Gene prioritization across ancestries uncovers distinct molecular pathophysiology and therapeutic landscape in polycystic ovary syndrome

Sindhuja Rajavelu[1] & Debojyoti De*[1]

[1]*Department of Biotechnology, National Institute of Technology Durgapur, West Bengal 713209, India.*

***Corresponding Author Details**

Dr. Debojyoti De

Assistant Professor

Department of Biotechnology

National Institute of Technology

Durgapur, West Bengal – 713209, India.

Email: dde.bt@nitdgp.ac.in

Phone: +91-9434789056

Contents

Supplementary figures

Supplementary note

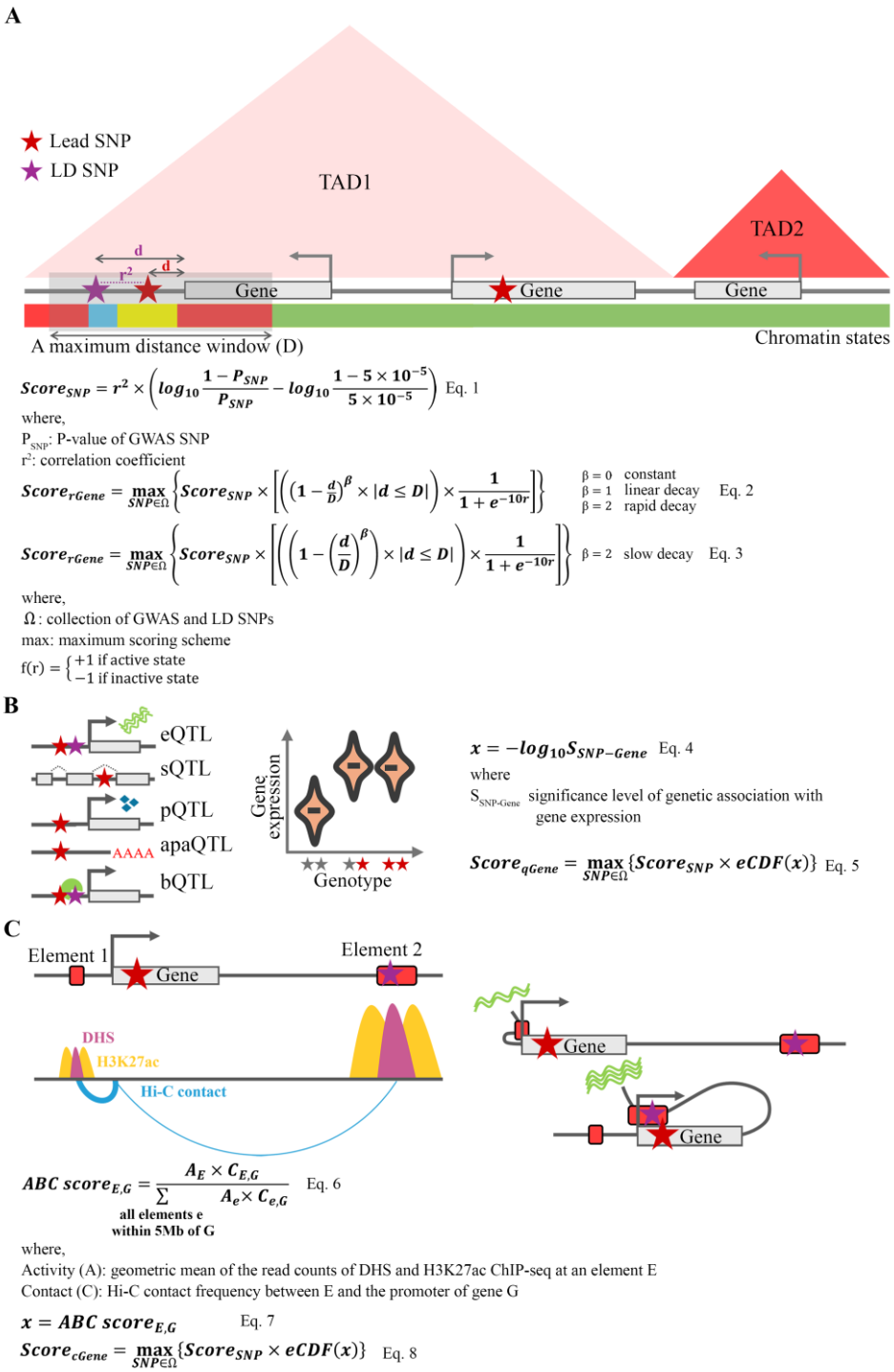

Figure S1: Schematic illustration of principles underlying the scoring method. A. Regulatory genes (rGene) were scored based on proximity to the Lead or its linked SNPs within a distance window (D) considering chromatin state of the region containing the SNP and constraining for genomic organization (has to be within the same topologically associated domain (TAD), Eqn 2 and 3). B. The qGenes are scored by incorporating information regarding the variation of various quantitative traits of genomic origin and utilizing significance of association with the framework of a percentile rank based scoring method (empirical cumulative distribution function, eCDF, Eqn 4 and 5). C. The cGenes are scored

considering the strength of Enhancer-Gene interaction due to chromatin conformation and activity and accessibility of the enhancer modelled through empirical cumulative distribution function (eCDF) on the interaction score (Eqn 6, 7 and 8). Details of process is further described in supplementary note.

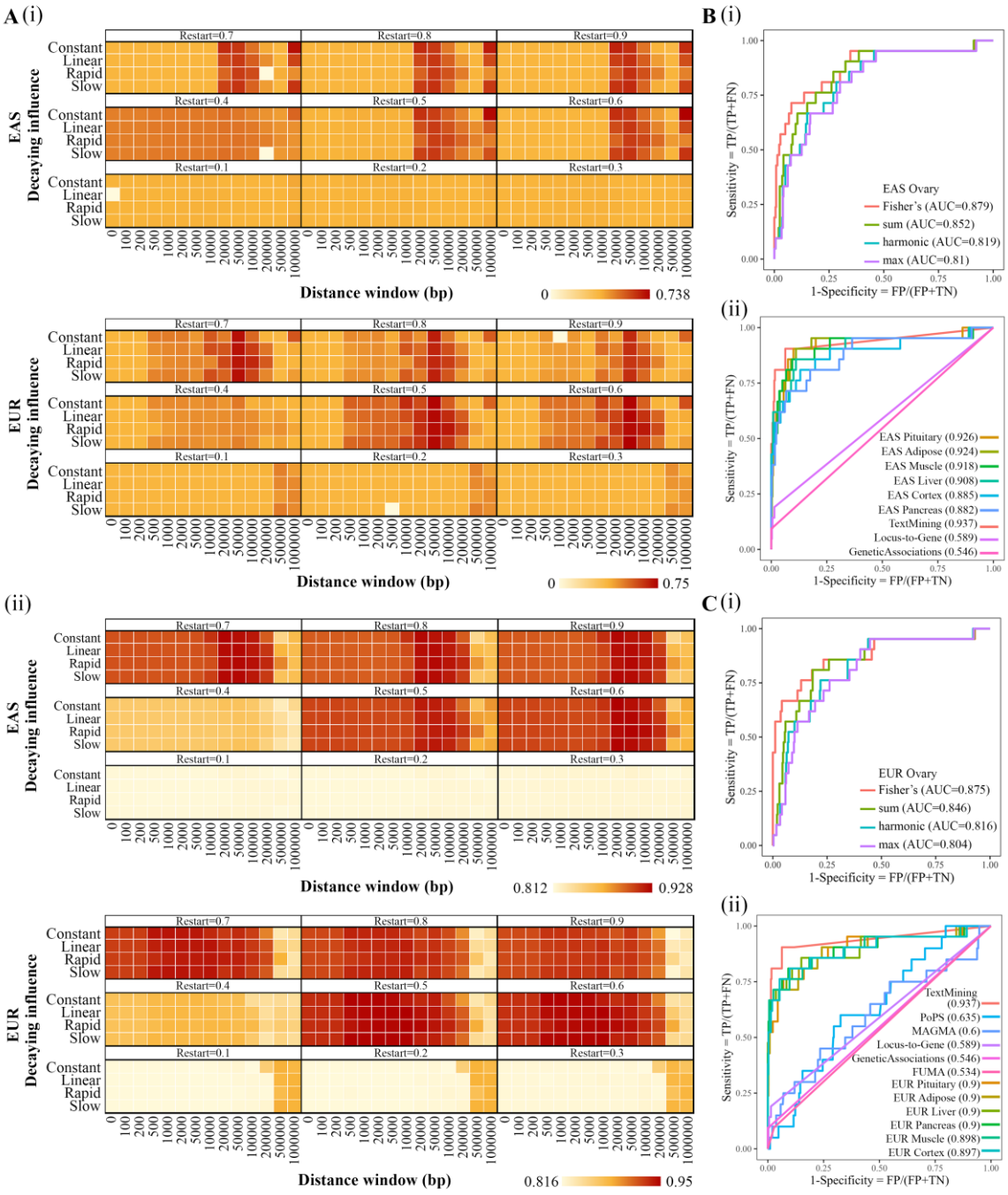

Figure S2: Simultaneous optimization of influential distance for rGene and protein interactome exploration parameters for random walk. A. Heatmaps displaying the area under the receiver operating characteristic curve (AUC) for gene prioritization performance under varying parameter settings as indicated below, evaluated separately for EAS and EUR populations. The optimization integrates genomic parameters—distance window and decay kernel (Constant, Linear, Rapid, Slow)—with network exploration parameters defined by restart probabilities (0.1 to 0.9) which were used for random walk with restart (RWR). The distance window denotes the genomic span within which SNP–gene associations are considered, while the decay kernel governs how SNP influence decays with distance. AUC is calculated based on the model's ability to distinguish gold standard positives (GSPs)

from gold standard negatives (GSNs). (i) AUC values upon exclusion of GSPs and their first-order neighbors (GSNs) from evaluation. (ii) AUC values when both first- and second neighbors of GSPs (GSNs) are excluded. **Benchmarking:** Performance of four integration strategies—Fisher's method (meta-analysis-like), sum of predictor scores, maximum score and harmonic prioritization methods evaluated for ovary assessed by comparing AUC obtained from different prioritization methods in its ability to distinguish GSPs and GSNs in EAS **(B (i)),** and EUR **(C (i))** respectively. (ii). Benchmarking of our gene prioritization methods evaluated in other tissues (used in our study) except ovary in comparison to Open Targets (Locus-to-Gene, Text Mining and Genetic Association) in EAS **(B (ii)),** and additionally with FUMA, MAGMA and PoPS in EUR **(C (ii))**.

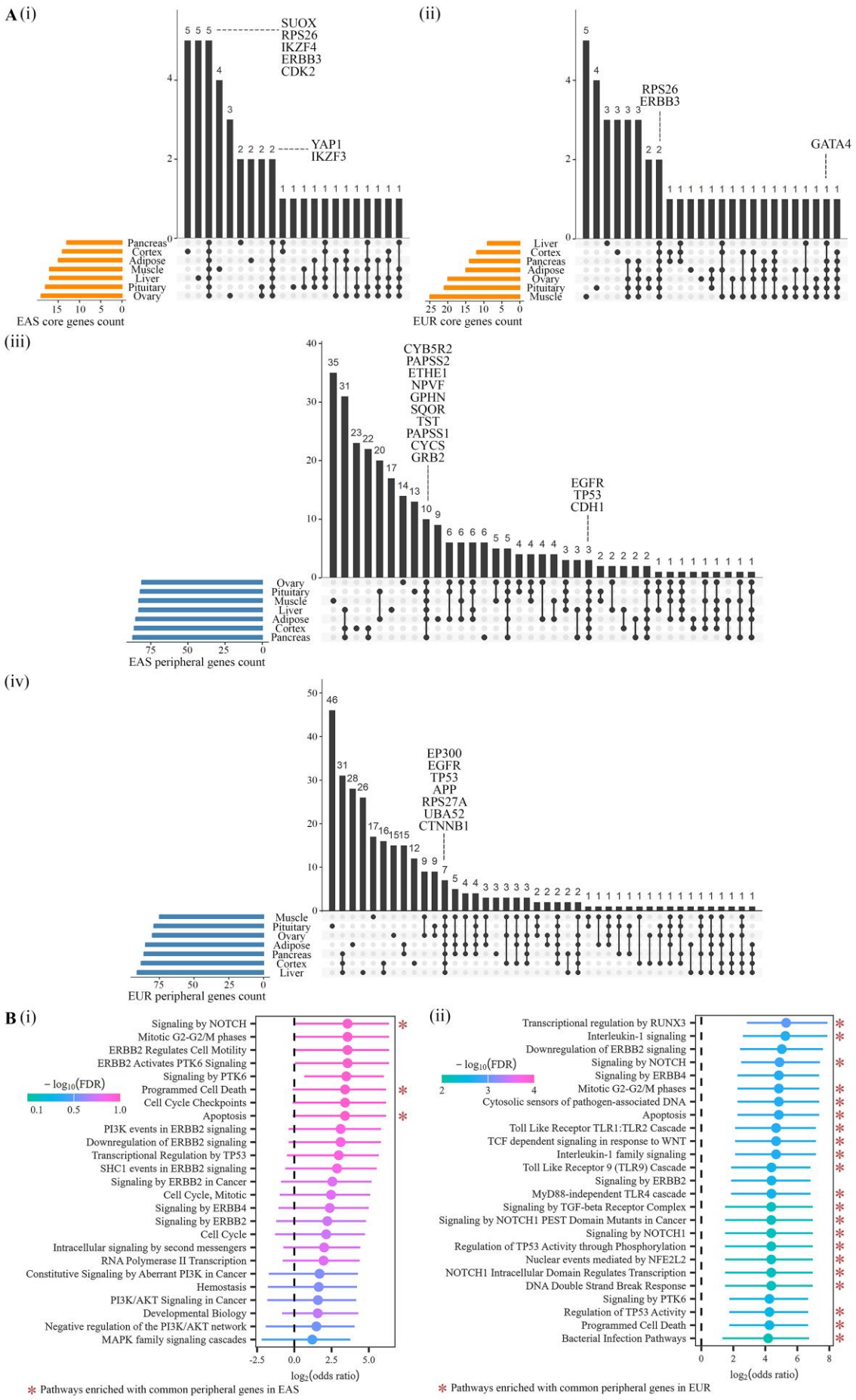
A (i)
SUOX
RPS26
IKZF4
ERBB3
CDK2
YAP1
IKZF3
Pancreas
Cortex
Adipose
Muscle
Liver
Pituitary
Ovary
EAS core genes count
(ii)
RPS26
ERBB3
GATA4
Liver
Cortex
Pancreas
Adipose
Ovary
Pituitary
Muscle
EUR core genes count
(iii)
CYB5R2
PAPSS2
ETHE1
NPVF
GPHN
SQOR
TST
PAPSS1
CYCS
GRB2
EGFR
TP53
CDH1
Ovary
Pituitary
Muscle
Liver
Adipose
Cortex
Pancreas
EAS peripheral genes count
(iv)
EP300
EGFR
TP53
APP
RPS27A
UBA52
CTNNB1
Muscle
Pituitary
Ovary
Adipose
Pancreas
Cortex
Liver
EUR peripheral genes count
B (i)
Signaling by NOTCH
Mitotic G2-G2/M phases
ERBB2 Regulates Cell Motility
ERBB2 Activates PTK6 Signaling
Signaling by PTK6
Programmed Cell Death
Cell Cycle Checkpoints
Apoptosis
PI3K events in ERBB2 signaling
Downregulation of ERBB2 signaling
Transcriptional Regulation by TP53
SHC1 events in ERBB2 signaling
Signaling by ERBB2 in Cancer
Cell Cycle, Mitotic
Signaling by ERBB4
Signaling by ERBB2
Cell Cycle
Intracellular signaling by second messengers
RNA Polymerase II Transcription
Constitutive Signaling by Aberrant PI3K in Cancer
Hemostasis
PI3K/AKT Signaling in Cancer
Developmental Biology
Negative regulation of the PI3K/AKT network
MAPK family signaling cascades
− log10(FDR)
log2(odds ratio)
* Pathways enriched with common peripheral genes in EAS
(ii)
Transcriptional regulation by RUNX3
Interleukin-1 signaling
Downregulation of ERBB2 signaling
Signaling by NOTCH
Signaling by ERBB4
Mitotic G2-G2/M phases
Cytosolic sensors of pathogen-associated DNA
Apoptosis
Toll Like Receptor TLR1:TLR2 Cascade
TCF dependent signaling in response to WNT
Interleukin-1 family signaling
Toll Like Receptor 9 (TLR9) Cascade
Signaling by ERBB2
MyD88-independent TLR4 cascade
Signaling by TGF-beta Receptor Complex
Signaling by NOTCH1 PEST Domain Mutants in Cancer
Signaling by NOTCH1
Regulation of TP53 Activity through Phosphorylation
Nuclear events mediated by NFE2L2
NOTCH1 Intracellular Domain Regulates Transcription
DNA Double Strand Break Response
Signaling by PTK6
Regulation of TP53 Activity
Programmed Cell Death
Bacterial Infection Pathways
− log10(FDR)
log2(odds ratio)
* Pathways enriched with common peripheral genes in EUR

Figure S3: Cross-tissue comparison and convergence of prioritized genes and pathways in PCOS. **(A)** UpSet plots illustrating the overlap of top 100 prioritized core and peripheral genes across tissues. (i) and (ii) represent core genes in EAS (EAS) and EUR (EUR) populations, respectively. (iii) and (iv) represent peripheral genes in EAS and EUR populations, respectively. **(B)** Pathway enrichment analysis of common core and peripheral genes across tissues, performed using one-sided Fisher's exact test and significance derived from Benjamini-Hochberg method. Pathways enriched with the common peripheral genes only were marked red asterisk.

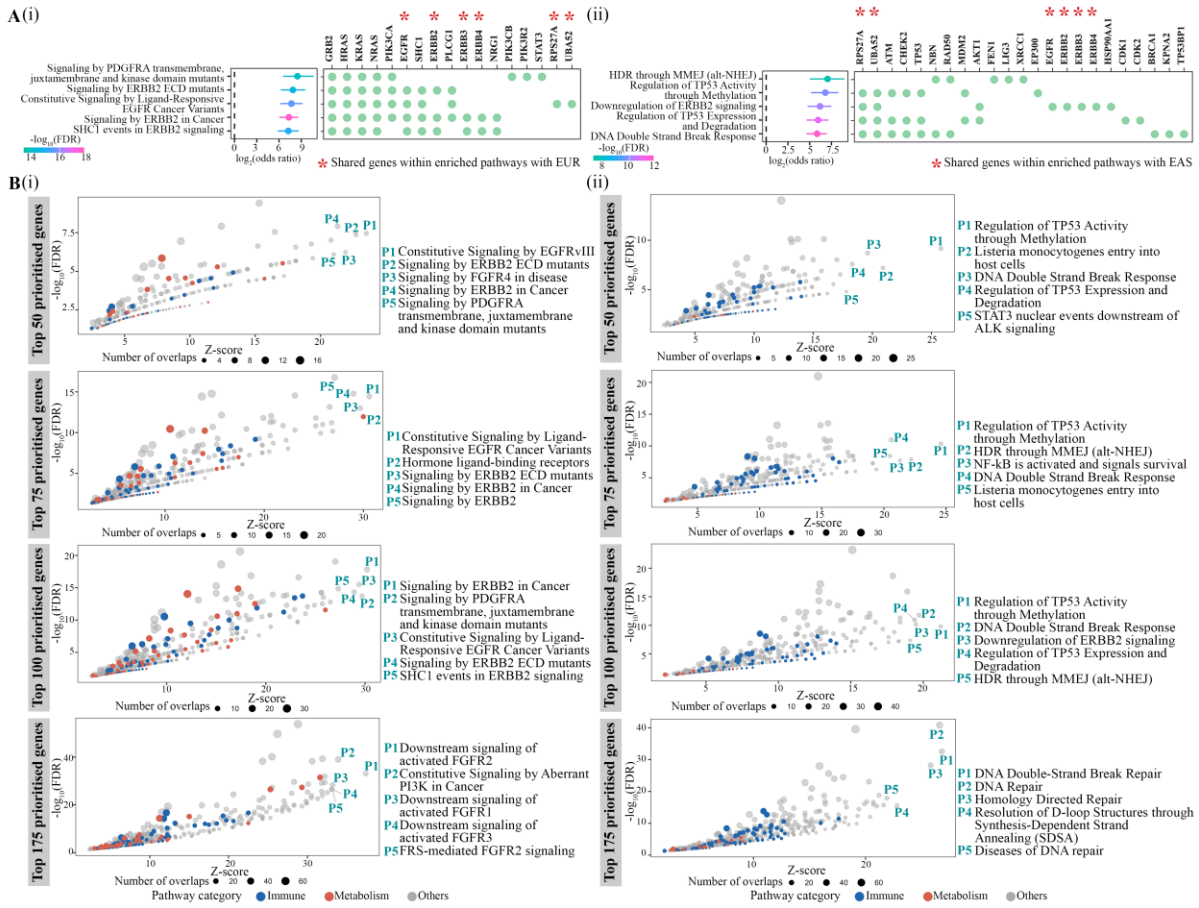

Figure S4: Comparison and consistency analysis of enriched pathways. Prioritized pathways are derived from the top 100 prioritized genes in ovary tissue for both populations. **(A)** Forest plot (Left) depicting the top 5 significantly enriched Reactome pathways in (i) EAS Ovary, (ii) EUR Ovary. The pathways are ranked by odds ratio, with enrichment significance indicated by $-\log_{10}$(FDR) from Benjamini-Hochberg method in one-sided Fisher's exact test. Right: Binary Heatmaps member genes corresponding to the enriched pathways, sorted by occurrence frequency recurrence across pathways. Genes common to both populations are marked with red asterisks. **(B) Consistent pathway prioritization with increasing numbers of top-ranked genes**. (i) EAS Ovary, (ii) EUR Ovary. Scatter plots showing enrichment of Reactome pathways for top 50, 75, 100, and 175 prioritized genes. Bubble size indicates the number of overlapping genes; the top 5 pathways by Z-score (P1–P5) are labelled in each panel. The enrichment z-score, the false discovery rate (FDR), odds ratio (OR), 95% confidence interval are all estimated using Fisher's exact test. Bubble color represents the functional category of the pathway involved in immune processes (marked in blue), pathways related to metabolism (highlighted in orange) and other pathways (grey).

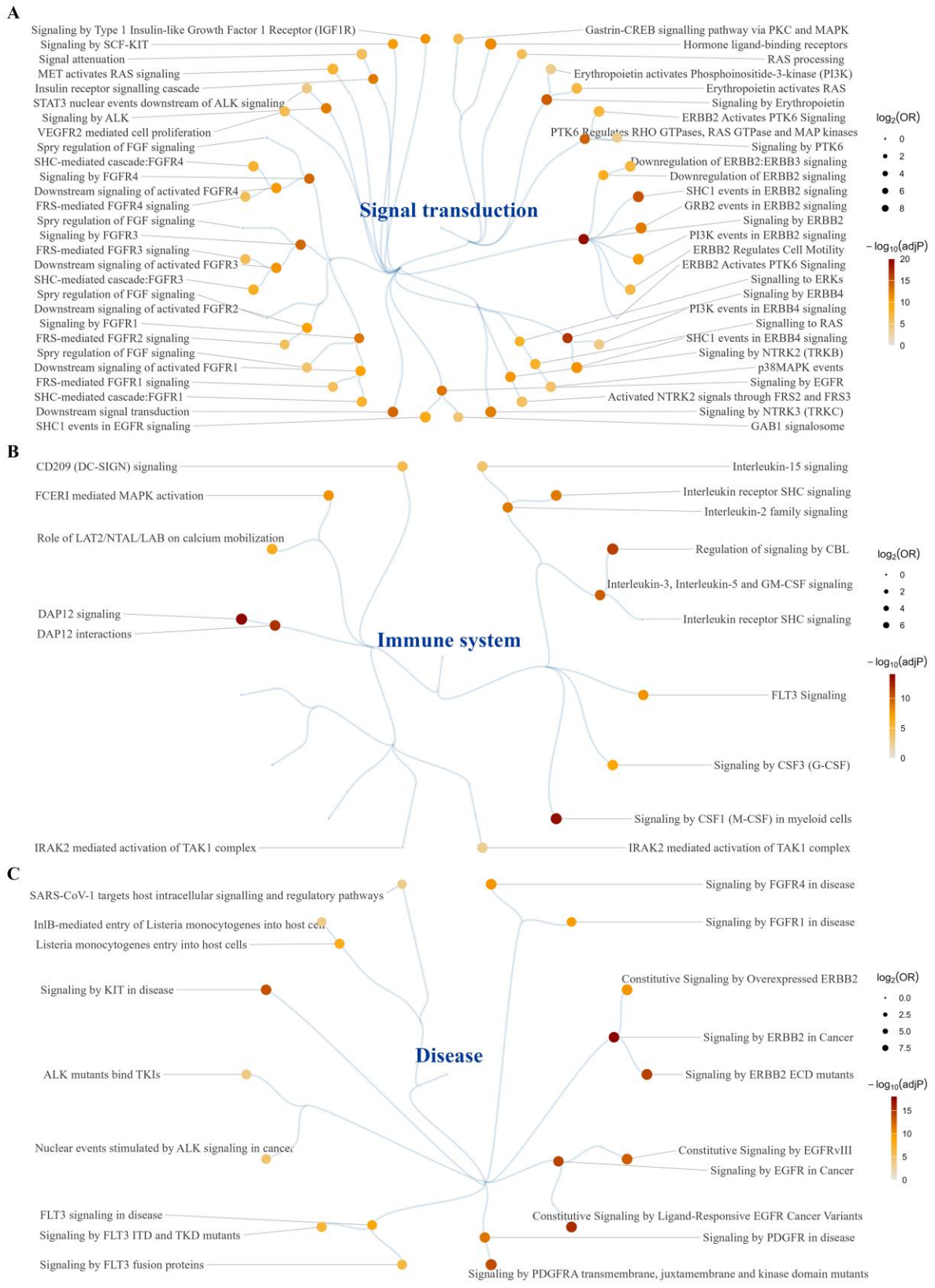

Figure S5: Grouping of enriched Reactome pathway into hierarchical subgraphs in EAS population: Circular overview of pathway enrichment among the top 100 prioritized genes in ovary tissue for EAS population, grouped by pathway hierarchy into **(A)** *Signal Transduction* **(B)** *Immune*

System **(C)** *Disease-associated* pathways. Each node represents a Reactome pathway, hierarchically organized under major biological categories (e.g., Signal Transduction, Immune System, and Disease). Node size corresponds to enrichment strength ($\log_2$ odds ratio). Node color intensity reflects statistical significance ($-\log_{10}$ adjusted P).

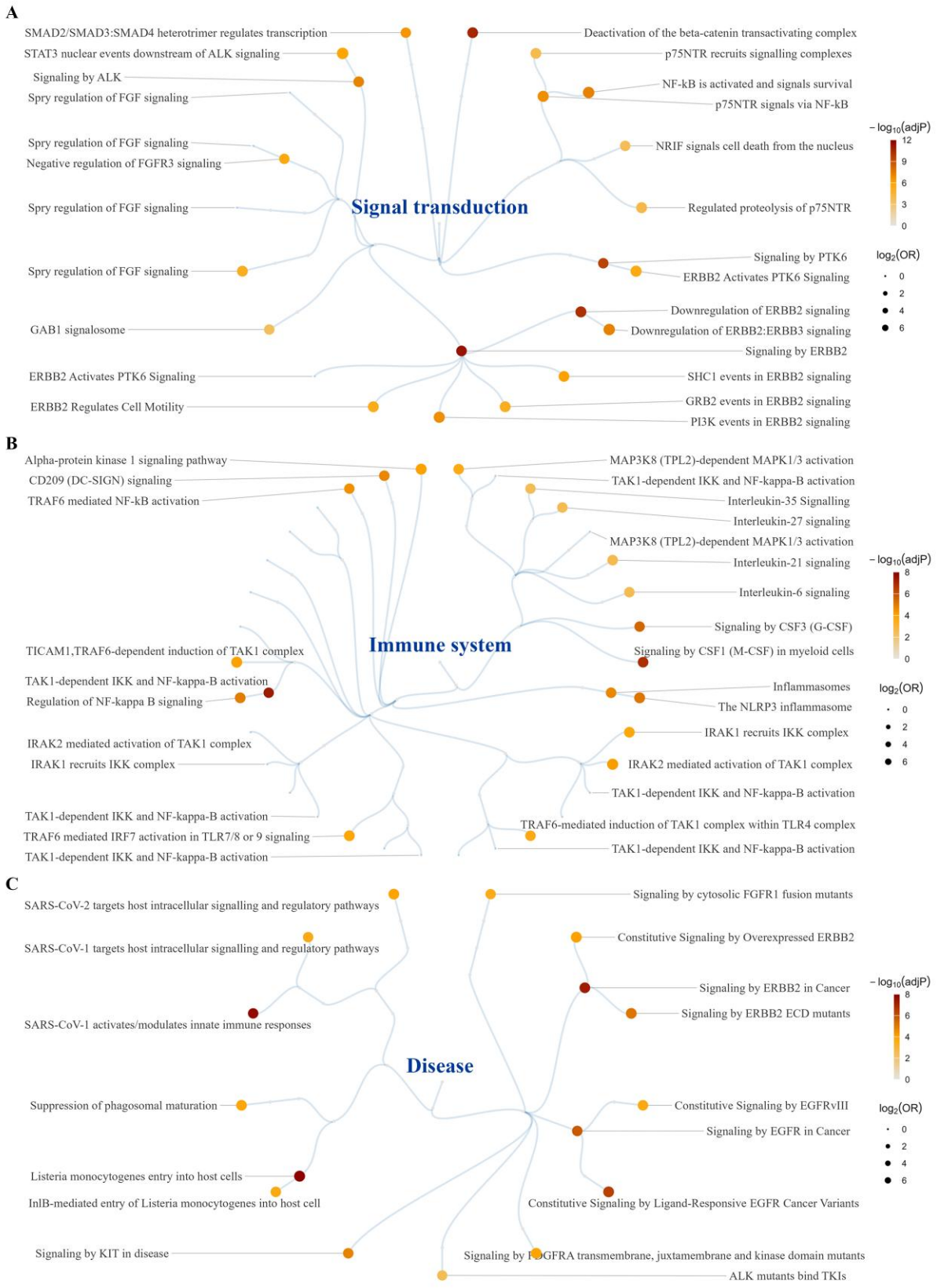

Figure S6: Grouping of enriched Reactome pathway into hierarchical subgraphs in EUR population: Circular overview of pathway enrichment among the top 100 prioritized genes in ovary tissue for EUR population, grouped by pathway hierarchy into **(A)** *Signal Transduction* **(B)** *Immune*

System **(C)** *Disease-associated* pathways. Each node represents a Reactome pathway, hierarchically organized under major biological categories (e.g., Signal Transduction, Immune System, and Disease). Node size corresponds to enrichment strength ($\log_2$ odds ratio). Node color intensity reflects statistical significance ($-\log_{10}$ adjusted P).

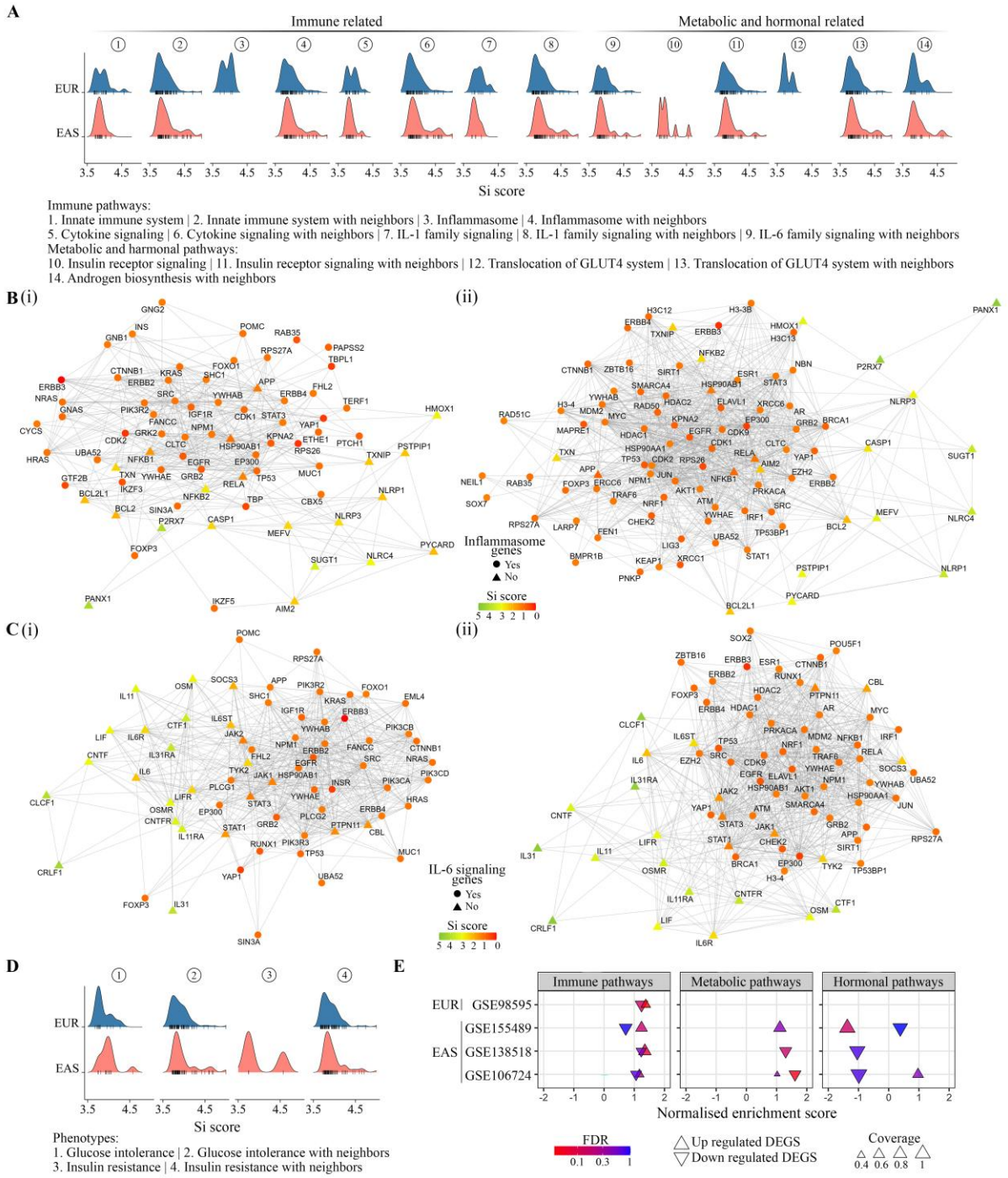

Figure S7: Genetic links of PCOS pathology with the immune, metabolic and androgen related Pathways: (A) Ridge plots displaying the density distribution and Si rating for the top 100 prioritized genes of ovary and their direct neighbors belonging to important functional pathways (immune, metabolic and androgen related). **(B&C)** Protein interaction networks of inflammasome (B) and interleukin-6 (C) centered on top 100 genes and their first neighbors. Gene nodes are color-coded by Si rating and shapes indicate whether they are functionally annotated for the pathway. (i) EAS Ovary (ii) EUR Ovary. **(D)** Ridge plots displaying the density distribution and Si rating for the top 100 prioritized genes of ovary and their direct neighbors belonging to pathways related to Insulin Resistance from

Human Phenotype Ontology (HPO). **(E)** Scatter plot illustrating the enrichment of immune, metabolic and hormonal pathways in DEGs of PCOS patients. Statistical significance of enrichment analysis were calculated using with 20,000 permutations.

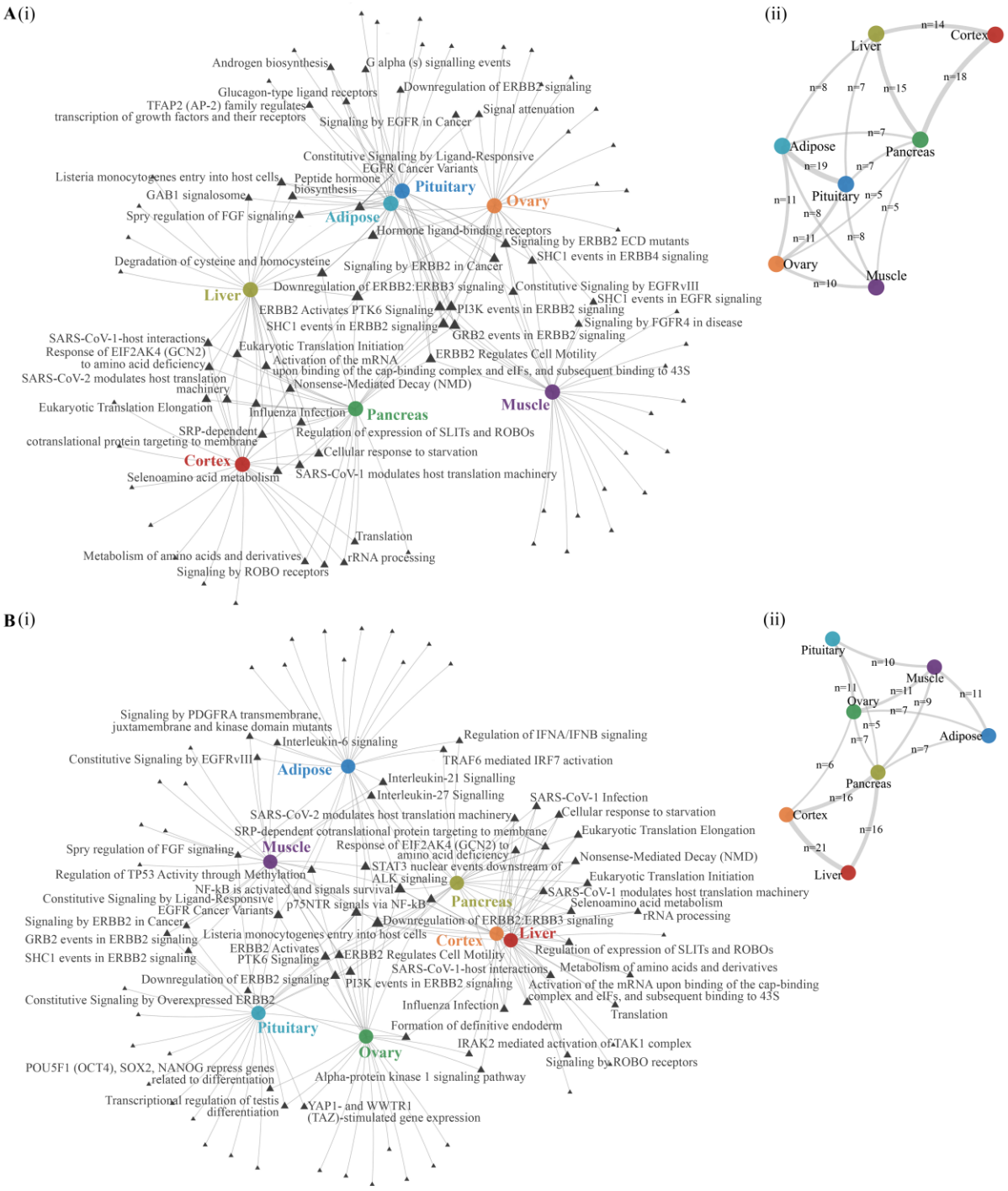

Figure S8: Representation of shared pathway among tissues: A (i) and B (i) Network-like representation of inter tissue (denoted as circular nodes) relationship based on overlap of top 25 (ranked by odds ratio) enriched Reactome pathways (denoted as triangular nodes) identified from the top 100 prioritized genes across tissues in EAS and EUR respectively. Pathway nodes are sized according to the number of tissues in which they are common. **A (ii) and B (ii)** Summary of the inter-tissue relationship of EAS and EUR respectively presented as connectivity network based on shared enriched pathways (>5). Edges thickness represents the number of shared pathways.

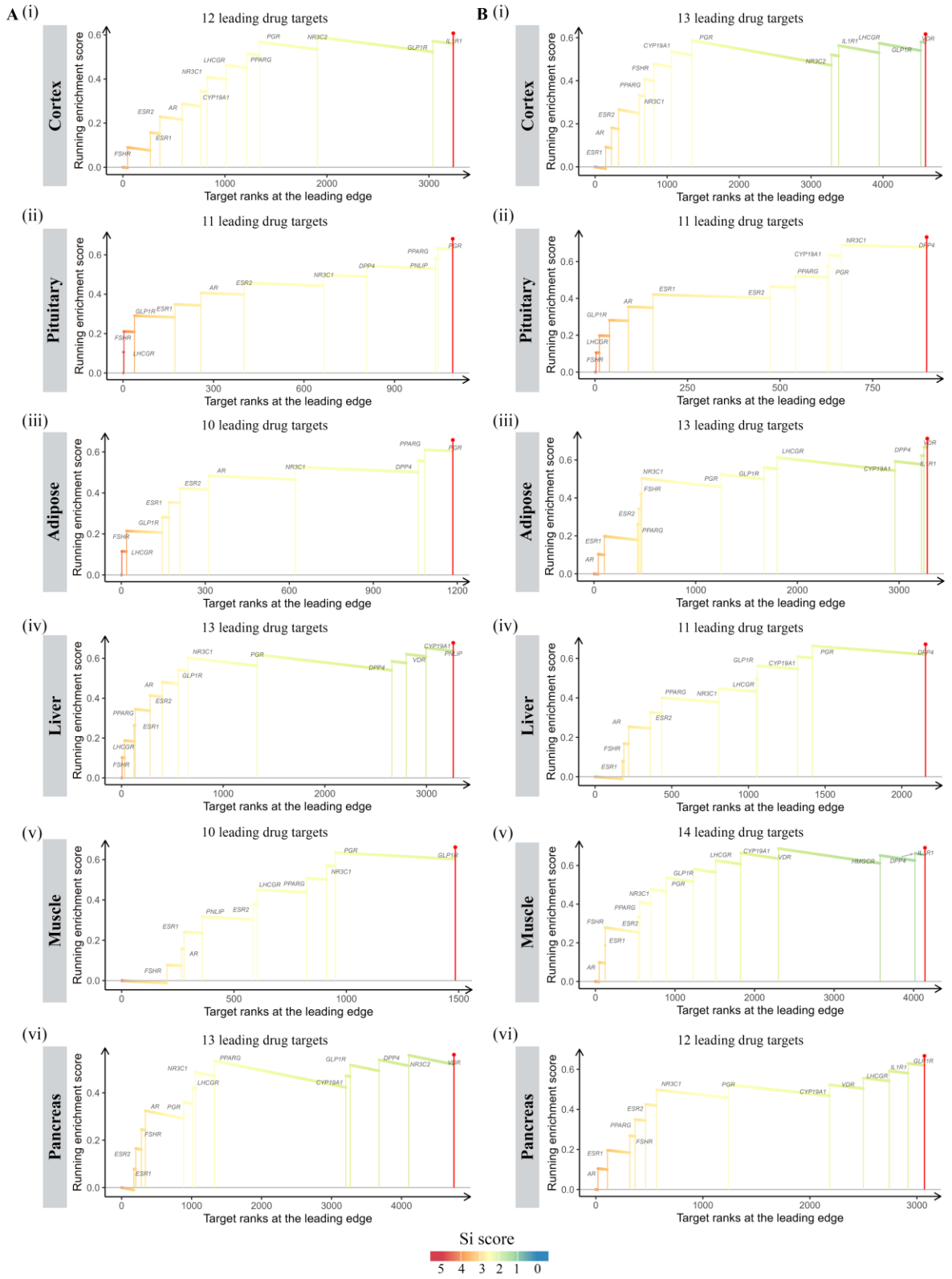

Figure S9: TSEA of PCOS drug targets in other disease relevant tissues: Each panel displays the leading edge of target set enrichment analysis in various tissues (i-vi) within EAS (A) and EUR (B) population. Known PCOS drug targets (Phase 2 and above) recovered within the leading edge of gene priority list across tissues are marked.

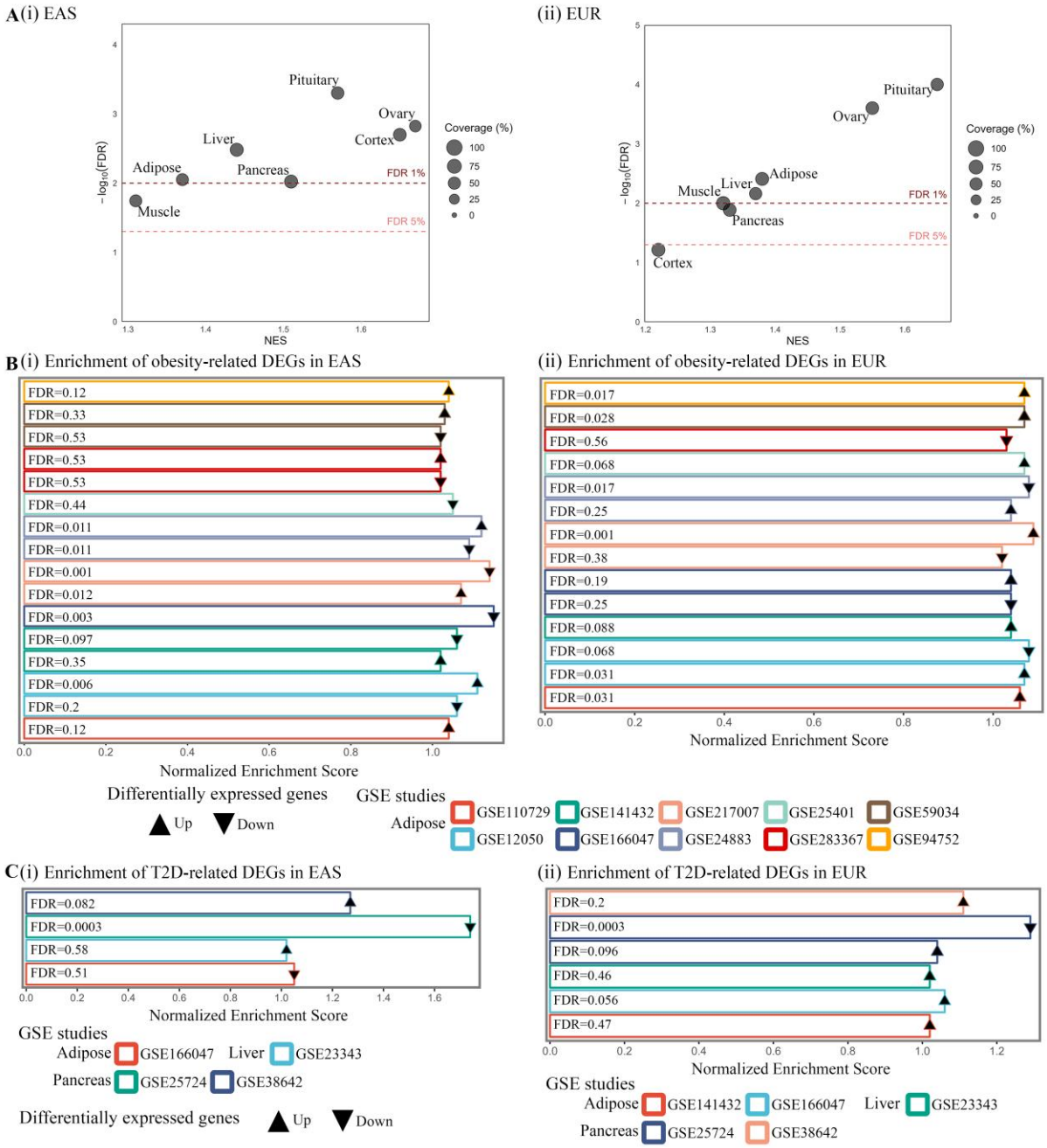

Figure S10: Therapeutic potential of known PCOS drug targets across tissue (A) Scatterplots summarizing TSEA-based therapeutic potential across all tissues for each ancestry group. Normalized Enrichment Score (NES), target coverage (that is, the total number within the leading edge for that tissue/total number of PCOS drug targets), and significance ($-\log_{10}$FDR) are shown for **(i)** EAS and **(ii)** EUR tissues. Enrichment of obesity (**B**) and T2D-related (**C**) genes in the prioritized gene list of PCOS for EAS (i), and (ii) EUR.

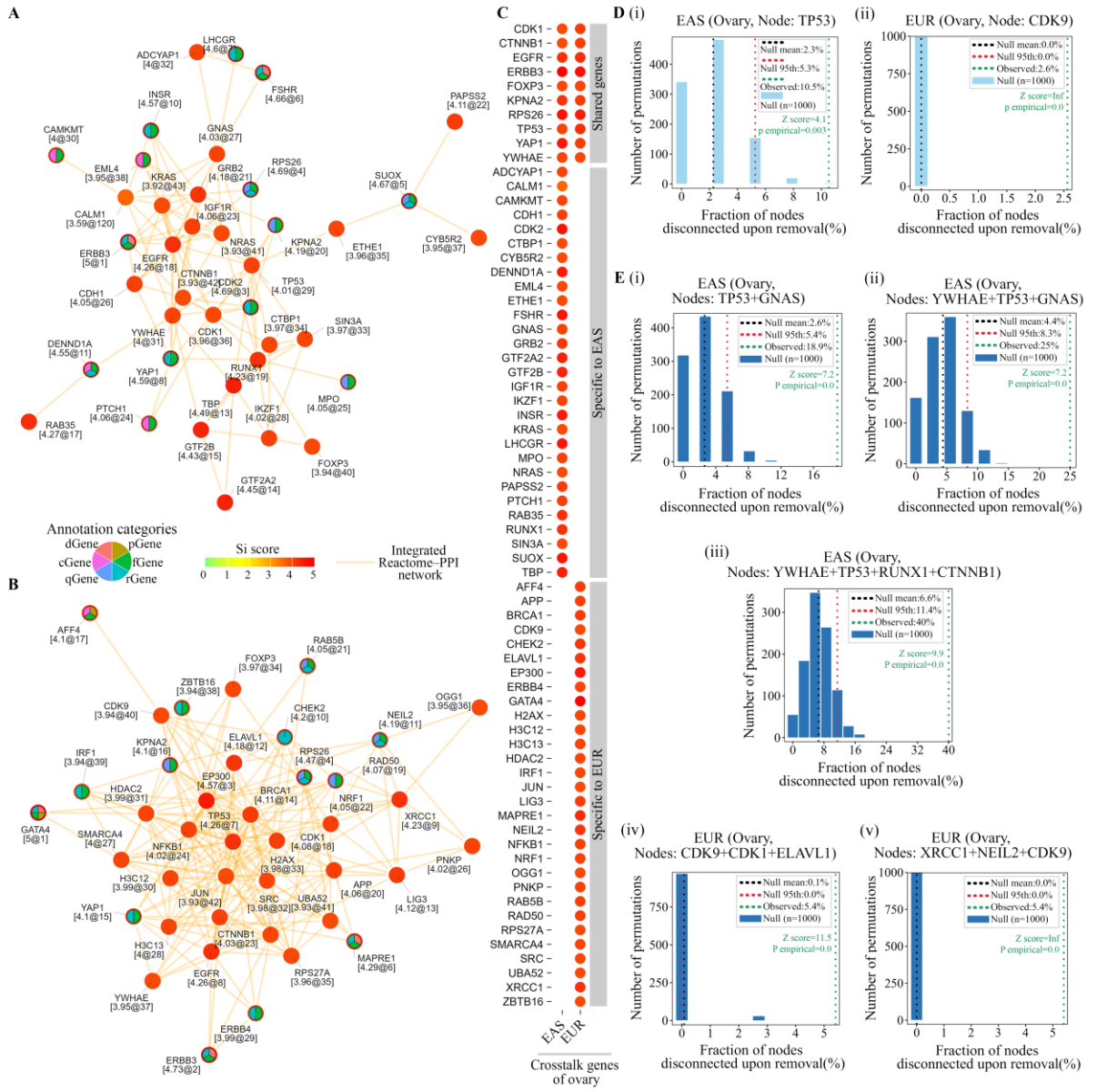

Figure S11: **Pathway crosstalk analysis:** Illustration of pathway crosstalk network of prioritized genes from ovary tissue in (A) EAS, (B) EUR. Each node represents a gene labeled with its symbol and Si score (formatted as "rating @ rank"), colored by magnitude. Edges indicate protein–protein interactions (PPIs). The core genes are designated according to the annotation categories. Both the networks were found to be statistically significant while performing degree preserving node permutation test. (C) Heatmap depicting the shared and unique genes of pathway crosstalk in both populations. (D) Histogram showing the null distribution of the fraction of nodes disconnected upon removal of single node from 1000 degree-preserving rewired networks generated from EAS Ovary (i) and EUR Ovary (ii) using the Maslov-Sneppen configuration model. The observed disconnected fraction in original network is indicated as green dotted line. (E) Combinatorial removal of nodes from 1000 degree-preserving rewired networks from EAS Ovary (i-iii) and EUR Ovary (iv-v).

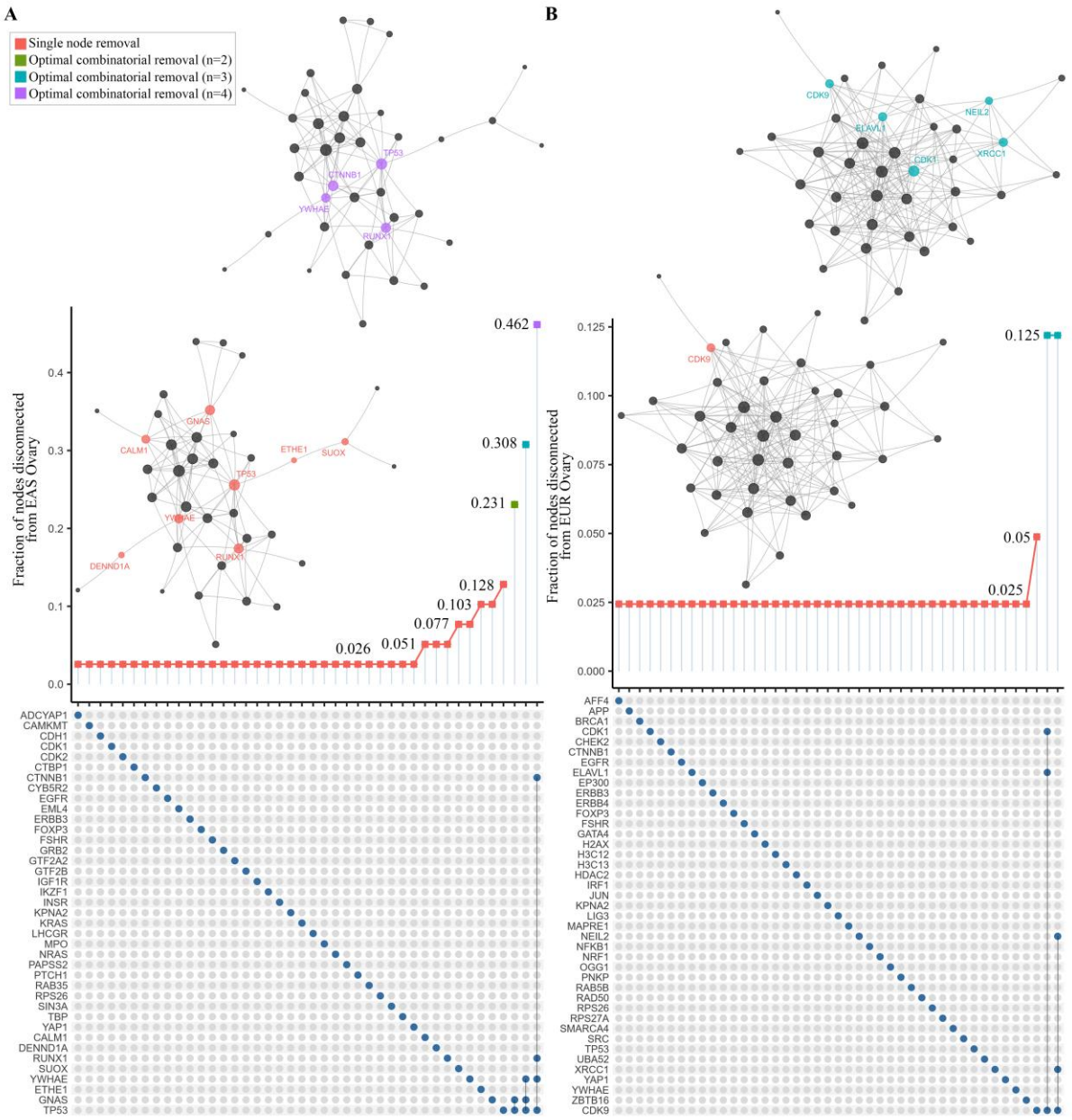

Figure S12: **Node Removal Analysis of pathway crosstalk in ovary:** Effect of single or combinatorial node removal on network robustness was assessed by determining the fraction of nodes disconnected from the largest connected component in EAS (A) and EUR population (B). The y-axis represents the fraction of nodes disconnected, while the x-axis denotes the sequential removal of nodes marked by blue circles in the UpSet plots presented below the x-axis. The plots also illustrates the node combination used for the removal analysis. Nodes having highest removal effect either as single or in combinations are labelled on the crosstalk networks (only the single node and 4 node combinatorial removal analysis on the network are shown).

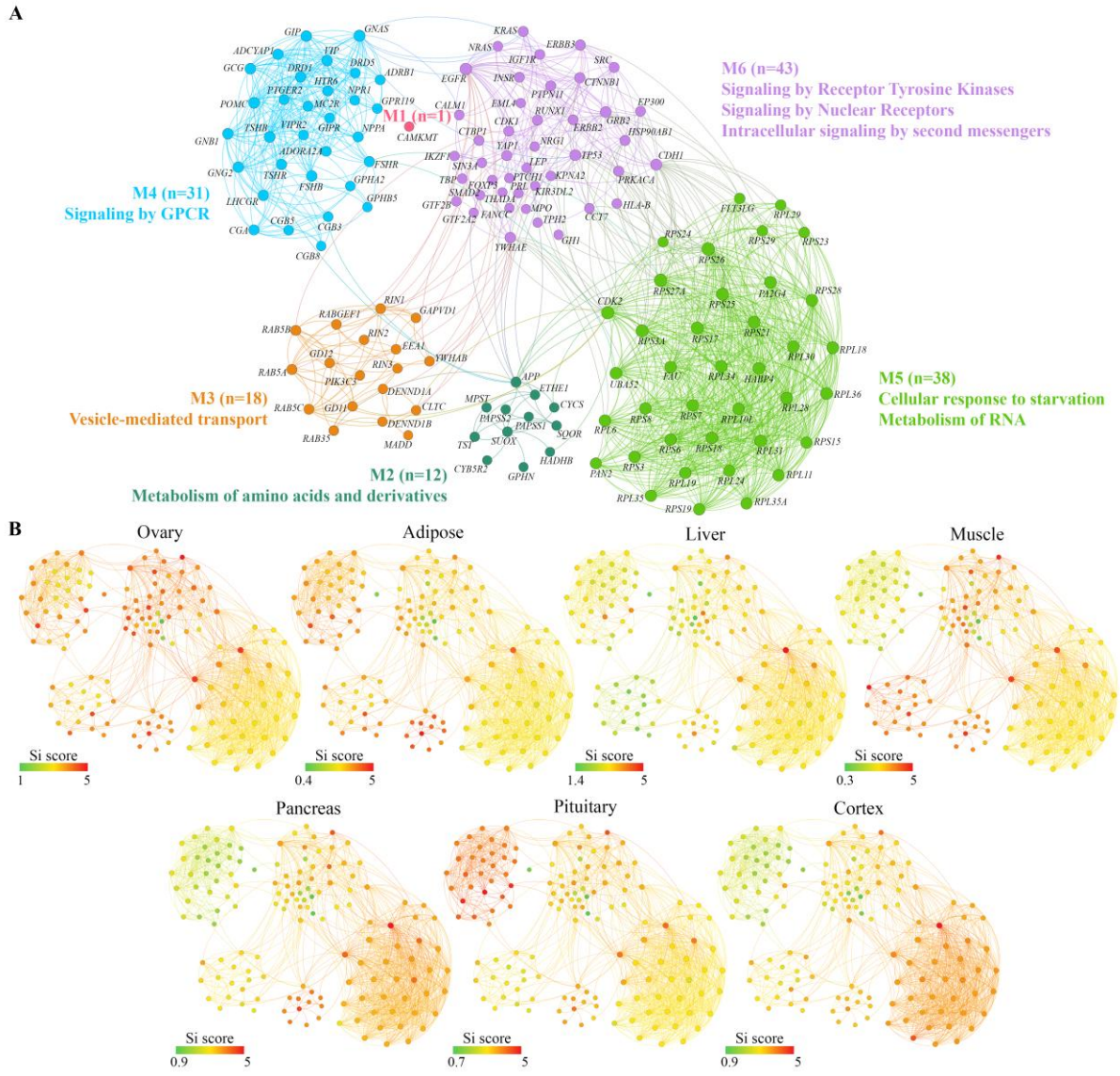

Figure S13: Modular Analysis of cross-tissue pathway crosstalk in EAS population: (A) Community detection analysis by measuring modularity of merged pathway crosstalk networks across tissues in EAS population revealed six modules, each annotated by functional enrichment using one-sided Fisher's exact test. Module-based visualization of pathway crosstalk genes from EAS. **(B)** The same modular network layout from panel (A) is shown with nodes colored by Si score of each individual tissue.

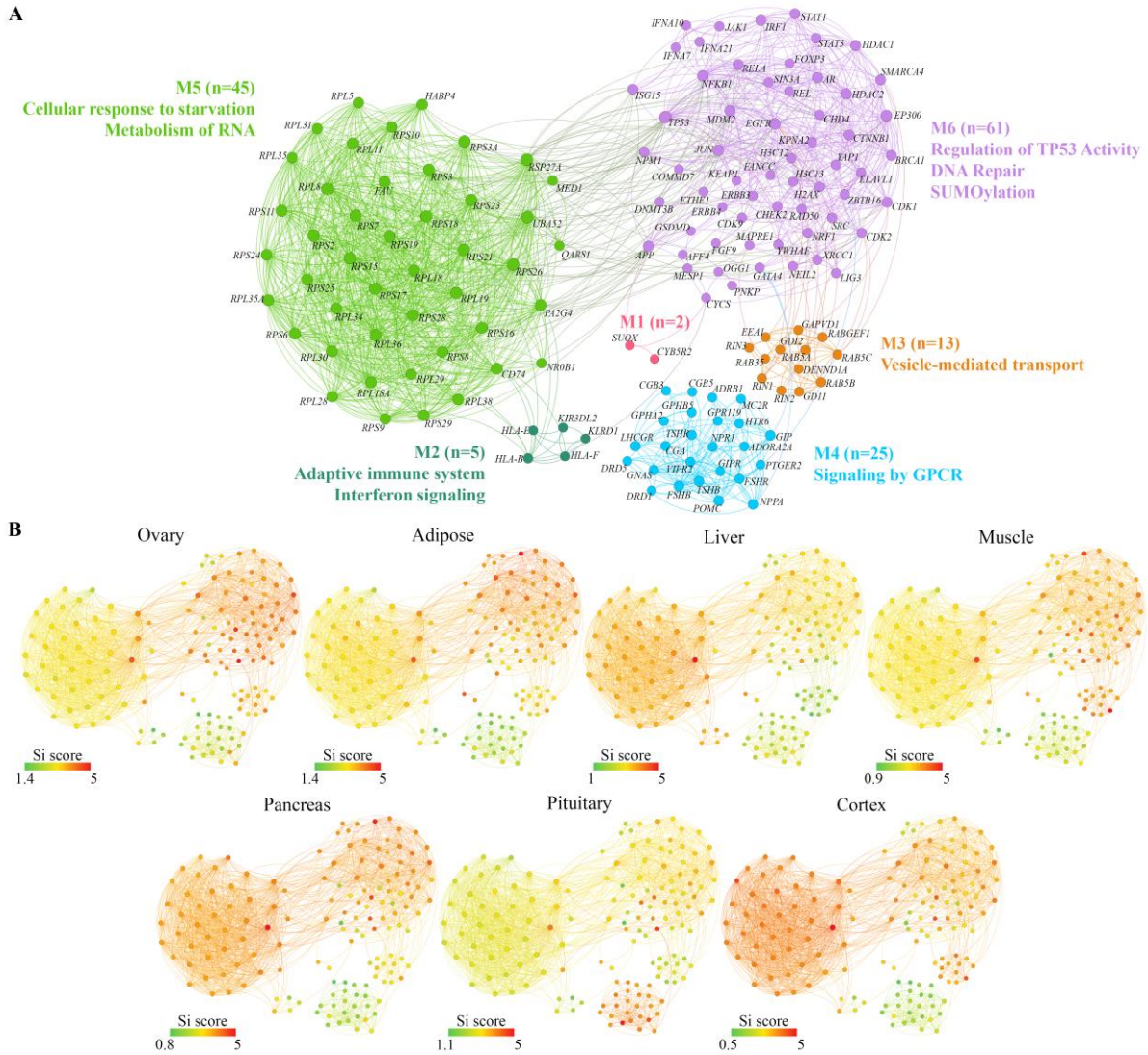

Figure S14: Modular Analysis of cross-tissue pathway crosstalk in EUR population: (A) Community detection analysis by measuring the modularity of merged pathway crosstalk networks across tissues in EUR population revealed six modules, each annotated by functional enrichment using one-sided Fisher's exact test. Module-based visualization of pathway crosstalk genes from EUR. **(B)** The same modular network layout from panel (A) is shown with nodes colored by Si score of each individual tissue.

A EAS


<H25>
ADCYAP1
GCG
GIP
GNG2
GPHB5
NPPA
TSHB
C2
<H26>
CGA
FSHB
GNAS
GPHA2
LHCGR
POMC
<H24>
TSHR
VIP
<H23>
ADORA2A
CGB3
CGB5
CGB8
DRD1
DRD5
GIPR
GPR119
HTR6
MC2R
NPR1
PTGER2
VIPR2
C1
<H27>
FSHR
<H12>
GNB1
TBP
<H11>
GTF2A2
GTF2B
MPO
<H22>
ADRB1
<H28>
CYB5R2
CYCS
ETHE1
GPHN
MPST
PAPSS1
PAPSS2
SQOR
TST
<H9>
CALM1
<H29>
SUOX
RAB5B
CDK2
RPS26
ERBB3
<H14>
DENND1A
YWHAE
<H5>
CLTC
RAB35
YWHAB
IKZF1
NRG1
SIN3A
<H1>
DENND1B
MADD
RAB5A
RAB5C
RIN1
<H2>
EEA1
GDI1
GDI2
PIK3C3
RABGEF1
RIN2
RIN3
THADA
<H20>
CAMKMT
EML4
KIR3DL2
LEP
PRL
PTCH1
TPH2
<H30>
PA2G4
EGFR
INSR
YAP1
<H15>
CDK1
FOXP3
<H6>
HADHB
CTBP1
FANCC
KRAS
NRAS
SMAD2
<H7>
GAPVD1
<H19>
PTPN11
<H37>
GH1
<H35>
HABP4
RPL10L
RPL11
RPL18
RPL19
RPL24
RPL28
RPL29
RPL30
RPL31
RPL34
RPL35
RPL35A
RPL36
RPL6
RPS15
RPS18
RPS23
RPS7
<H34>
FAU
FLT3LG
RPS17
RPS19
RPS21
RPS24
RPS25
RPS28
RPS29
RPS3
RPS3A
RPS6
RPS8
C3
<H31>
APP
CDH1
CTNNB1
ERBB2
GRB2
IGF1R
RUNX1
SRC
TP53
<H17>
PRKACA
<H36>
PAN2
CCT7
HLA-B
<H32>
RPS27A
UBA52
EP300
HSP90AB1
KPNA2
C4

B EUR

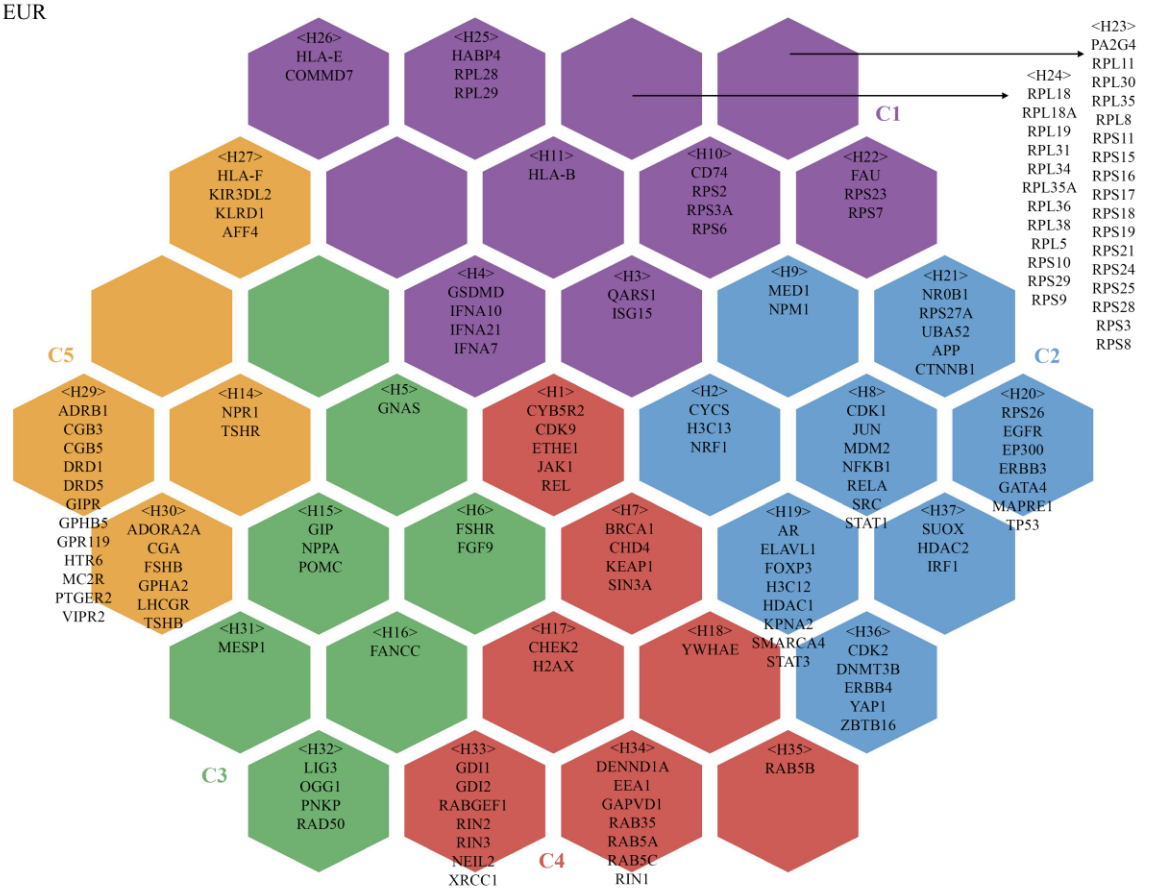
<H26>
HLA-E
COMMD7
<H25>
HABP4
RPL28
RPL29
<H23>
PA2G4
RPL11
RPL30
RPL35
RPL8
RPS11
RPS15
RPS16
RPS17
RPS18
RPS19
RPS21
RPS24
RPS25
RPS28
RPS3
RPS8
<H24>
RPL18
RPL18A
RPL19
RPL31
RPL34
RPL35A
RPL36
RPL38
RPL5
RPS10
RPS29
RPS9
C1
<H27>
HLA-F
KIR3DL2
KLRD1
AFF4
<H11>
HLA-B
<H10>
CD74
RPS2
RPS3A
RPS6
<H22>
FAU
RPS23
RPS7
<H4>
GSDMD
IFNA10
IFNA21
IFNA7
<H3>
QARS1
ISG15
<H9>
MED1
NPM1
<H21>
NR0B1
RPS27A
UBA52
APP
CTNNB1
C5
C2
<H29>
ADRB1
CGB3
CGB5
DRD1
DRD5
GIPR
GPHB5
GPR119
HTR6
MC2R
PTGER2
VIPR2
<H14>
NPR1
TSHR
<H5>
GNAS
<H1>
CYB5R2
CDK9
ETHE1
JAK1
REL
<H2>
CYCS
H3C13
NRF1
<H8>
CDK1
JUN
MDM2
NFKB1
RELA
SRC
STAT1
<H20>
RPS26
EGFR
EP300
ERBB3
GATA4
MAPRE1
TP53
<H30>
ADORA2A
CGA
FSHB
GPHA2
LHCGR
TSHB
<H15>
GIP
NPPA
POMC
<H6>
FSHR
FGF9
<H7>
BRCA1
CHD4
KEAP1
SIN3A
<H19>
AR
ELAVL1
FOXP3
H3C12
HDAC1
KPNA2
SMARCA4
STAT3
<H37>
SUOX
HDAC2
IRF1
<H31>
MESP1
<H16>
FANCC
<H17>
CHEK2
H2AX
<H18>
YWHAE
<H36>
CDK2
DNMT3B
ERBB4
YAP1
ZBTB16
C3
<H32>
LIG3
OGG1
PNKP
RAD50
<H33>
GDI1
GDI2
RABGEF1
RIN2
RIN3
NEIL2
XRCC1
C4
<H34>
DENND1A
EEA1
GAPVD1
RAB35
RAB5A
RAB5C
RIN1
<H35>
RAB5B

Figure S15: Cross-tissue prioritization gene clustering: The clustered 2D hexagonal map with target genes listed per hexagon (H1 - H37) indexed circularly outward from the center in **(A)** EAS and **(B)** EUR.

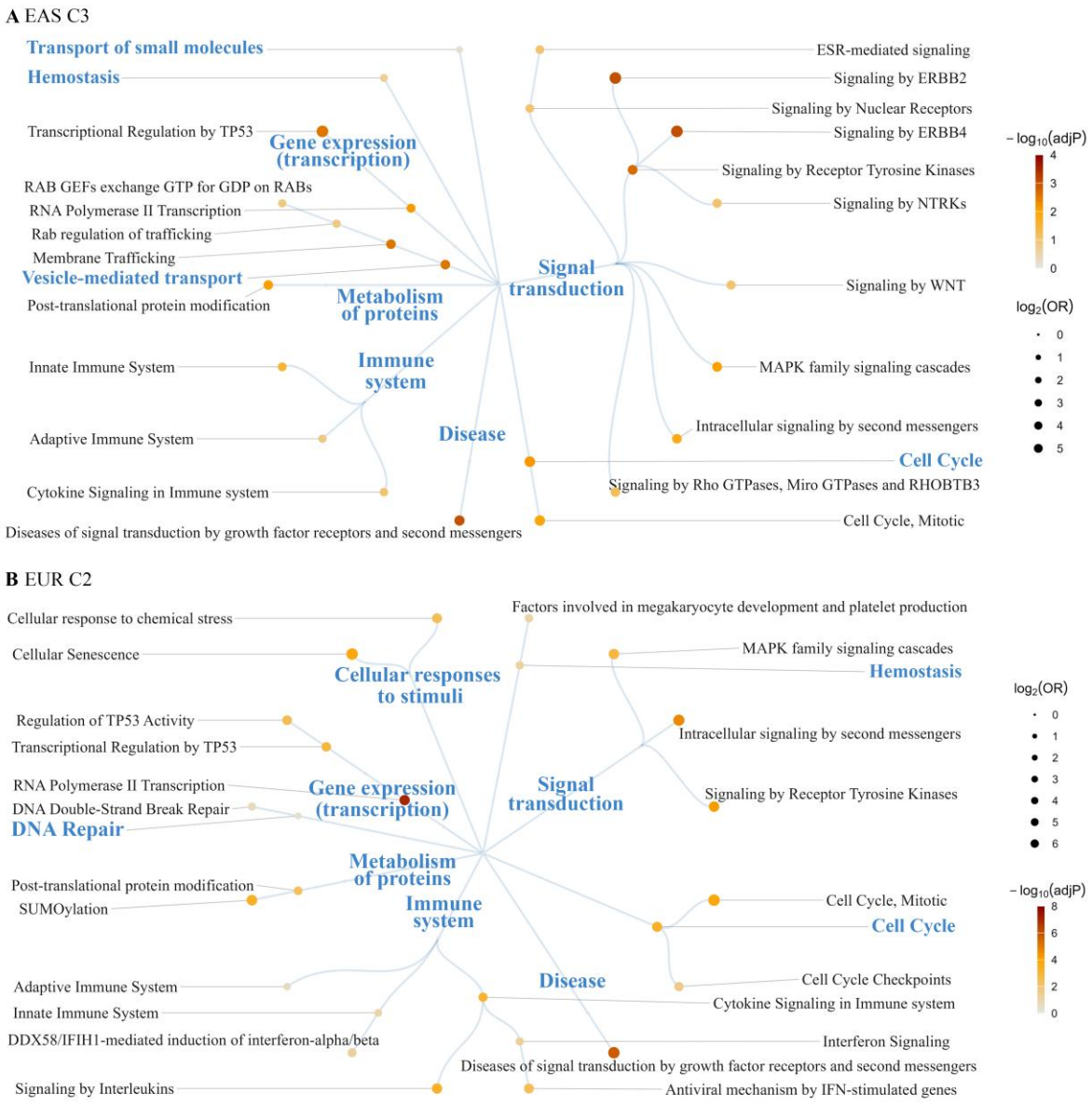

Figure S16: Pathway enrichments of target genes in the high scored and druggable clusters in populations: Reactome pathway enrichment analysis of high scored and druggable cluster; cluster 3 in EAS **(A)** and cluster 2 EUR population **(B)** are presented. Enrichment was performed using one-sided Fisher's exact test. Pathways are hierarchically organized in a circular bubble layout where the bubble size indicates the odds ratio (OR) and color intensity corresponds to the false discovery rate.

Supplementary Note

GWAS summary-level data collection and processing

To identify genetic variants associated with polycystic ovary syndrome (PCOS), we accessed summary-level GWAS data from the NHGRI-EBI GWAS Catalog (www.ebi.ac.uk/gwas), using the Experimental Factor Ontology (EFO) term for PCOS "EFO:0000660" as our primary search criterion. This query yielded 12 studies reporting genome-wide significant associations with PCOS, of which 6 studies were from Han Chinese and Korean (which comprised our EAS (EAS) Group)[1–5] and 5 studies were from individuals of EUR origin (comprising our EUR (EUR) group)[6–10]. One study[11] was not considered because it originated from mixed population. In addition to these GWAS, one more study[12] identified through literature search was included in the EAS group. For each selected study, we extracted lead SNPs that met the genome-wide association significance (P-value < 5×10^{-5}). This yielded a total of 107 and 53 SNPs in EAS and in EUR respectively (Supplementary Data 1, I). To capture additional genetic variants, potentially associated to the lead SNPs due to linkage disequilibrium, we conducted linkage disequilibrium (LD) expansion and enlisted linked SNPs with $r^2 \geq 0.5$ within a genomic window of ±500 kb in a population specific manner. LD calculations were based on Phase 3 of the 1000 Genomes Project, using the LDProxy tool[13] and selected ancestry-matched reference populations (EUR or EAS) accordingly. This process yielded LD-expanded SNP sets of 2,013 and 1,872 SNPs for EAS and EUR populations respectively (Supplementary Data 1, I). To evaluate the relative contribution of each SNP to PCOS, we computed a composite SNP score (Eq1, Figure S1) that integrates both association significance (P-value) of lead SNP and corresponding LD strength (r^2) of linked SNP from LDProxy (as mentioned in the earlier study[14]). We have used the threshold of significance $=5 \times 10^{-5}$.

Optimization for rGene Annotation

Before rGene annotation, an optimization step to determine two critical parameters (Genomic influential range and Network influential range) required for SNP-to-nearest gene mapping was done:
1) Genomic Influential Range: The regulatory impact of a genomic region containing SNPs typically decays with increasing distance from a gene. To model this phenomenon, we used a similar method reported previously[14]. Briefly, we evaluated combinations of distance windows (0 kb to 1 Mb) with different decay functions (constant, linear, slow decay and rapid decay) to quantify how the influence of SNPs decay of on nearby genes within defined window by accessing the ability to discriminate Gold standard positives drug targets of PCOS (GSPs) from Gold standard negatives (GSNs) from the ChEMBL version 34 database[15].

- ***Defining Gold Standard Positives (GSPs)*:** Genes targeted by drugs that are in clinical development phase 2 or above, indicating possible evidence of their efficacy in treating PCOS, were considered positive targets (Supplementary Data 1, II). To reduce bias, we excluded drugs

with more than 5 targets, as it becomes challenging to determine which among these targets are of clinical significance pertaining to the disease of our interest and can bias our down-stream analysis.

- ***Simulating Gold Standard Negatives (GSNs)*:** Negative targets were simulated based on the GSPs. The first step involved forming a druggable landscape by including all drug targets reported for *Homo sapiens*, regardless of the diseases or drug development phases. We constructed a druggable space by extracting a network by overlaying these targets on the background PPI network. GSNs were derived by excluding GSPs and their interacting neighbors (1st and 2nd degree neighbors), ensuring that the remaining genes represented true negatives. The 1st and 2nd degree neighbors were defined according to the integrated PPI network (as mentioned below).
- **PPI Network used:** We have combined protein-protein interaction information from various sources. The said network was constructed by integrating two high-confidence interaction information sources: the human interactome compiled by Cheng, F. *et al*[16] and the STRING database (v12.0)[17] . The interactome by Cheng, F. *et al* was derived from the integration of 15 publicly available databases and their in-house resource, collectively encompassing diverse categories of experimentally supported interactions. These included binary yeast two-hybrid (Y2H) interactions, kinase–substrate relationships, affinity purification followed by mass spectrometry (AP-MS), interactions inferred from 3D protein structural data, and curated signaling pathways. UBC was removed from the Cheng, F. *et al.* network due to its overrepresentation, and to reduce bias. In parallel, we retrieved PPI data from the STRING database and filtered it to retain only those interactions that were either experimentally validated or database-supported, applying a confidence score threshold of ≥ 700. The combined network comprised approximately 17,500 unique genes.

2. Network Influential Range: To propagate the influence of proximal genes within protein-protein interactome (PPI) in order to identify additional genes relevant with respect to the proximal gene list, depending on network connectivity, we performed Random Walk with Restart (RWR) (described in details below) with restart probabilities ranging between 0.1 and 0.9. This generated a ranked gene list based on RWR affinity scores for each combination of distance windows and decay functions. The performance of each parameter combination was evaluated based on its ability in discriminating GSPs from GSNs (defined above). The parameter set that maximally separates GSPs from GSNs was selected as the optimal genomic influential range and network influential range. Both of these influence information are needed to identify clinically relevant rGenes.

Pathway enrichment analysis

Enrichment analysis was conducted using human Reactome pathways derived from the Molecular

Signatures Database (MSigDB) version 2023.2, specifically the curated gene sets under category C2[18]. The analysis was performed using the xEnricher function from the Pi package in R (version 2.10.0), which implements a one-sided Fisher's exact test (hypergeometric test) to assess statistical significance. Enrichment results were represented with three metrics: Z-score, odds ratio (OR) and 95% confidence interval (CI), and p-value. The p-values obtained were adjusted for multiple hypothesis testing using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR). Z-scores and ORs were used to rank the enriched terms. To visualize the broader biological context of the enriched pathways, the Reactome pathway hierarchy was utilized. The "Pathway hierarchy relationship" file (version 90), containing parent–child relationships, was downloaded from the Reactome database. This file includes two columns: the first corresponding to the parent pathway and the second to the child pathway. A hierarchical tree of pathway relationships was constructed from this file, while only relevant enriched pathways (based on OR) were extracted and used for visualizations. The background, positive and check sets were tailored based on the objective of each analysis. For the enrichment of top 50, 75, 100 and 175 prioritized genes in reactome pathways, we utilized ~17500 rated genes as background set for each tissue. While evaluating enrichment of common core and peripheral genes identified within the 100 prioritized genes across all tissues, the background was restricted to the pooled top 100 genes across all tissues.

Benchmarking the scoring strategy

Benchmarking of this approach was done by comparing the performance of Si approach (this study) with other competing methods. The process was also based on evaluating AUC to determine how well our approach separates 'clinical proof-of-concept targets' (Gold standard Positives) of PCOS from negative controls) (See above). Specifically, this approach was compared with 'Open Targets[19] text mining' (an approach of evaluating the importance of a gene associated to a disease by using Natural Language Processing based text mining from scientific literature and 'Open Targets Genetic association' (based on curated genetic associations to a disease from literature, UK biobank and FinnGen. Additionally, two publicly available European-ancestry GWAS summary statistics for PCOS were obtained from the NHGRI-EBI GWAS Catalog: Day *et al*.[7] (GCST007089, excluding 23andMe participants; 4,890 cases, 20,405 controls), and Tyrmi *et al*.[10] (GCST90044902; 3,609 cases, 229,788 controls). Analyses were restricted to autosomal variants (chromosomes 1–22). Both datasets were filtered to retain unique, monoallelic SNPs with MAF > 0.01. To prevent strand ambiguity, palindromic SNPs with EAFs of 0.40–0.60 were additionally excluded from the Day *et al*. dataset, this filter was not applied to Tyrmi *et al*. as EAF data was unavailable.

- **FUMA**

  Filtered summary statistics were uploaded to SNP2GENE[20] of FUMA (v1.8.3) for gene mapping. FUMA defined lead SNPs ($p<1\times10^{-5}$), identified all SNPs that were in linkage disequilibrium within a window of ±500 kb (1000G Phase3 EUR) and mapped gene through three complementary strategies: positional mapping within ±10 kb window for each SNP

filtered based on default RegulomeDB and 15-core chromatin state annotations; eQTL mapping across all available data; chromatin interaction mapping using all builtin chromatin interaction datasets to capture SNPs overlapping enhancer regions interacting with promoters. From this, genes mapped by SNPs with gwas $p< 1\times10^{-5}$ and supported by at least one mapping strategy were used as input for RWR on PPI network and the prioritized genes were used for benchmarking.

- **MAGMA**

  Gene-level association statistics were computed using MAGMA[21] (v1.10) within the FUMA pipeline, with gene window of ±50 kb. Both the .out and .raw output files were generated for downstream PoPS analysis. PoPS was implemented using a RidgeCV regression framework with 25 regularization alphas, optimized via generalized leave-one-out cross-validation using normalized mean squared error (NMSE) as the scoring metric. Each PCOS dataset was processed independently. To obtain a single gene-level association statistic per gene, the maximum z-statistic across the two datasets was retained.

- **PoPS**

  MAGMA output files were used and input for PoPS[22] (Polygenic Priority Score) to get causal genes share functional characteristics. Gene features were derived from default GTEx expression feature data across 53 tissues. No feature filtering was applied (FDR threshold set to 1), retaining all 53 GTEx tissue features in the model. PoPS was run independently for each dataset. All genes with computed maximum PoPS scores across two datasets were carried forward for benchmarking.

Credible set locus-to-gene (L2G)[23] scores > 0.05 were obtained from the Open Targets Platform for PCOS (EFO:0000660), providing an orthogonal machine learning-based measure of gene-to-variant assignment confidence. The resulting 190 genes were subjected to RWR on the PPI network for prioritization. The resulting prioritized genes were carried forward for benchmarking analysis.

Cross-tissue concordance analysis

To formally get an idea about the consistency of gene prioritization across seven disease-relevant tissues within each population, and to see that the Si score rankings across tissues converge beyond expectation by chance, we used three complementary non-parametric concordance analysis (Mean pairwise Spearman Rank correlation, Top-K Jaccard overlap with null permutation and Kendall's coefficient of concordance) which we performed on 14 population-tissue ranking matrices (7 tissues × 2 populations). In case of **Mean pairwise Spearman rank correlation**, the correlation coefficient (ρ) was computed for all 21 tissue pairs in each population using 'spearmanr' function from the scipy.stats module (SciPy v1.11). Analysis was restricted to the top 500 genes, identified by their mean rank across the tissues, as we expected that the tissue-specific ranking differences are most informative within the high ranks. To test whether the set of highest-priority genes is shared across the tissues beyond chance, the **Jaccard**

similarity coefficient was computed for each of the 21 tissue pairs using the top 100 gene (ranked by Si Score) for each tissue; these top 100 gene sets served as the reference set for all subsequent permutation. The mean Jaccard coefficient across all 21 pairs constituted the observed test statistics. A null distribution was generated by independently shuffling the rank assignments within each tissue 1000 times and computing the mean pairwise Jaccard coefficient at each iteration. This step preserves the within –tissue rank structure while destroying any cross-tissue signal. A one-sided empirical p-value was defined as the proportion of permuted statistics meeting or exceeding the observed value. Fold enrichment over the null was computed relative to the analytically expected Jaccard under full independence, given by $k^2/N^2/ (2k/N- k^2/N^2)$, where k=100 and N is the total number of genes in the matrix. All the permutation analysis were implemented in Python using NumPy (v1.24) and pandas (v2.0). The **Kendall's coefficient** of concordance (W) (the canonical coefficient of concordance for k independent rankers applied to n common subjects) was computed treating the 7 tissues as rankers and genes as subjects. Analysis were restricted to the top 500 genes by mean rank. The genes were re-ranked within this mean ranking subset prior to computation. The concordance coefficient was calculated as:

$$W = \frac{12S}{k^2(n^3-n)} \quad \text{Eq.15}$$

where S is the sum of the squared deviations of each gene's rank sum from the grand mean rank sum, k denotes the number of tissue and n is the number of genes in the subset. The statistical significance was assessed via a chi-squared approximation with the test statistic χ^2=k (n-1) W and n-1 degrees of freedom, which provides a reliable approximation of the sample size used here. W ranges from 0 (for complete disagreement) to 1 (for perfect concordance). All computations were performed using Scipy (v1.11) and NumPy (v1.24).

Target set enrichment analysis (Validation 1)

To quantify the extent to which known PCOS drug targets at different phases of clinical development are enriched at the top of the significance matrix (Si Matrix) obtained in the previous stage, we conducted a rank-based gene set enrichment analysis (GSEA) using the xPierGSEA function from the Pi package in R. This enrichment was visually represented as the leftmost region of the peak (leading edge) in the running enrichment plot generated by the analysis. We additionally calculated the Normalized Enrichment Score (NES) using the fgsea package in R (version 1.24.0). The NES was determined by dividing the observed running enrichment score by the mean of the null distribution of enrichment scores, which was generated through permutation testing. We performed 20,000 permutations to estimate the null distribution and assess the statistical significance of the observed enrichment. The resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg (BH) procedure to control the false discovery rate (FDR). To evaluate the contribution of network connectivity, we analyzed the enrichment analysis of core and peripheral genes from leading edge against drug targets for PCOS and other diseases (Phase 2 and above), utilizing a background of ~17500

prioritized genes. Furthermore, to assess repurposing potential, we conducted a subsequent enrichment analysis testing whether the PCOS drug targets recovered in the leading edge were significantly enriched for approved drug targets of other diseases, using all rated genes as the test background.

Genetics-to-Current-Therapeutics (G2CT) potential

To assess and quantitate the importance of the prioritized genes in revealing current therapeutics for a specific tissue, we used a composite metric termed Genetics-to-Current-Therapeutics (G2CT) potential (as described previously[14]. Briefly, this metric integrates three key aspects of enrichment analysis: (i) Normalized enrichment score (NES), (ii) Enrichment score significance assessed through adjusted p-value and (iii) Enrichment coverage defined as the fraction of GSPs (F) located within the "leading leftmost peak". Together, these components reflect how well prioritized genes capture known therapeutic targets.

$$G2CT = log_{10}\frac{NES \times F}{FDR} \quad \text{Eq.16}$$

This formulation ensures that tissues with stronger enrichment signals, higher statistical confidence, and broader leading-edge representation receive higher G2CT potential scores, reflecting their greater relevance for therapeutic prioritization.

CREEDS (Validation 2)

We utilized disease-specific gene signatures v1.0 sourced from CREEDS[24] , a crowdsourced repository that curates and identifies gene signatures from the Gene Expression Omnibus (GEO). For our analysis, we specifically focused on human PCOS-associated GEO identifiers. Further, we used the GEO identifiers to access the metadata including tissue and population information. To check the enrichment of tissue specific differentially expressed genes, we performed Gene Set Enrichment Analysis (GSEA). Tissues considered for the analyses were adipose tissue and skeletal muscle, as these were the only relevant tissues available. The GEO datasets used for these tissues are GSE5090 (adipose tissue), GSE6798 (skeletal muscle), and GSE8157 (skeletal muscle).We performed 20,000 permutations to estimate the null distribution and assess the statistical significance.

Assessment of Obesity and T2D DEGs in ranked list

In addition to PCOS-associated signatures, we also extended our analysis to relevant metabolic disorders including Obesity and type 2 diabetes (T2D) with gene signatures obtained from GEO datasets. We utilized GEO2R from the GEO data repository to analyze available datasets from these conditions. For obesity, we processed adipose tissue datasets from EAS region (GSE217007 and GSE283367) and EUR region (GSE24883, GSE25401, GSE94752, GSE59034, GSE12050, GSE110729, GSE141432 and GSE166047). For T2D we analyzed GSE23343 of liver from EAS region, adipose tissue datasets (GSE141432 and GSE166047) and pancreas tissue datasets (GSE38642 and

GSE25724) from EUR region. $|\log_2 FC| > 0.5$ and P-value < 0.05 were applied as cutoff criteria to identify differentially expressed genes (DEGs). Subsequently, we performed gene set enrichment analysis with the tissue-specific prioritized gene list of adipose, pancreas and liver. We performed 20,000 permutations to estimate the null distribution and assess the statistical significance.

Validation with patient data

We used the GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r/) from GEO data repository[25] to analyze microarrays data (EUR: GSE98595) and RNA-seq (EAS: GSE155489, GSE138518, GSE106724) to identify differentially expressed genes (DEGs) in PCOS patient derived tissues with respect to and healthy control. Microarray data was processed with Limma package and RNA-seq was processed with DESeq2.| $\log_2 FC| > 0.5$ and P-value < 0.05 were used as the cut-off to obtain DEGs. DEGs were further categorized into upregulated (log2 FC > 0.5) and downregulated (log2FC < -0.5) genes. Subsequently, we filtered out the gene members of the following Reactome pathways (Immune: Innate immune system (R-HSA-168249), Cytokine signaling in immune system (R-HSA-1280215); Metabolic:Signaling by insulin receptor (R-HSA-74752), Translocation of SLC2A4 (GLUT4) to the plasma membrane (R-HSA-1445148); Hormone: Peptide hormone metabolism (R-HSA-2980736), Metabolism of steroids (R-HSA-8957322)) from the DEGs as well as from the prioritized matrix of ovary (containing the ranked leading-edge genes with Significance score greater than 0). Finally we checked for the enrichment of the filtered DEGs in the filtered priority Matrix by using Gene Set Enrichment Analysis.

Construction of Pathway-centric connectivity map (minimum spanning tree)

In addition to conventional gene-level network visualization, the identified crosstalk was also represented as a pathway-centric connectivity map. In this representation, each node corresponds to a significantly enriched pathway, and edges denote inferred connections between these pathways. Only pathways that were significantly overrepresented among the crosstalk genes (based on enrichment analysis) were retained as nodes. Initially, edges were introduced between pathways that shared genes. To refine the map, a minimum spanning tree was computed using the igraph package, and only the edges present in the resulting tree were retained. Edge thickness in the final visualization was proportional to the number of shared genes between the connected pathways, reflecting the strength of their functional association.

Repurposing of drugs

We utilized the ChEMBL database as the primary source of information on therapeutic agents, including details on drugs, their target genes, clinical development phases, and associated disease indications. ChEMBL curates data from authoritative sources such as ATC classification, ClinicalTrials.gov, DailyMed, and the U.S. FDA. For each disease indication under investigation, we retrieved drug and its targets, provided that the mechanism of action of the target gene is available. For

each target, we retrieved the drug that had achieved the highest clinical development phase. The corresponding disease indication for this highest-phase drug was also recorded. PCOS drug targets were removed from the list.

Description of Supplementary Data

Supplementary Data 1. Sources and datasets used for PCOS variant-to-gene mapping

Supplementary Data 2. Top prioritized list of target genes in each tissue along with pathway enrichment and concordance analysis across all tissues

Supplementary Data 3. Enrichment of Reactome pathways and HPO phenotypes among ovary-prioritized genes and their first neighbors

Supplementary Data 4. PCOS drug target-set enrichment scores and repurposing of other disease drug targets for PCOS

Supplementary Data 5. Pathway enrichment of tissue-specific crosstalk genes

Supplementary Data 6. Impact of single-node and combinatorial node removal in each tissue pathway crosstalk network

Supplementary Data 7. Network modules of crosstalk genes across all tissues and their module-specific pathway enrichment

Supplementary Data 8. SupraHex clustering of crosstalk genes and cluster-level pathway enrichment

Supplementary Data 9. Orthogonal clinical trial evidences for drug repurposing in PCOS

Supplementary References

1. Chen, Z. J. *et al.* Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat. Genet.* (2011) doi:10.1038/ng.732.
2. Shi, Y. *et al.* Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat. Genet.* (2012) doi:10.1038/ng.2384.
3. Lee, H. *et al.* Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome. *Hum. Reprod.* **30**, 723–731 (2015).
4. Kim, S. H., Liu, M., Jin, H. S. & Park, S. High Genetic Risk Scores of ASIC2, MACROD2, CHRM3, and C2orf83 Genetic Variants Associated with Polycystic Ovary Syndrome Impair Insulin Sensitivity and Interact with Energy Intake in Korean Women. *Gynecol. Obstet. Invest.* **84**, 225–236 (2019).
5. Yan, J. *et al.* A genome-wide association study identifies FSHR rs2300441 associated with follicle-stimulating hormone levels. *Clin. Genet.* **97**, 869–877 (2020).
6. Hayes, M. G. *et al.* Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat. Commun.* (2015)

doi:10.1038/ncomms8502.

7. Day, F. *et al.* Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* (2018) doi:10.1371/journal.pgen.1007813.
8. Day, F. R. *et al.* Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat. Commun.* (2015) doi:10.1038/ncomms9464.
9. Dapas, M. *et al.* Distinct subtypes of polycystic ovary syndrome with novel genetic associations: An unsupervised, phenotypic clustering analysis. *PLOS Med.* **17**, e1003132 (2020).
10. Tyrmi, J. S. *et al.* Leveraging Northern European population history: novel low-frequency variants for polycystic ovary syndrome. *Hum. Reprod.* **37**, 352–365 (2022).
11. Zhang, Y. *et al.* A genome-wide association study of polycystic ovary syndrome identified from electronic health records. *Am. J. Obstet. Gynecol.* **223**, 559.e1-559.e21 (2020).
12. Tian, Y. *et al.* Variants in FSHB Are Associated With Polycystic Ovary Syndrome and Luteinizing Hormone Level in Han Chinese Women. *J. Clin. Endocrinol. Metab.* **101**, 2178–2184 (2016).
13. Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btv402.
14. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).
15. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).
16. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun. 2018 91* **9**, 1–12 (2018).
17. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).
18. Subramanian, A., Tamayo, P. & Mootha, V. GSEA : Gene set enrichment analysis Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide expression profiles. *Pnsa* (2014).
19. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).
20. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* (2017) doi:10.1038/s41467-017-01261-5.
21. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set

Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

22. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nat. Genet. 2023 558* **55**, 1267–1276 (2023).
23. Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat. Genet. 2021 5311* **53**, 1527–1533 (2021).
24. Wang, Z. *et al.* Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* **7**, 1–11 (2016).
25. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).