*D1032–D1035* *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
*doi: 10.1093/nar/gkv1146*

# PCOSKB: A KnowledgeBase on genes, diseases, ontology terms and biochemical pathways associated with PolyCystic Ovary Syndrome

**Shaini Joseph[1,†], Ram Shankar Barai[1,†], Rasika Bhujbalrao[2] and Susan Idicula-Thomas[1,*]**

[1]Biomedical Informatics Center of Indian Council of Medical Research, National Institute for Research in Reproductive Health, Mumbai-400012, India and [2]G.N. Khalsa College, University of Mumbai, India

Received August 14, 2015; Revised October 05, 2015; Accepted October 19, 2015

## ABSTRACT

**Polycystic ovary syndrome (PCOS) is one of the major causes of female subfertility worldwide and ≈7–10% of women in reproductive age are affected by it. The affected individuals exhibit varying types and levels of comorbid conditions, along with the classical PCOS symptoms. Extensive studies on PCOS across diverse ethnic populations have resulted in a plethora of information on dysregulated genes, gene polymorphisms and diseases linked to PCOS. However, efforts have not been taken to collate and link these data. Our group, for the first time, has compiled PCOS-related information available through scientific literature; cross-linked it with molecular, biochemical and clinical databases and presented it as a user-friendly, web-based online knowledgebase for the benefit of the scientific and clinical community. Manually curated information on associated genes, single nucleotide polymorphisms, diseases, gene ontology terms and pathways along with supporting reference literature has been collated and included in PCOSKB (http://pcoskb.bicnirrh.res.in).**

## INTRODUCTION

Polycystic ovary syndrome (PCOS) is a multi-factorial reproductive disorder affecting 7–10% of women globally (1,2). It is one of the major causes of female subfertility (3,4). The complexity of the syndrome is contributed by both the genes and the diseases associated with it. Several factors like gene interactions, environmental and ethnic influences and lifestyle play a role in the clinical manifestation of PCOS (5-7).

Genetic studies on PCOS has revealed that genes having dissimilar functions and affecting varied biochemical pathways like ovarian and adrenal steroidogenesis (e.g. *CYP11A, CYP21*), gonadotropin action and regulation (e.g. *FST*, *LHCGR* and *FSHR*), insulin action and secretion (e.g. *INS, IRS-1, IRS-2*), inflammation (e.g. *IL-6*), complement and coagulation cascades (e.g. *VWF*), signalling (e.g. *ADIPOQ, INS, LHCGR, AMH*), cancer (e.g. *INS, AR, MMP1*) etc. are associated with PCOS (7,8). Single Nucleotide Polymorphisms (SNPs)/mutations in one or more of these genes have been linked to diverse comorbid conditions associated with PCOS such as obesity, diabetes, dyslipidemia, cardiovascular diseases, cancer and subfertility (6).

Thus, depending on the causal factors, the phenotypic manifestations vary widely making diagnosis a daunting task. PCOS diagnosis is currently based on the NIH or the Rotterdam criteria formulated in 2003. For a disease condition to be diagnosed as PCOS, presence of any two of the following symptoms is essential: clinical or biochemical hyperandrogenism, oligo-anovulation and/or polycystic ovaries, excluding other endocrinopathies as per the Rotterdam criteria (1,5).

Being a highly researched disorder, there is abundant information available in literature on the various aspects of PCOS such as its genetics, diverse phenotypic manifestations, associated comorbidities and even inconsistencies in the study results across different population groups. Electronic databases and computational tools are required to integrate this information from disparate sources and to analyse the complex data arising out of genetic heterogeneity for PCOS. Use of such computational approaches that involve collating and mining information on genes, their function/ontology terms, expression profiles, tissue specificity and pathway information is expected to lead to a better understanding of the genetic aetiology of PCOS (9,10).

Several gene-disease networks and gene prioritization methods have been applied to identify novel candidate genes or the most crucial genes involved in pathophysiology of multi-factorial diseases like Alzheimer's, Cancer etc. (11–13). This approach has not yet been used for PCOS; an important contributing factor being absence of comprehensive and curated repositories catering to PCOS-related genomic,

*To whom correspondence should be addressed. Tel: +91 22 24192107/04; Fax: +91 22 24139412; Email: thomass@nirrh.res.in
†These authors contributed equally to the paper as first authors.

proteomic, metabolomic and clinical information. To facilitate and augment high-throughput research on PCOS including gene network and prioritization studies, we have created a manually curated knowledgebase with complete information on genes associated with PCOS collated from literature references and reviewed databases. It currently holds information on 241 genes and their corresponding proteins, 3D structures, SNPs/mutations, ontology terms, pathways and diseases.

## DATA COLLECTION AND ORGANIZATION

PubMed, was searched using relevant keywords namely 'PCOS', 'Polycystic ovarian syndrome', 'PCO', 'PCO1', 'anovulation', 'Stein Leventhal', 'Ovarian Syndrome', 'polycystic ovaries', 'polycystic ovary disease', 'POS', 'PCOD-Polycystic Ovarian Disease', 'hyperandrogenic chronic anovula', 'hyperandrogenic chronic anovulation', 'ovarian hyperthecosis', 'Sclerocystic Ovarian Disease', 'sclerocystic ovary syndrome', 'Bilateral PCOS' and 'functional ovarian hyperandrogen'. The genes cited in the retrieved literature were acquired from the NCBI Gene database. The reference literature was reviewed to ascertain the association of each of these genes to PCOS. Genes and SNPs, whose association with PCOS could not be confirmed based on the reference, were discarded.

Relevant information on the validated genes such as nature of the study population, ethnicity, mutations/ SNPs, gene ontology terms, protein structural information, biochemical pathway and disease related information were retrieved from online databases such as NCBI PubMed (14), Gene (15), dbSNP (14), Ensembl (16), UniProt (17), Pfam (18), InterPro (19), GO (20), KEGG (21), GeneCards (22) and OMIM (23).

The diseases associated with each of the validated genes were retrieved from GeneCards database. These diseases were grouped into non-redundant categories such as cancers/tumors, cardiac disorders, reproductive disorders, immune disorders etc. Each of the gene-disease associations was manually verified by reviewing the reference literature. Several gene-disease associations, retrieved using GeneCards, could not be established based on reference literature and were therefore discarded.

The data collected on genes associated with PCOS have been organised into seven tables in PCOSKB.

## DATA ARCHITECTURE

PCOSKB is built on Apache HTTP Server 2.0.59. The database was created using MySQL Server 5.0 and the web interfaces were designed using PHP 5.2.9, HTML and JavaScript. These are platform independent, open source software and they support multithreading and multiuser environment. The analysis graphs and charts were generated using the JpGraph library (http://jpgraph.net/).

## WEB INTERFACES

PCOSKB has a user-friendly interface, as described below:

(i) Home: a brief introduction on PCOS and information accessible in the database is provided.
(ii) Search: two search options: quick and advanced are available to users.
(iii) Browse: this tab can be used to browse the database for:
  (a) Genes: this page provides description of the genes, their genomic location and links to the complete gene record.
  (b) SNPs: the detailed information on PCOS associated SNPs such as their upstream/downstream sequence, functional significance and links to relevant databases such as dbSNP and PubMed are provided.
  (c) Associated diseases: the unique gene-disease associations in PCOSKB can be browsed
  (d) Pathways: the pathways involving PCOS-related genes hyperlinked to the KEGG pathway database is provided.
  (e) Ontology terms: the unique ontology terms associated with genes along with their associated pathways and diseases can be analysed.
  (f) Tools: this section enables users to identify the significant gene ontology terms, pathways and disease phenotypes corresponding to a user-defined gene list. The gene list can be obtained either by directly selecting the genes present in PCOSKB or through their associated diseases.
(iv) Help: the features present in the database are explained with examples to assist users.
(v) Analysis: the relative frequencies of the disease phenotypes, gene ontology terms and pathways predominantly associated with genes involved in PCOS pathophysiology are displayed.
(vi) Statistics: it gives information on the total number of genes, associated diseases, ontology terms, pathways and SNPs present in the database. Users can directly view the associated diseases, gene ontology terms, SNPs for genes using the '*Diseases associated with genes'* link. The number of genes that are associated with each of the diseases can be accessed through *'Genes associated with disease'* link.

## DATA ANALYSIS

The data in PCOSKB were analysed to identify the significant comorbid conditions, pathways and ontology terms associated with PCOS. The results of the analysis can be viewed in Figure 1.

Reproductive and cardiac disorders were the most prevalent diseases associated with PCOS, as expected. Hyperlinks are provided in the pie-chart for users to view the gene-disease associations (Figure 1A).

Information available in KEGG pathway database was used to cluster the PCOS-related genes and retrieve the biochemical pathways. The distribution of the pathways, based on the gene members is depicted as bar graph. Each of the bars, represent a pathway and are hyperlinked to the respective KEGG pathway. The position (role) of the PCOS-associated gene/s in the pathway can be identified by their red highlights. The number of gene members for each of the pathway is indicated adjacent to the bar. Of the 175 unique pathways, ≈30% had five or more gene members.

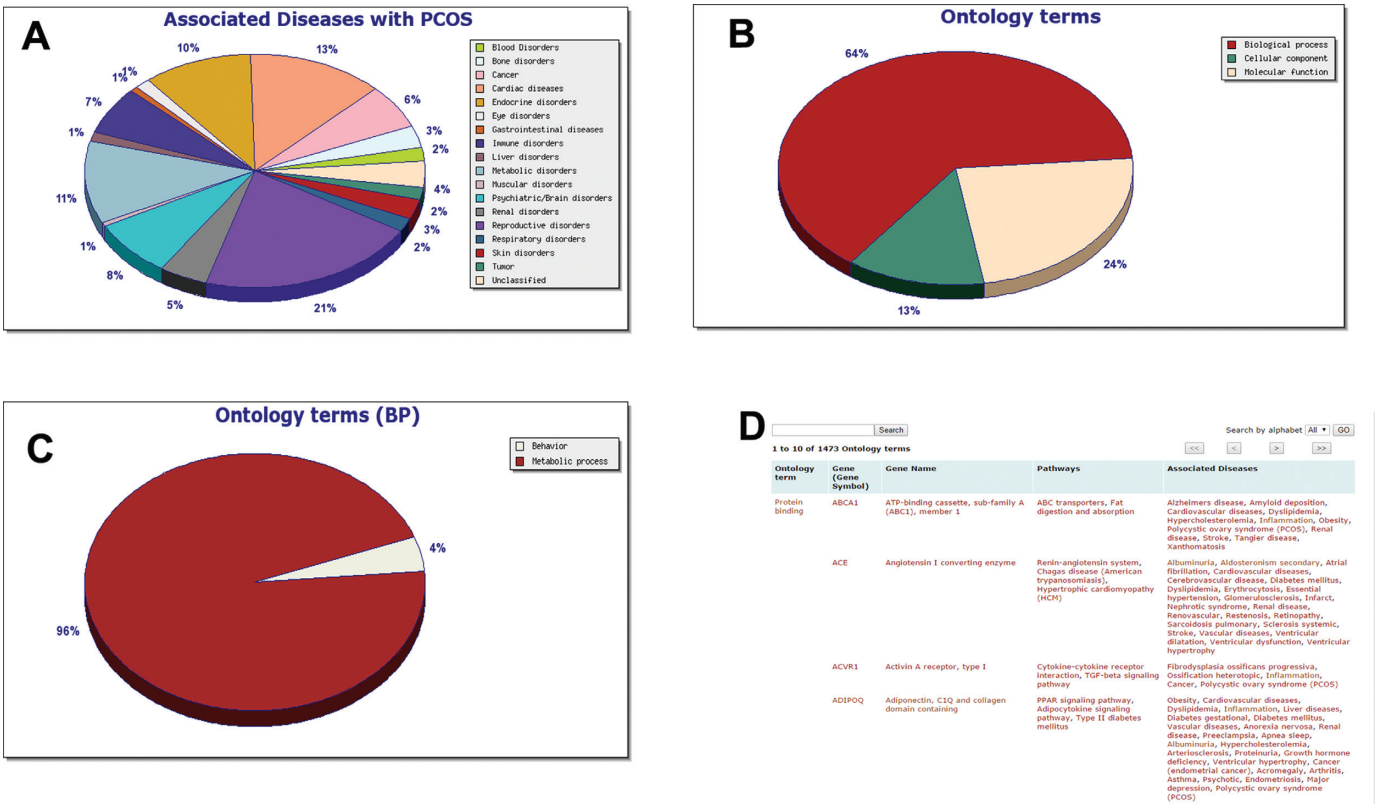


**Figure 1.** (**A**) Pie chart of associated diseases, (**B**) Pie chart of ontology terms, (**C**) Pie chart of ontology terms (BP), (**D**) Snapshot of an associated ontology terms page in PCOSKB.

The ontology terms for the PCOS-related genes were retrieved from the GO database (AmiGO 2). The gene ontology annotations in the GO database are organized in a hierarchical manner with '*parent*' terms being broad and 'child' terms being more specific to a biological process (BP), molecular function (MF) and cellular component (CC). Parents of the PCOS-related GO terms were retrieved using the GOOSE tool (20). These data were further used to group the genes based on the three broad categories namely, MF, BP and CC. The relative share of these three components in the total ontology terms is represented as a pie chart in the Analysis tab. Each of these components were further analysed exclusively for relative proportion of the ontology terms at the subsequent level of distance from the parent. The data associated with the ontology term such as genes, pathways, diseases can be accessed in a tabular format using the hyperlinks provided in the pie-charts (Figure 1B, C, D).

## DATABASE SUMMARY

- *Genes and SNPs:* PCOSKB holds information on 241 genes and 114 SNPs associated with PCOS. These genes and SNPs have been validated for their association with PCOS by reviewing the reference literature. The information is organised into six sections namely gene related, protein related, gene ontology, validated SNPs, associated diseases and references. Information mined from PCOS-related references (e.g. population size, ethnicity, SNPs/ mutation and associated conditions) can be viewed in the reference section.
- *Associated diseases:* the database has 1905 unique gene-disease associations and 500 unique PCOS-associated diseases. These diseases have been validated for their association with the genes/PCOS by reviewing the reference literature. The diseases are broadly divided into 18 categories. Reproductive (21%) and cardiac disorders (13%) feature in the top disease associations of PCOS.
- *Biochemical pathways and gene ontology terms:* analysis of the validated genes revealed its association with 175 unique pathways and 1473 unique ontology terms. These data highlight the diverse biological role of the genes associated with PCOS; an observation which corroborates the known complexity and genetic heterogeneity of PCOS.

## COMPARISON WITH EXISTING DATABASES

Few gene-based databases like GeneCards (22) and DisGeNET (24) provide information on genes and associated disorders. The information present in these databases is not entirely manually curated and derived from several primary and secondary databases. A curated data set of genes and associated diseases is essential to understand the pathophysiology of multi-factorial diseases. There are few disease-specific databases like EpilepsyGene (25), AutismKB (26) which are manually curated. However, till date no database exists with information on genes, their as-

sociated ontology terms, pathways and disorders for reproductive disorders like PCOS. In PCOSKB, we have made an attempt to integrate complete information on genes associated with PCOS. Each gene has been manually verified with reference literature for its role in PCOS and details of the study population have been incorporated in the database. Each gene-disease association present in the database has been manually curated with published literature.

## CONCLUSION

Although there are few online databases catering to gene disease associations, there are no manually curated databases dedicated to PCOS that focuses on observed human gene-disease associations along with information on gene ontology terms and biochemical pathways. Clinicians and researchers can leverage the well-integrated information on molecular and clinical aspects of PCOS available in PCOSKB. PCOSKB would be updated annually.

## ACKNOWLEDGEMENTS

The authors are grateful to Dr Smita D. Mahale, Director, NIRRH and PI of Biomedical Informatics Centre for all the help and support. The authors also acknowledge the assistance provided by Ms Sapna Iyer in data collection.

## FUNDING

This work (RA/285/08-2015) was supported by Indian Council of Medical Research (63/128/2001-BMS). The open access publication charge for this paper has been waived by Oxford University Press—*NAR*.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Brady,C., Mousa,S.S. and Mousa,S.A. (2009) Polycystic ovary syndrome and its impact on women's quality of life: More than just an endocrine disorder. *Drug Healthc Patient Saf.*, **1**, 9–15.
2. Manco,M., Castagneto-Gissey,L., Arrighi,E., Carnicelli,A., Brufani,C., Luciano,R. and Mingrone,G. (2014) Insulin dynamics in young women with polycystic ovary syndrome and normal glucose tolerance across categories of body mass index. *PLoS One.*, **9**, e92995.
3. Arain,F., Arif,N. and Halepota,H. (2015) Frequency and outcome of treatment in polycystic ovaries related infertility. *Pak J Med Sci.*, **31**, 694–699.
4. Kousta,E., White,D.M., Cela,E., McCarthy,M.I. and Franks,S. (1999) The prevalence of polycystic ovaries in women with infertility. *Hum Reprod.*, **14**, 2720–2723.
5. Lee,J.Y., Baw,C., Gupta,S., Aziz,N. and Agarwal,A., (2010) Role of oxidative stress in polycystic ovary syndrome. *Curr Womens Health Rev.*, **6**, 96–107.
6. Barthelmess,E.K. and Naz,R.K. (2015) Polycystic ovary syndrome: current status and future perspective. *Front Biosci (Elite Ed)*, **6**, 104–119.
7. Prapas,N., Karkanaki,A., Prapas,I., Kalogiannidis,I., Katsikis,I. and Panidis,D. (2009) Genetics of polycystic ovary syndrome. *Hippokratia*, **13**, 216–223.
8. Welt,C.K. and Duran,J.M. (2014) The genetics of polycystic ovary syndrome. *Semin Reprod Med.*, **32**, 177–182.
9. Mukherjee,S. and Maitra,A. (2010) Molecular & genetic factors contributing to insulin resistance in polycystic ovary syndrome. *Indian J Med Res*, **131**, 743–760.
10. Kosova,G. and Urbanek,M. (2013) Genetics of the polycystic ovary syndrome. *Mol Cell Endocrinol.*, **373**, 29–38.
11. Bromberg,Y. (2013) Disease gene prioritization. *PLoS Comput Biol.*, **9**, e1002902.
12. Talwar,P., Silla,Y., Grover,S., Gupta,M., Agarwal,R., Kushwaha,S. and Kukreti,R. (2014) Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC Genomics*, **15**, 199.
13. Chuang,H.Y., Lee,E., Liu,Y.T., Lee,D. and Ideker,T. (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol.*, **3**, 140.
14. NCBI Resource Coordinators. (2015) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **43**, D6–D17.
15. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R. *et al.* (2015) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
16. Cunningham,F., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2015) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
17. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
18. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
19. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
20. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
21. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
22. Safran,M., Dalah,I., Alexander,J., Rosen,N., Iny Stein,T., Shmoish,M., Nativ,N., Bahir,I., Doniger,T., Krug,H. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)*, doi: 10.1093/database/baq020.
23. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
24. Pinero,J., Queralt-Rosinach,N., Bravo,A., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, doi:10.1093/database/bav028.
25. Ran,X., Li,J., Shao,Q., Chen,H., Lin,Z., Sun,Z.S. and Wu,J. (2015) EpilepsyGene: a genetic resource for genes and mutations related to epilepsy. *Nucleic Acids Res.*, **43**, D893–D899.
26. Xu,L.M., Li,J.R., Huang,Y., Zhao,M., Tang,X. and Wei,L. (2012) AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res.*, **40**, D1016–D1022.