



Genetics of PCOS: A systematic bioinformatics approach to unveil the proteins responsible for PCOS



Pritam Kumar Panda^{a,*}, Riya Rane^a, Rahul Ravichandran^a, Shrinkhla Singh^a, Hetalkumar Panchal^b

^a School of Biotechnology and Bioinformatics, D. Y. Patil University, CBD Belapur, Navi Mumbai, Maharashtra, India

^b Gujarat Agricultural Biotechnology Institute, Navsari Agricultural University, Athwa Farm, Ghod Dod Road, Surat, 395007, Gujarat, India

ARTICLE INFO

Article history:

Received 5 February 2016

Received in revised form 22 March 2016

Accepted 23 March 2016

Available online 31 March 2016

Keywords:

PCOS

Phylogenetic analysis

R

MeV

String

Modeller

RaptorX

Microarray

ABSTRACT

Polycystic ovary syndrome (PCOS) is a hormonal imbalance in women, which causes problems during menstrual cycle and in pregnancy that sometimes results in fatality. Though the genetics of PCOS is not fully understood, early diagnosis and treatment can prevent long-term effects. In this study, we have studied the proteins involved in PCOS and the structural aspects of the proteins that are taken into consideration using computational tools. The proteins involved are modeled using Modeller 9v14 and Ab-initio programs. All the 43 proteins responsible for PCOS were subjected to phylogenetic analysis to identify the relatedness of the proteins. Further, microarray data analysis of PCOS datasets was analyzed that was downloaded from GEO datasets to find the significant protein-coding genes responsible for PCOS, which is an addition to the reported protein-coding genes. Various statistical analyses were done using R programming to get an insight into the structural aspects of PCOS that can be used as drug targets to treat PCOS and other related reproductive diseases.

© 2016 Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Polycystic ovary syndrome (PCOS) is a most common endocrine disorder prevailing in reproductive age women across the world. It is identified with chronic anovulation and hyperandrogenism [1]. The symptoms associated with PCOS are obesity and several dermatological features. PCOS are found to have a significant reproductive and metabolic impact, which may result in type-2-diabetes or cardiovascular disease [2]. A mutation in PCOS proteins interacts with multiple inherited and environmental factors. Multiple inherited genes are responsible for the occurrence of PCOS. In India, the prevalence of PCOS was 22.5% by Rotterdam and 10.7% by Androgen Excess Society criteria. Mild PCOS is one of the most common phenotype occurring in about 52.6% of women [3]. The following study deals with in-depth genetic and phylogenetic analyses of all the genes related to PCOS. Microarray analyses were performed using 'R' to identify the highly expressed genes among all the selected ones for the study.

2. Genetics of PCOS

2.1. Ovary and adrenal steroidogenesis

2.1.1. CYP11A

The conversion of cholesterol into progesterone is the first step of steroidogenesis. It is catalyzed by P450. A gene known as CYP11A that is located at 15q24 encodes P450. CYP11A gene also shows the association in the serum testosterone levels. The CYP11A alleles also show the association with 5' untranslated region (5' UTR) [4].

2.1.2. CYP17A1

CYP17A1 converts pregnenolone and progesterone into 17-hydroxypregnenolone and 17-hydroxyprogesterone and invert these steroids into dihydroepiandrosterone (DHEA) and Δ 4-Androstendione (Δ 4A), which is catalyzed by the P450c17 enzyme. The P450c17 enzyme has activities such as 17,20-lyase and 17-hydroxylase activities. P450c17 is encoded by CYP17A, which is located at 10q27.3. It was proclaimed that the P450c17 enzyme activity and expression increases in ovary theca cells from women with PCOS [4].

2.1.3. CYP19

CYP19 converts androgen to estrogen. The enzyme complex is composed of cytochrome P450 aromatase and NADPH cytochrome P450 reductase, and P450arom is encoded by CYP19 located at 15p21.1.

* Corresponding author.

E-mail address: pritampkp15@gmail.com (P.K. Panda).

Aromatase deficiency has been reported in a number of people having hyperandrogenism. With comparison to the control follicles all the PCOS follicles have estradiol, bioactivity of lower aromatase stimulation

[4]. This shows that the aromatase activity might be decreased in PCOS follicles and the excess androgen leads to the improper follicle development.

Table 1

List of 43 Genes involved in PCOS with UNIPROT-ID, PDB-ID and chromosomal locations.

Gene name	Locus	Organism	Uniprot ID	Accession number	PDB-ID
CYP11A-cytochrome P450 side-chain cleavage enzyme (Marker Locus: D15S520)	Chromosome 15: 74,367,990-74,368,179	<i>Homo Sapiens</i>	P05108	P05108	3N9Y
Sex hormone binding globulin	Chromosome 17: 7,613,946-7,633,383	<i>Homo Sapiens</i>	P04278	ABY68008	1D2S
Luteinizing hormone choriogonadotropin receptor	Chromosome 2: 48,708,770-48,723,514	<i>Homo Sapiens</i>	P22888.4	P22888	1LUT
Follicle-stimulating hormone receptor	Chromosome 2: 48,962,157-49,154,537	<i>Homo Sapiens</i>	P23945	AAB26480	1XUN
Mothers against decapentaplegic homolog 4	Chromosome 18: 51,028,394-51,085,045	<i>Homo Sapiens</i>	Q13485	NP_005350	1DD1
Leptin	Chromosome 7: 128,241,284-128,257,628	<i>Homo Sapiens</i>	P41159	AAH69452	1AX8
Leptin receptor	Chromosome 1: 65,420,652-65,641,559	<i>Homo Sapiens</i>	P48357	AAB09673	3V60
Inhibin beta-B	Chromosome 2: 120,346,143-120,351,808	<i>Homo Sapiens</i>	P09529	AAH30029	–
Inhibin beta-A	Chromosome 7: 41,685,114-41,703,108	<i>Homo Sapiens</i>	P08476	AAH07858	1NYS
Inhibin A (gene symbol: INHA)	Chromosome 2: 219,569,162-219,575,713	<i>Homo Sapiens</i>	P05111	P05111	–
Pro-opiomelanocortin (gene symbol: POMC)	Chromosome 2: 25,160,853-25,168,903	<i>Homo Sapiens</i>	P01189	P01189	–
Uncoupling protein 213 (gene symbol: UCP2 + 3)	Chromosome 11: 73,974,667-73,983,307	<i>Homo Sapiens</i>	P55851	P55851	–
Melanocortin 4 receptor	Chromosome 18: 60,371,110-60,372,775	<i>Homo Sapiens</i>	P32245	AAO92061	2IQP
Insulin-like growth factor I	Chromosome 12: 102,395,867-102,480,645	<i>Homo Sapiens</i>	P05019	CAA40342	1B9G
Insulin-like growth factor I receptor	Chromosome 15: 98,648,971-98,964,530	<i>Homo Sapiens</i>	P08069	P08069	1IGR
Insulin-like growth factor binding protein1 + 3 (gene symbol: IGFBP1 + 3)	Chromosome 7: 45,888,357-45,893,668	<i>Homo Sapiens</i>	P08833	P08833	1ZT3
Insulin gene VNTR (gene symbol: INS VNTR)	Chromosome 11: 2,159,779-2,161,341	<i>Homo Sapiens</i>	P01308	P01308	1A7F
Insulin receptor (Marker Locus: INSR)	Chromosome 19: 7,112,255-7,294,034	<i>Homo Sapiens</i>	P06213	AAA59452	1GAG
Insulin receptor (Marker Locus: D19S216)	Chromosome 19: 4,903,080-4,962,154	<i>Homo Sapiens</i>	Q96T88	Q96T88	2FAZ
Insulin receptor (Marker Locus: D19S905)	Chromosome 19: 7,595,902-7,618,304	<i>Homo Sapiens</i>	Q9P1Y5.2	Q9P1Y5	–
Insulin receptor (Marker Locus: D19S884)	Chromosome 19: 8,065,402-8,149,846	<i>Homo Sapiens</i>	Q75N90	Q75N90	–
Insulin receptor (Marker Locus: D19S391)	Chromosome 19: 8,520,790-8,577,577	<i>Homo Sapiens</i>	O00160	O00160	–
Insulin receptor (Marker Locus: D19S906)	Chromosome 19: 8,580,242-8,610,735	<i>Homo Sapiens</i>	Q9H324.2	Q9H324	–
Insulin receptor (Marker Locus: D19S840)	Chromosome 19: 13,906,201-13,930,879	<i>Homo Sapiens</i>	Q6P1N0	Q6P1N0	–
Insulin receptor (Marker Locus: D19S212)	Chromosome 19: 18,207,961-18,255,419	<i>Homo Sapiens</i>	Q08493	Q08493	1LXU
Insulin receptor (Marker locus: D19S410)	Chromosome 19: 17,281,645-17,287,646	<i>Homo sapiens</i>	Q8NAG6	Q8NAG6	–
Insulin receptor substrate 1	Chromosome 2: 226,731,317-226,799,759	<i>Homo Sapiens</i>	P35568	NP_005535	1IRS
Peroxisome proliferator-activated receptor-gamma	Chromosome 3: 12,287,368-12,434,356	<i>Homo Sapiens</i>	P37231	P37231	1FM6
Mothers against decapentaplegic homolog 4 (gene symbol: MADH4)	Chromosome 18: 51,028,394-51,085,045	<i>Homo sapiens</i>	Q13485	Q13485.1	1DD1
Androgen receptor (Marker Locus: AR)	Chromosome X: 67,544,032-67,730,619	<i>Homo Sapiens</i>	P10275	P10275	1E3G
CYP11A-cytochrome P450 side-chain cleavage enzyme Marker Locus: (D15S519)	Chromosome 15: 74,337,759-74,367,740	<i>Homo Sapiens</i>	P05108	P05108	3N9Y
CYP17-cytochrome P450 17a-hydroxylase/17,20-desmolase Marker Locus: (D10S192)	Chromosome 10: 102,830,531-102,837,533	<i>Homo Sapiens</i>	P05093	P05093	2C17
CYP19-cytochrome P450 aromatase Marker Locus: (CYP19)	Chromosome 15: 51,208,057-51,338,610	<i>Homo Sapiens</i>	P11511	P11511	1TQA
17 b-hydroxysteroid dehydrogenase, type I Marker Locus: (D17S934)	Chromosome 17: 42,549,214-42,555,213	<i>Homo Sapiens</i>	P14061	P14061	1A27
17 b-hydroxysteroid dehydrogenase, type II Marker Locus: (HSD17B2)	Chromosome 16: 82,035,004-82,098,534	<i>Homo Sapiens</i>	P37059	P37059	–
17 b-hydroxysteroid dehydrogenase, type III Marker Locus: (D9S1809)	Chromosome 9: 96,235,306-96,302,152	<i>Homo Sapiens</i>	P37058	P37058	–
3 b-hydroxysteroid dehydrogenase, type I and II Marker Locus: (D1S514)	Chromosome 1: 119,507,198-119,515,054	<i>Homo Sapiens</i>	P14060	P14060	–
Steroidogenic acute regulatory protein Marker Locus: (D8S1821)	Chromosome 8: 38,143,649-38,151,265	<i>Homo Sapiens</i>	P49675	P49675	1IMG
Activin receptor 1 Marker Locus: (D12S347)	Chromosome 10: 102,479,229-102,502,711	<i>Homo Sapiens</i>	P61163	P61163	–
Activin receptor 2A Marker Locus: (D2S2335)	Chromosome 2: 147,844,517-147,930,826	<i>Homo Sapiens</i>	P27037	P27037	3Q4T
Activin receptor 2B Marker Locus: (D3S1298)	Chromosome 3: 38,453,851-38,493,142	<i>Homo Sapiens</i>	Q13705	Q13705	2H6U
Inhibin C Marker Locus: (D12S1691)	Chromosome 12, NC_000012.12 (57434685.57452062)	<i>Homo Sapiens</i>	P55103	P55103	–
Follistatin Marker Locus: (D5S474)	Chromosome 5: 53,480,409-53,487,134	<i>Homo Sapiens</i>	P19883	P19883	–

2.1.4. HSD17B1 & HSD17B2

They belong to the group of alcohol oxidoreductase, which catalyze the dehydrogenation of 17-hydroxysteroids in steroidogenesis. They include the interconversion of androstenedione and testosterone, DHEA and androstenediol and estrone and estradiol. A higher level of expression of mRNA synthesizing and inactivating enzyme in women has been reported without PCOS endometrial treatment [5].

2.1.5. HSD3B1 & HSD3B2

The type I 3β -HSD isoenzyme is expressed in placenta and peripheral tissues, whereas the type II 3β -HSD isoenzyme is expressed in the adrenal gland, ovary, and testis. (HSD3B) deficiency in hyperandrogenic females (HF) is related to insulin-resistant polycystic ovary syndrome (PCOS) [5].

2.1.6. StAR

The Steroidogenic Acute Regulatory protein is also known as StAR. StAR is a kind of a transport protein, which transports cholesterol within the mitochondria [6]. In some patients, PCOS is caused maybe due to the defect of steroidogenesis, which results in the increased level of the ovary and adrenal androgen productions. The steroidogenesis process is initiated by the action of steroidogenic acute regulatory protein (StAR) [6].

2.2. Steroid hormone actions

2.2.1. Androgen receptor

Hypersecretion of androgen is one of the most common characteristics in PCOS. This situation results in an excess of androgen production by the ovary. It is known as Hyper-Androgenism, which is the second

important characteristic leading to PCOS. 17% to 83% of women are in prevalence to this disorder [7].

2.2.2. SHBG

Patients with hyperandrogenism and PCOS have a low level of Serum Sex Hormone-Binding Globulin (SHBG). PCOS results in Hyperinsulinemia, which causes a decrease in the levels of SHBG. The synthesis of SHBG in the liver is suppressed [9].

2.3. Gonadotropin action and regulation

2.3.1. LH, FSH

Inappropriate gonadotropin secretion is an important characteristic of PCOS. An elevated level of LH is one of the common reasons of PCOS. Females affected with PCOS have high LH secretion and low FSH secretion. The ratio used to indicate abnormal gonadotropin secretion is normally 2–3/1 [8]. PCOS is usually characterized by low levels of follicle stimulating hormone (FSH). It is responsible for stimulating the growth of follicles in the ovaries. They contain maturing eggs. If FSH were absent for a longer period of time then the follicle would not mature and would not release eggs. Thus, this would result in infertility. The immature follicles in ovaries will lead to the production of small cysts [10].

2.3.2. Inhibin β A and Inhibin β B

PCOS is a syndrome associated with insulin resistance. Inhibin is a heterodimer responsible for the regulation of FSH secretion [11]. The increase in FSH concentrations can be suppressed by the release of inhibin. It has got two variants Inhibin A and Inhibin B. Gonads, pituitary gland, placenta etc. secrete it. Inhibin B is more important than Inhibin

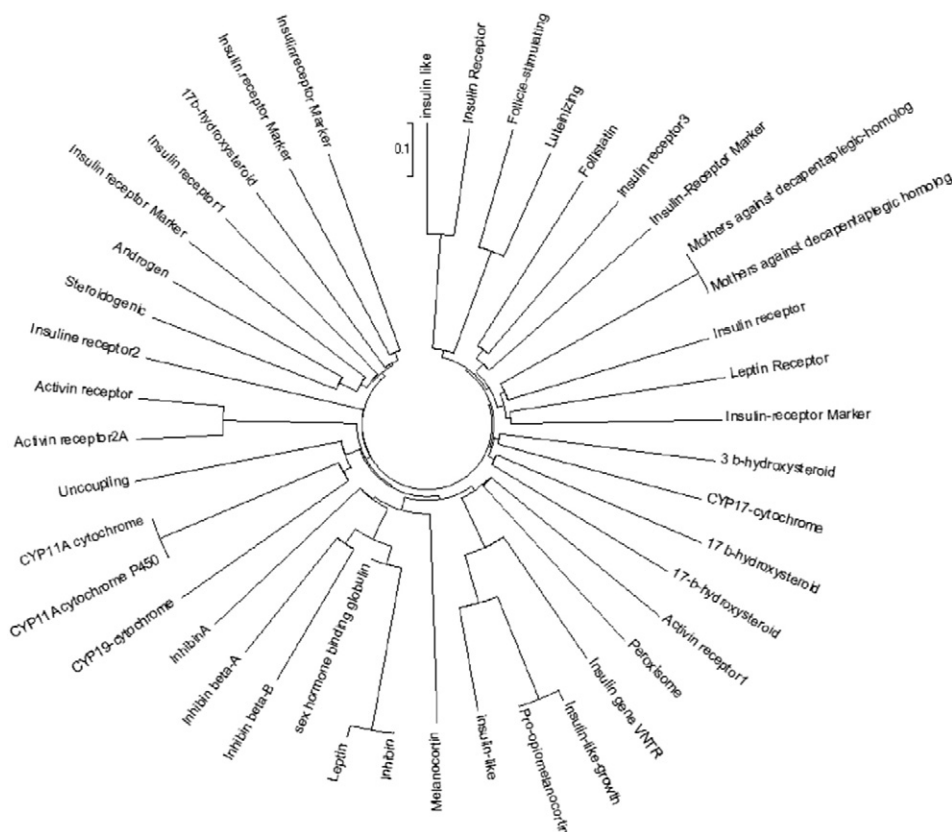


Fig. 1.1. Unrooted phylogenetic tree representing 43 genes.

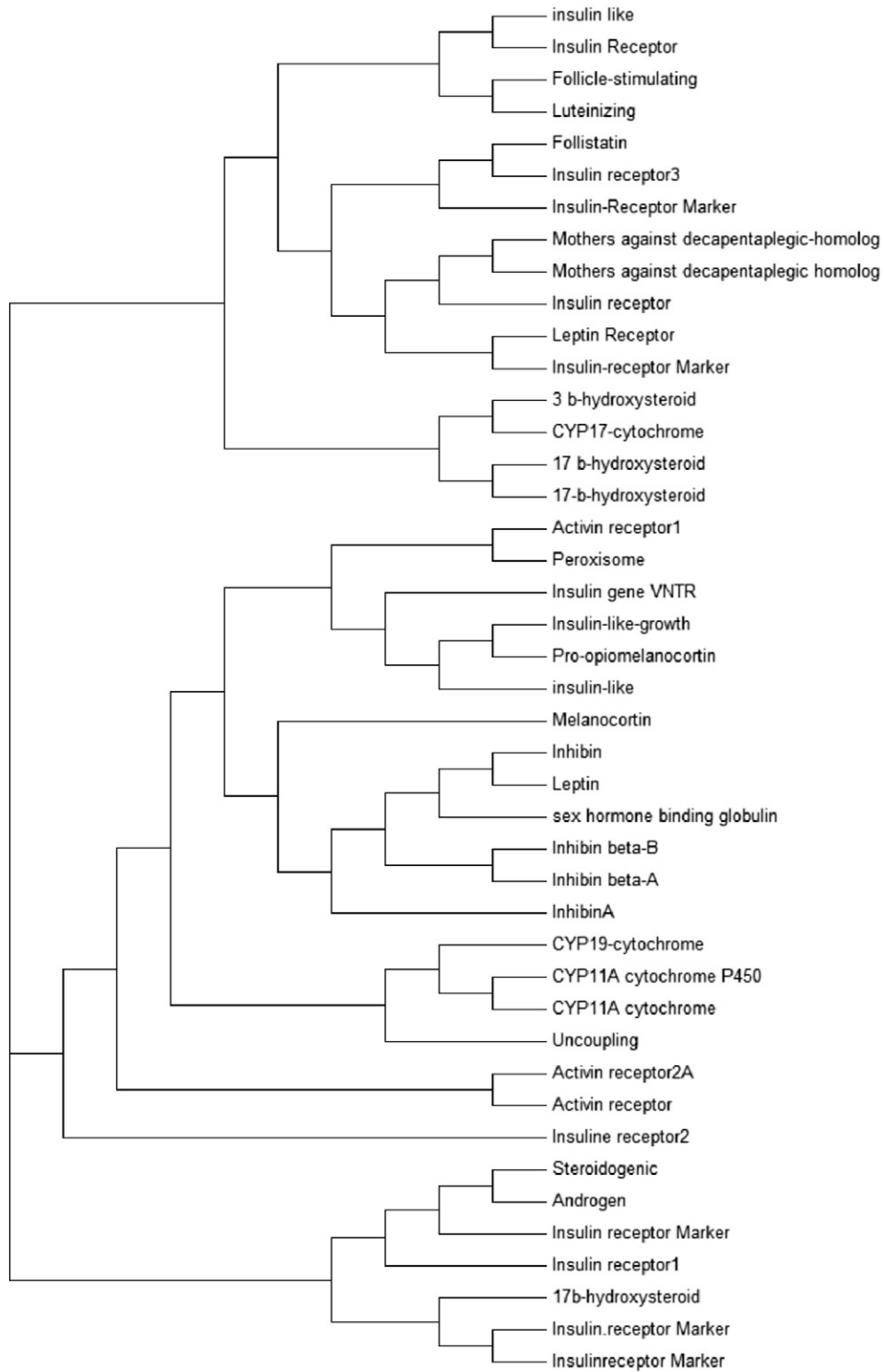


Fig. 12. Rooted phylogenetic tree of 43 genes involved in PCOS.

A during the follicular phase. Woman with PCOS possesses high inhibin level than the normal ones [10].

2.3.3. MADH4

Mothers against decapentaplegic homolog 4 are a protein involved in signaling in mammals. The protein belongs to SAMD family. SAMD4 has two functional domains MH1 and MH2 consist of a tridimensional structure (Regions M and H represent MAD

homology). This resembles similarity between SAMD4 in mammals and Drosophila protein.

2.4. Insulin action and secretion

2.4.1. Insulin and IGF-1

Insulin and IGF-1 are responsible for simulating ovary growth. They increase the action of gonadotropins on ovary steroid synthesis. Insulin

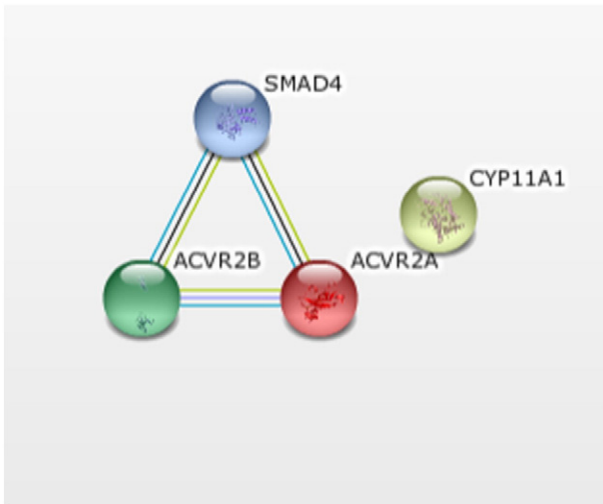


Fig. 2. String analysis of closely identified proteins.

is responsible for augmenting the concentration of IGF-I and androgens. It does this by regulating the synthesis of IGFBP-1 and SHBG in the liver. Resistance to insulin is one of the common symptoms of PCOS. Increase in insulin level and IGFBP-1 activity can be considered as the important reasons resulting in PCOS [12].

2.4.2. Insulin VNTR gene, IGF-II, insulin receptor gene and insulin receptor substrate

The insulin gene consists of variable number tandem repeat (VNTR) present at the 5' regulatory region. Polymorphism of VNTR is responsible for regulation of the transcriptional rate of insulin. It also regulates the gene encoding IGF-II. The Class-I alleles are made up of a length of 40 repeats and Class-II alleles are composed of 80 repeats. PCOS results in insulin resistance and may directly have an effect on pancreatic β -cell. VNTR polymorphism has a wide effect on the insulin resistance in some PCOS phenotypes. SNP at tyrosine kinase domain of INSR is found to be associated with PCOS [10].

2.4.3. PPAR- γ

PPAR γ is also known as Peroxisome proliferator-activated receptor gamma. It is an important nuclear transcription factor. It is involved in regulation of glucose, lipid metabolism, and ovary steroidogenesis. The most extensive findings on polymorphism in PPAR γ encompass Proline and Alanine in exon B. One more polymorphism studied in PPAR γ gene is C1431T in exon6. This variation is associated with PCOS. PPAR γ gene has an influence on insulin resistance path physiology in women affected with PCOS [13].

2.5. Obesity and energy regulation

2.5.1. Leptin and leptin receptor

Leptin plays an essential role in the pathological process of PCOS. Obese PCOS affected females the leptin and the free leptin index are higher than thin PCOS subjects. PCOS results in elevation of free leptin index and decrease in leptin receptor. Leptin and leptin receptors are associated with PCOS, which are dependent on BMI [14–15].

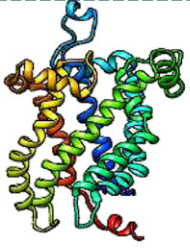


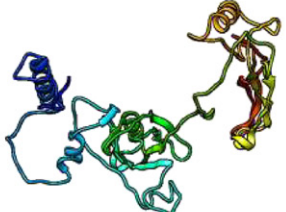

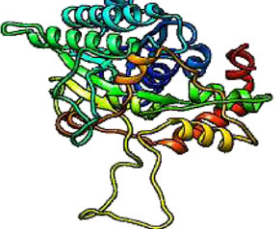
2.5.2. POMC

POMC is also known as Pro-opiomelanocortin fragments. They are used for identification of factors resulting in excessive adrenal androgen secretion. POMC is a 16K fragment [16].

2.5.3. UCP2 + 3

UCP2 stands for uncoupling protein, which is responsible for androgen synthesis of granulosa cells from PCOS, affected patients. The increase in expression of ovary UCP2 was identified when treated with

Table 2
Modeled structures of proteins related to PCOS using Modeller 9v14 and Raptor X.

UniprotKB ID	Modeled structures
P55851	 Modeller 9v14
Q75N90	 Modeller 9v14
P01189	 Raptor X
P05111	 Raptor X
P09529	 Raptor X
P14060	 Raptor X

(continued on next page)

Table 2 (continued)


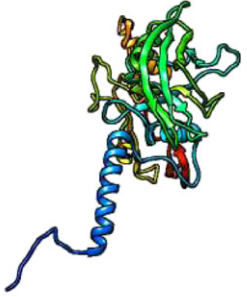







P37059		Raptor X
P55103		Raptor X
Q6PN10		Raptor X
Q8NAG6		Raptor X
Q9H324		Raptor X

Table 2 (continued)

Q9P1Y5		Raptor X
O00160		Modeller9v14
P37058		Modeller9v14
P61163		Modeller9v14

T3 (triiodothyronine) in PCOS. This may alter pregnenolone synthesis resulting in influencing P450scc expression. This thus affects testosterone production [17–18].

3. Collection of genes related to PCOS

Table 1 consists of 43 genes that are related to the polycystic ovary syndrome (PCOS) also known as Hyperandrogenic Ovulation (HA).

PCOS is a disorder that affects women due to the chronic ovulation & hyperandrogenism; PCOS is also a major cause of infertility that is observed in women [18]. Chromosomal location is retrieved from the Ensembl database [19]. The UniprotKB ID is taken from UniprotKB [20]. The retrieval of protein sequences & the accession number is done from National Center for Biotechnology Information (NCBI) [21].

4. Phylogenetic analysis

All the 43 protein sequences of genes related to PCOS were aligned using Clustal Omega and an unrooted tree is generated as shown in (Fig. 1.1) [22]. Phylogenetic analysis was done from the cladogram generated by the alignment (Fig. 1.2). Visualizing the cladogram using tree viewer identified closely related genes. A cladogram is used to show

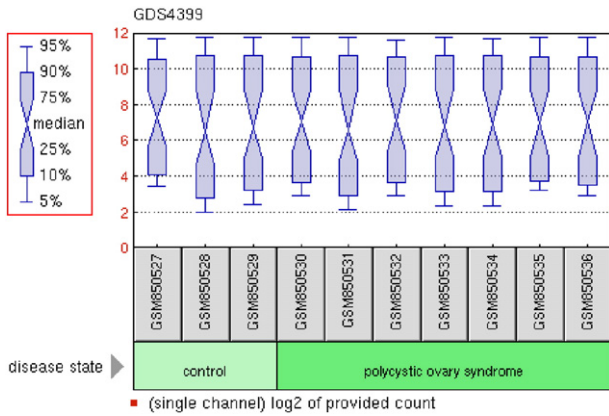


Fig. 3. GDS4399 polycystic ovary syndrome: granulosa cell samples.

relatedness between all the responsible genes for PCOS. A clade is a group of species used in cladogram to represent genes having the same ancestral origin. The closely related proteins identified were CYP11A-cytochrome P450 side-chain cleavage enzyme (P05108), Sex hormone binding globulin (ABY68008), Mothers against decapentaplegic homolog 4 (NP_005350), Mothers against decapentaplegic homolog 4 (Q13485.1), Activin receptor 2A Marker Locus (P27037), and Activin receptor 2B Marker (Q13705), Mothers against decapentaplegic homolog 4 (NP_005350), and Mothers against decapentaplegic homolog 4 (Q13485.1) are found to be equidistant from a root node and hence can be considered as related to each other. Similarly for P27037, Q13705, P05108, ABY68008. The relatedness between all the identified genes can be further studied using string analysis.

5. String analysis

The closely related genes that we identified from Clustal Omega were analyzed using string analysis server and the string network was obtained (Fig. 2) [23]. In the above string network analysis, all the nodes like ACV2A represent Activin A receptor, type IIA, ACVR2B represents Activin A receptor, type IIB, CYP11A1 represents cytochrome P450, family 11, subfamily A, polypeptide 1, SMAD4 represents SMAD family member 4. This network forms a triangular interaction between SMAD4, ACVR2A and ACVR2B within the nodes of the different colors of interaction lines representing a different form of evidence such as the yellow line represents text-mining evidence, the red line indicates the presence of fusion evidence, light blue line represents database evidence. CYP11A1 is a protein, which does not have any interaction with the other proteins.

6. Tertiary structures

All the 43 genes listed in (Table 1) were searched for existing 3D structures in protein data bank [24] and listed with their PDB ID in

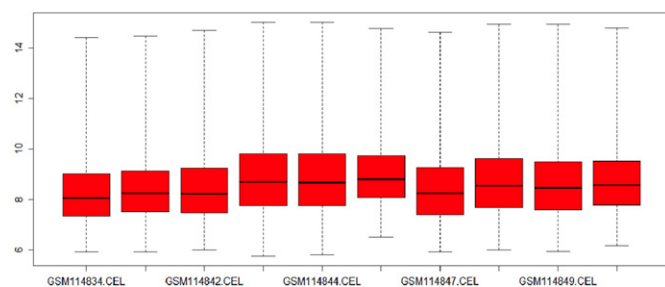


Fig. 3.1. Box plot of dataset GSE34526 before normalization.

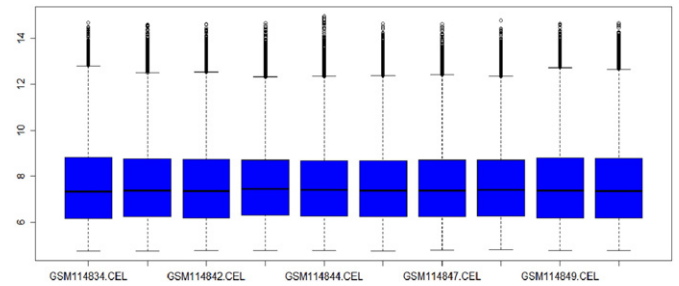


Fig. 3.2. Box plot of dataset GSE34526 after normalization.

tabular format. Out of the 43 genes, 15 genes had no significant structures present in PDB and hence they were modeled using different modeling techniques. Based on the BLASTp [25] score obtained for individual genes the way of modeling the 3D structure was done using Modeller and Ab-initio programs. Sequences having low identity and query coverage were modeled using Raptor X [26], a server for ab-initio modeling of protein tertiary structures. While those sequences having good blast score were modeled using Modeller v9.14 [27]. (Table 2) consists of genes with their modeled structures along with the tool used for modeling.

7. Microarray analysis of PCOS genes

7.1. Gene datasets

The Gene Expression Omnibus (GEO) [31] is a database consists of microarray, next generation sequencing and other forms of high-throughput genomic data that archives and freely available for scientific research purposes. PCOS gene dataset was obtained from GEO viz. GSE34526 (polycystic ovary syndrome: granulosa cells) [28] sample count is equal to 10 as shown in (Fig. 3). The platform of this dataset is GPL570: [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. This database stores curated gene expression datasets, as well as original series and platform records in the Gene Expression Omnibus (GEO) repository. Dataset records contain additional resources including cluster tools and differential expression queries.

7.2. Analyzing the datasets

R is a freely available software environment, which runs on wide variety of UNIX platforms, Windows and MacOS [29]. It is used to perform statistical computing and graphical techniques. The various techniques that can be performed using R are linear and nonlinear modeling,

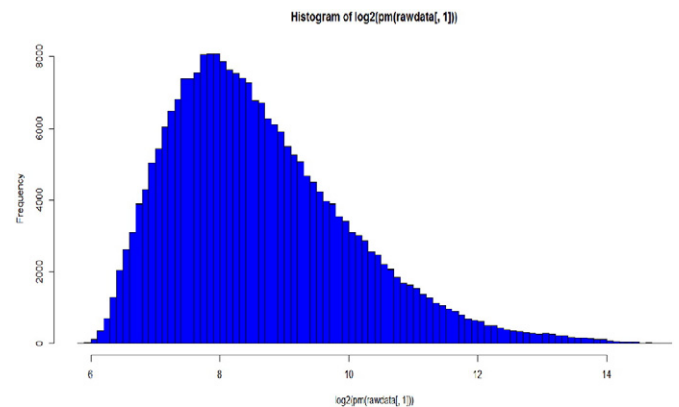


Fig. 3.3. Histogram plot of GSE34526.

Table 3Microarray analysis results: significant genes found from Volcano Plot using *t*-test.

Significant genes annotation ID	Gene names	Gene	Chromosomal location	Regulation
200914_x_at	Kinectin 1 (kinesin receptor)	KTN1	chr14q22.1	(+)
211126_s_at	Cysteine and glycine-rich protein 2	CSRP2	chr12q21.1	(-)
208852_s_at	Calnexin	CANX	chr5q35	(+)
203951_at	Calponin 1, basic, smooth muscle	CNN1	chr19p13.2-p13.1	(+)
210087_s_at	Myelin protein zero-like 1	MPZL1	chr1q24.2	(+)
212008_at	UBX domain protein 4	UBXN4	chr2q21.3	(+)
217889_s_at	Cytochrome b reductase 1	CYBRD1	chr2q31.1	(+)
212713_at	Microfibrillar-associated protein 4	MFAP4	chr17p11.2	(-)
202216_x_at	Nuclear transcription factor Y, gamma	NFYC	chr1p32 1p32	(+)
215346_at	CD40 molecule, TNF receptor superfamily member 5	CD40	chr20q12-q13.2	(+)
219868_s_at	Ankyrin repeat and FYVE domain containing 1	ANKFY1	chr17p13.3	(-)
214544_s_at	Synaptosomal-associated protein, 23kDa	SNAP23	chr15q14	(+)
210756_s_at	Notch 2	NOTCH2	chr1p13-p11	(-)
209196_at	WD repeat domain 46	WDR46	chr6p21.3	(-)

classical statistical tests, time-series analysis, classification, clustering, and others.

Bioconductor is free open source software, which is widely used in order to analyze genomic data, generated from wet lab experiments in molecular biology. It is based on statistical R programming. Multi Experiment Viewer (MeV) [30] is an application, which is used to view processed microarray slide representations. It also identifies genes and expression pattern of interest. The validation of genes obtained from SAM and *t*-test are back validated using this crude analysis. Up-regulated and down-regulated genes were selected from the data obtained from the tests performed.

7.3. Box plot analysis

Box plot analysis is a widely used method that provides information about the spread and skewness in the dataset. It is used for visualizing distribution of data values throughout the dataset. Fig. 3.1 and Fig. 3.2 represent the box plot analysis of the gene dataset GSE34526 as before and after normalization.

7.4. Histogram analysis

A histogram is a way of representing a statistical graph having vertical rectangles of different heights, which are proportionate to the corresponding frequencies. It is a graph used for the frequency distribution. Fig. 3.3 represents the histogram curves of the dataset. The microarray data shows histogram as frequency on the y-axis and log transform of pixel intensities along the x-axis.

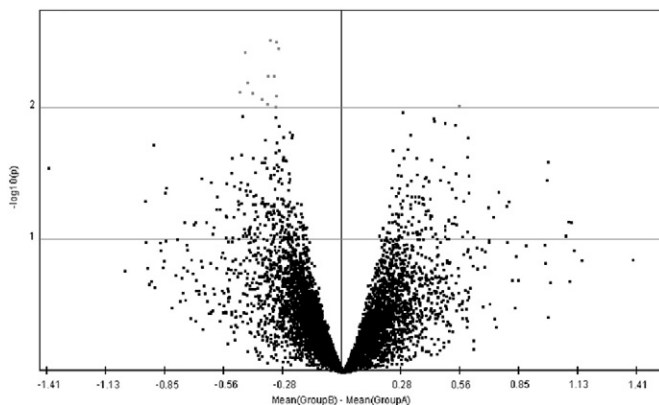


Fig. 3.4. Volcano plot for finding significant and non-significant genes using *t*-test. Red color represents the significant genes and black color represents non-significant genes.

7.5. SAM results & *t*-test analysis

The SAM test analysis (Fig. 3.5) and *t*-test analysis (Fig. 3.6) of the properly normalized datasets indicated that there are some up-regulated as well as down-regulated genes in the datasets. A list of up-regulated and down-regulated genes, which were common in all the datasets were represented in a tabular format (Table 3). The volcano plot was generated using the normalized data to differentiate the significant with non-significant genes as shown in Fig. 3.4.

8. Conclusions

Although the genetics and mechanism of PCOS are not yet understood, computational tools may be helpful in finding the cause of this syndrome using various structural aspects. Here in our work, the proteins responsible as reported earlier were subjected to computational modeling using various bioinformatics tools e.g. Modeller. The proteins whose structural homologous were not significant, were subjected to Ab-initio modeling. 43 protein-coding genes were analyzed in structural viewpoint, which could assist computational biologists to carry out further aspects of research. The relatedness of the 43 protein-coding genes was analyzed using Clustal Omega, a phylogeny tool to find out the closely related proteins responsible for PCOS. The work was not restricted to the above said aspects, but statistical analysis of PCOS dataset retrieved from GEO datasets from NCBI (10 samples 3 control and 7 diseased states) were taken into consideration to have an in-sight into the genes that are responsible for PCOS apart from 43 reported proteins. Sam test and *t*-test differentiate the significant from the non-significant

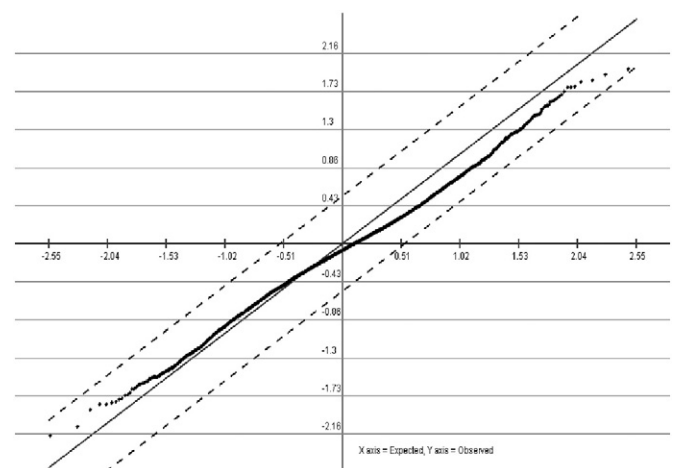


Fig. 3.5. Sam test from MeV using normalized PCOS dataset.

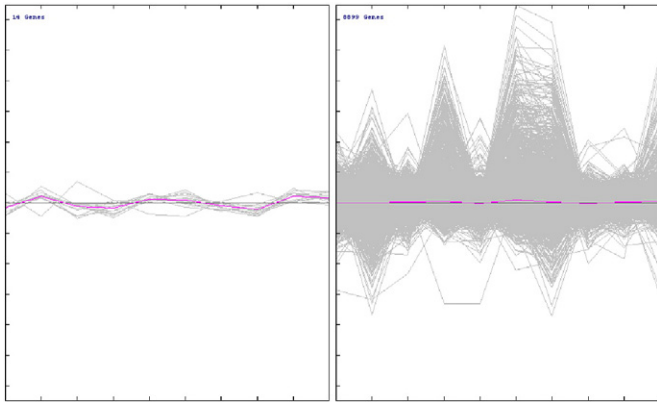


Fig. 3.6. Plotting of significant and non-significant genes using MeV using normalized PCOS dataset.

genes from the dataset. 14 significant genes were identified and were retrieved from NCBI and Uniprot. Out of the 14 genes, 8 genes were up-regulated genes and the rest 6 genes were reported as down-regulated genes as seen in Table 3. Thus, in conclusion, the computational study of genetic disorders like PCOS can be scrutinized in both sequential and structural aspects to cognize the mechanism of such type of genetic syndromes.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- R. Azziz, D.A. Dumesic, M.O. Goodarzi, Polycystic ovary syndrome: an ancient disorder? *Fertil. Steril.* 95 (5) (2011) 1544–1548, <http://dx.doi.org/10.1016/j.fertnstert.2010.09.032>.
- U.S. Department of Health and Human Services, Office of Women's Health, Office of Women's Health, Polycystic ovary syndrome (PCOS) fact sheet. 2010 Retrieved April 24, 2012, from <http://www.womenshealth.gov/publications/our-publications/fact-sheet/polycystic-ovary-syndrome.html>.
- B. Joshi, S. Mukherjee, A. Patil, A. Purandare, S. Chauhan, R. Vaidya, A cross-sectional study of polycystic ovarian syndrome among adolescent and young girls in Mumbai, India. *Indian J. Endocrinol. Metab.* 18 (3) (2014) 317–324, <http://dx.doi.org/10.4103/2230-8210.131162>.
- N. Prapas, A. Karkanaki, I. Prapas, I. Kalogiannidis, I. Katsikis, D. Panidis, Genetics of polycystic ovary syndrome. *Hippokratia* 13 (4) (2009) 216–223.
- C.H. Blomquist, Kinetic analysis of enzymic activities: prediction of multiple forms of 17 beta-hydroxysteroid dehydrogenase. *J. Steroid Biochem. Mol. Biol.* 55 (5–6) (1995) 515–524, [http://dx.doi.org/10.1016/0960-0760\(95\)00200-6](http://dx.doi.org/10.1016/0960-0760(95)00200-6).
- Kahsar-Miller, B.A. Conway-Myers, L.R. Boots, R. Azziz, Steroidogenic acute regulatory protein (StAR) in the ovaries of healthy women and those with polycystic ovary syndrome. *Am. J. Obstet. Gynecol.* 185 (6) (2001 Dec) 1381–1387.
- Jennifer R. Wood, Velen L. Nelson, Clement Ho, Erik Jansen, Clare Y. Wang, Margrit Urbanek, Jan M. McAllister, Sietse Mosselman, Jerome F. Strauss, The molecular phenotype of polycystic ovary syndrome (PCOS) theca cells and new candidate PCOS genes defined by microarray analysis. *J. Biol. Chem.* 278 (2003) 26380–26390, <http://dx.doi.org/10.1074/jbc.M300688200>.
- Mohammad Hasan Sheikhha, Seyed Mehdi Kalantar, Nasrin Ghasemi, Genetics of polycystic ovary syndrome. *Iran. J. Reprod. Med.* 5 (1) (Winter 2007) 1–5.
- V. Jayagopal, E.S. Kilpatrick, P.E. Jennings, D.A. Hepburn, S.L. Atkin, The biological variation of testosterone and sex hormone-binding globulin (SHBG) in polycystic ovary syndrome: implications for SHBG as a surrogate marker of insulin resistance. *J. Clin. Endocrinol. Metab.* 88 (4) (2003 Apr) 1528–1533, <http://dx.doi.org/10.1210/jc.2002-020557>.
- Melissa H. Hunter, James J. Sterrett, Polycystic ovary syndrome: it's not just infertility. *Am. Fam. Physician* 62 (5) (Sep 1, 2000) 1079–1088.
- Pascal Pigny, Rachel Desailoud, Christine Cortet-Rudelli, Alain Duhamel, Delphine Deroubaix-Allard, André Racadot, Didier Dewailly, Serum α -inhibin levels in polycystic ovary syndrome: relationship to the serum androstenedione level. *J. Clin. Endocrinol. Metab.* 82 (6) (Jun 1997) 1939–1943, <http://dx.doi.org/10.1210/jcem.82.6.4015>.
- F. Nobels, D. Dewailly, Puberty and polycystic ovary syndrome: the insulin/insulin-like growth factor I hypothesis. *Fertil. Steril.* 58 (4) (Oct 1992) 655–666 PMID: 1426306.
- Nuzhat Shaikh, Roshan Dadachanji, and Srabani Mukherjee, Genetic markers of polycystic ovary syndrome: emphasis on insulin resistance. *Int. J. Med. Genet.*, vol. 2014 (2014), Article ID 478972. DOI: <http://dx.doi.org/10.1155/2014/478972>.
- Yatap-Dong, Bundang-Gu, Seongnam-Si, Gyeonggi-Do. Department of Biomedical Science, CHA University, Bundang CHA Hospital, 502 463-840, Republic of Korea, *Gene* 527 (1) (Sep 15, 2013) 71–74, <http://dx.doi.org/10.1016/j.gene.2013.05.074>.
- Nasser M. Rizk and Elham Sharif, Leptin as well as free leptin receptor is associated with polycystic ovary syndrome in young women, *Int. J. Endocrinol.*, vol. 2015 (2015), Article ID 927805. <http://dx.doi.org/10.1155/2015/927805>.
- S.K. Cunningham, T. Loughlin, X. Bertagna, F. Girard, T.J. McKenna, Plasma pro-opiomelanocortin fragments and adrenal steroids following administration of metyrapone to normal and hirsute women. *J. Endocrinol. Investig.* 11 (4) (1988 Apr) 247–253.
- Liu Y 1, Jiang H, Xing FQ, Huang WJ, Mao LH, He LY, Uncoupling protein 2 expression affects androgen synthesis in polycystic ovary syndrome, *Endocrine*. 2013 Jun; 43(3):714–23. doi: <http://dx.doi.org/10.1007/s12020-012-9802-0>.
- Margrit Urbanek, Richard S. Legro, Deborah A. Driscoll, Ricardo Azziz, David A. Ehrmann, Robert J. Norman, Jerome F. Strauss III, Richard S. Spielman, Andrea Dunaif, Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. *Proc. Natl. Acad. Sci. USA* 96 (July 1999) 8573–8578.
- A. Yates, K. Beal, S. Keenan, et al., The Ensembl REST API: ensembl data for any language. *Bioinformatics* 31 (1) (2015) 143–145, <http://dx.doi.org/10.1093/bioinformatics/btu613>.
- The UniProt Consortium, UniProt: a hub for protein information. *Nucleic Acids Res.* 43 (2015) D204–D212.
- The NCBI handbook [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2002 Oct. Chapter 18, The Reference Sequence
- F. Sievers, A. Wilm, D. Dineen, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7 (2011) 539, <http://dx.doi.org/10.1038/msb.2011.75>.
- D. Szklarczyk, A. Franceschini, S. Wyder, et al., STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43 (Database issue) (2015) D447–D452, <http://dx.doi.org/10.1093/nar/gku1003>.
- H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank. *Nucleic Acids Res.* 28 (2000) 235–242.
- S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* 215 (1990) 403–410.
- Morten Kallberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, Jinbo Xu, Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7 (8) (2012) 1511–1522.
- N. Eswar, M.A. Marti-Renom, B. Webb, M.S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, Comparative protein structure modeling with Modeller. *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc. 2006, pp. 5.6.1–5.6.30 Suppl. 15.
- S. Kaur, K.J. Archer, M.G. Devi, A. Kriplani, et al., Differential gene expression in granulosa cells from polycystic ovary syndrome patients with and without insulin resistance: identification of susceptibility gene sets through network analysis. *J. Clin. Endocrinol. Metab.* 97 (10) (2012 Oct) E2016–E2021 PMID: 22904171.
- R Development Core Team (2008). R: A Language and Environment For Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL (<http://www.R-project.org>).
- A.I. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Z. Liu, A. Vinsavich, V. Trush, J. Quackenbush, TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34 (2) (2003 Feb) 374–378.
- R. Edgar, M. Domrachev, A.E. Lash, Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30 (1) (2002 Jan 1) 207–210.