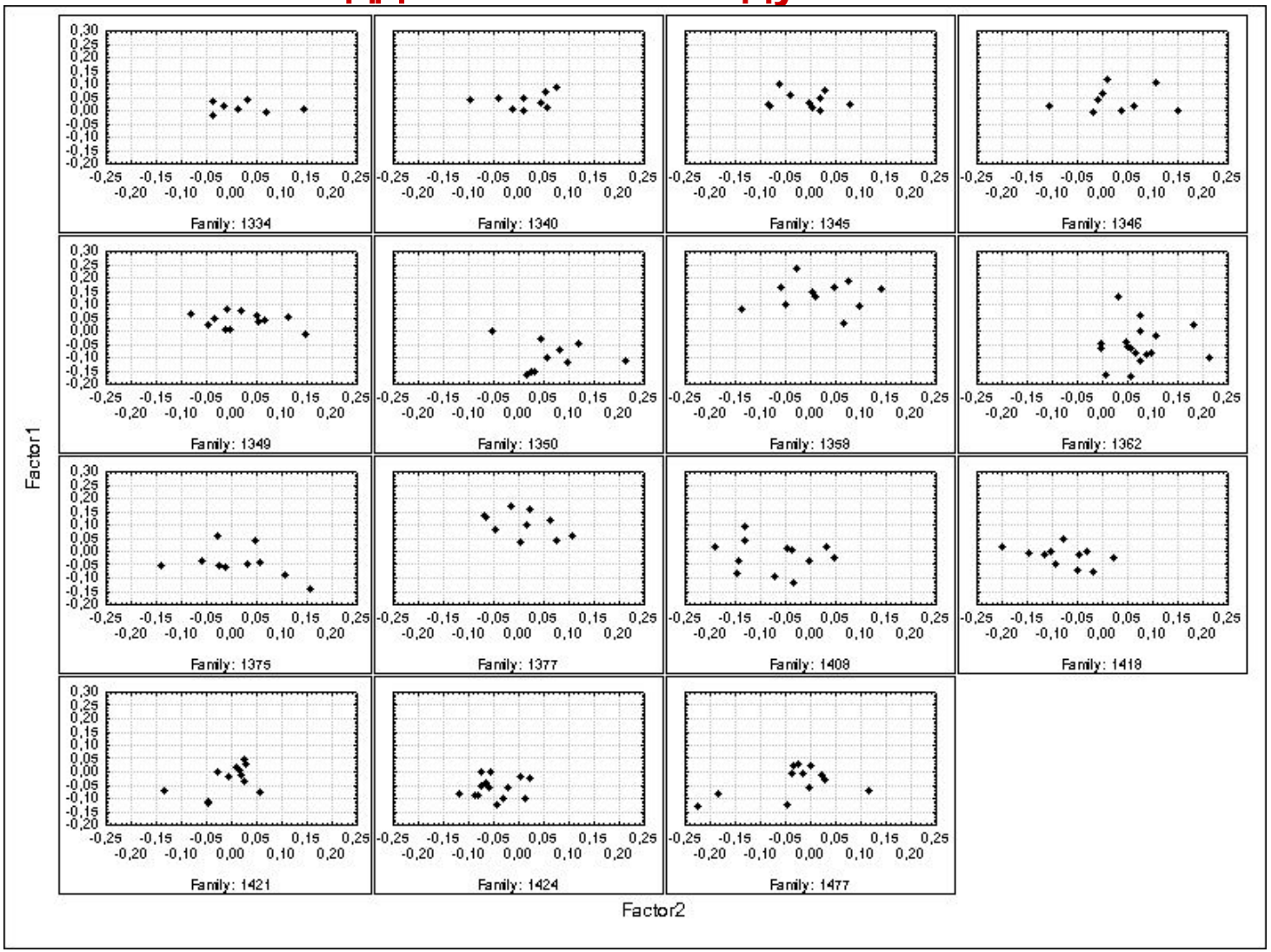
(в случае отсутствия ассортативности)

$$h^2 = \frac{2r}{1+\rho_a}$$

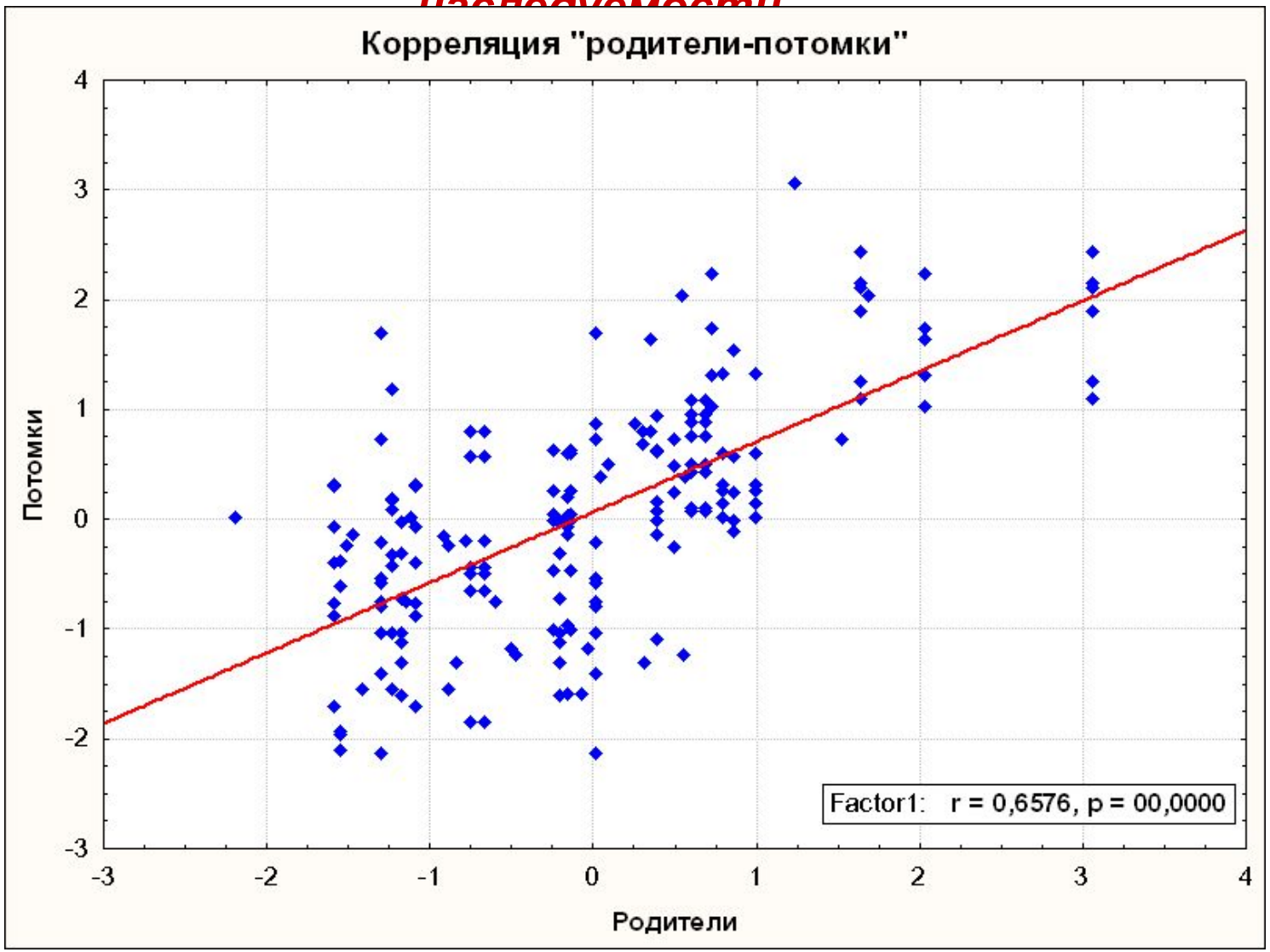
один родитель - потомок

(при ассортативности)

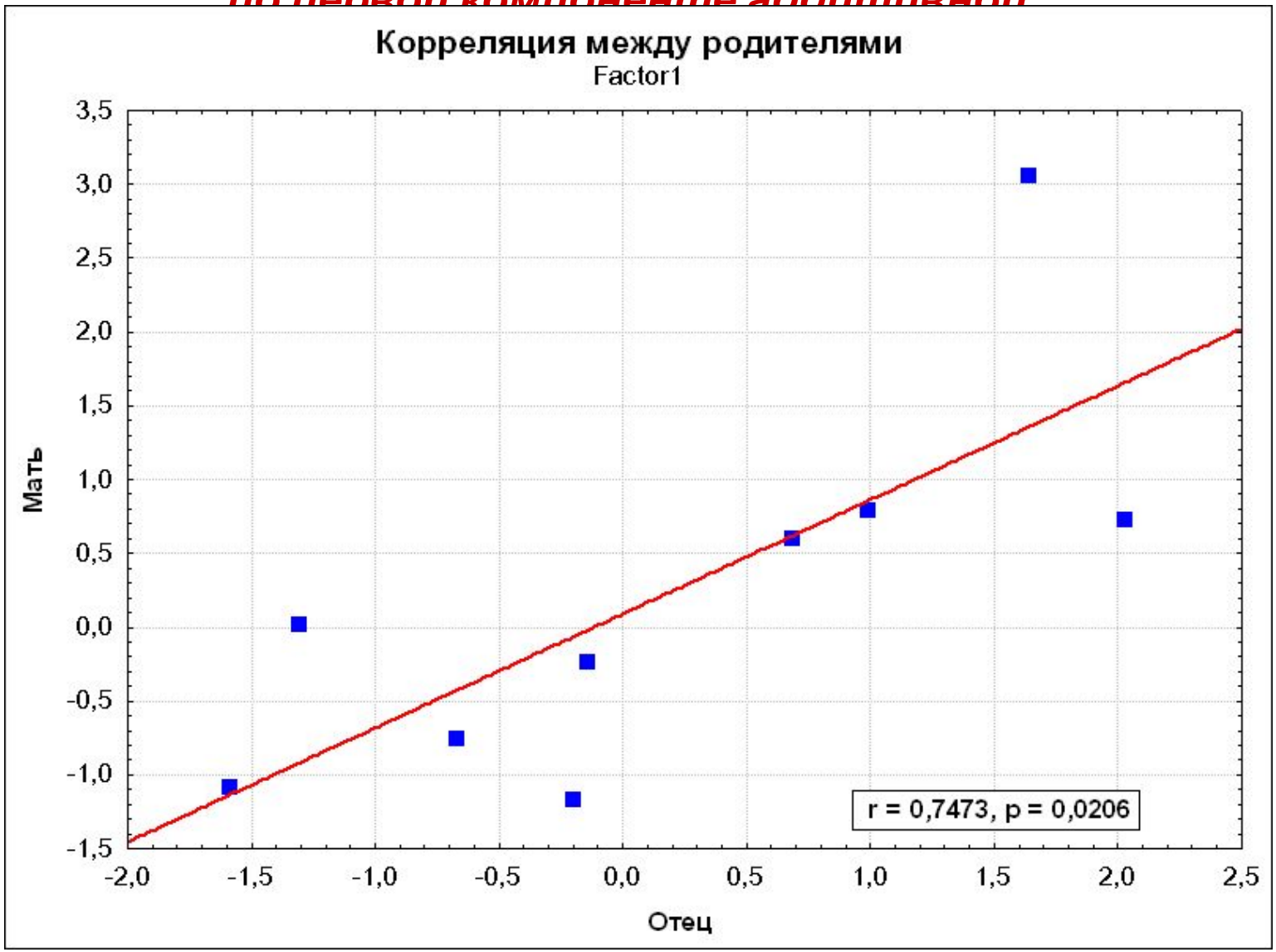
Расположение семей на плоскости первых двух компонент аддитивной наследуемости



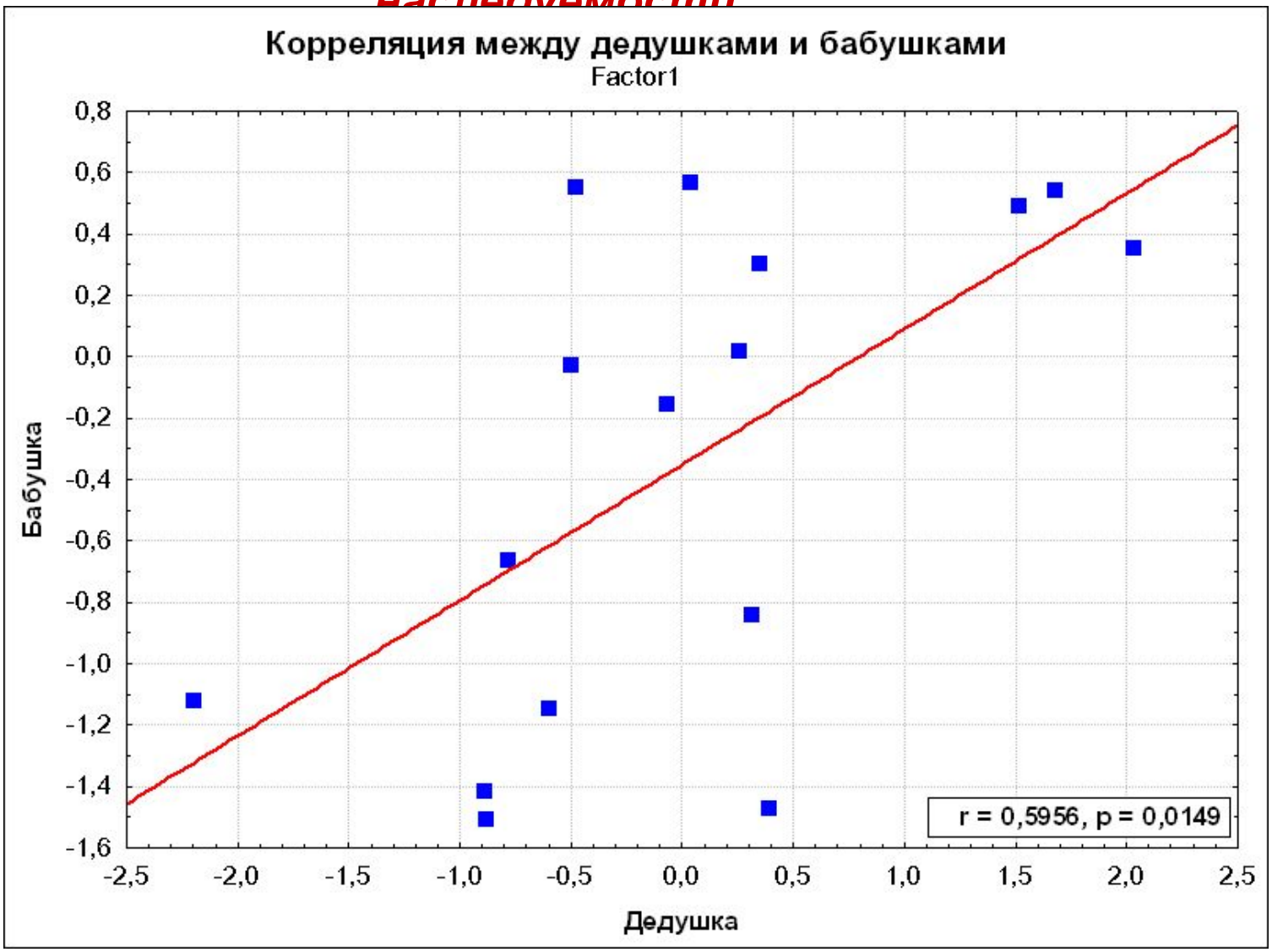
Корреляция между родителями и детьми по первой компоненте аддитивной наследуемости



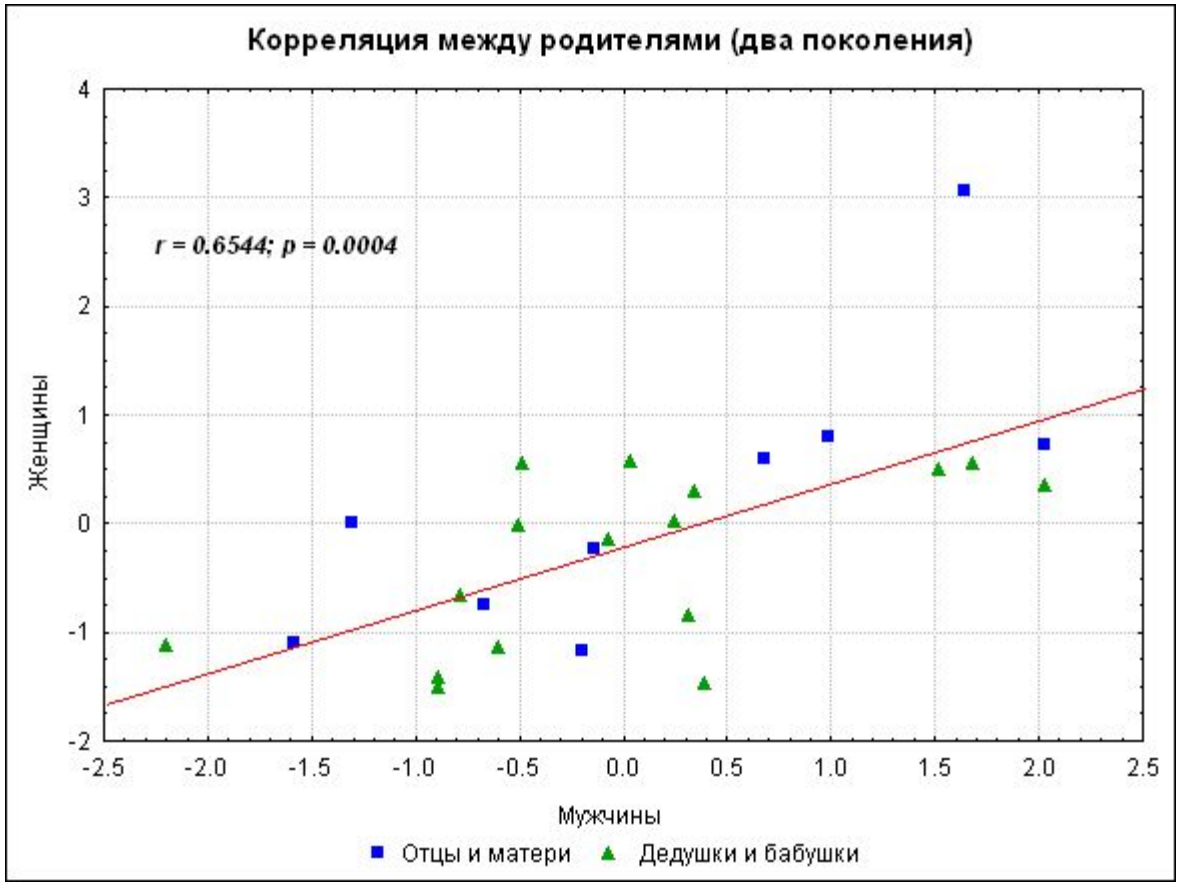
Корреляция между родителями (ассортативность) по первой компоненте аддитивной



Корреляция между дедушками и бабушками по первой компоненте аддитивной наследуемости



Корреляция между родителями (два поколения)



После учета ассортативности выявилось, что четыре компоненты шкалирования имеют значимые коэффициенты наследуемости

	Factor1	Factor2	Factor3	Factor4	Factor5
r	0.66	0.35	0.32	0.14	0.03
ρ_a	0.75	0.42	0.29	-0.3	0.75
h^2	0.75	0.49	0.50	0.40	0.03