“Genetic Architecture of Non-Syndromic Premature Ovarian Insufficiency:

A Systematic Review.”

Amy V. Kaucher[1], Evaggelia Alexopoulou[2], Isa A. Olofsson[1], Kevin Bourge[1], Cathrine Voigt Dalgaard[1,2], L. Kevin Lie[1], Nathalie Friis Wang[2], Morten Dunø[3], Anja Pinborg[2,4], Eva R. Hoffmann[1]

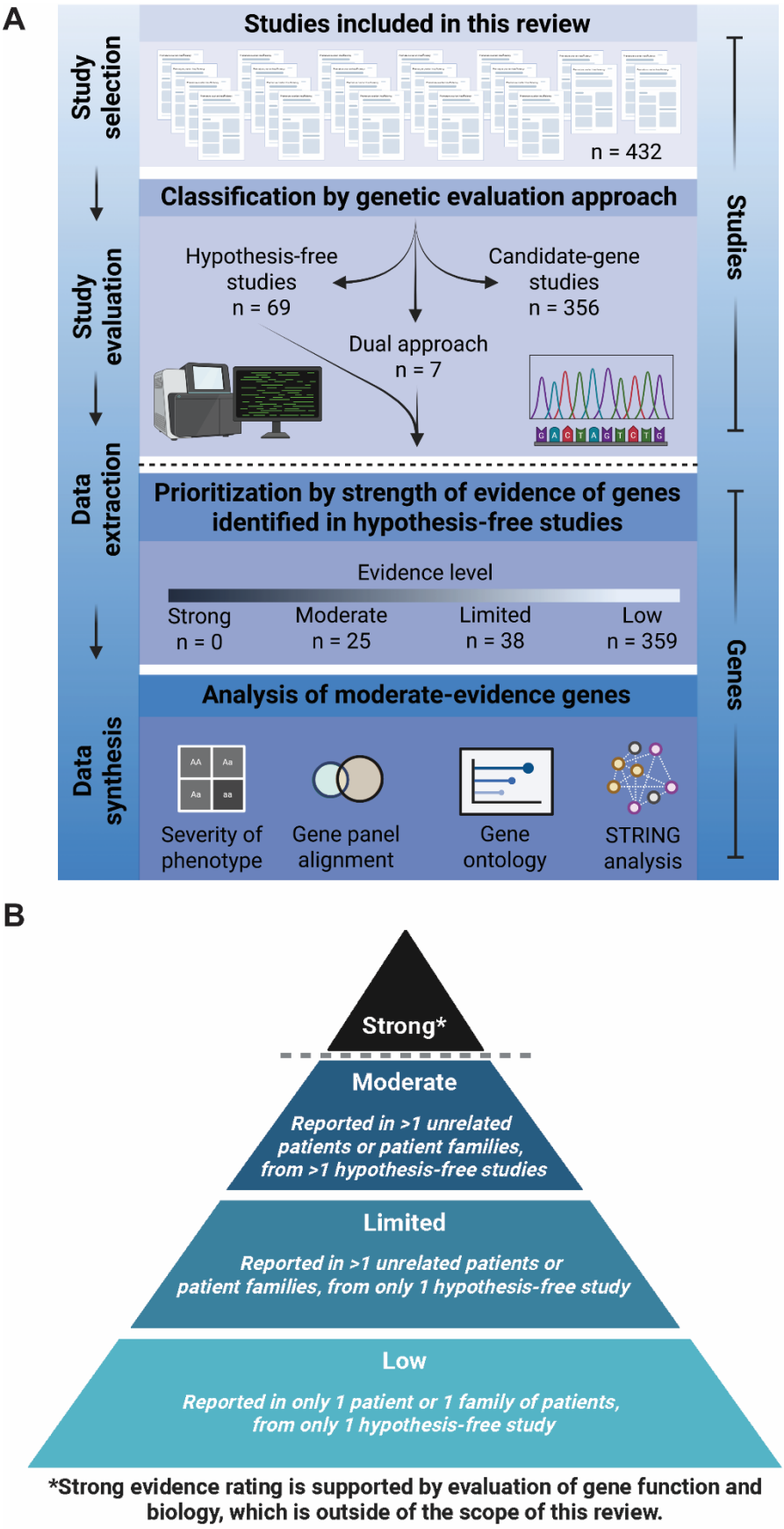
[1] Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

[2] The Fertility Clinic, Department of Gynecology, Fertility and Obstetrics, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark.

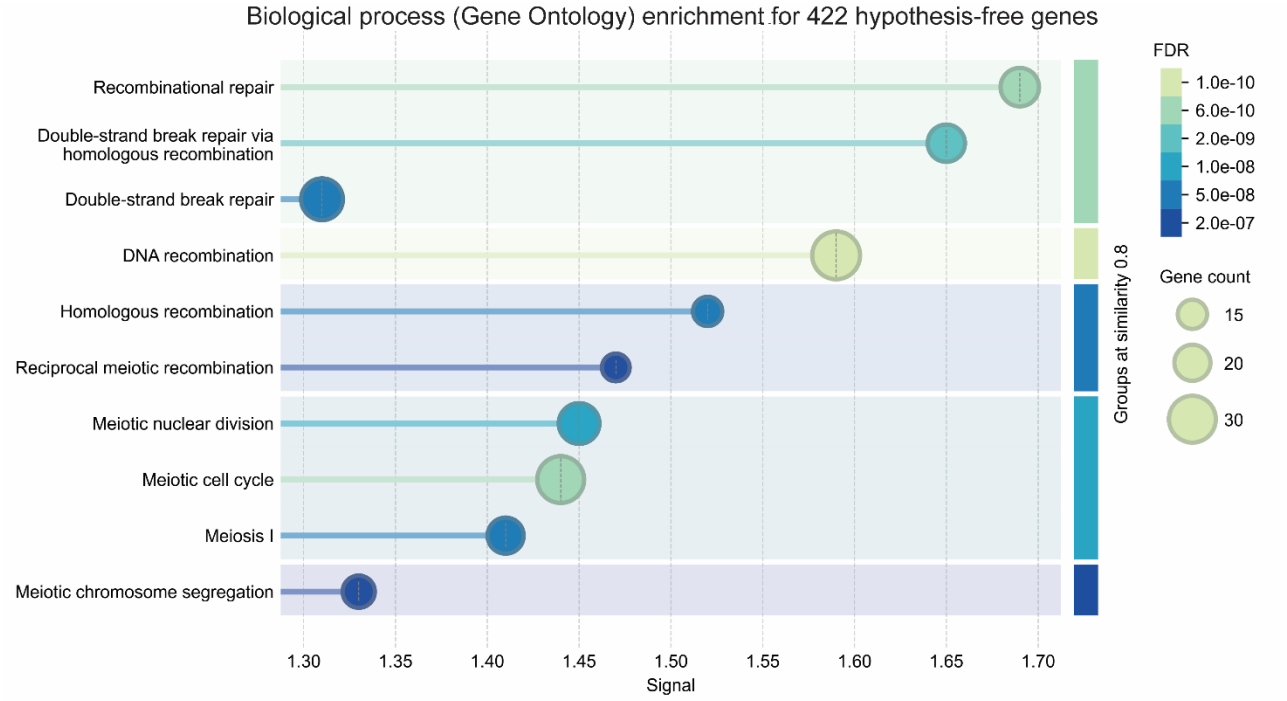
[3] Department of Clinical Genetics, Copenhagen University Hospital – Rigshospitalet, Copenhagen, Denmark.

[4] Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

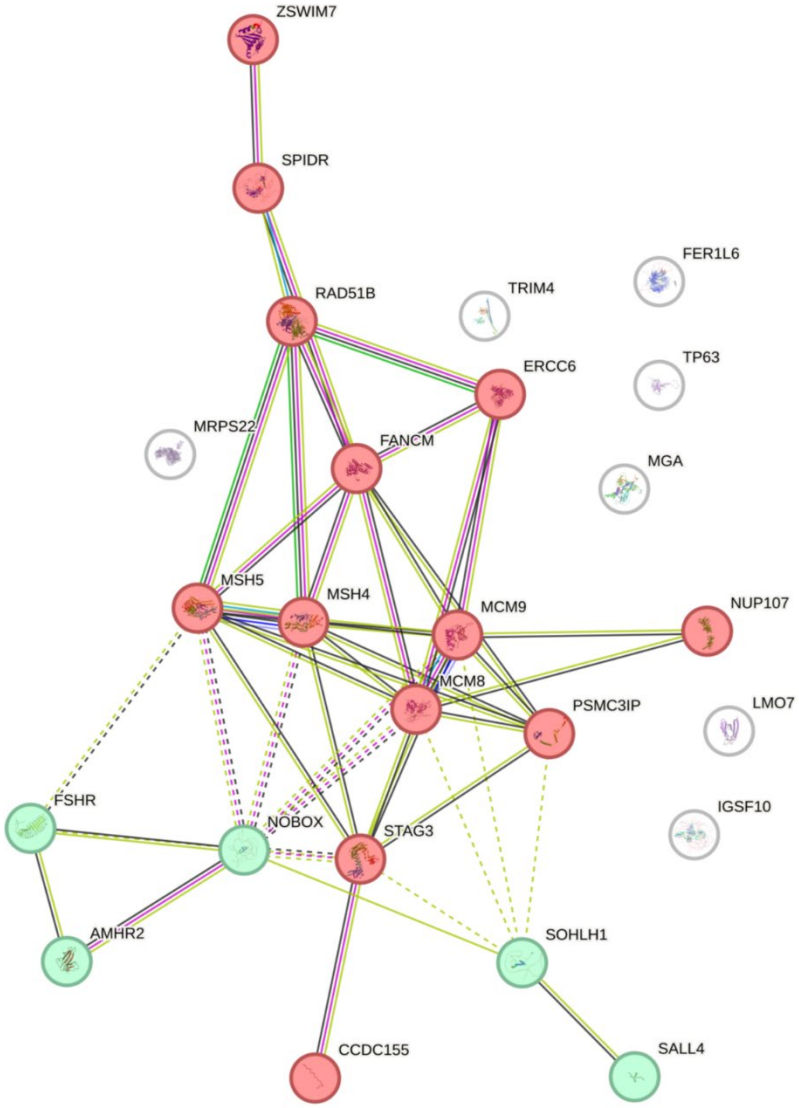
Table of Contents



Supplementary Figure S1 Systematic review workflow and prioritization framework for hypothesis-free studies. (A) Workflow of systematic review. (B) Prioritization scheme for hypothesis-free studies. Created in https://BioRender.com



Supplementary Figure S2 Gene ontology enrichment of all 422 genes associated with non-syndromic POI in hypothesis-free studies. Biological process gene ontology enrichment of our hypothesis-free gene list generated using the STRING database (version 12.0) (Szklarczyk et al., 2025). For the enrichment analysis, the whole genome is used for statistical background. The Y-axis lists the enriched processes, and the X-axis represents the signal strength. The circles depict the gene count for each process, and the colour of the items represents the false discovery rate, which ranges from 1.0×10^{-10} to 2.0×10^{-7}.



Supplementary Figure S3 K-means STRING clustering analysis identifies two distinct clusters within our moderate-evidence gene list. Strings were analysed with K-means clustering (k = 2) to group genes based on similarity in textual patterns using STRING database (version 12.0) (Szklarczyk et al., 2025). The resulting two clusters represent distinct similarity groups within the dataset. Colour shading indicates cluster membership. Cluster 1 (red group) is "double strand break repair via homologous recombination" and comprises 13 genes. Cluster 2 (green group) is "translational attenuation and 46,XX gonadal dysgenesis" and comprises five genes. Dotted lines represent edges between clusters. Known interactions are signified by solid lines in teal (from curated databases) or magenta (experimentally determined. Predicted interactions are represented in solid lines in green (gene neighbourhood), red (gene fusions), or blue (gene co-occurrence). Other interactions include lines of yellow (textmining), black (co-expression), or lavender (protein homology).

Supplementary Data File S1 Modified Q-Genie assessment tool for non-GWA studies - Original assessment tool referenced in Sohani et al., 2015, 2016.

Template accessed from https://fhs.mcmaster.ca/pgp/documents/Q-Geniev.1.1.pdf (accessed 20 September 2023).

1. Rationale for study

Please rate the study on the adequacy of the presented hypothesis and rationale.

When rating the study, please consider the following:

- Was a scientific rationale for chosen genes presented to avoid selective reporting of positive *results?*

  Where a hypothesis-free approach is taken, a rationale for selecting this design should be presented.

☐	☐	☐	☐	☐	☐	☐
1	**2**	**3**	**4**	**5**	**6**	**7**
Poor		Good		Very Good		Excellent

2. Selection and definition of outcome of interest. The outcome can be cases/disease status or a quantitative trait.

Please rate the study on the classification of the outcome (e.g. disease status or quantitative trait).

When rating the study, please consider the following:

- Were the cases appropriately defined?

  Outcome definitions will vary from independent adjudication or reliable laboratory measures (strong) to self-report (moderate) to no-description (poor)

- Were participants appropriately sampled?

  Participants should be sampled in a way to avoid selection bias as appropriate to the study objectives (e.g. such as selecting the most sick cases if the objective is not to enrich cases).

Included participants should reflect the entire population of interest.

☐	☐	☐	☐	☐	☐	☐
1	**2**	**3**	**4**	**5**	**6**	**7**
Poor		Good		Very Good		Excellent

3. Selection and comparability of comparison groups (if applicable)

Please rate the study on appropriateness of comparison groups (e.g. control groups).

When rating the study, please consider the following:

- Were the controls appropriately defined?
- Were the controls sampled in a way to minimize selection bias?
- Was a detailed description of selection procedure (i.e. eligibility criteria, sources and methods of ascertainment, methods of matching if applicable) outlined or referenced?

Please note: In multi-ethnic studies, allele frequencies and disease risks may differ. Consequently, confounding may occur if these sub-populations are unevenly distributed across exposure groups (or between cases and controls); therefore, details of the sub-populations (e.g. ethnicity) should be reported.

☐ ☐ ☐ ☐ ☐ ☐ ☐

1 **2** **3** **4** **5** **6** **7**

Poor Good Very Good Excellent

4. Technical classification of the exposure

Please rate the study on the technical classification of the genetic variant.

When rating the study, please consider the following:

- Was the source (e.g. buffy coat) and method of storage for the DNA sample appropriate?
- Were the methods of DNA ascertainment similar for comparison groups (if applicable)?
- Was the genotyping platform and alignment method appropriate?
- Were the sequencing depth & coverage appropriate?
- Were the sequencing depth & coverage similar between the comparison groups?

Please note: if genotypes are imputed, authors should describe methods and rationale for imputing

☐ ☐ ☐ ☐ ☐ ☐ ☐

1 **2** **3** **4** **5** **6** **7**

Poor Good Very Good Excellent

5. Non-technical classification of the exposure

Please rate the study on the non-technical classification of the genetic variant.

When rating the study, please consider the following:

- Did a blinded assessor conduct the genotyping?
- Was genotyping conducted in all the participants from the study simultaneously or in smaller *batches? If so, were methods across batches same?*
- If applicable, were samples randomized prior to genotyping (e.g. not all controls on one plate *and cases on another)?*

☐ ☐ ☐ ☐ ☐ ☐ ☐

1 2 3 4 5 6 7

Poor Good Very Good Excellent

6. Other sources of bias

Please rate the study on the disclosure and discussion of sources of bias.

In addition to selection and classification bias previously discussed, many other potential sources of bias exist (e.g. time-lag bias, attrition bias, et cetera). Please consider whether all sources of bias were disclosed and their effect on the results discussed.

- Things to consider: description of study limitation, inherent bias of methods/assumptions

☐ ☐ ☐ ☐ ☐ ☐ ☐

1 2 3 4 5 6 7

Poor Good Very Good Excellent

7. Sample size and power

Please rate whether the study was adequately powered.

- Was the sample size appropriate?
    - Things to consider: number of generations (# of mieoses) per family; # of POI patients per generation; or if more than 1 family is included
- Was an a priori power analysis conducted?
- Is a replication (control) cohort included?
- Is a replication (case) cohort included?

☐ ☐ ☐ ☐ ☐ ☐ ☐

1 2 3 4 5 6 7

Poor Good Very Good Excellent

8. A priori planning of analyses

Please rate the study on the planned analyses.

- Was the analysis plan appropriate and sufficiently described?
- Was selective and/or inappropriate reporting avoided (i.e. all results from tests conducted were reported)? Authors should identify where additional results can be found if not included in the primary paper (e.g. supplementary tables).
- Were the tested subgroups, interactions, and sensitivity analyses described and reported?
- Was the statistical software used identified?
- Was variant filtering method explained?

1	2	3	4	5	6	7
Poor		Good		Very Good		Excellent

10. Testing of assumptions and inferences for genetic analyses

Please rate the study on the description and test of all assumptions and inferences.

- Were all assumptions concerning the genetic analysis tested?

  Specifically,

  - In non-family based studies, some individuals may be distantly related or part of a consanguineous group, which may lead to inaccurate results and should be tested with appropriate measures.
  - Reported sex and ethnicity should also be checked prior to conducting analyses.

1	2	3	4	5	6	7
Poor		Good		Very Good		Excellent

11. Appropriateness of inferences drawn from results

Please rate the study on whether conclusions drawn by the authors were supported by the results and appropriate methods.

- Things to consider: functional follow-up (in silico, in vitro, etc.); monogenic vs oligogenic consideration; proper literature review (what is known about the biology)

1	2	3	4	5	6	7
Poor		Good		Very Good		Excellent

Scoring

Please add the total score from each question. _____

For studies with control groups: Scores ≤32 indicate poor quality studies, >32 and ≤40 indicate studies of moderate quality, and >40 indicate good quality studies.

For studies without control groups: Scores ≤29 indicate poor quality studies, >29 and ≤36 indicate studies of moderate quality, and >36 indicate good quality studies.