Πρόγραμμα Μεταπτυχιακών Σπουδών

«ΜΕΤΑΦΡΑΣΤΙΚΗ ΕΡΕΥΝΑ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ»

Μοριακή Διαγνωστική, Βιοδείκτες και Στοχευμένες Θεραπείες

2023-2024

Ταυτοποίηση μοριακών δεικτών

20/04/2024

Μαριάνθη Γεωργίτση, PhD
Επίκουρη Καθηγήτρια Γενετικής και Μοριακής Βάσης Ασθενειών

Εργαστήριο Γονιδιωματικής Ποικιλότητας και Γενετικής Επιδημιολογίας
Τμήμα Μοριακής Βιολογίας και Γενετικής - Σχολή Επιστημών Υγείας
Δημοκρίτειο Πανεπιστήμιο Θράκης

E-mail: mgeorgit@mbg.duth.gr

ΔΙΑΛΕΞΗ

«Προοπτικές και περιορισμοί στη χρήση γενετικών δεικτών στη μεταφραστική έρευνα πολυπαραγοντικών νοσημάτων»

ΠΕΡΙΕΧΟΜΕΝΑ

- **Λεξιλόγιο όρων**
- **Εισαγωγή (Βασικές έννοιες, προσεγγίσεις ανάλυσης δεδομένων και εργαλεία)**
- **Προοπτικές και περιορισμοί στη χρήση γενετικών δεικτών ως βιοδείκτες**
- **Το παράδειγμα του Σακχαρώδους Διαβήτη Τύπου 2 (ΣΔΤ2)**

Published: 11 June 2013

Pleiotropy in complex traits: challenges and strategies

Nadia Solovieff, Chris Cotsapas, Phil H. Lee, Shaun M. Purcell & Jordan W. Smoller

Nature Reviews Genetics **14**, 483–495(2013) | Cite this article

Glossary

Genome-wide association studies	(GWASs). Studies in which hundreds of thousands (or millions) of genetic markers are tested for association with a phenotypic trait; they are an unbiased approach to survey the entire genome for disease-associated regions using common variation.
Genome-wide-significant	A term describing the statistical significance threshold that accounts for multiple testing in GWASs.
Complex traits	Traits controlled by a combination of many genes and environmental factors.
Pleiotropy	A gene or genetic variant that affects more than one phenotypic trait.
Heritability	The proportion of phenotypic variance attributed to genetic differences among individuals in a population.
Colocalizing	Different genetic variants in high linkage disequilibrium located in the same gene that affect different phenotypes.
Single-nucleotide polymorphisms	(SNPs) Single-nucleotides in the genome that vary across individuals in the population.
Linkage disequilibrium	(LD). The correlation between genetic markers owing to limited recombination.
Copy number variants	Regions of the genome in which the copy number is polymorphic (for example, deletions and duplications) across individuals.
Polygenic	Controlled by many genes.
Population stratification	A source of bias in genome-wide association studies that occurs when a phenotype and the allele frequency of a single-nucleotide polymorphism vary owing to ancestral differences.
Batch effect	Systematic biases in the data that arise from differences in sample handling.
Genotype imputation	Inference of missing genotypes or untyped single-nucleotide polymorphisms using statistical techniques.
Ascertainment bias	A consequence of collecting a nonrandom subsample with a systematic bias so that results based on the subsample are not representative of the entire sample.
Tag SNPs	(tagSNP) Single-nucleotide polymorphisms (SNPs) chosen to represent a region of the genome owing to strong linkage disequilibrium.
Multivariate analyses	The simultaneous inclusion of two or more phenotypes in one analysis when testing the association with a genetic variant.
Univariate analyses	Tests of association between one phenotype and a genetic variant.
Polygenic scoring	(PRS) A score that aggregates the number of risk alleles a subject carries weighted by the effect size of the allele for a particular trait. The risk allele and effect size for each single-nucleotide polymorphism is generally taken from a genome-wide association study of an independent study.
Linear mixed-effect model	A linear model that contains both fixed and random effects. This type of model can be used to estimate genetic correlation between traits using a genome-wide set of single-nucleotide polymorphisms.
Cohort studies	Observational studies in which defined groups of people (the cohorts) are

	followed over time and outcomes are compared in subsets of the cohort who were exposed to different levels of factors of interest. These studies can either be prospectively or retrospectively carried out from historical records.
Cross-sectional studies	Studies in which data are collected on subjects at one specific point in time and subjects are not selected for a particular trait or exposure.
Case–control study	Compares cases (that is, a selected group of individuals: for example, those diagnosed with a disorder) with controls (that is, a comparison group of individuals: for example, those who are not diagnosed with the disorder). Genome-wide association case–control studies test whether genetic marker allele frequencies differ between cases and controls.
Non-parametric approach	A statistical analysis method that does not rely on specific distributional assumptions (for example, normality) for the variables being analysed.
Principal components analysis	(PCA) A statistical method used to simplify data sets by transforming a series of correlated variables into a smaller number of uncorrelated factors. It is also commonly used to infer continuous axes of variation in genetic data, often representing genetic ancestry.
Summary statistics	A statistic that summarizes a set of observations. In the context of genome-wide association studies, meta-analyses can be carried out solely by using summary statistics and typically include estimates of the effect size (for example, odds ratio) and standard error.
Effect heterogeneity	Different effect sizes across phenotypes.
Expression quantitative trait loci	(eQTLs) Loci at which genetic allelic variation is associated with variation in gene expression.
Fine mapping	Extensively genotyping or sequencing a region of the genome that was identified in genome-wide association studies to identify the causal variant.
Confounding factor	A variable (for example, batch effects or population structure) that is associated with both the genotype and the phenotype of interest and can give rise to a spurious association.
Genetic architecture	A genetic model (that is, the number of single-nucleotide polymorphisms, effect sizes, allele frequency, and so on) underlying a phenotypic trait.

ΣΗΜΕΙΩΣΕΙΣ ΣΤΙΣ ΔΙΑΦΑΝΕΙΕΣ ΤΗΣ ΠΑΡΟΥΣΙΑΣΗΣ

Διαφάνεια	Σημειώσεις/Βιβλιογραφία
3	https://reporter.nih.gov/search/-8S94HETAE-9hbOmmbj3bg/projects/charts
11	Narrow-sense heritability (h^2) is an important genetic parameter that quantifies the proportion of phenotypic variance in a trait attributable to the additive genetic variation generated by all causal variants. Estimation of h^2 previously relied on closely related individuals, but recent developments allow estimation of the variance explained by all SNPs used in a genome-wide association study (GWAS) in conventionally unrelated individuals, that is, the SNP-based heritability (h^2_{SNP}), which is part of the narrow-sense heritability.
12	The heritability gap in six major psychiatric disorders. Heritability (the proportion of causation attributable to genetic factors) estimated from family and twin studies (family/twin estimates, black bars) and from GWAS of SNPs and CNVs across the human genome (molecular estimates, red bars). The large difference between the family/twin and molecular estimates is indicated as the “heritability gap” (gray bars).
17	Genome-wide association studies help identify candidate genes for complex disorders, using existing tools and also novel methodology, which allow us to correlate: a) genetic variation to disease susceptibility, b) genetic variation to disease subphenotypes/subcategories, or c) environmental to genetic factors, as well as clusters of genes to specific phenotypes.

	Quantitative versus Categorical phenotypes: Quantitative traits offer improved statistical power to detect a genetic effect (eg variants that influence the levels of a quantitative trait can have a clear interpretation – HDL/LDL levels for instance in an allele- or genotype-specific manner).
18	QC: percentage of genotype calls per individual (>90% accepted), SNP call rates (% of missing SNP data, >95% accepted), Hardy-Weinberg Equilibrium, cryptic relatedness, ancestry, MAFs, population stratification, concordance rates if using duplicate samples. Q-Q plots: a diagnostic plot that compares distribution of observed test statistics with the distribution expected under the null. Haplotype-based methods, imputation methods The causal variant will only occasionally be among those directly typed by GWAS scans and the interval within which the aetiological variant(s) are expected to lie may contain several genes. In the case of regulatory elements, susceptibility genes could lie well beyond the interval of association.
23	Limitations of GWAS: - Low to moderate risk (10-40%) - Markers not directly associated with trait/disease (tagSNPs – not necessarily pathogenic) - Markers lying outside gene regions (indirect or unclear association) - Missing heritability - Varying risk contribution across populations (most GWAS on Caucasians) - Sex chromosomes omitted from analysis (until recently) - Genes x Environment interactions (unclear)
25	The solid X indicates a permanent limitation. The dotted Xs represent limitations that have the potential to be overcome, at least to some extent, in the future (for example, with larger sample sizes, technological and methodological advancements, and a shift from the use of single-nucleotide polymorphism (SNP) arrays to whole-genome sequencing).
33	Polygenic Risk scores (PRSs) = is a number based on variation in multiple genomic loci and their associated weights. It serves as the best prediction for the trait that can be made when taking into account variation in multiple genetic variants. PRSs are essentially a count of the number of the risk variants present in the person's DNA, weighted so that the presence of some risk variants is considered more important than others. The identities of the specific risk variants, and the basic information about how to weigh them, comes from the allele frequency differences between cases and controls identified in genome-wide association studies (GWAS). The optimal selection of variants and the weights associated with them is an active area of research. Notably, risk prediction does not need knowledge of causal variants and can tolerate inclusion of some false-positive variants. PRSs are validated by application in cohorts with already known case/control status. If the PRS are found to be predictive of the disease, then the PRS can be applied to an individual with unknown disease status. Ideally, at this stage, the PRS should be further validated for utility through formal clinical trials (Wray et al., JAMA Psychiatry, 2020).
35	The orange dashed line in the graph represents the threshold for genome-wide significance in a GWAS study. The filled red dots in the rsPS and gePS sections represent genetic variants reaching genome-wide significance, and the filled blue dots variants that have not reached genome-wide significance. In the pPS section, open dots reflect variants that have been assigned to one of the four groups of partitioned loci (Udler *et al*., *Endocr Rev*, 2019).
37	A polygenic risk score can only explain the **relative** risk for a disease. Why relative? The data used for generating a polygenic risk score comes from large scale genomic studies. These studies find genomic variants by comparing groups with a certain disease to a group without the disease. A polygenic risk score tells you how a person's risk compares to others with a different genetic constitution. However, polygenic scores do not provide a baseline or timeframe for the progression of a disease. For example, consider two people with high polygenic risk scores for having coronary heart disease. The first person is 22 years old, while the latter is 98. Although they have the same polygenic risk score, they will have different lifetime risks of the disease. Polygenic risk scores only show correlations, not causations.

	Absolute risk is different. Absolute risk shows the likelihood of a disease occurring. Women who carry a *BRCA1* mutation have a 60-80% absolute risk of breast cancer. This would be true even without any comparison to any groups of people. (https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores#four)
39	Receiver Operating Characteristic (ROC) curve/AUC area under the curve. The AUC provides an estimate of the probability a randomly selected subject with the condition has a test result indicating greater than that of a randomly chosen individual without the cancer. The solid line represents a receiver operator curve based on polygenic risk score from known risk SNPs based on reference. An AUC of 0.5 (dashed line) indicates that the classifier does not provide any useful information in discriminating cases from controls.
40	There were 218.754 individuals with 55.917 cases of hypertension. DBP: diastolic blood pressure; SBP: systolic blood pressure. Vaura *et al, Hypertension,* 2021
41	Contrasting and combining clinical risk factors and polygenic risk. The relative risk conveyed to individuals via commonly measured clinical risk factors (left panel) and polygenic risk estimation (middle panel) for coronary artery disease (CAD) is comparable and, when combined, can lead to different action recommendations (right panel). Relative risks for commonly measured clinical risk factors (left panel) can vary across populations and are approximated here. For polygenic risk (middle panel), the black sigmoidal curve represents the estimated CAD risk relative to average polygenic risk (at population incidence) based on 74 genome-wide association study (GWAS)-significant single nucleotide polymorphisms (SNPs). The bars represent the percentile thresholds typically used to define low (<20th percentile), medium (20th–80th percentile) and high (>80th percentile) polygenic risk. The combination of clinical and polygenic risk estimates (right panel) can lead to combined risk estimates that exceed the appropriate thresholds of risk versus benefit that justify certain medical interventions (action threshold). In this example, an individual with estimated clinical risk near the action threshold in the absence of polygenic risk information (bottom bar) could clarify their total risk with the addition of a polygenic risk estimate to decide against (low polygenic risk) or for (high polygenic risk) taking clinical action. PRS, polygenic risk score (Torkamani *et al., Nat Rev Genet,* 2018).
43	CRC: Colorectal cancer, BrCA: Breast Cancer, CKD: Chronic Kidney Disease, D-t-C: Direct-to-Consumer (testing), WGS: Whole-Genome Sequencing Polygenic risk scores can only explain part of the genetic aspect of a condition. Because nongenetic factors also contribute to risk, the maximum accuracy of genetic predictor (PRS) is limited by the heritability of the disorder, where heritability is the proportion of the variance between people in their liability to a disease that is attributed to genetic factors. However, construction of PRS is, to date, limited to DNA risk variants that have frequency of at least 1% in the population (and in some applications, variants are only included if they have a frequency of more than 10%, owing to greater instability in PRS using low-frequency variants [currently]). Hence, PRS are not designed to capture all genetic variation only tagged by common single nucleotide variants (SNVs or SNPs). Therefore, the so-called SNP-based heritability (h^2_{SNP}) gives the upper limit of the variance between people in their liability to a disease that can be explained by PRS and represents the variance explained by common DNA variants. As GWAS sample sizes increase, the variance explained by PRS will also increase and approach the SNP-based heritability. The h^2_{SNP} estimates vary across diseases, but an approximate upper limit is approximately 30% (Wray et al., JAMA Psychiatry, 2020).
44	The heritability gap in six major psychiatric disorders. Heritability (the proportion of causation attributable to genetic factors) estimated from family and twin studies (family/twin estimates, black bars) and from GWAS of SNPs and CNVs across the human genome (molecular estimates, red bars). The large difference between the family/twin and molecular estimates is indicated as the “heritability gap” (gray bars).
50-51	Τα γονίδια αυτά επιλέχθηκαν ως υποψήφια προς μελέτη είτε: α) εξαιτίας του προφανή βιολογικού τους ρόλου στην ομοιόσταση της γλυκόζης, στην (ηπατική ή περιφερική) αντίσταση στη δράση της ινσουλίνης, στη φυσιολογική τους λειτουργία στα β-παγκρεατικά κύτταρα, ή στο ρόλο τους στη δημιουργία λιπώδους ιστού (αδιπογένεση) (functional candidates), ή β) εξαιτίας της θέσης τους κοντά σε σήματα συσχέτισης που προέκυψαν σε προγενέστερες μελέτες ανάλυσης σύνδεσης (positional candidates).

	Βέβαια, η συγκεκριμένη προσέγγιση έχει τα μειονεκτήματα ότι, αφενός, μόνο ένα ή ένας μικρός αριθμός γονιδίων δύναται να μελετηθεί κάθε φορά και αφετέρου, ενδέχεται να αγνοούνται από την ανάλυση γονίδια, των οποίων ο βιολογικός ρόλος δεν είναι τόσο προφανής. Τελικά, παρά το μεγάλο αριθμό μελετών, που βασίσθηκαν στην επιλογή υποψήφιων γονιδίων, μόνο για μερικά από αυτά επιβεβαιώθηκαν σε μετέπειτα μελέτες συσχέτισης τα αρχικά ευρήματα, σε επίπεδο ολικού γονιδιώματος και σε μεγαλύτερα πληθυσμιακά δείγματα. Το γονίδιο *KCNJ11* κωδικοποιεί την υπομονάδα KIR6.2 του ATP-εξαρτώμενου διαύλου καλίου των β-παγκρεατικών κυττάρων, που διαδραματίζει καθοριστικό ρόλο στην εξωκυττάρωση των κυστιδίων ινσουλίνης μέσω της εκπόλωσης της μεμβράνης των παγκρεατικών κυττάρων, ενώ σπάνιες μεταλλάξεις στο γονίδιο αυτό προκαλούν νεογνικό σακχαρώδη διαβήτη και διαβήτη τύπου MODY. Το γονίδιο *PPARG* κωδικοποιεί τον αντίστοιχο πυρηνικό υποδοχέα PPARγ, ο οποίος παίζει σημαντικότατο ρόλο στη διαφοροποίηση και λειτουργία του λιπώδους ιστού, ρυθμίζοντας την έκφραση πολλών γονιδίων, ενώ μεταλλάξεις στο γονίδιο αυτό προκαλούν σοβαρή αντίσταση στην ινσουλίνη με λιποδυστροφία και ηπατική νόσο, που εξελίσσεται προοδευτικά σε ΣΔΤ2 σε νεαρή ηλικία. Τα γονίδια *PPARG* και *KCNJ11* επιλέχτηκαν ως υποψήφια γονίδια, και για έναν ακόμη λόγο. Τα προϊόντα τους αποτελούν στόχους φαρμάκων, που χρησιμοποιούνται για την αντιμετώπιση του ΣΔΤ2, όπως ο παράγοντας ευαισθητοποίησης ινσουλίνης θειαζολιδινεδιόνη, που δρα στον πυρηνικό υποδοχέα PPARγ και οι σουλφονυλουρίες, που αναγνωρίζουν και συνδέονται με τους δίαυλους καλίου με την υπομονάδα KIR6.2.
52	Simplified schematic of the processes involved in genetic predisposition to type 2 diabetes. Assignments of loci to particular processes are based on current knowledge of the presumed function of the best candidates within each signal and human physiological studies. These assignments should be considered provisional until the causal variants have been identified and the molecular mechanisms through which they act are established. Current evidence shows, however, that the majority of genes implicated in diabetes susceptibility act through effects on β-cell function and/or mass (McCarthy & Hattersley, *Diabetes*, 2008). During the past decade, T2D-associated variants have been shown to modulate T2D risk through diverse mechanisms: some increase T2D risk through an impact on obesity (e.g., FTO), others reduce insulin sensitivity (e.g., PPARG, IRS1), whereas others compromise insulin secretion, either through direct effects on islet function (e.g., KCNJ11) or development (e.g., HNF1A) or indirectly through impact on incretin signaling (e.g., GLP1R). The various classes of T2D therapeutics operate through the same range of mechanisms to reverse the diabetic phenotype or control its glycemic consequences. The weight of evidence indicating that the genetic contribution to T2D predisposition mostly arises from common variants of limited individual effect emphasizes the need to think in terms of a gradation of polygenic risk across individuals, rather than a classification based around rigid, discrete subtypes (Udler *et al*., *Endocr Rev*, 2019)
54	>130 loci associate, >400 genetic signals have been associated with T2D in adults (>50 for T1D). The biggest effects for T2D modulate risk by no more than 40% per allele (*TCF7L2*), and most have much smaller effects.
56	Venn diagrams. The variants are represented by gene names here, which could indicate that the location is present either in the gene, or in the vicinity of the gene. The black circle represents T2D, and the gene names in black in this represent variants only associated with T2D. The overlapping circles indicate additional reporting associations for that variant for instance, *TCF7L2*, *KCNQ1*, *MTNR1B etc.*, are associated with T2D and also with beta-cell dysfunction. An *ADCY5* variant is associated with 2 h insulin adjusted for 2 h glucose; 2 h glucose/T2D (in brown) *** variants in TMEM163 are also associated with fasting insulin, *TCF7L2*—associated with fasting and 2 h glucose and *MADD* variants associated with fasting proinsulin, fasting glucose and HOMA-B (Prasad & Groop*, Genes,* 2015).
57	Through sequencing and genotyping of ~150,000 individuals across 5 ancestry groups, a spectrum of 12 rare predicted protein-truncating variants was identified in *SLC30A8*. Shown for each variant (Table 1) are ancestry group, cohort, number of genotyped cases and controls (*N*), number of cases and controls observed to carry the variant, and observed allele frequencies in cases and controls. ORs and *P* values were computed separately for three groups of variants: p.Arg138*, p.Lys34Serfs*50 and the remaining variants.

58	ZnT8, encoded by *SLC30A8*, is highly expressed in the membrane of insulin granules within pancreatic β-cells, where it transports zinc ions for crystallization and storage of insulin. Expression of genes associated with β-cell maturation and development were also influenced by *SLC30A8* knockdown with decreased expression of NKX6.1 and PDX1 and increased expression of SOX4, SOX6 and SOX11. A pathway enrichment analysis of differentially expressed genes showed enrichment of genes involved in the WNT signaling and insulin secretion pathways. A global gene set enrichment analysis (GSEA) of all expressed genes (N = 12,956) using a gene ontology database showed enrichment of genes involved in positive regulation of TOR signaling. Collectively, data demonstrate a link between *SLC30A8* expression and transcriptional networks involved in cell development, cell fate and plasma membrane polarization (Dwivedi *et al., Nat Genet*, 2019).
59	SLC16A11 haplotypes constructed from the synonymous and four missense SNPs associated to T2D, with haplotype frequencies derived from the 1,000 Genomes Project and SIGMA samples. AFR, Africa (*n*=185); EUR, European (*n*=379); ASN, East Asian (*n*=286); MXL, Mexican samples from Los Angeles (*n*=66). Frequencies from SIGMA samples are calculated from genotypes and represent either the entire dataset ("All") or only samples estimated to have ≥95% Native American ancestry ("≥95 N.A.", *n*=290). Haplotypes with population frequency <1% are not depicted. (**e**) Predicted membrane topology of human SLC16A11. Locations of SNPs carried by the T2D-associated haplotype are indicated. (**f**) Forest plot depicting odds ratio estimates at rs75493593 from the four SIGMA cohorts, the SIGMA pooled mega-analysis, the replication cohorts, replication-only meta-analysis, and the overall meta-analysis (including all replication cohorts and the SIGMA mega-analysis). (A) Expression-QTL (eQTL) analyses in liver. Boxplots depict the log2 of the relative expression level for RNASEK, BCL6B, SLC16A11, and SLC16A13 according to genotype at rs13342692. n=21 homozygous reference (REF), 16 heterozygous (HET), and 10 homozygous T2D risk (T2D).
60	SLC16A11 affects cellular fatty acid and lipid metabolism, with the increase in acylcarnitines, indicating an effect on fatty acid β-oxidation by the mitochondria and with the increases in DAGs (diacylglycerols) and TAGs (triacylglycerols), indicating a shift toward energy storage in the form of glycerolipids, metabolic changes that match those seen in the pathophysiology of insulin resistance and T2D (Rusu, Hoch *et al., Cell*, 2017).
61	rs = restricted to significant SNPs only PRS ge = global extended PRS