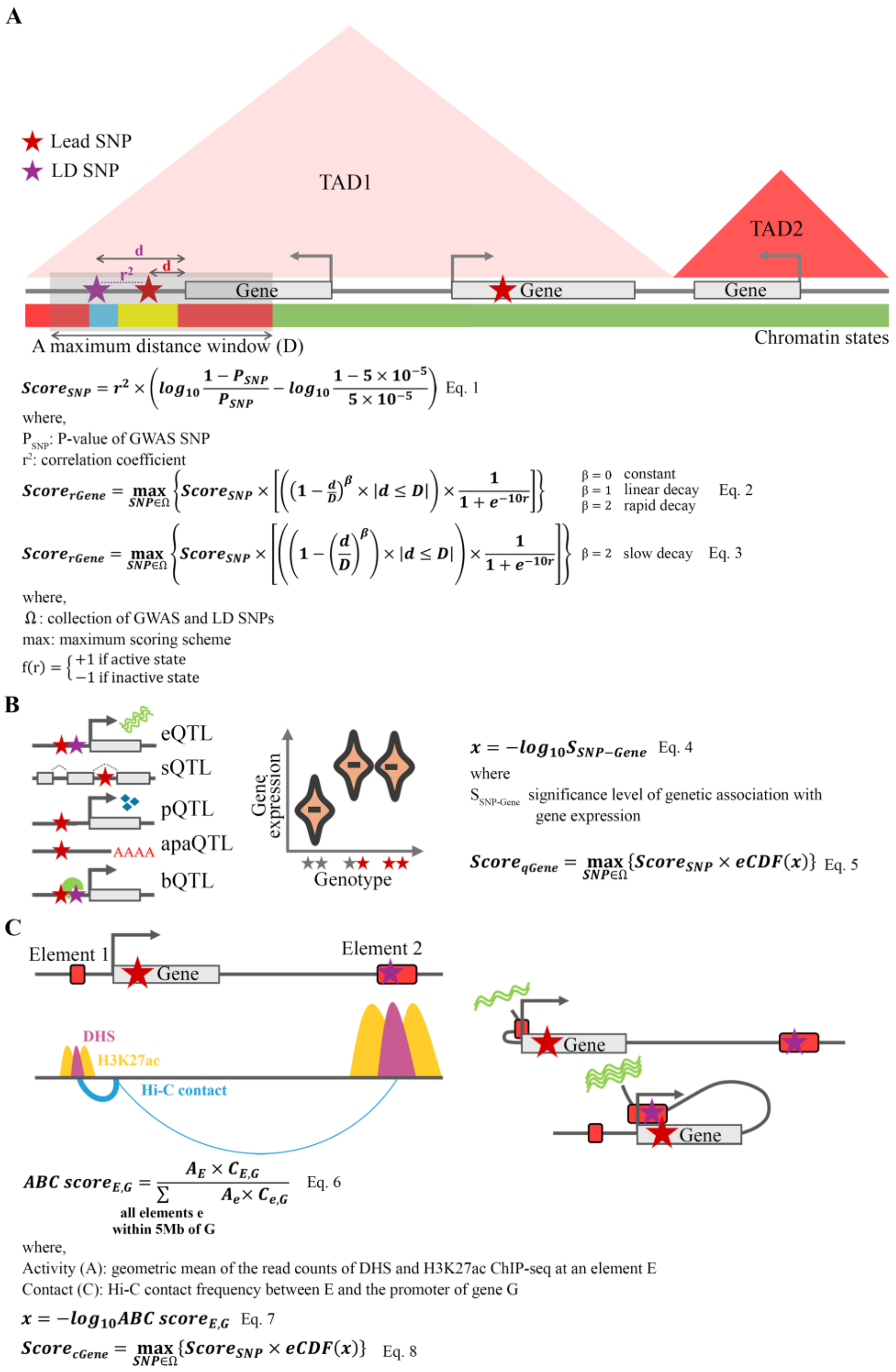
Evidence-based genetic variants to gene mapping and prioritization uncovers distinct molecular pathophysiology and therapeutic landscape in polycystic ovary syndrome patients of different ethnicity.

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



$$Score_{SNP} = r^2 \times \left(log_{10} \frac{1 - P_{SNP}}{P_{SNP}} - log_{10} \frac{1 - 5 \times 10^{-5}}{5 \times 10^{-5}} \right) \quad \text{Eq. 1}$$

where,
P_{SNP}: P-value of GWAS SNP
r^2: correlation coefficient

$$Score_{rGene} = \max_{SNP \in \Omega} \left\{ Score_{SNP} \times \left[\left((1 - \tfrac{d}{D})^{\beta} \times |d \leq D| \right) \times \frac{1}{1 + e^{-10r}} \right] \right\} \quad \begin{matrix} \beta = 0 & \text{constant} \\ \beta = 1 & \text{linear decay} \\ \beta = 2 & \text{rapid decay} \end{matrix} \quad \text{Eq. 2}$$

$$Score_{rGene} = \max_{SNP \in \Omega} \left\{ Score_{SNP} \times \left[\left(\left(1 - \left(\frac{d}{D} \right)^{\beta} \right) \times |d \leq D| \right) \times \frac{1}{1 + e^{-10r}} \right] \right\} \quad \beta = 2 \quad \text{slow decay} \quad \text{Eq. 3}$$

where,
Ω: collection of GWAS and LD SNPs
max: maximum scoring scheme
$f(r) = \begin{cases} +1 \text{ if active state} \\ -1 \text{ if inactive state} \end{cases}$

$$x = -log_{10} S_{SNP-Gene} \quad \text{Eq. 4}$$

where
$S_{SNP\text{-}Gene}$ significance level of genetic association with gene expression

$$Score_{qGene} = \max_{SNP \in \Omega} \{ Score_{SNP} \times eCDF(x) \} \quad \text{Eq. 5}$$

$$ABC\ score_{E,G} = \frac{A_E \times C_{E,G}}{\sum_{\substack{\text{all elements e} \\ \text{within 5Mb of G}}} A_e \times C_{e,G}} \quad \text{Eq. 6}$$

where,
Activity (A): geometric mean of the read counts of DHS and H3K27ac ChIP-seq at an element E
Contact (C): Hi-C contact frequency between E and the promoter of gene G

$$x = -log_{10} ABC\ score_{E,G} \quad \text{Eq. 7}$$

$$Score_{cGene} = \max_{SNP \in \Omega} \{ Score_{SNP} \times eCDF(x) \} \quad \text{Eq. 8}$$

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 though empirical cumulative distribution function (eCDF) on the interaction score (Eqn 6, 7 and 8). Details of process is further described in supplementary note.

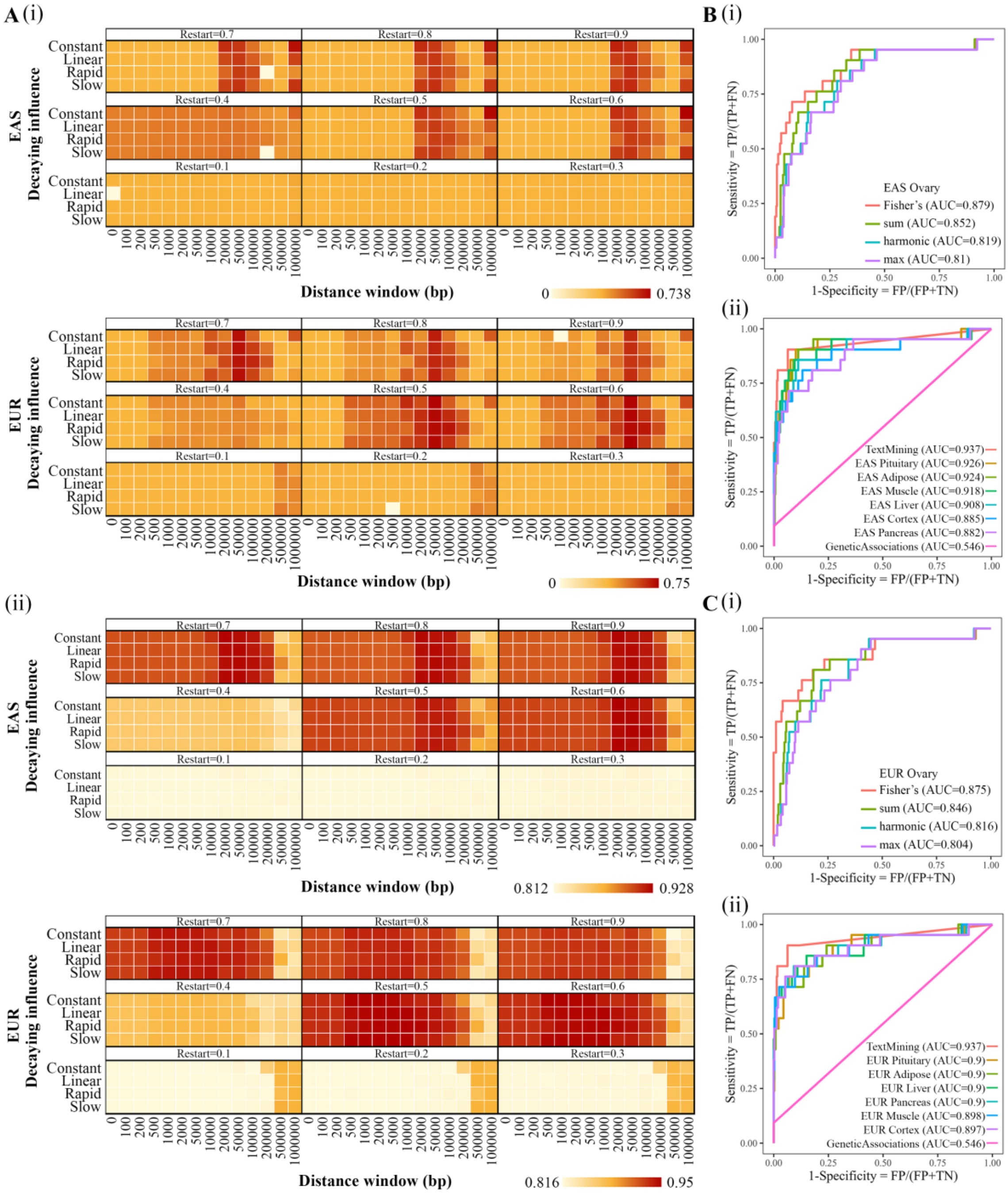


Figure S2: Simultaneous optimization of influential distance for rGene and protein interactome exploration parameters for random walk. A. Heatmaps displaying the area under the receiver operating characteristic curve (AUC) for gene prioritization performance under varying parameter settings as indicated 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 (GSN) from evaluation. (ii) AUC values when both first- and second neighbors of GSPs (GSN) 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 (Text Mining and Genetic Association) in EAS **(B (ii)),** and EUR **(C (ii))** .

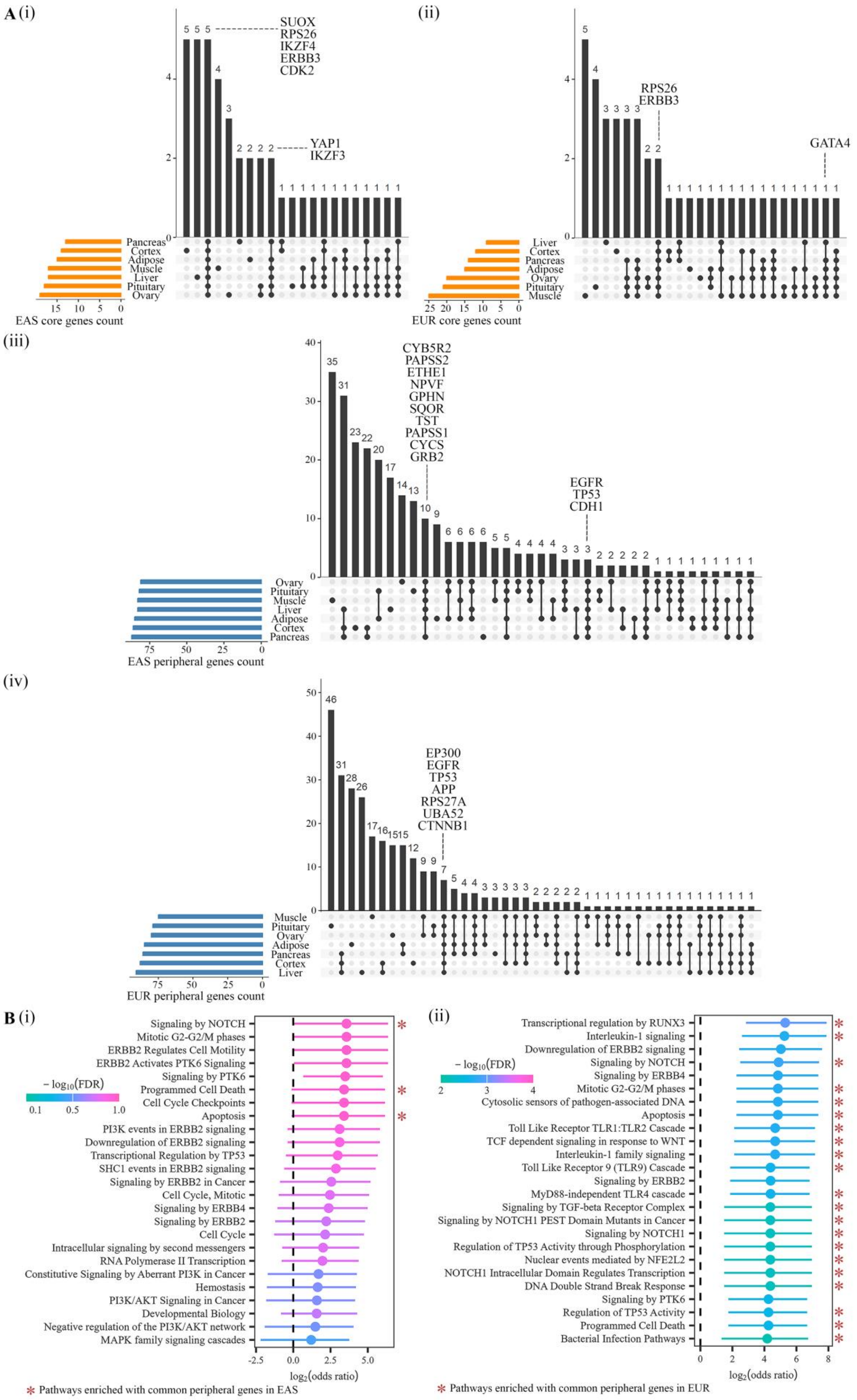

A (i)
5 5 5
SUOX
RPS26
IKZF4
ERBB3
CDK2
4
3
2 2 2 2
YAP1
IKZF3
1 1 1 1 1 1 1 1 1 1 1 1
Pancreas
Cortex
Adipose
Muscle
Liver
Pituitary
Ovary
15 10 5 0
EAS core genes count
(ii)
5
4
3 3 3 3
RPS26
ERBB3
2 2
GATA4
Liver
Cortex
Pancreas
Adipose
Ovary
Pituitary
Muscle
25 20 15 10 5 0
EUR core genes count
(iii)
35 31 23 22 20 17 14 13 10 9 6 6 6 6 5 5 4 4 4 4 3 3 3 2 2 2 2 2 1 1 1 1 1 1 1 1 1
CYB5R2
PAPSS2
ETHE1
NPVF
GPHN
SQOR
TST
PAPSS1
CYCS
GRB2
EGFR
TP53
CDH1
Ovary
Pituitary
Muscle
Liver
Adipose
Cortex
Pancreas
75 50 25 0
EAS peripheral genes count
(iv)
46 31 28 26 17 16 15 15 12 9 9 7 5 4 4 3 3 3 3 3 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
EP300
EGFR
TP53
APP
RPS27A
UBA52
CTNNB1
Muscle
Pituitary
Ovary
Adipose
Pancreas
Cortex
Liver
75 50 25 0
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)
0.1 0.5 1.0
-2.5 0.0 2.5 5.0
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)
2 3 4
0 2 4 6 8
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.

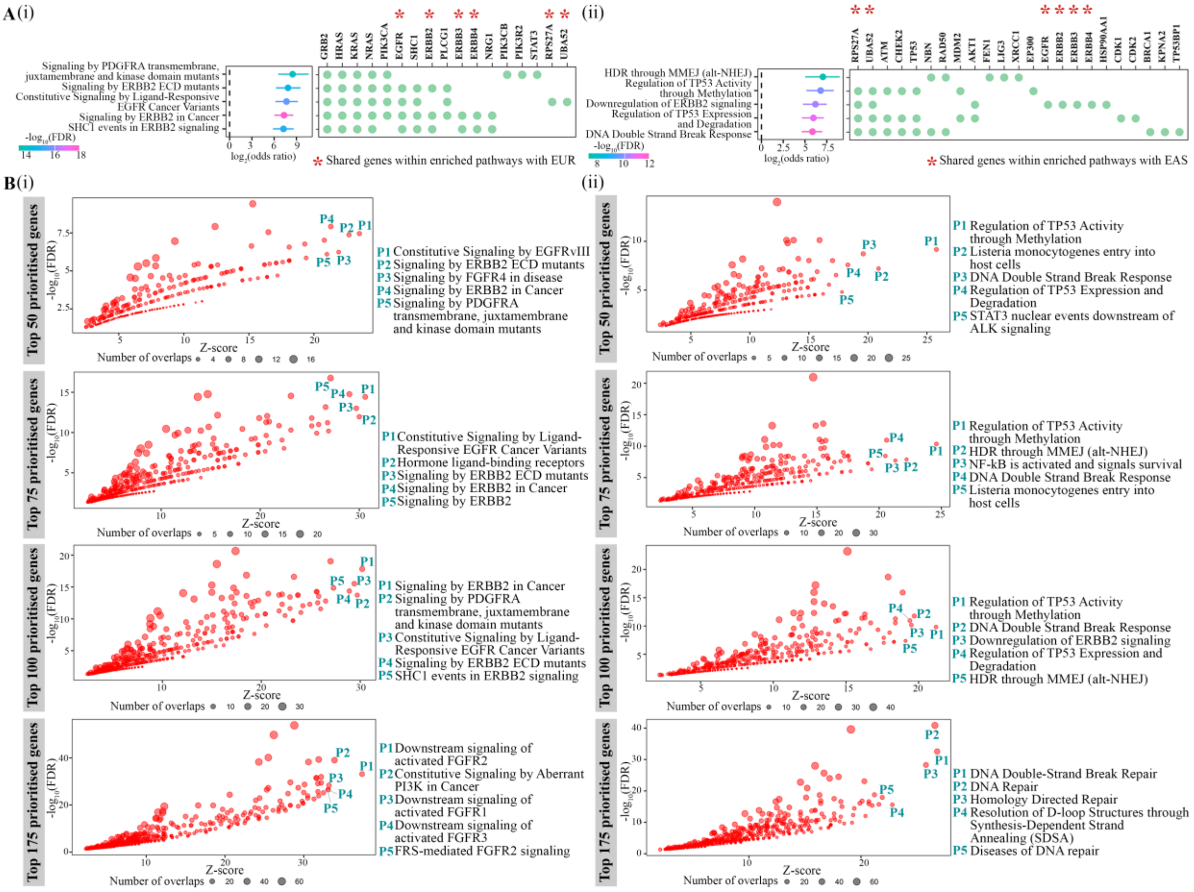


Figure S4: Comparison and consistency analysis of enriched pathways. Prioritized pathways are derived from the top 100 prioritized genes in ovary tissue for both the populations. **(A)** Forest plot (Left) depicting the top 5 significantly enriched Reactome pathways in (i) EAS Ovary, (ii) EUR Ovary. 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.

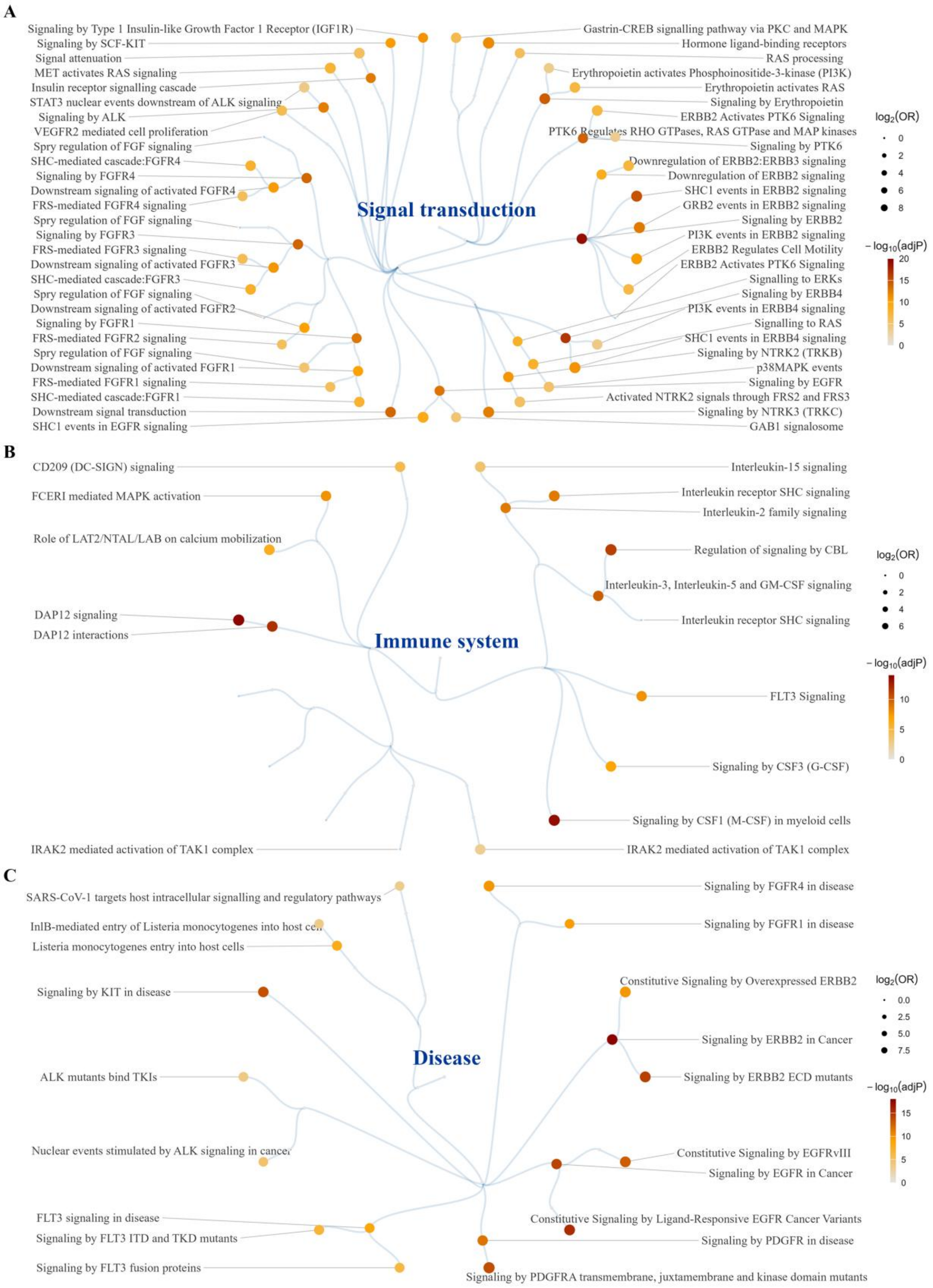


Figure S5: Grouping of enriched Reactome pathway in to hierarchical subgraphs in EAS population: Circular overview of pathway enrichment among the top 100 prioritized genes in ovary tissue for EAS population, grouped by pathway hierarchy into **(A)** *Signal Transduction* **(B)** *Immune*

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).

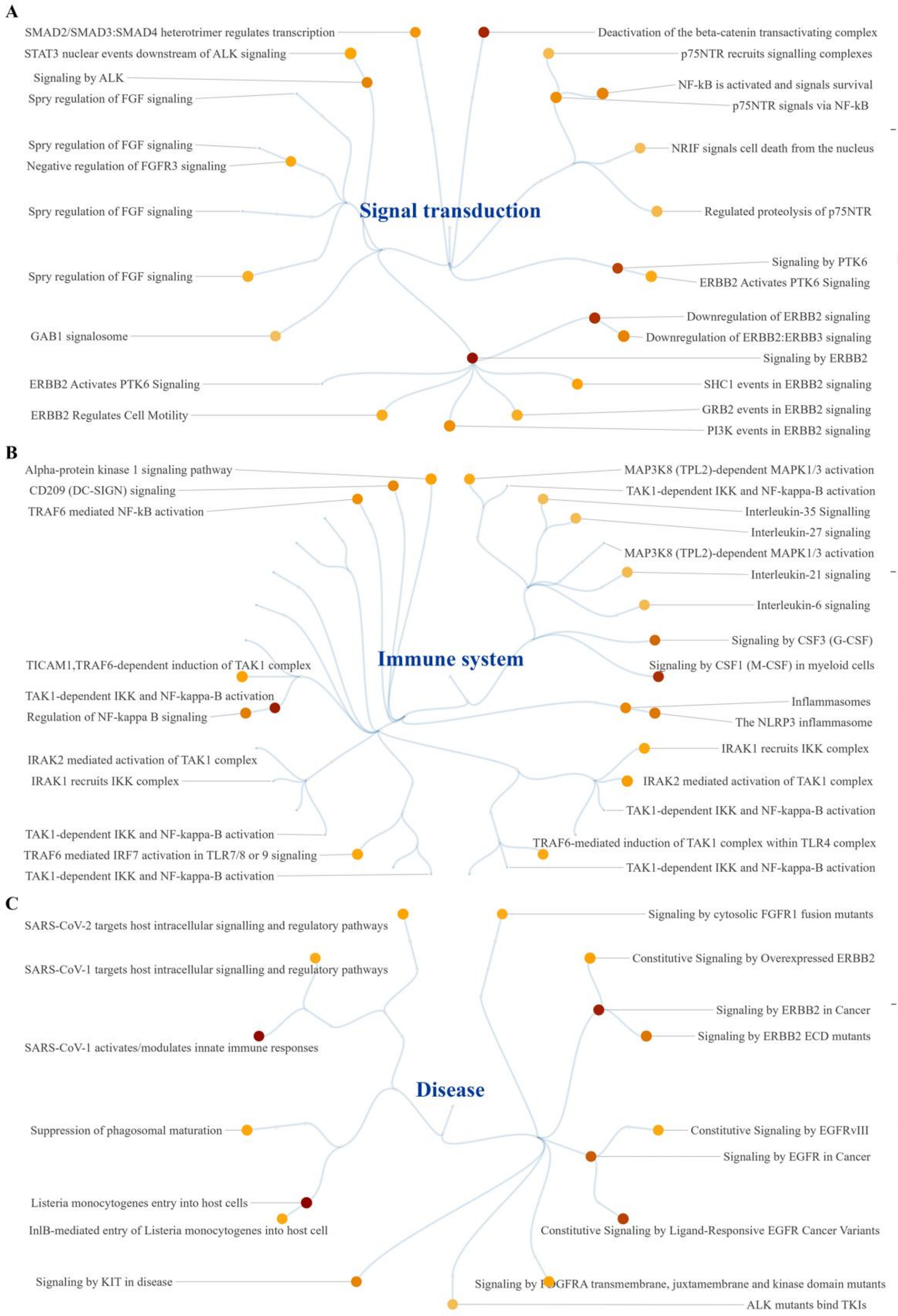


Figure S6: Grouping of enriched Reactome pathway in to hierarchical subgraphs in EUR population: Circular overview of pathway enrichment among the top 100 prioritized genes in ovary tissue for EUR population, grouped by pathway hierarchy into **(A)** *Signal Transduction* **(B)** *Immune*

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).

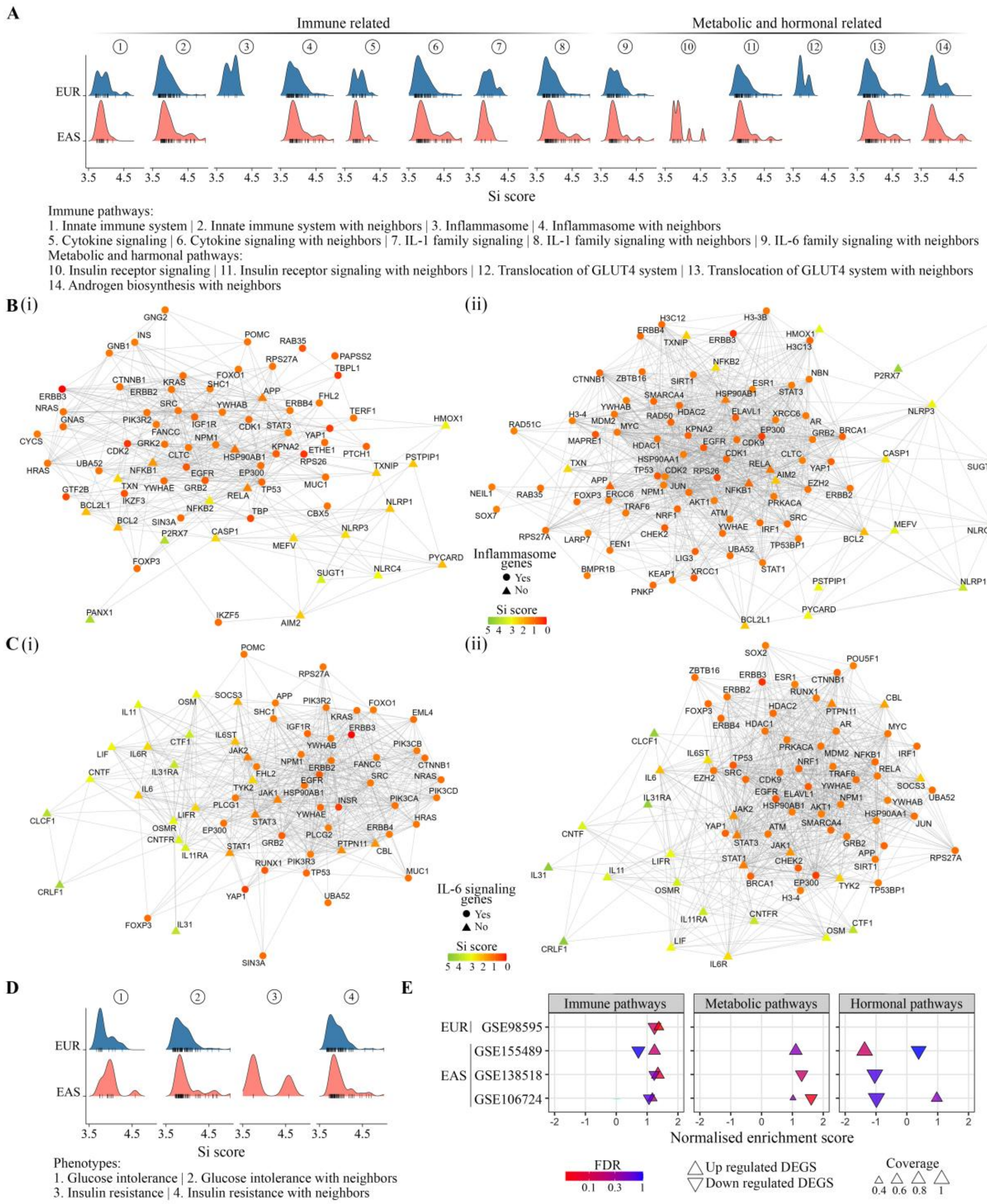


Figure S7: Genetic links of PCOS pathology with the immune, metabolic and androgen related Pathways: (A) Ridge plots displaying the density distribution and Si rating for the top 100 prioritized genes of ovary and their direct neighbors belonging to important functional pathways (immune, metabolic and 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 Pathway 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 20000 permutations.

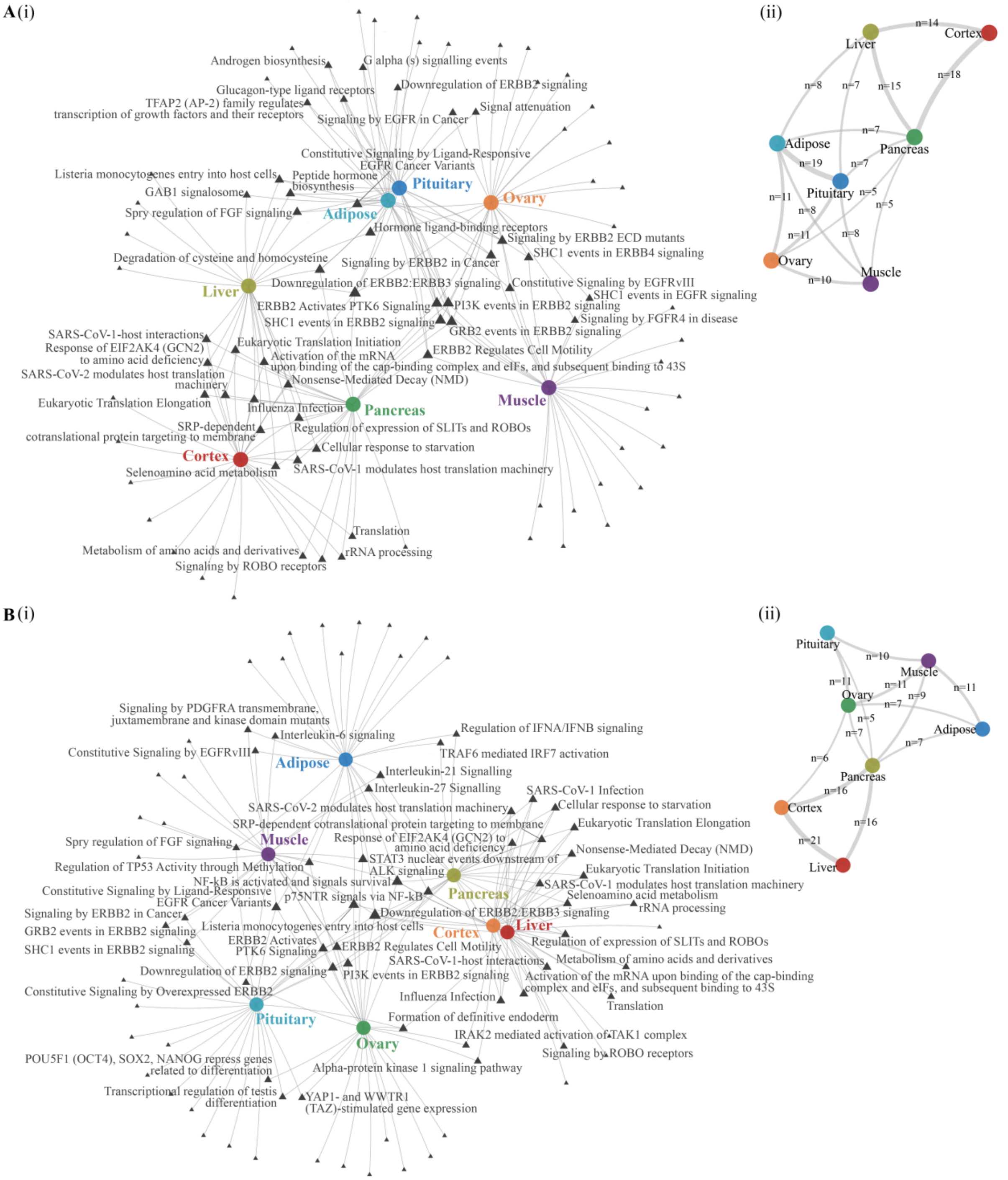

A(i)
Androgen biosynthesis
G alpha (s) signalling events
Glucagon-type ligand receptors
Downregulation of ERBB2 signaling
TFAP2 (AP-2) family regulates transcription of growth factors and their receptors
Signaling by EGFR in Cancer
Signal attenuation
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
Listeria monocytogenes entry into host cells
Peptide hormone biosynthesis
Pituitary
GAB1 signalosome
Ovary
Adipose
Spry regulation of FGF signaling
Hormone ligand-binding receptors
Signaling by ERBB2 ECD mutants
SHC1 events in ERBB4 signaling
Degradation of cysteine and homocysteine
Signaling by ERBB2 in Cancer
Downregulation of ERBB2:ERBB3 signaling
Constitutive Signaling by EGFRvIII
Liver
SHC1 events in EGFR signaling
ERBB2 Activates PTK6 Signaling
PI3K events in ERBB2 signaling
SHC1 events in ERBB2 signaling
Signaling by FGFR4 in disease
GRB2 events in ERBB2 signaling
SARS-CoV-1-host interactions
Eukaryotic Translation Initiation
Response of EIF2AK4 (GCN2) to amino acid deficiency
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S
ERBB2 Regulates Cell Motility
SARS-CoV-2 modulates host translation machinery
Nonsense-Mediated Decay (NMD)
Eukaryotic Translation Elongation
Influenza Infection
Pancreas
Muscle
SRP-dependent cotranslational protein targeting to membrane
Regulation of expression of SLITs and ROBOs
Cellular response to starvation
Cortex
Selenoamino acid metabolism
SARS-CoV-1 modulates host translation machinery
Translation
Metabolism of amino acids and derivatives
rRNA processing
Signaling by ROBO receptors
(ii)
Liver
n=14
Cortex
n=18
n=8
n=7
n=15
n=7
Adipose
Pancreas
n=19
n=7
Pituitary
n=5
n=11
n=5
n=8
n=11
n=8
Ovary
Muscle
n=10
B(i)
Signaling by PDGFRA transmembrane, juxtamembrane and kinase domain mutants
Interleukin-6 signaling
Regulation of IFNA/IFNB signaling
Constitutive Signaling by EGFRvIII
Adipose
TRAF6 mediated IRF7 activation
Interleukin-21 Signalling
Interleukin-27 Signalling
SARS-CoV-1 Infection
SARS-CoV-2 modulates host translation machinery
Cellular response to starvation
SRP-dependent cotranslational protein targeting to membrane
Eukaryotic Translation Elongation
Spry regulation of FGF signaling
Muscle
Response of EIF2AK4 (GCN2) to amino acid deficiency
Nonsense-Mediated Decay (NMD)
STAT3 nuclear events downstream of ALK signaling
Regulation of TP53 Activity through Methylation
Eukaryotic Translation Initiation
NF-kB is activated and signals survival
SARS-CoV-1 modulates host translation machinery
Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants
Pancreas
p75NTR signals via NF-kB
Selenoamino acid metabolism
Signaling by ERBB2 in Cancer
rRNA processing
Downregulation of ERBB2:ERBB3 signaling
GRB2 events in ERBB2 signaling
Listeria monocytogenes entry into host cells
Cortex
Liver
SHC1 events in ERBB2 signaling
ERBB2 Activates PTK6 Signaling
ERBB2 Regulates Cell Motility
Regulation of expression of SLITs and ROBOs
SARS-CoV-1-host interactions
Metabolism of amino acids and derivatives
Downregulation of ERBB2 signaling
PI3K events in ERBB2 signaling
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S
Constitutive Signaling by Overexpressed ERBB2
Influenza Infection
Translation
Pituitary
Formation of definitive endoderm
Ovary
IRAK2 mediated activation of TAK1 complex
Signaling by ROBO receptors
POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation
Alpha-protein kinase 1 signaling pathway
Transcriptional regulation of testis differentiation
YAP1- and WWTR1 (TAZ)-stimulated gene expression
(ii)
Pituitary
n=10
Muscle
n=11
n=11
n=11
Ovary
n=9
n=7
n=5
Adipose
n=7
n=7
n=6
Pancreas
n=16
Cortex
n=16
n=21
Liver

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 (ii) B** 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.

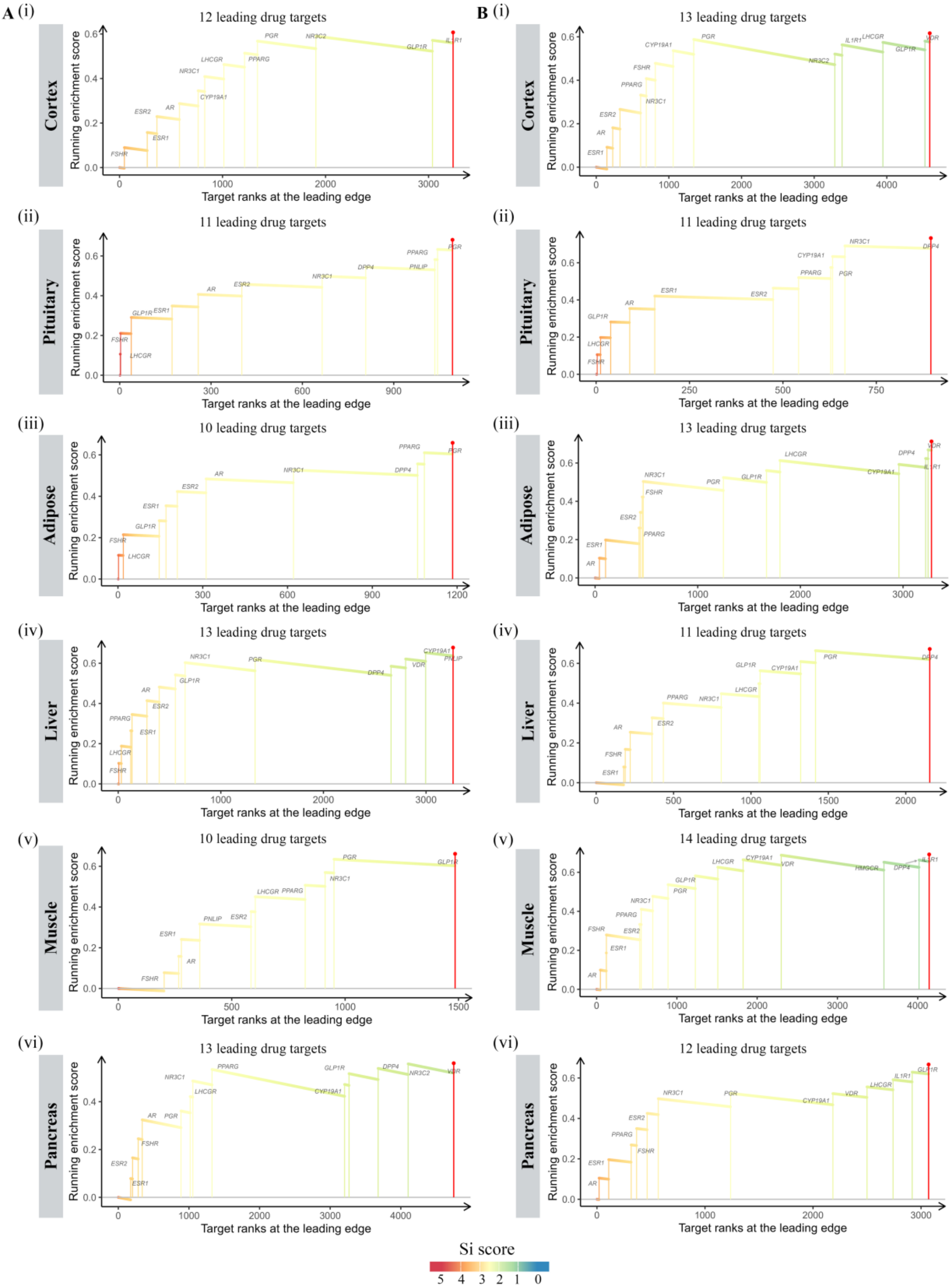


Figure S9: TSEA of PCOS drug targets in other disease relevant tissues: Each panel displays the leading edge of target set enrichment analysis in various tissues (i-vi) within EAS (A) and EUR (B) population. Known PCOS drug targets (Phase 2 and above) recovered within the leading edge of gene priority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

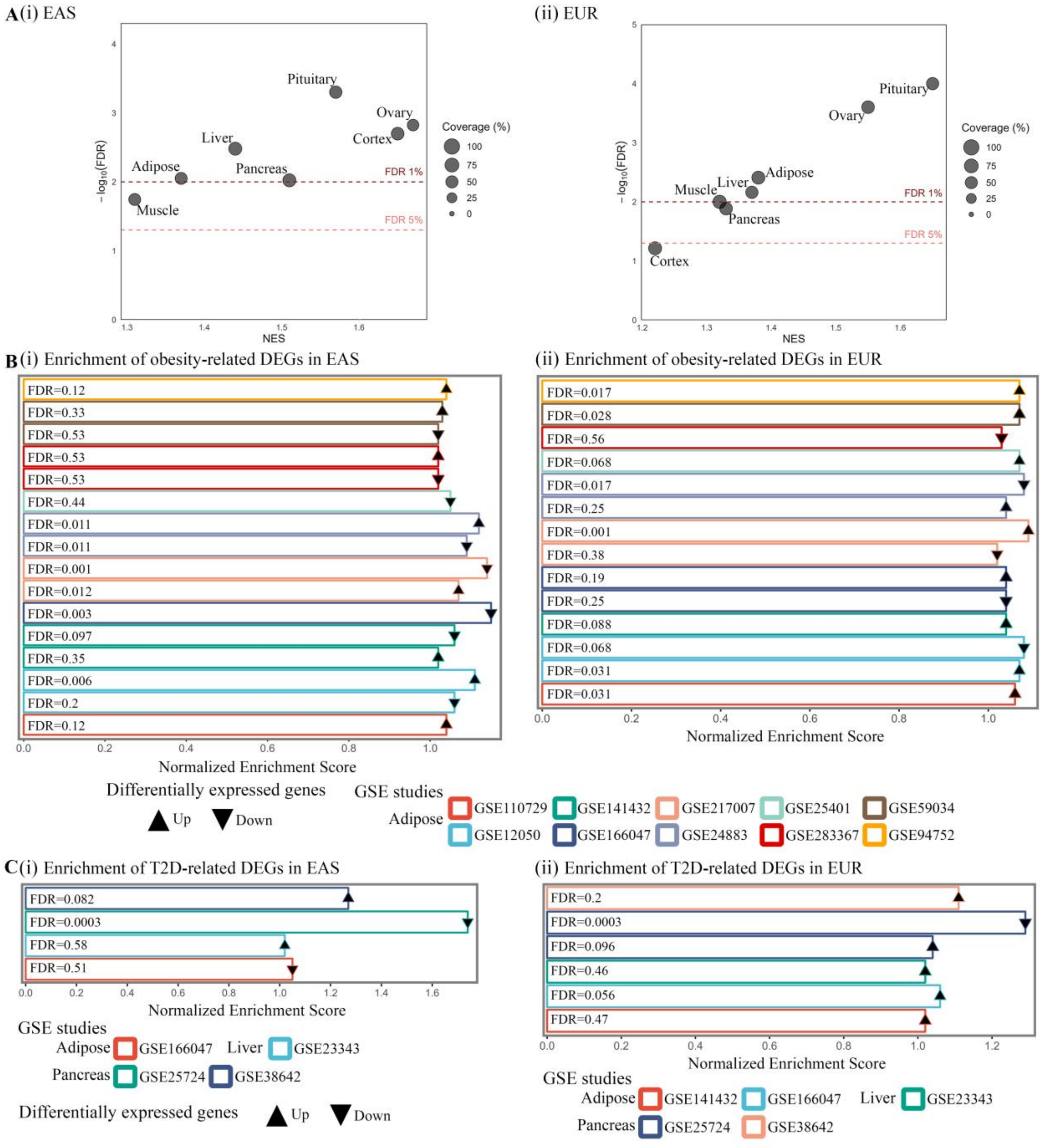


Figure S10: Therapeutic potential of know PCOS drug targets across tissue (A) Scatterplots summarizing TSEA-based therapeutic potential across all tissues for each ancestry group. Normalized Enrichment Score (NES), target coverage (that is, the total number within the leading edge for that tissue/total 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.

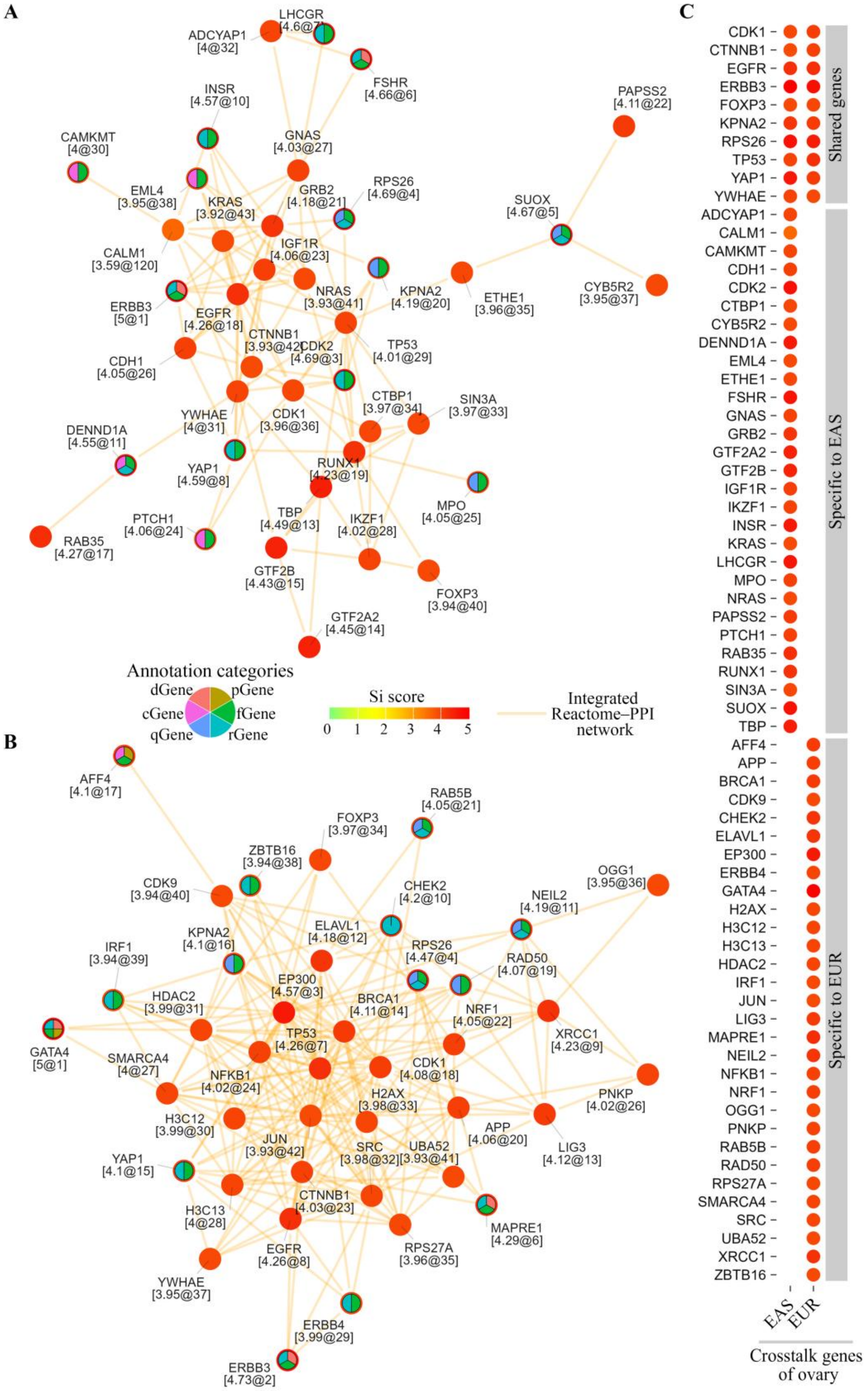


Figure S11: **Pathway crosstalk analysis:** Illustration of pathway cross-talk network of prioritized genes from ovary tissue in (A) EAS, (B) EUR. Each node represents a gene labeled with its symbol and Si score (formatted as “rating @ rank”), colored by magnitude. Edges indicate protein–protein interactions (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 cross talk in both the population.

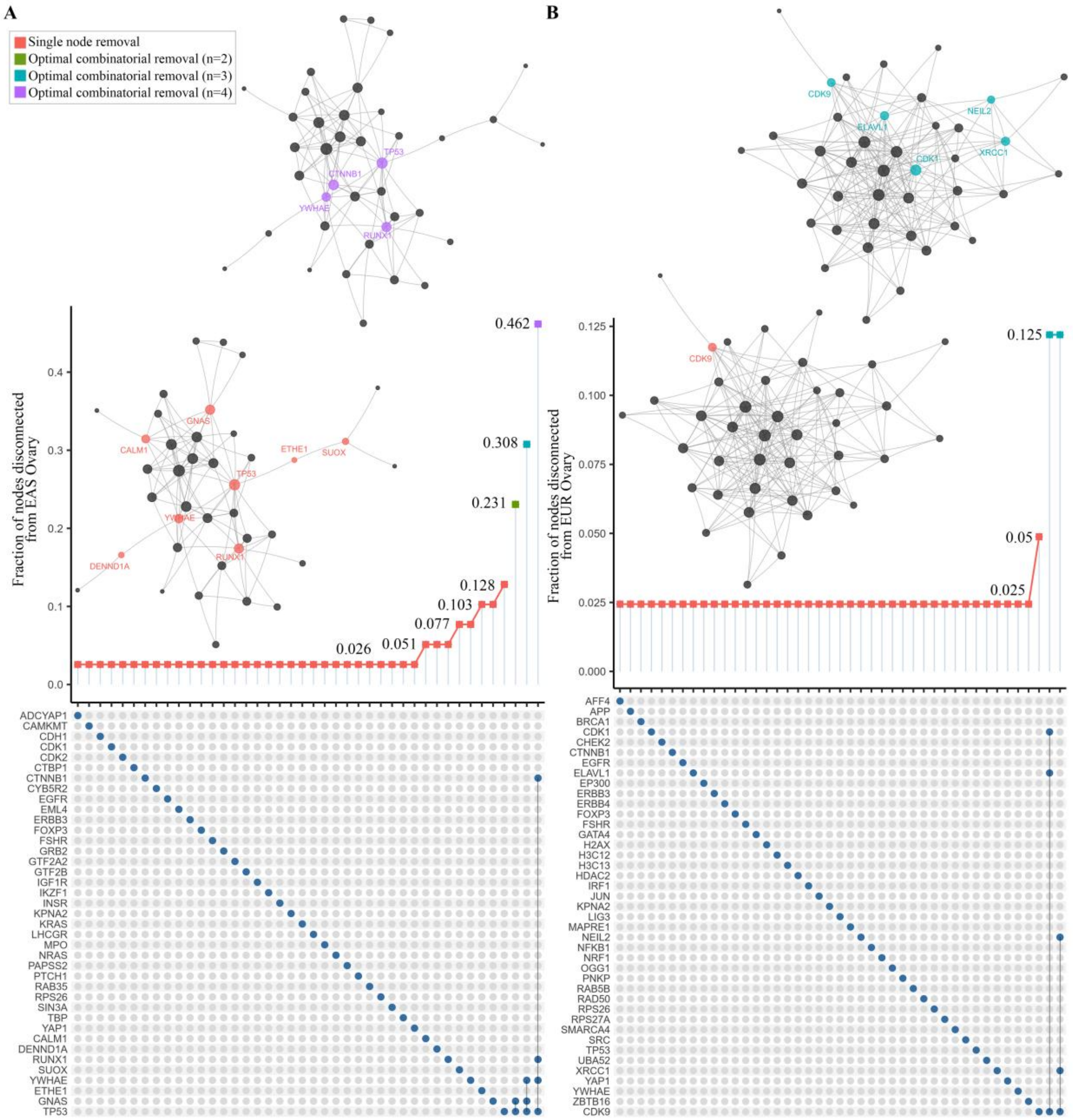


Figure S12: **Node Removal Analysis of pathway cross-talk in ovary:** Effect of single or combinatorial node removal on network robustness was assessed by determining the fraction of nodes disconnected from the largest connected component in EAS (A) and EUR population (B). The y-axis represents the fraction 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 cross-talk networks (only the single node and 4 node combinatorial removal analysis on the network are shown).

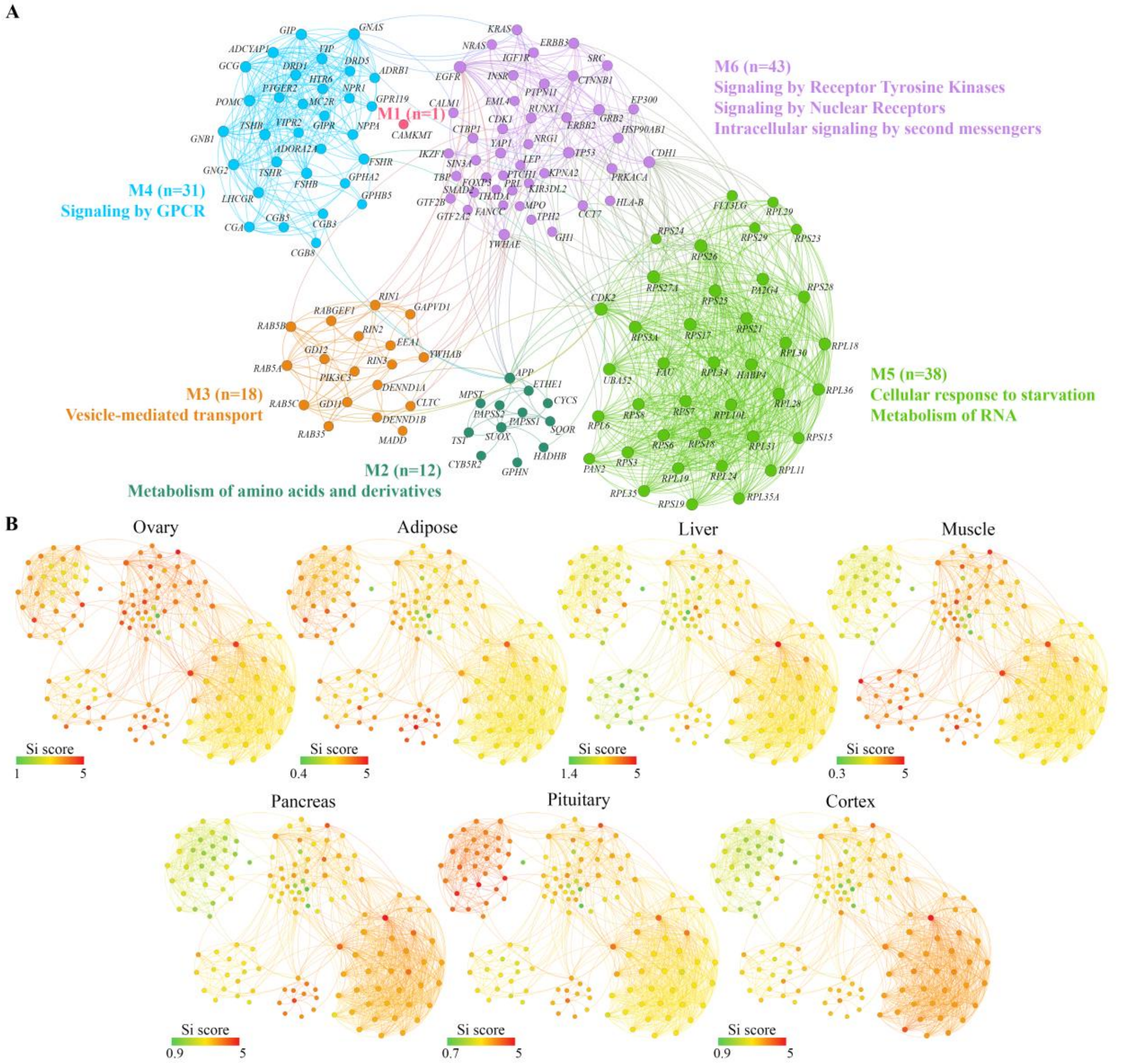


Figure S13: Modular Analysis of cross-tissue pathway crosstalk in EAS population: (A) Community detection analysis by measuring modularity of merged pathway crosstalk networks across tissues in EAS population revealed six modules, each annotated by functional enrichment using one-sided Fisher's exact 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.

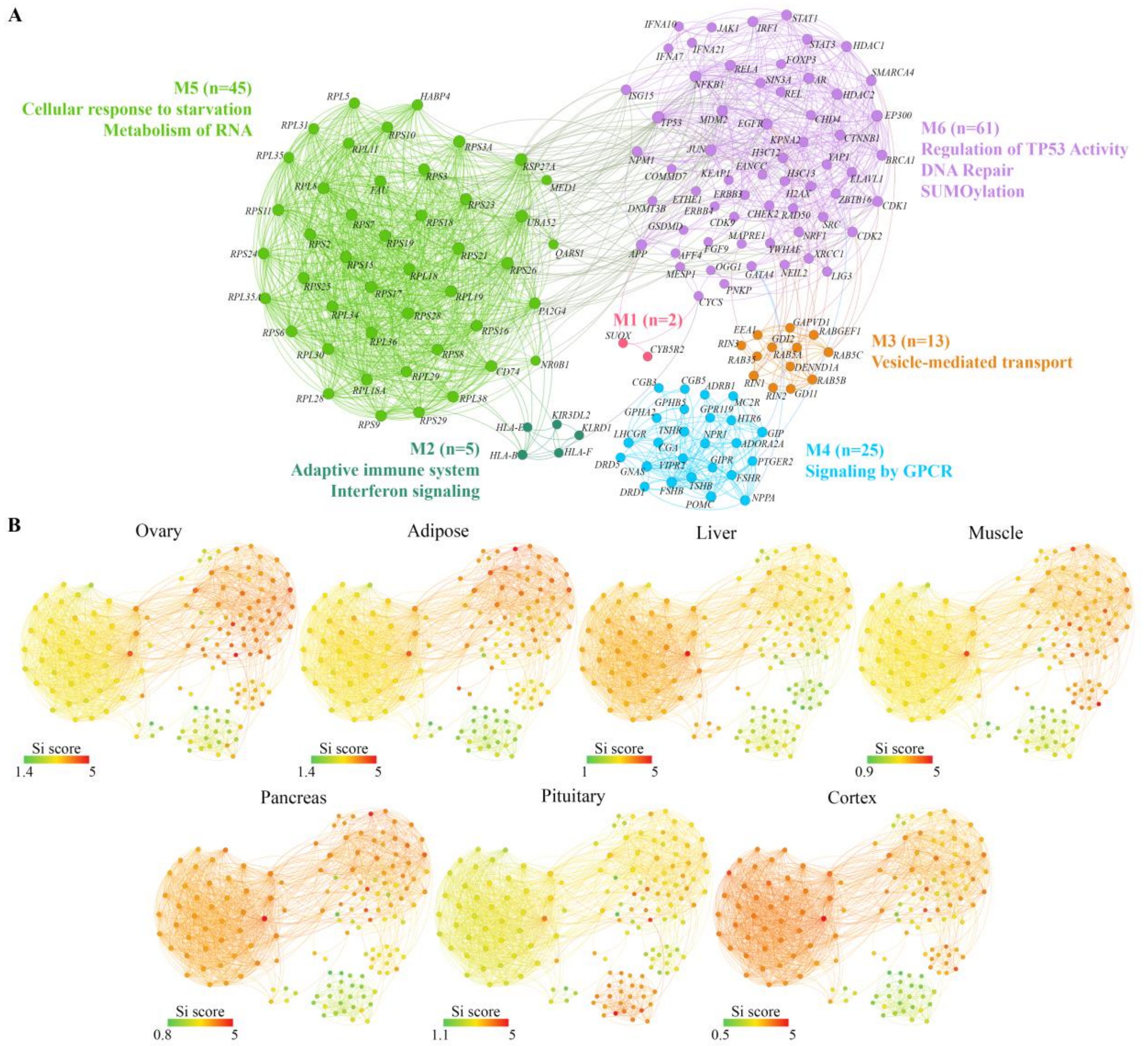


Figure S14: Modular Analysis of cross-tissue pathway crosstalk in EUR population: (A) Community detection analysis by measuring the modularity of merged pathway crosstalk networks across tissues in EUR population revealed six modules, each annotated by functional enrichment using one-sided Fisher's exact 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.

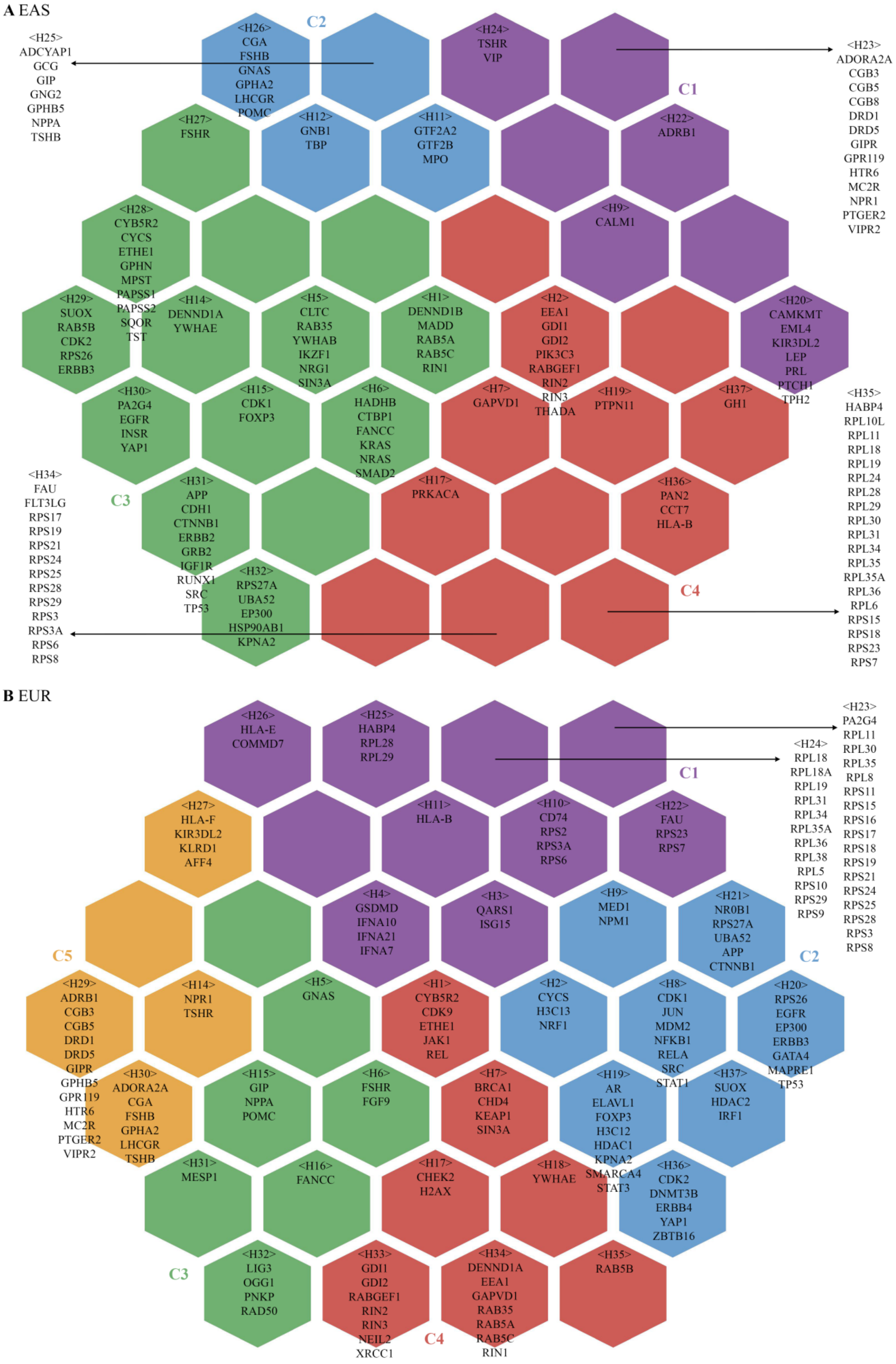

A EAS
<H25>
ADCYAP1
GCG
GIP
GNG2
GPHB5
NPPA
TSHB
<H26>
CGA
FSHB
GNAS
GPHA2
LHCGR
POMC
C2
<H24>
TSHR
VIP
<H23>
ADORA2A
CGB3
CGB5
CGB8
DRD1
DRD5
GIPR
GPR119
HTR6
MC2R
NPR1
PTGER2
VIPR2
C1
<H27>
FSHR
<H12>
GNB1
TBP
<H11>
GTF2A2
GTF2B
MPO
<H22>
ADRB1
<H28>
CYB5R2
CYCS
ETHE1
GPHN
MPST
PAPSS1
PAPSS2
SQOR
TST
<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.

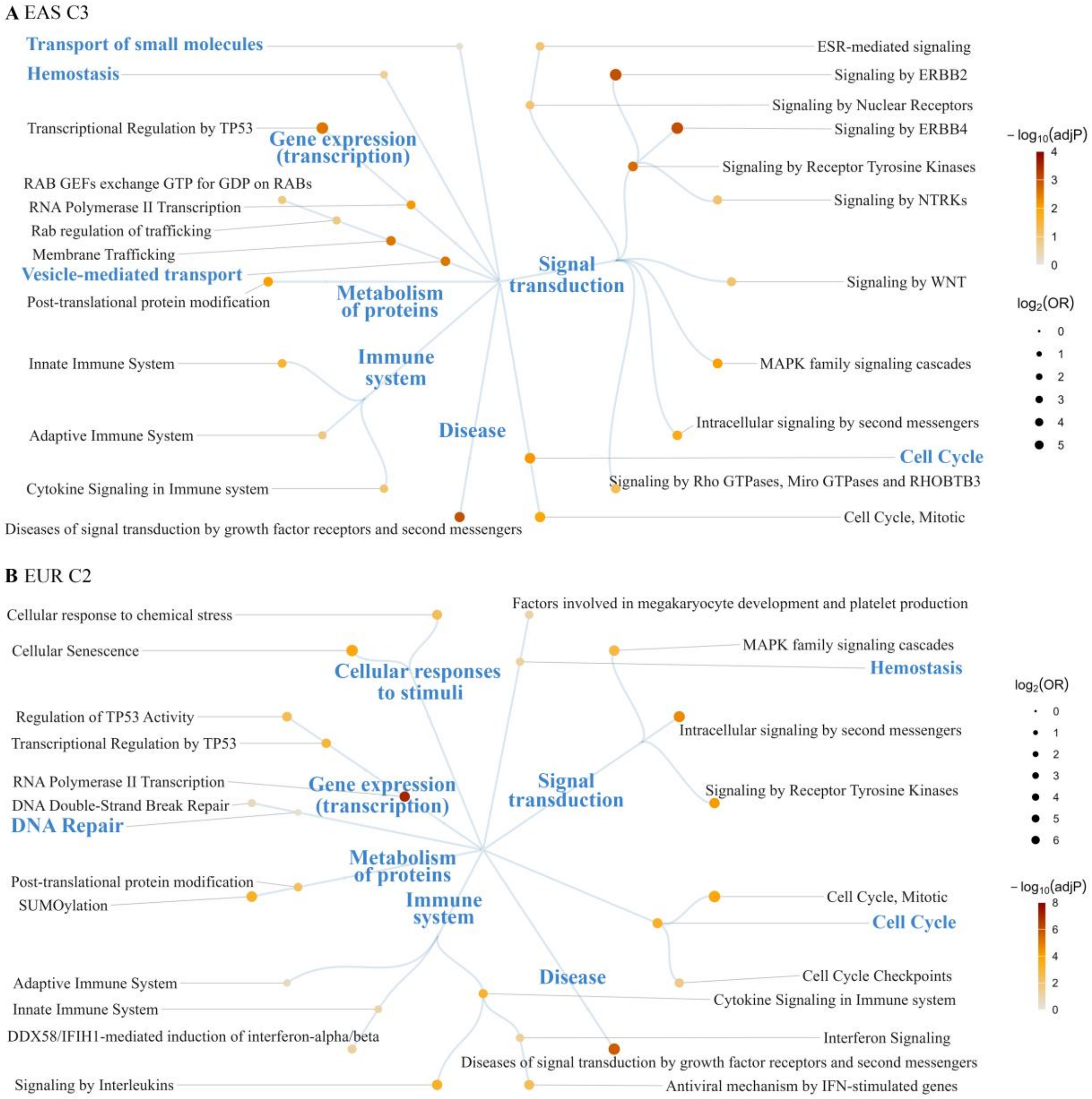


Figure S16: Pathway enrichments of target genes in the high scored and druggable clusters in populations: Reactome pathway enrichment analysis of high scored and druggable cluster; cluster 3 in EAS **(A)** and cluster 2 EUR population **(B)** are presented. Enrichment was performed using one-sided Fisher's 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 with 107 and 53 SNPs in EAS and in EUR respectively (Table S1, 1st Tab). 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 2013 and 1873 SNPs for EAS and EUR populations respectively (Table S1, 1st Tab). 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 $=5 \times 10^{-5}$. Codes used in entire analysis were obtained from publicly available repositories, including https://github.com/r-forge/pi314/tree/master/pkg and https://github.com/hfang-bristol.

Identification of core genes under genomic influence

To systematically identify core genes involved in PCOS pathophysiology, we integrated scored SNPs with functional genomic regulatory data (see below). A comprehensive scoring framework was adapted from a previous study[14] and used to capture and evaluate the regulatory impact of SNPs across relevant tissues both linear genomic influence and influence on higher-order chromatin architecture. Briefly, the scoring was primarily based upon three types of genomic evidence: (i) genomic proximity with regulatory implication (rGene): determined by the distance between SNP-to-gene within an influential range window combined with the regulatory information from ENCODE and REP[15] (detailed below); (ii) Quantitative Trait Loci (QTLs), obtained from the quantitative associations of SNPs with different genetic/molecular traits like protein abundance (pQTLs), gene expression (eQTLs) etc. (detailed list of QTLs studied are presented in Table S1, 3rd Tab). Collectively they are termed as qGenes (iii) cGenes obtained from the evidence measured from the strength of Enhancer regions containing SNPs, interacting physically with gene promoters using ABC scoring method[16]. The details of this identification are accounted below.

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[17].

- ***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 (Table S1, 2nd Tab). 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*[18] and the STRING database (v12.0)[19] . 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.

rGene identification:

Quantification of genomic proximity evidence: To evaluate the regulatory potential of PCOS-associated variants on nearby genes, we developed a scoring framework based on chromatin states, topological architecture, and linear genomic distance. For the EAS dataset, we applied a constant decay model with a distance threshold of 20,000 for fetching core genes and a restart probability of 0.7 to capture peripheral genes within the network. Similarly, for the EUR dataset, we employed a linear decay model with a distance threshold of 10,000 for identifying core genes and a restart probability of 0.6 for retrieving peripheral genes (Figure S2A). All these parameters were optimized from the previous step. Tissue specific chromatin states were sourced form Roadmap Epigenomics Project using ChromHMM[15] and TAD (topologically associating domain) data is sourced from Schmitt, A. D. *et al* [20] . List of disease relevant tissues and the corresponding datasets are presented in Table S1, 3rd Tab. All the chromatin states were categorized into two functional types—active, and inactive (repressed and quiescent)—and assigned weights reflecting transcriptional activity. +1 weights were assigned to active, -1 to repressed and quiescent states. Active states includes the epigenetic signature associated to: Active TSS, Flanking Active TSS, Transcribed state observed at gene 5' and 3', Strong transcription, Weak transcription, Genic enhancers, Enhancers and ZNF genes & repeats. Inactive states includes epigenetic signature associated to: Heterochromatin, Bivalent/Poised TSS, Flanking Bivalent TSS/Enh, Bivalent Enhancer, Repressed PolyComb, Weak Repressed PolyComb and Quiescent/Low. A sigmoid function (Eqn 2 and 3, Figure S1A) was applied to integrate these weights into a proximity-based scoring. For all the tissues except pituitary, for which epigenetic data is not available, we used equation 2 and 3. Whereas, for pituitary the equations used were the same as those used in Fang, H. *et al*[14]. rGene score ranging from 0 to 1 were computed for each gene. A higher score indicates a greater likelihood that the gene is functionally influenced by a nearby PCOS-associated SNP. An additional filtering was done constraining only the SNP-gene pair which were present within the same topologically associating domain (TAD) boundary. Adipose and pituitary were excluded from this exercise due to unavailability of data.

qGene Annotation:

Quantification of QTL evidence: To assess associations of PCOS-associated variants with various quantitative genomic traits, we collected the significance score from the statistically significant, population and tissue specific quantitative trait loci (QTL) data from GTEx[21] and QTLbase[22] without disease or drug treatment) (Table S1, 3rd Tab). qGene Scoring: SNP-gene associations were normalized within each population group for a given tissue using an empirical cumulative distribution function (eCDF, Eqn 4, Figure S1B). This transformation normalized the significance scores to a 0–1 range (Eqn 5, Figure S1B).

cGene Annotation:

To enlist the PCOS variants overlapping with an Enhancer element and can potentially interact with gene promoters and thus modulate the regulation of genes, we collected the strength of interactions from tissue specific Hi-C datasets (Table S1, 3rd Tab) which forms the basis of our scoring. The entire method is adapted from the Activity-by-Contact (ABC) scoring schema using default parameters[16] . ABC model uses chromatin accessibility (ATAC-seq or DNase-seq), histone modifications (H3K27ac ChIP–seq to predict enhancer–gene connections for a tissue and interaction frequency for the contact. We have used ABC scores for each element–gene pair, where the enhancer element is constrained within 5 Mb from the TSS of a gene. Candidate enhancers were defined from ENCODE DNase-seq data and H3K27ac using MACS2 peak calling with a P-value cut-off of 0.1[16]. The top 150,000 peaks (by read count) were resized to 500 bp around summits. Only genes with measurable expression from polyA plus RNA-seq from ENCODE were included in this step. This combined Enhancer/TSS list comprised our candidate elements. Candidate enhancer activity denoted as A_E in Eqn 6, Figure S1C was quantified by calculating the geometric mean of DNase-seq and H3K27ac ChIP-seq signals across each region. We next computed ABC scores by integrating enhancer activity (A_E) with Hi-C contact frequencies obtained from tissue specific Hi-C datasets ($C_{E,G,}$ where E and G denotes an enhancer and a gene respectively). The ABC score is expressed as the relative contribution of an element on gene expression to the total effect of all elements within 5Mb. ABC scores were computed for ovary, adipose, liver, muscle, and pancreas. Next we used xSNP2cGene, https://github.com/r-forge/pi314/tree/master/pkg) to include those ABC scores for which a PCOS-associated SNP position overlaps with either of an enhancer or a promoter (Eqn 7 and 8, Figure S1C). For brain, in the absence of chromatin accessibility and H3K27ac data, we have used only Hi-C contact frequencies for cGene annotation (accordingly the final equation is modified). No information was available for pituitary. So this tissue was excluded from cGene scoring.

Annotation of core genes with functional evidences

To add additional weightage to these core genes with respect to disease relevance we incorporated various functional evidences. Specifically, core genes were annotated based on three distinct categories of prior knowledge: (i) *function evidence*, genes (fGene) that have been shown to be associated to PCOS through expression studies; (ii) *phenotype evidence*, genes (pGene) associated with specific associated phenotypes frequently observed in individuals with PCOS; and (iii) *disease evidence*, genes (dGene) annotated to PCOS manifestation. In brief, functional annotations were sourced from PCOSKB[23], a manually curated knowledgebase for PCOS-associated genes, from where we have enlisted the genes reported only from expression studies. Phenotype–genes were curated from Human Phenotype Ontology (HPO)[24] based on phenotype searches including polycystic ovarian morphology, hyperandrogenism and ovulatory dysfunction. Specific HPO terms like polycystic ovaries (HP:0000147), increased circulating androgen concentration (HP:0030348), oligomenorrhea (HP:0000876), amenorrhea (HP:0000141), hirsutism (HP:0001007), alopecia of scalp (HP:0002293), insulin resistance (HP:0000855), acanthosis nigricans (HP:0000956), hyperinsulinemia (HP:0000842), hyperglycemia (HP:0003074), abnormal circulating luteinizing hormone concentration (HP:0030345), abnormal circulating follicle-stimulating hormone concentration (HP:0030346) and abnormal circulating testosterone concentration (HP:0030087) were used. Disease-related genes of PCOS were annotated using data from DisGeNet[25] (ID: C0032460), a platform that integrates gene-disease association information from multiple sources.

Identification of peripheral genes with network evidence

To identify peripheral genes (non-core) that are influenced by core genes (core) through Protein-Protein interaction network connectivity with a concept of "guilt by association", we employed Random Walk with Restart (RWR)[26] , an algorithm based on network diffusion with the core genes obtained from the above described categories (rGene, qGene, cGene, dGene, pGene and fGene), using an optimized restarting probability (as described above). For each category, we initialized a random walker at the core nodes. Genes within the PPI network more frequently visited by the walker starting from the cores, due to higher connectivity to core, are assigned higher affinity scores (Eqn 9), highlighting their potential functional relevance despite lacking direct genomic evidence.

The probability vector at iteration t, denoted by P_t, was updated according to the following equation:

$$\overrightarrow{P_t} = (1-r) \times nadjM \times \overrightarrow{P_{t-1}} + r \times \overrightarrow{P_0} \qquad \text{Eq.9}$$

Where P_0 represents the initial probability vector, where core genes are assigned their respective scores (as defined in earlier scoring schemes), and all non-cores are set to zero. The matrix nadjM denotes the normalized Laplacian adjacency matrix of the network, which encodes the transition probabilities between nodes. The parameter r is the restart probability, governing the likelihood of returning to a core node at each step, while 1-r corresponds to the probability of continuing the random walk to neighboring nodes. The process is iterated until convergence, defined as when the difference between P_t and P_{t-1} falls below a predefined threshold. The resulting steady-state vector captures the affinity of each gene to the core gene set, reflecting both functional and topological relevance within the interaction network. Following this step, we constructed a gene-predictor matrix comprising all the categories of core genes (a total of six categories), wherein each row corresponds to a gene (core or peripheral), and each column corresponds to a specific category or predictor.

Gene-predictor matrix to gene prioritization

To combine affinity scores from both genomic annotations and functional evidences (Six categories) to get a single consolidated score for a gene (Core or peripheral), the scores from gene-predictor matrix, we employed direct and indirect combining methods. The direct combining methods include sum (summing up evidences), max (taking maximum of evidences) and harmonic (sequentially weighting evidence), while indirect combination is a method similar to Fisher's combined meta-analysis (as described before[14]. Briefly, first, for gene-predictor matrix belonging to each category, we transformed the raw affinity scores into P-like values by applying the eCDF [Eq. 10]. These P-like values represent the relative rank of a gene for that particular category, ensuring comparability across heterogeneous categories. Next, for each gene, these P-like values across all categories were combined using Fisher's method to derive a single combined P-like value (Eq. 11 -13). This integrated metric was then rescaled into a Significance Index (Si), ranging from 0 to 5 (using Eq. 14), to yield a standardized prioritization score for each gene.

$$P_i^j = eCDF\left(AF_i^j\right) \qquad \text{Eq.10}$$

Where $\mathrm{AF}_\mathrm{i}^\mathrm{j}$ denotes the affinity score for the *i*th gene concerning the *j*th predictor, $\mathrm{P}_\mathrm{i}^\mathrm{j}$ is the corresponding converted *P*-value, and *eCDF* is estimated based on all genes

$$x = -2\sum_j^J log\left(P_i^j\right) \qquad \text{Eq.11}$$

$$x \sim \chi^2(2J) \qquad \text{Eq.12}$$

$$CP_i = CDF(x) \qquad \text{Eq.13}$$

where J is the number of the predictors, $\chi^2(2J)$ denotes Chi-Squared distribution with *2*J degrees of freedom, CP$_\mathrm{i}$ is the final combined *P*-value for the *i*th gene (i.e., CDF of Chi-Squared distribution valued at x).

$$x_i = -log(CP_i)$$

$$Si_i = 5 \times \frac{x_i - MIN(x_k)}{MAX(x_k) - MIN(x_k)} \qquad \text{Eq.14}$$

Where Si$_\mathrm{i}$ is the Si score (Significance Score) for the i$^{\mathrm{th}}$ gene, MIN and MAX for the minimum and maximum value respectively.

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[27]. 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.

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 is also 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[28] 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.

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).

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 Gene-to-Current-Therapy (G2CT) potential (as described previously[14]. Breifly, 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.

$$G2CT = log_{10} \frac{NES \times F}{FDR} \qquad \text{Eq.15}$$

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[29] , 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. $|log_2FC| > 0.5$ and $P\text{-value} < 0.05$ 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[30] 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 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.

Pathway crosstalk

To explore the potential interaction and crosstalk between different pathways to which the prioritized genes affiliate to, we constructed a subset of the PPI network using MSigDB version 2023.2[31] comprising of curated human genes derived from the reactome pathway database. The genes from all the pathways in MSigDB were pooled, and duplicate genes were removed. The resulting gene set was then overlaid onto the PPI network and the largest connected subnetwork was extracted. We then used the dnet package in R (version 1.1.7) to extract a subnetwork overlaying the prioritized genes from the Si matrix on the subnetwork from the previous step, while allowing a limited number of lower-priority genes to serve as linkers to maintain network connectivity. The approach begins by converting significance score to P values (xPierSubnet), assigning positive scores to nodes with P values below a user-defined threshold (0.01), indicating nodes of interest, while nodes with P values exceeding this threshold are given negative scores. Following score transformation, the pipeline searches for a connected subgraph enriched in positively scored nodes, allowing a limited number of negatively scored nodes to act as linkers. This subgraph search is facilitated through a minimum spanning tree algorithm, which serves as a heuristic solution to the prize-collecting Steiner tree problem[32]. To evaluate the statistical significance of the observed subnetwork, we conducted a degree-preserving node permutation test. This procedure was repeated 100 times to generate a null distribution of subnetworks that are expected by chance. The statistical significance of the subnetwork was calculated based on how often a subnetwork of equal or higher prioritization score appeared in the null distribution.

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.

Node removal analysis

To assess the importance of individual genes in the crosstalk network, we performed node removal analysis in two modes (single-node removal and combinatorial node removal). In both the cases, single node and nodes in different combinations were removed from the network, and the resulting fragmentation was evaluated by quantifying the fraction of nodes disconnected from the largest connected component. . The combinatorial node removal was done to model the potential synergistic effects of targeting multiple genes. The effect of both single and combinatorial node removal was visualized using upset plots generated with the ggupset package.

Determining inter-tissue cross-talk network modularity:

To investigate inter-tissue communication at the molecular level, we constructed a tissue–tissue crosstalk network based on crosstalk genes of all tissues pulled together. To uncover modular organization within the crosstalk network, we used clusterspinglass function of igraph package in R (version 2.0.3) to identify network modules by simulating annealing. The spin-glass model[33] is a method used in community detection, a process of grouping nodes in a network, based on their internal connections, to find clusters that are more densely connected internally. For each module, we conducted enrichment analyses.

Cluster analysis

We employed the supraHex[34] R package to construct a cross-tissue topological map to cluster cross-talk genes having similar Significance score or drugability across tissues. A supra-hexagonal map consisting of 37 hexagons was trained on Significance scores of the crosstalk genes across tissues. The map was generated using a self-organizing learning algorithm. These trained maps represented tissue-specific cross-talk gene prioritization profiles, with different tissues organized in 2D axis in such a way that tissues with similar Si score distribution were placed close to each other. Subsequently, a single integrated hexagonal map was generated to capture this similarity in Si scoring across tissues with each cluster capturing a set of cross-talk genes with similar prioritization trends across tissues. To assess druggability, binary druggable pocket (presence or absence) annotations from Bao, C. *et al*[35] were overlaid on the map. Genes were classified as tractable if their corresponding protein structures (sourced from the Protein Data Bank, PDB) contained predicted drug-like binding sites, as identified using fpocket software[36]. We identified the cluster exhibiting high prioritization scores and high druggability. Genes within this cluster were subjected to pathway enrichment analysis.

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. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).

16. Fulco, C. P. *et al.* Activity-by-Contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664 (2019).

17. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).

18. 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).

19. 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).

20. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveal Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042 (2016).

21. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

22. Zheng, Z. *et al.* QTLbase: an integrative resource for quantitative trait loci across multiple human molecular phenotypes. *Nucleic Acids Res.* **48**, D983–D991 (2020).

23. Joseph, S., Barai, R. S., Bhujbalrao, R. & Idicula-Thomas, S. PCOSKB: A KnowledgeBase on genes, diseases, ontology terms and biochemical pathways associated with PolyCystic Ovary Syndrome. *Nucleic Acids Res.* **44**, D1032–D1035 (2016).

24. Gargano, M. A. *et al.* The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res.* **52**, D1333–D1346 (2024).

25. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update.

Nucleic Acids Res. **48**, D845–D855 (2020).

26. Grady, L. Random walks for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 1768–1783 (2006).

27. 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).

28. Ochoa, D. *et al.* Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic Acids Res.* **49**, D1302–D1310 (2021).

29. Wang, Z. *et al.* Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd. *Nat. Commun.* **7**, 1–11 (2016).

30. 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).

31. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

32. Fang, H. & Gough, J. The ‘dnet’ approach promotes emerging research on cancer patient survival. *Genome Med.* **6**, 1–16 (2014).

33. Reichardt, J. & Bornholdt, S. Statistical mechanics of community detection. *Phys. Rev. E* **74**, 016110 (2006).

34. Fang, H. & Gough, J. supraHex: An R/Bioconductor package for tabular omics data analysis using a supra-hexagonal map. *Biochem. Biophys. Res. Commun.* **443**, 285–289 (2014).

35. Bao, C. *et al.* A cross-disease, pleiotropy-driven approach for therapeutic target prioritization and evaluation. (2024) doi:10.1016/j.crmeth.2024.100757.

36. Schmidtke, P. & Barril, X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J. Med. Chem.* **53**, 5858–5867 (2010).