Advancing Polygenic Risk Scores: Challenges and Opportunities

AUTUMN 2025

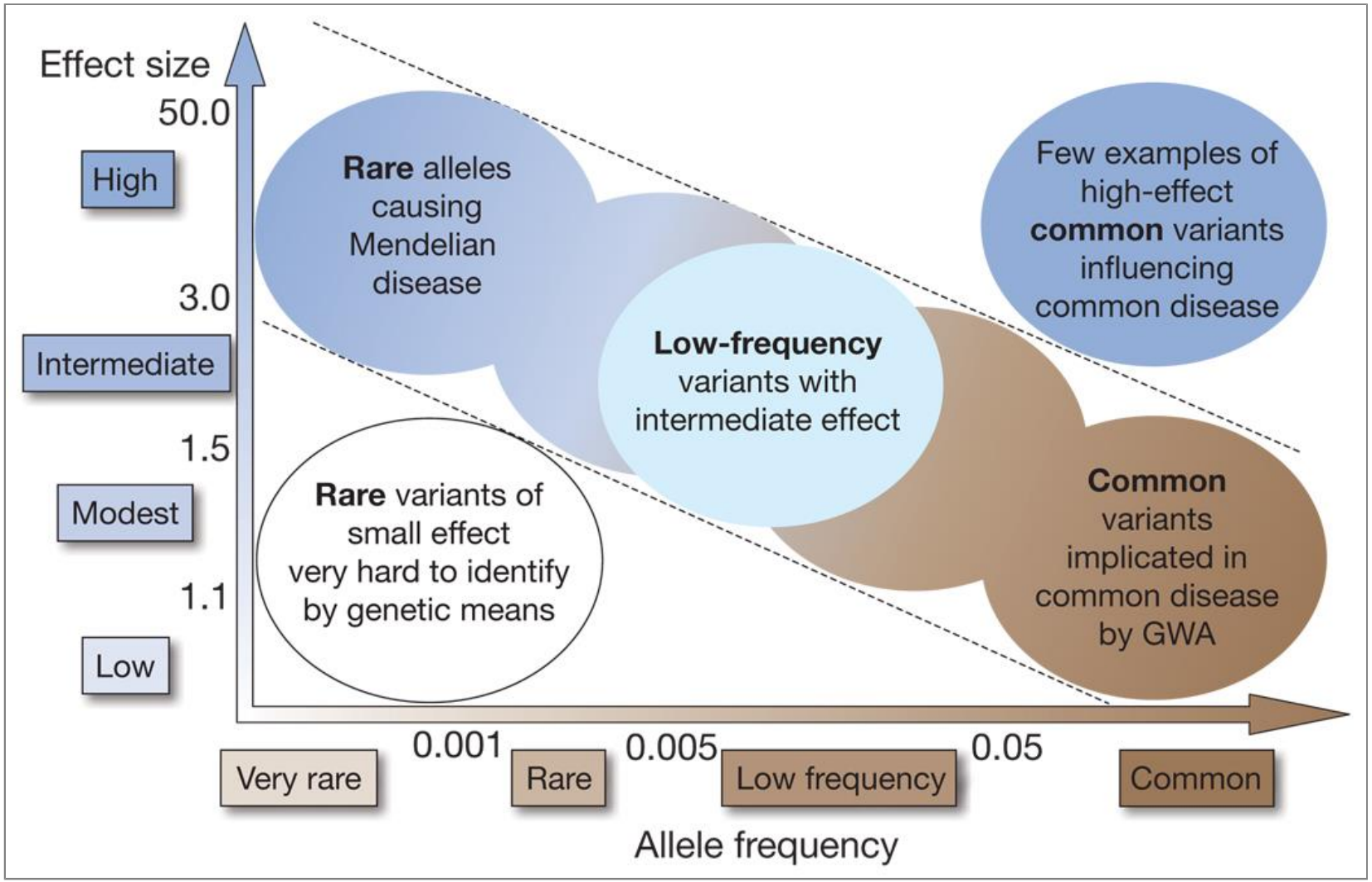
Post GWAS Era – Missing Heritability

Table 1

Estimates of heritability and number of loci for several complex traits

Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual*phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

Post GWAS Era – Miniscule Effect Sizes



The Hunt for Missing Heritability – Two Directions

- Set of rare variants (collapsing or kernel-based methods) (2008-)
- Polygenic risk scores (PRS: testing the classic theory of polygenic inheritance) (2009-)
  - Classic — major gene + polygenic component via mixed effect modeling
  - Contemporary — combined mean effects

nature Vol 460|6 August 2009|doi:10.1038/nature08185

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Schizophrenia is a severe mental disorder with a lifetime risk of about 1%, characterized by hallucinations, delusions and cognitive deficits, with heritability estimated at up to 80%[1,2]. We performed a genome-wide association study of 3,322 European individuals with schizophrenia and 3,587 controls. Here we show, using two analytic approaches, the extent to which common genetic variation underlies the risk of schizophrenia. First, we implicate the major histocompatibility complex. Second, we provide molecular genetic evidence for a substantial polygenic component to the risk of schizophrenia involving thousands of common alleles of very small effect. We show that this component also contributes to the risk of bipolar disorder, but not to several non-psychiatric diseases.

Table 2, Supplementary Fig. 2 and section 5 and 6 in Supplementary Information).

The best imputed SNP, which reached genome-wide significance (rs3130297, $P = 4.79 \times 10^{-8}$, T allele odds ratio = 0.747, minor allele frequency (MAF) = 0.114, 32.3 megabases (Mb)), was also in the MHC, 7 kilobases (kb) from *NOTCH4*, a gene with previously reported associations with schizophrenia[4]. We imputed classical human leukocyte antigen (HLA) alleles; six were significant at $P < 10^{-3}$, found on the ancestral European haplotype[5] (Table 1, Supplementary Table 3 and section 3 in Supplementary Information). However, it was not possible to ascribe the association to a specific HLA allele, haplotype or region (Supplementary Table 3 and

Rationale

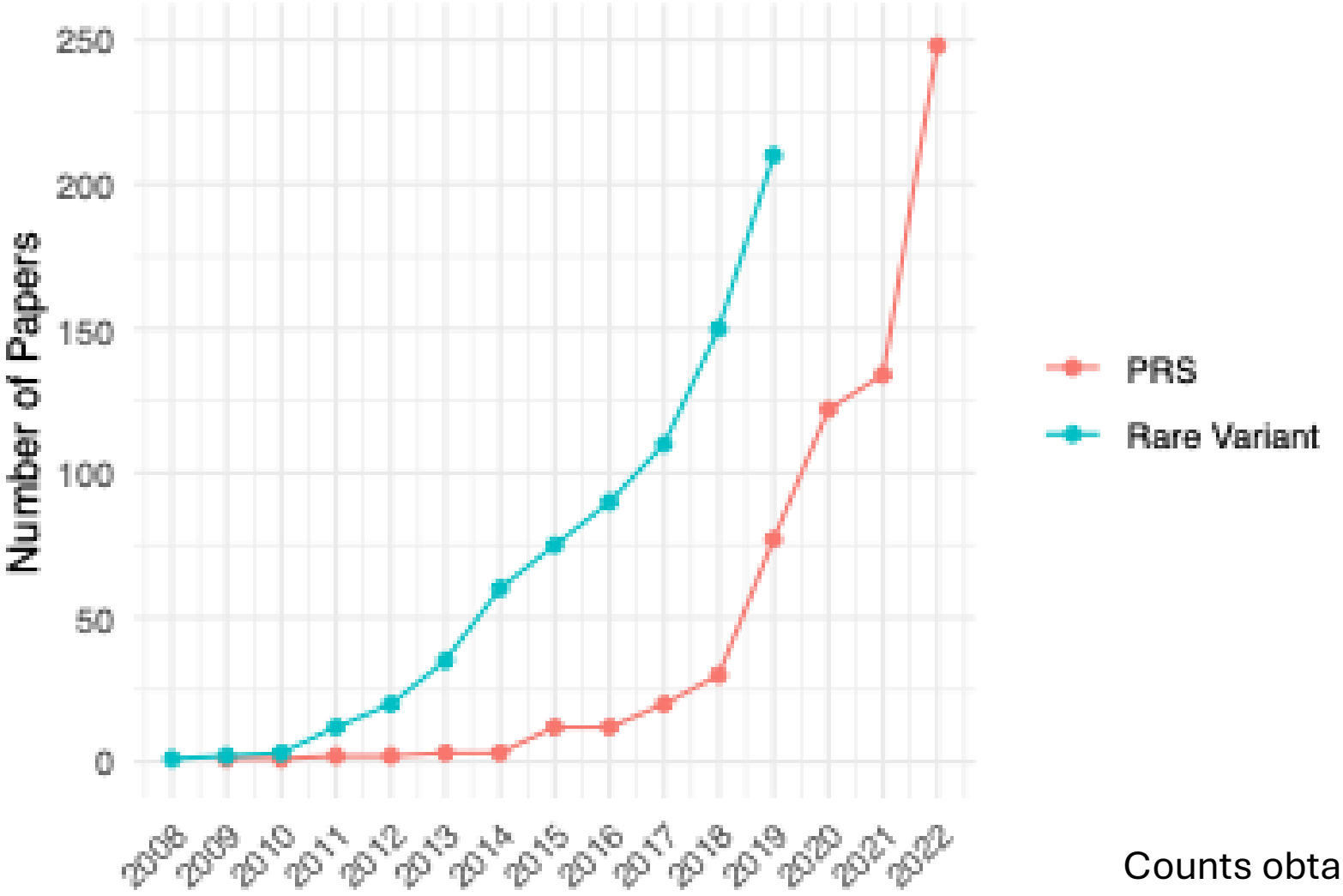
- “Thousands of very small individual effects that collectively account for a substantial proportion of variation in risk.”

- Summarize variation across “nominally associated loci” (p value may be as large as 0.5 – very weak signals) into quantitative scores.

Workflow

- Discovery (base) sample
  - Find SNPs that are “associated” with a trait.
  - Filtering step to reduce the number of SNPs: thresholding, LD clumping, MAF, genotyping rate.
- Target sample
  - For each individual, sum of number of “score alleles” weighted by the log odds ratio estimated from the discovery sample $\mathrm{PRS} = \sum_{j=1}^{M} G_j \hat{\beta}_j$
  - Testing for a significant difference between the mean PRS of the cases and that of the controls.

Two Directions: PRS vs. Rare Variants



Counts obtained using ChatGPT

Three (Typical) Sets of Data

- Reference samples: ancestry-matched or diverse (e.g. 1000 Genomes)
  - LD matrix
- Discovery (base) samples (UK Biobank)
  - GWAS summary statistics
  - Sometimes also use to estimate the effect size (especially if same “ancestries”)
- Target samples
  - Individual-level data
  - May be split into training and testing if evaluating and comparing among methods

Categories of PRS Methods (1)

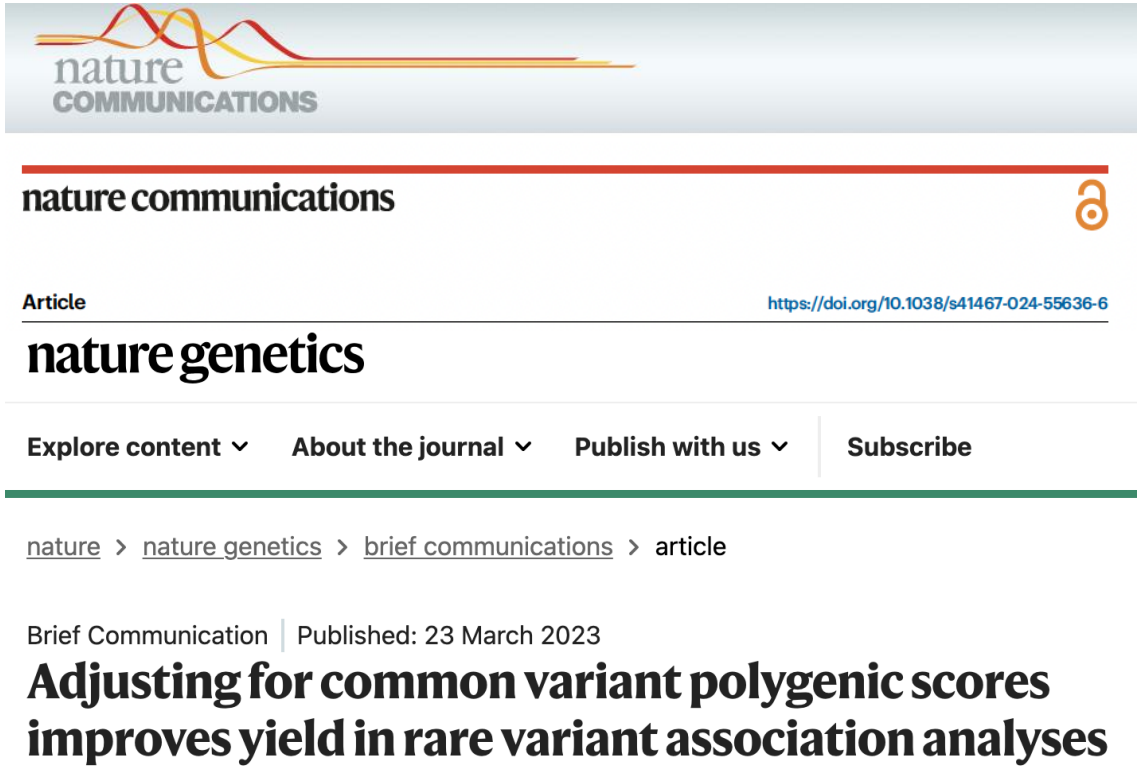
- Clumping and Thresholding (C+T)
  - Filters SNPs by significance (p-value thresholding) and LD pruning (PRSice, PLINK)
- Penalized Regression (Regularization-Based)
  - Applies penalties to avoid overfitting and select informative SNPs (Lassosum, SBLUP)
- Bayesian Methods
  - Use prior distributions and shrinkage based on LD structure (LDpred/LDpred2, PRS-CS/PRS-CSx)

Categories of PRS Methods (2)

- Multi-Population / Transfer Learning Models
  - Account for cross-ancestry correlation or limited training data in non-European populations (JointPRS, PRS-PGx-TL)
- Machine Learning–Based Approaches
  - Utilize data-driven techniques for non-linear or interaction modeling (XGBoost-PRS, Elastic Net PRS)
- Deep Learning / AI-Augmented Approaches
  - Model complex non-linear genotype-phenotype relationships; some explore image/genomic feature fusion (DeepPRS, EIR)

Future Outlook

- Emerging Themes and Opportunities
  - Longitudinal PRS
  - PRS integration with EHR and biomarkers
  - Rare variant PRS (coalescence)
- Challenges
  - Portability across ancestries
  - Clinical translation (thresholds, interpretability)
  - Ethical concerns: discrimination, insurance, and data privacy



Summary and Discussion

- PRS methods are maturing—clear shift toward inclusivity, interpretability, and clinical relevance.
- Innovations like mixed-effect models, cross-population and family-based methods, pharmacogenomics-specific modeling, machine/deep learning and AI models push boundaries.
- Broad future outlook for translational research:
  - How can PRS be made equitable across populations?
  - What's needed for broader PRS clinical utility?
  - How about the role of regulatory or clinical guidelines?