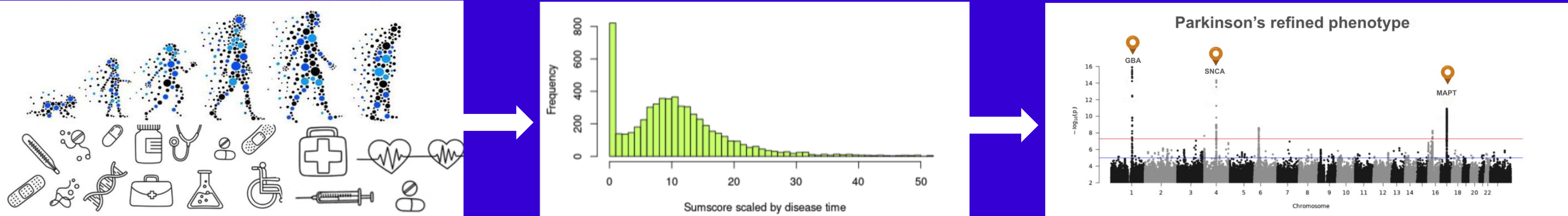


Genetic Signal as an External Benchmark for Phenotype Algorithms

FinnGen Phenotype Team

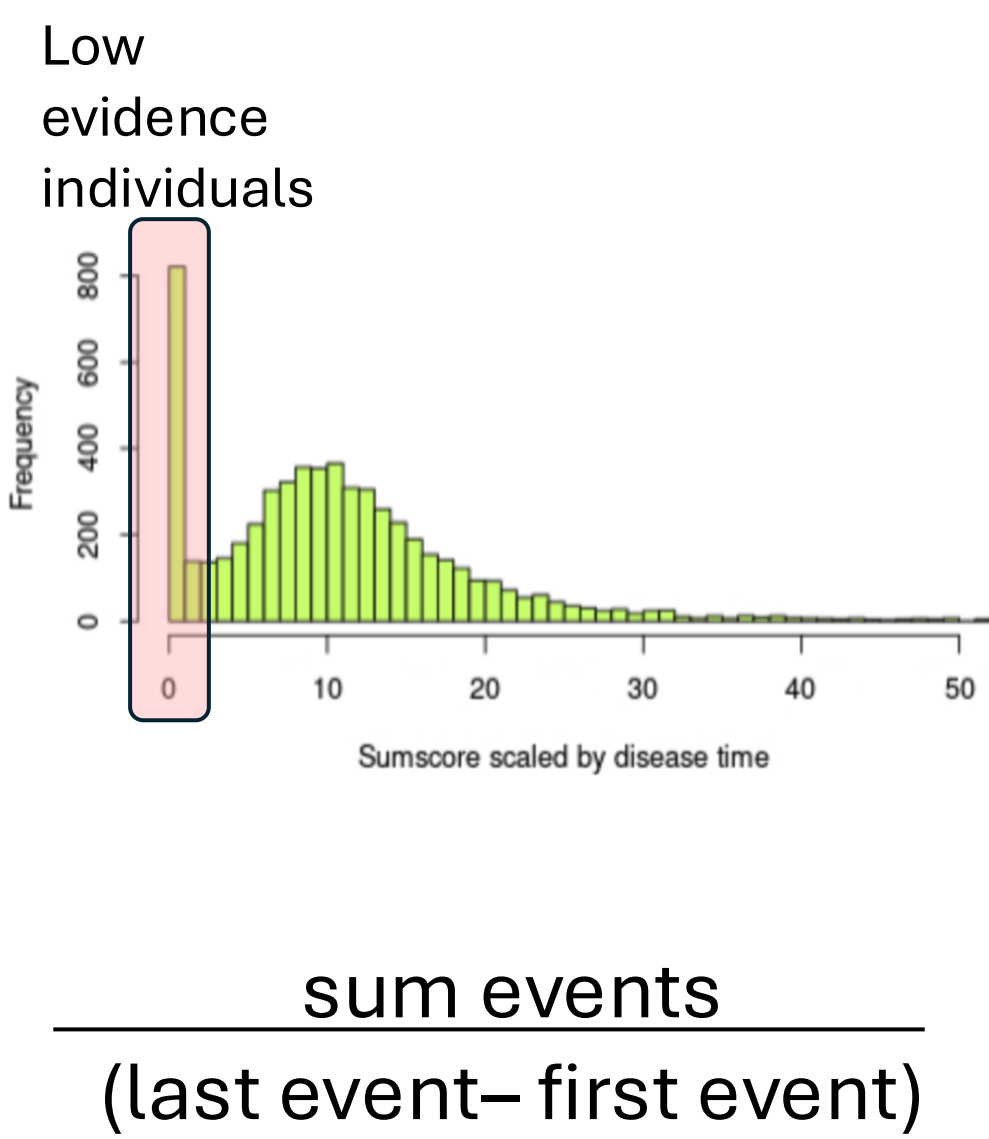


Golden genetic loci let us quantify phenotype algorithm improvement - no chart review needed

- GWAS provides two independent measures of refinement: p-value and effect size
- Weighting evidence across data domains adapts naturally to changing standard-of-care
- Cleaner cohorts improve trajectory, progression, and age-of-onset analyses

1 Unsupervised EHR code enrichments enable a scaled evidence score.

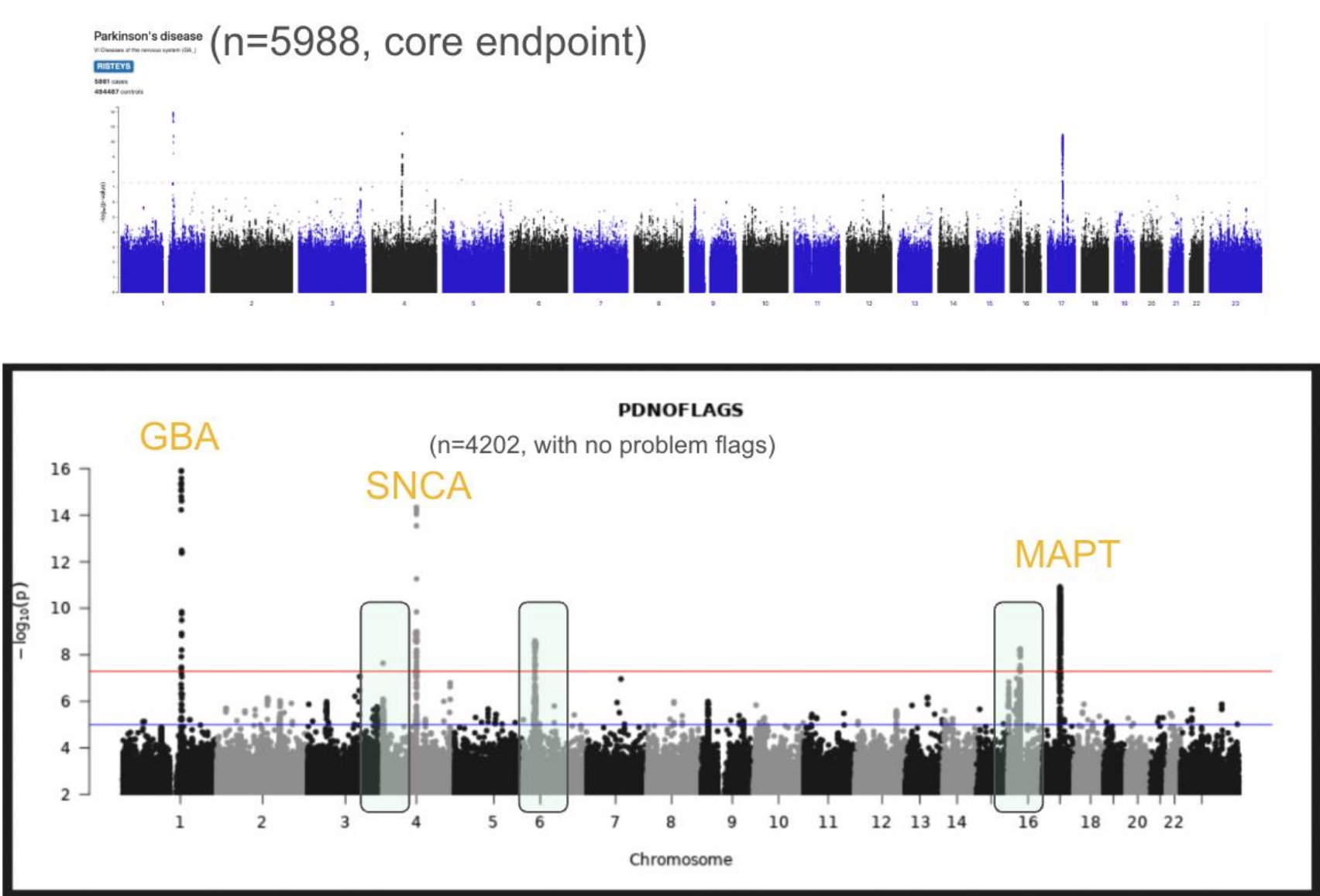
CodeWAS						
CodeWAS is a tool for exploring the associations between an endpoint and all of the medical codes and drug codes. This is a new tool, please reach out using the contact form for feedback and improvement ideas.						
First, a cohort is built by matching controls to the endpoint cases using year of birth and sex. Then, a Fisher test is done for all the medical codes and drug codes between the cases and controls of this cohort. Codes are reported in the table below if they have $-\log_{10}(p\text{-value}) \geq 6$.						
Matched cohort						
Matched cases 6458						
Matched controls 64576						
Code	Vocabulary	Description	Odds Ratio	$-\log_{10}(p)$	N matched cases	N matched controls
type to filter	type to filter					
N04BA01	ATC	amantadine;oral	16.1	307.7	442	293
N04BA02	ATC	levodopa and decarboxylase inhibitor;oral	613.1	307.7	5266	462
P02_39a	ICD-10 Finland	Dementia in Parkinson's disease without intracranial implantation of electrode	8031.8	307.7	804	2
C23.3	ICD-10 Finland	Dysphagia	87.6	307.7	184	28
A04A08	NOMESCO Finl...	Adjustment of implanted neurostimulator	82.0	307.7	191	24
N04BA02	ATC	rasagiline;oral	5103.0	307.7	1088	2
N04BA05	ATC	pramipexole;oral	12.3	307.7	2243	2671
A02C08	NOMESCO Finl...	Stereotactic intracranial implantation of electrode	171.9	307.7	248	15
C26	ICD-10 Finland	Parkinson disease	9126.3	307.7	5167	59
N04AA02	ATC	levodopa;systemic	20.6	307.7	588	313
N04BA02	ATC	entacapone;oral	5234.6	307.7	495	2
A05D01	NOMESCO Finl...	Brain receptor or transport protein distribution...	54.2	307.7	220	42
N04BA09	ATC	rotigotine;transdermal	182.8	307.7	443	26



We weight evidence rather than collapse it, preserving the signal gradient that LASSO-style approaches discard. CodeWAS is similar to CohortDiagnostics Characterization but we use birth year and sex matched controls. Birth year matched controls help adjust for changes in standard-of-care over decades.

3 Genetic association studies independently quantitate phenotype improvement with effect size and p-value

Known loci from disease meta-analyses - GBA, SNCA, MAPT for Parkinson's - serve as biological anchors. Signal strength at these loci measures cohort quality, independent of any coding practice.



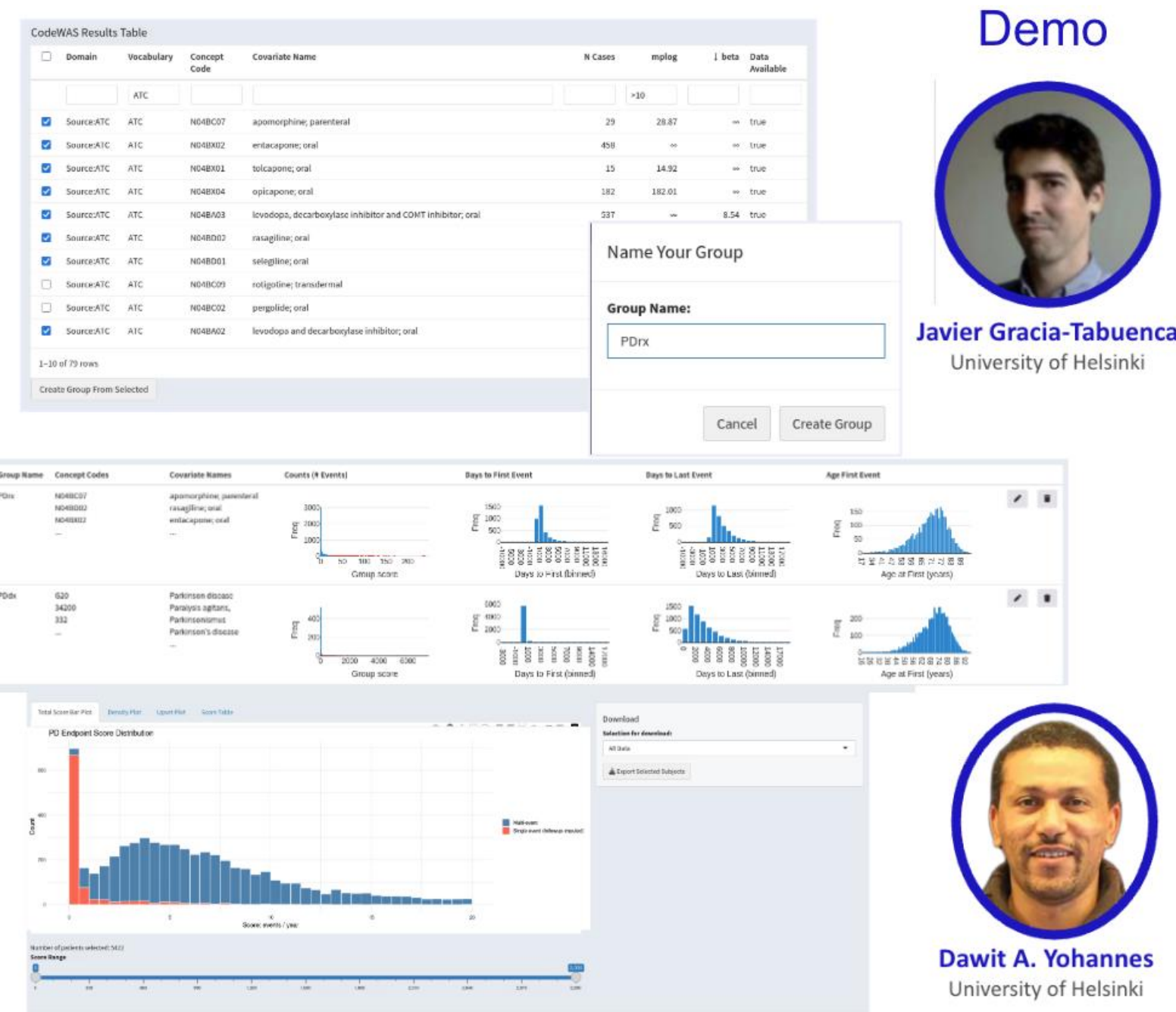
2 Polygenic risk scores (PGS) can also validate phenotyping algorithms

Low-evidence individuals show no polygenic risk for Parkinson's disease, confirming they are not true cases. Retained individuals show strong PGS association.



4 PhenotypeScoring: automating evidence-based cohort refinement on the OMOP CDM (CohortOperations tool)

Starting from CodeWAS enrichments, users select and group features, build a scoring algorithm, and flag low-evidence individuals for exclusion



Marika Kaakinen¹, Javier Gracia-Tabuenca^{1,2}, Dawit A. Yohannes¹, FinnGen¹, Mary Pat Reeve^{1,3}
¹ FinnGen, University of Helsinki, Helsinki, Finland; ² Tampere University, Tampere, Finland; ³ Broad Institute of MIT and Harvard, Cambridge, MA, USA



FinnGen consists of 520,000 genotyped individuals linked to national health registries on the OMOP CDM, providing decades of longitudinal EHR data across 4,000+ disease endpoints. On average, each person has 1,020 events including diagnoses, procedures, medications and lab tests. The combination of whole-population registry coverage and universal genotyping makes it uniquely suited for phenotype-GWAS feedback loops.