RESEARCH ARTICLE

In silico characterization of functional SNP within the oestrogen receptor gene

MAHA REBAÏ* and AHMED REBAÏ*

Laboratory of Molecular and Cellular Screening Processes, Centre of Biotechnology of Sfax, University of Sfax, Route Sidi Mansour, P.O. Box 1177, 3018 Sfax, Tunisia

Abstract

Single-nucleotide polymorphism (SNP) association studies have become crucial in uncovering the genetic correlations of genomic variants with complex diseases, quantitative traits and physiological responses to drugs. However, the identification of SNPs responsible for specific phenotypes is a difficult problem to solve, requiring multiple testing of hundreds or thousands of SNPs in candidate genes. In this study, we performed an analysis of the genetic variations that can alter the structure and function of oestrogen receptor α using different computational tools. Among the nonsynonymous SNPs, a total of four SNPs were found to be damaging by both a sequence homology-based tool (SIFT) and a structural homology-based method (polyphen-2, SNAP), as well as by the ESEfinder program, and one nonsense nsSNP was found. For noncoding SNPs, we found that one SNP in 5′UTR may potentially change protein expression level, nine SNPs were found to affect miRNA binding site and 28 SNPs might affect transcriptional regulation of the *ESR1* gene. Reviewing the literature, 89 SNPs were found to be functional among which only four were located in exons.

[Rebaï M. and Rebaï A. 2016 *In silico* characterization of functional SNP within the oestrogen receptor gene. *J. Genet.* **95**, 865–874]

Introduction

Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variation in the human genome that are widely used in associations studies with complex diseases and quantitative traits. In fact, SNPs occur in every 100–300 bases along the human genome and represent 90% of all genetic variations. They are found in both coding (gene) and noncoding regions of the genome with different densities (Lee *et al.* 2005).

In the human genome, SNPs are abundant in noncoding regions, including regulatory and untranslated (UTR) regions as well as introns. Single-nucleotide variations may affect phenotypic functions and eventually contribute to disease development. SNPs in the regulatory regions may influence gene expression or transcription factor binding (Stenson *et al.* 2009; Kimura-Kataoka *et al.* 2012), while SNPs of the UTR regions may modify the transcriptional activity (Milanese *et al.* 2007), RNA stability and ribosomal translation of mRNA (Boffa *et al.* 2008). Polymorphisms in coding DNA, especially those that lead to an amino acid residue change in the protein product (nonsynonumous, nsSNP) are of particular importance as they are responsible for nearly half of the known genetic variations related to human inherited disease (Hampe *et al.* 2007).

SNP association studies are useful in the recognition of genomic variants that are responsible for specific phenotypes, such as complex diseases, quantitative traits and physiological responses to drugs (Andrawiss 2005; Takahashi *et al.* 2012). However, identifying these SNPs among the thousands of SNPs in candidate genes is a difficult problem to solve (Zhernakova *et al.* 2009). Thus, it is essential to make a preselection of the SNPs to be analysed according to their functional significance (Ilhan and Tezel 2013; Patnala *et al.* 2013) by using bioinformatics prediction tools, which allows to differentiate between neutral SNPs and those of likely functional importance.

Oestrogen receptor α (ERα) is a prototype of nuclear transcription factors that regulate the expression of target genes. This protein controls a diverse set of essential functions such as reproduction, development and proliferation (Knobil and Neill 1994). ERα is encoded by the *ESR1* gene and different promoters have been identified in this gene (Kos *et al.* 2001). Their resulting transcripts differ in their 5′UTR but not in their coding regions. Among the

*For correspondence. E-mail: Maha Rebaï, maha.rebai@cbs.rnrt.tn; Ahmed Rebaï, ahmed.rebai@cbs.rnrt.tn.

Keywords. *ESR1* gene; polymorphism; prediction; bioinformatic tools; association studies.

described variants, four lead to a functional protein ERα1 (NM_000125.3, NM_001122740.1, NM_001122741.1 and NM_001122742.1).

In the recent years, a multitude of SNPs in *ESR1* gene have been identified, and many associations studies have determined the importance of some SNPs in different hormone-dependent diseases (Ding *et al.* 2010; Kim *et al.* 2011; Paskulin *et al.* 2013). Yet, it has been reported that some polymorphisms of the *ESR1* gene may affect the enhancer activity of gene (Maruyama *et al.* 2000) or gene regulation (Albagha *et al.* 2001). However, few studies have been able to determine the mechanism by which these polymorphisms may act in disease development. Also, there are a large number of SNPs in *ESR1* gene which have not been studied and which may be implicated in the aetiology of several diseases. Therefore, there is a need to achieve SNPs prioritization of the *ESR1* gene and to determine the functional significance of these polymorphisms.

In this study, we present an *in silico* characterization of SNPs that can affect the function of *ESR1* gene. Nine different computational tools (Polyphen-2, SIFT, PROVEAN, SNAP, INPS, Net-O-Glyc, ESEfinder, UTRscan and TFsearch) with good prediction performance (sensitivity and specificity) are applied to identify candidate SNPs that are likely to affect the protein function and that can be selected in priority for future association studies. These programs are based on different methods and use different data and information including protein sequence, protein structure and protein stability. The use of multiple bioinformatics tools allow a more reliable annotation and classification of variants since those that are similarly predicted by many tools will have stronger evidence about their function.

Materials and methods

Datasets

The NCBI database SNP, dbSNP available at www.ncbi.nim.nih.gov (bluid 141, May 2014) was used to recover the list of candidate SNPs analysed in this study. The Medline database was used to review literature until May 2014 and to extract the SNP association studies regarding the *ESR1* gene.

Prediction of phenotypic effects of nsSNP

The effect of the amino acid substitution on protein function was predicted using Polyphen-2 (http://www.bork.embl-heidelberg.de/PolyPhen/), SIFT (http://blocks.fhcrc.org/sift/SIFT.html), PROVEAN (http://provean.jcvi.org), SNAP (https://www.rostlab.org/services/snap/) and INPS (http://inps.biocomp.unibo.it/inps).

Also, the Net-O-Glyc server (http://www.cbs.dtu.dk/services/NetOGlyc) was used to determine the effect of nsSNPs on glycosylation sites of the ERα protein.

Polyphen-2 tool: Polyphen-2 (polymorphism phenotyping v2) uses a combination of structural and sequence homology analyses to predict whether an amino acid change is likely to be deleterious for the structure and function of the protein (Adzhubei *et al.* 2010). The program searches for 3D protein structures, performs multiple alignments of homologous sequences and analyses in several protein structure databases, the information of amino acid contact. Thus, position-specific independent count (PSIC) scores will be calculated for each of the two amino acid residues and then the PSIC score difference will be computed between them. The higher a PSIC score difference is, the higher is the functional impact of a particular amino acid substitution is likely to have.

SIFT tool: SIFT program is a sequence homology-based tools that predicts the possible impact of an amino acid substitution on protein function. This program uses multiple sequence alignment conservation approach to determine if the substitution is either tolerated by the protein or damaging to the protein. Thus, a normalized tolerance score is calculated. The higher the score, the less functional impact a particular amino acid substitution is likely to occur (Sim *et al.* 2012).

Protein variation effect analyser (PROVEAN) tool: PROVEAN measures the damaging effect of an amino acid change in protein sequences (Choi *et al.* 2012). The software uses a delta alignment score of the protein sequence without and with variation with respect to the alignment of homologous sequences. A score ≤ -2.5, indicates that the change is deleterious.

Screening for nonacceptable polymorphisms (SNAP) tool: SNAP tool evaluates the effects of nonsynonymous substitutions on protein function based on a neural-network method that uses *in silico* derived protein information (e.g. secondary structure, evolutionary information for residue conservation, annotation, solvent accessibility, etc.). The network calculates a score for each substitution and then the scores will be translated into binary predictions of neutral or nonneutral effect (Bromberg and Rost 2007).

Impact of nonsynonymous mutations on protein stability (INPS): INPS allows to determine the possible impact of nonsynonymous polymorphisms on the protein stability from its sequence. INPS predicts the thermodynamic free energy change induced by the amino acid residue change in protein sequences by using two types of characteristics concerning the mutation type (mutability, molecular weight, hydrophobicity, etc.) and the evolutionary information (Fariselli *et al.* 2015).

Net-O-glyc: Net-O-glyc 4.0 server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins. The program analyses the reference sequence as well as those carry the amino acid change. The substitution is predicted to have functional significance,

if different functional patterns are found for each of the sequences analysed. The sites are predicted as glycosylated if they have scores higher than 0.5 (Steentoft *et al.* 2013).

Prediction of SNP effect on splicing events

ESEfinder (http://exon.cshl.edu/ESE/) analyse exonic sequences to predict putative exonic splicing enhancer (ESE) reactive with the four human SR proteins (SF2/ASF, SC35, SRp40 and SRp55) (Cartegni *et al.* 2003). ESEfinder uses weight matrices corresponding to the different human SR protein motifs to identify putative ESE in query sequences. The matrices are based on frequency values obtained by functional systematic evolution of ligands by exponential enrichment (SELEX) experiments and the initial SELEX library, which was made by chemical synthesis.

Prediction of SNP effect in regulatory region

The impact of SNPs in regulatory region was assessed using UTRscan (http://itbtools.ba.itb.cnr.it/utrscan), TFsearch (http://diyhpl.us/~bryan/irc/protocol-online/protocol-cache/TFSEARCH.html) and MicroSNiPER (http://epicenter.ie-freiburg.mpg.de/services/microsniper/).

UTRscan tool: UTRscan was used to predict the functional significance of noncoding SNPs in the UTR regions (Pesole and Liuni 1999). The program compares the query sequences (without and with nucleotide change) with the functional sequence patterns collection (located in 5′UTR and 3′UTR sequences) of the UTR site (Grillo *et al.* 2010). If the UTR SNP has different functional patterns, this polymorphism is predicted to have functional significance.

TFsearch tool: TFsearch was used to predict the effect of SNPs on a putative DNA transcription factor-binding site of the 'TRANSFAC' databases (Heinemeyer *et al.* 1998).

MicroSNiPER tools: MicroSNiPER is a web-based tool that predicts the possible impact of SNPs on putative microRNA (miRNA) binding site (Barenboim *et al.* 2010). This application uses a straightforward method to identify if a 3′UTR SNP will disrupt or create a microRNA binding site. Micro SNiPer offers a high flexibility and simple graphical representation of the results.

Results

The list of candidate SNPs of the *ESR1* gene that was investigated in this work was retrieved from dbSNP database. A total of 362 SNPs were surveyed and prioritized. Among these SNPs, 87 were located in exons (45 were found to be nonsynonymous and 42 to be synonymous) and 275 were located in UTR and regulatory regions.

SNPs of the coding region of the *ESR1* gene were analysed by different programs to select SNPs that have functional significance on the protein structure or function. First, we used the SIFT program to prioritize the 43 nsSNPs of the *ESR1* gene. The protein sequence with mutational position and amino acid residue variants associated to these missense nsSNPs were submitted to the SIFT program. Among the SNPs analysed, 18 variations (41.9%) were found to be deleterious with a tolerance index of 0.05. The highly deleterious indexes were observed for the following SNPs: rs369520220, rs200924028, rs104893956, rs148034868, rs188957694, rs121913044, rs374786087, rs121913043, rs138891155 and rs397509428. It is interesting to note that the SNP rs104893956 is a nonsense nsSNP which leads to a premature stop codon (table 1).

Subsequently, we used the Polyphen-2 program and after computing the PISC score for each nsSNP, 18 nsSNPs (41.9%) were evaluated to significantly affect protein structure and exhibited a range PSIC score difference between 1.5 and 2.6 (table 1). When comparing the results obtained from the evolutionary-based approach realized by the SIFT program and the structural approach effectuated by the Polyphen program, we observed a significant correlation between the two programs for 14 SNPs (32.5%).

SNAP program was also used to assess the effect on protein sequence of the 43 nsSNPs. Among the SNPs analysed, nine (20.9%) were found to be nonneutral (table 1). It is interesting to note that when these three programs were used together, we observed 53% agreement between them.

Nonsynonymous SNPs of the *ESR1* gene were investigated by the PROVEAN program and eight nsSNPs (18.6%) were found to be deleterious. Seven of the eight nsSNPs were predicted to be functional SNPs by both SIFT and Polyphen approaches. An agreement between the four programs used was observed in three nsSNPs (rs121913044, rs121913043 and rs182943916) (table 1).

Then, the stability change of the ERα protein caused by nsSNPs was determined by the IPNS server. Twenty-two nsSNPs (51.16%) showed an increase in energy (DDG $<$ 0 kcal/mol, less favourable change) in comparison with the native structure. The most destabilizing nsSNPs were rs121913044 (V364E, DDG $=$ -2.398 kcal/mol) and rs121913043 (C447A, DDG $=$ -2.343 kcal/mol). This result correlates with the program results of PolyPhen-2, SIFT, SNAP and PROVEAN, which have predicted that these two polymorphic sites to be damaging.

The exonic SNPs of the *ESR1* gene (87 SNPs) were analysed by the ESEfinder program to predict the effect of these variations on the splicing phenomenon. After comparing the functional element of each SNP, 17 SNPs (19.5%) were found to change the functional pattern of putative splicing site (table 2). Among these polymorphisms, eight SNPs (9.2%) induce a splicing site abolition, while the other nine SNPs (10.3%) yield to the creation of an exonic splicing site.

By combining the results of the SIFT, Polyphen-2, SNAP and ESEfinder programs, the following SNPs (4.6%) were found to have a functional significance by all programs at

Table 1. nsSNPs of the *ESR1* gene predicted to have functional significance by Polyphen-2/SIFT/PROVEAN/SNAP.

SNP ID	Genomic position	Nucleotide change	Variation	Polyphen-2[1]	SIFT[2]	PROVEAN[3]	SNAP
rs139960913	152129063	C/T	H6Y	Prob. damaging	Intolerant	Neutral	Nonneutral
rs369520220	152129088	C/T	L14P	Prob. damaging	Intolerant	Neutral	Neutral
rs200924028	152129143	G/T	K32N	Prob. damaging	Intolerant	Neutral	Nonneutral
rs200075329	152129399	C/T	S118P	Benign	Tolerant	Deleterious	Neutral
rs185717042	152129456	A/C/G	S137R	Benign	Intolerant	Neutral	Neutral
rs17847065	152129484	A/C	P146Q	Poss. damaging	Tolerant	Neutral	Neutral
rs139548761	152163733	C/G	P152A	Poss. damaging	Intolerant	Neutral	Neutral
rs104893956	152163748	C/T	R157Ter	–	Intolerant	–	–
rs148034868	152163821	A/C	E181A	Poss. damaging	Intolerant	Deleterious	Neutral
rs373558014	152163827	A/G	R183H	Prob. damaging	Tolerant	Neutral	Neutral
rs142712646	152265352	C/T	R269C	Prob. damaging	Intolerant	Deleterious	Neutral
rs188957694	152265353	A/G	R269H	Prob. damaging	Intolerant	Neutral	Nonneutral
rs77797873	152265443	A/G	K299R	Poss. damaging	Tolerant	Neutral	Neutral
rs149490424	152265445	C/T	R300C	Poss. damaging	Tolerant	Neutral	Neutral
rs121913044*	152265638	A/T	V364E	Prob. damaging	Intolerant	Deleterious	Nonneutral
rs397509428	152332819	G/T	Q375H	Prob. damaging	Intolerant	Neutral	Nonneutral
rs374786087	152382206	A/G	N439S	Benign	Intolerant	Neutral	Neutral
rs121913043*	152382229	GC/TG	C447A	Poss. damaging	Intolerant	Deleterious	Nonneutral
rs79374934	152415590	A/C	D480E	Poss. damaging	Intolerant	Deleterious	Neutral
rs201562714	152419948	C/G	D545E	Benign	Intolerant	Deleterious	Neutral
rs182943916*	152419955	C/T	R548C	Prob. damaging	Intolerant	Deleterious	Nonneutral
rs141662120	152420067	C/T	T585M	Prob. damaging	Intolerant	Neutral	Nonneutral
rs138891155	152420094	C/T	T594M	Prob. damaging	Intolerant	Neutral	Nonneutral

SNPs found to have functional significance at least by three programs are in bold.
*More damaging SNPs (SNPs found to have functional significance by the four programs).
[1]Polyphen-2 impact; 0–0.5, benign; 1.00–1.24, borderline; 1.25–1.49, potentially damaging; 1.5–1.75, possibly damaging; ≥2, probably damaging.
[2]SIFT impact: 0.05, intolerant; 0.051–0.1, potentially intolerant; 0.0101–0.2, borderline; 0.201–1, tolerant.
[3]PROVEAN impact: ≤ -2.5, deleterious; > -2.5, neutral.

Table 2. ESEfinder results.

SNP ID	Nucleotide change	Variation	Genomic position	Site change
rs2077647	A/G	S10S	152129077	SC35 → no pattern
rs146774945	C/T	N27N	152129128	No pattern → Srp40, SRp55
rs200924028	G/T	K32N	152129143	SF2/ASF, SC35 → no pattern
rs199867565	A/G	E56E	152129215	No pattern → SF2/ASF, SC35
rs146586199	C/T	N69N	152129254	No pattern → SC35, Srp40, SRp55
rs104893956	C/T	R157Ter	152163748	Srp40 → no pattern
rs4986934	C/T	R243R	152201875	No pattern → SC35, SRp55
rs371170665	A/G	D285N	152265400	No pattern → SF2/ASF
rs148773555	A/C	D313E	152265486	No pattern → SF2/ASF, Srp40
rs367647625	A/G	R363K	152265635	No pattern → SF2/ASF
rs121913044	A/T	V364E	152265638	No pattern → Srp40
rs199620236	C/T	L507L	152415671	No pattern → Srp40
rs376984075	C/T	H516H	152415698	SF2/ASF, SC35 → no pattern
rs200128829	C/T	A551V	152419965	SRp55 → no pattern
rs146924427	A/G	T563A	152420000	SF2/ASF → no pattern
rs141662120	C/T	T585M	152420067	SF2/ASF, Srp40 → no pattern
rs138891155	C/T	T594M	152420094	SF2/ASF, Srp40 → no pattern

the same time: rs200924028, rs121913044, rs141662120 and rs138891155 (table 3).

Finally, protein sequence and the substitutions of the 43 nsSNPs were loaded to the Net-O-Glyc server. After comparing the functional element for each SNP, four SNPs (9.3%) were found to change a putative site of O-glycosylation, three of which created new glycosylation sites (table 4). Three different programs were applied to survey and prioritize the SNP in the UTR and regulatory regions of the *ESR1* gene. Polymorphisms in the UTR affect the gene expression by affecting the ribosomal translation of mRNA or by influencing the RNA half-life. The UTRscan program

Table 3. List of the most important SNPs found by prediction and/or association studies results.

SNP	Aa change	SIFT	PROVEAN	Polyphen-2	SNAP	INPS (DDG: kcal/mol)	ESEfinder	TFsearch	Association studies	Heterozygosity	Validation status
rs2077647	S10S						Change		Yes	0.498	V
rs200924028	K32N	Intolerant	Neutral	Prob. damag.	NN	−0.170	Change			NA	
rs104893956	R157*	Intolerant	Neutral				Change			NA	V
rs121913044	V364E	Intolerant	Deletetious	Prob. damag.	NN	−2.398	Change			NA	
rs121913043	C447A	Intolerant	Deleterious	Poss. damag.	NN	−2.343	No			NA	
rs182943916	R548C	Intolerant	Deleterious	Prob. damag.	NN	−0.061	No			NA	V
rs141662120	T585M	Intolerant	Neutral	Prob. damag.	NN	0.587	Change			NA	
rs138891155	T594M	Intolerant	Neutral	Prob. damag.	NN	0.587	Change		Yes	0.001	V
rs851993								Change	Yes	0.354	V
rs2941740								Change	Yes	0.415	V
rs2881766								Change	Yes	NA	V
rs6903180								Change	Yes	0.233	V
rs9478245								Change	Yes	0.208	V
rs2234693									Yes	0.494	V
rs9340799									Yes	NA	V

DDG, free energy change value; NN, nonneutral; NA, not assigned; Prob. damag., probably damaging; V, validated.

was used to predict the effect of SNPs of UTR region on the mRNA UTR of functional significance. Among the 110 SNPs in the mRNA UTR of the four variants encoding the isoform ERα1 (NM_000125.3, NM_001122740.1 and NM_001122741.1, NM_001122742.1), one variation (rs9340771) in the 5′UTR of NM_001122740.1 was related to functional pattern change of GY-BOX (table 5).

Polymorphisms mapping to miRNA binding sites have been shown to disrupt the ability of miRNA to target genes resulting in differential mRNA and protein expression. MicroSNiPER program was applied to predict the effect of 3′UTR SNP on miRNA binding site. Seventy nine SNPs were analysed and the results showed that nine (11.4%) were found to affect miRNA binding (table 5). Among these polymorphisms, there are four SNPs (5.1%) that potentially abolished miRNA target sites, while the other (6.3%) created a putative miRNA target sites.

Eventually, the TFsearch program was applied to predict the SNP effect of the regulatory region of the *ESR1* gene on putative transcription factor-binding sites. For this, a total of 167 SNPs were analysed and the results revealed that 28 SNPs (16.8%) change the transcription factor-binding sites. Most of these variations were related to the variant NM_001122742.1 (22 SNPs, 20.8% of the SNPs of the variant NM_001122742.1), while 13 SNPs (27.1% of the SNPs of the variant NM_001122741.1) were related to the variant NM_001122741.1 and 16 SNPs (35.5 and 34% of the SNPs of the variant NM_001122740.1 and NM_000125.3, respectively) were related to the variants NM_001122740.1 and NM_000125.3 (table 6).

The Medline database was reviewed to determine a list of functional SNPs from the data of association studies performed on the *ESR1* gene. Polymorphisms of the *ESR1* gene were associated with numerous diseases (coronary artery disease, cancer, diabetes, etc.) and physiological parameters such as cholesterol level, body height, total fat mass, height at menarche, testosterone level, etc. A high number of associations were observed with cancers, bone diseases and coronary artery disease (table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/).

Reviewing the literature, 89 SNPs of the *ESR1* gene were selected to be functional SNPs. Among these variations, only four SNPs (4.5%) are located in exons, while the others were in noncoding regions (95.5%). However, reviewing the number of positive studies associated to each of these functional SNPs, six SNPs (rs2077647, rs1801132, rs2228480, rs2234693, rs3020314 and rs9340799) were found to be the most important and where the SNPs rs2234693 and rs9340799 were in the lead of this list.

By combining the results of SNP prioritization and association studies, six SNPs (1.6%) among the 362 SNPs analysed were found to be shared (rs2077647, rs851993, rs291740, rs2881766, rs6903180 and rs9478245) (table 3). All these SNPs, except rs2077647, were located in regulatory region and were found to change putative-binding site for transcription factors. While, the rs2077647 is a nsSNP and it was found to change putative splicing site.

Table 4. nsSNPs of the *ESR1* gene predicted to have functional significance by Net O-glyc.

SNP ID	Genomic position	Nucleotide change	Variation	Glycosylation changement
rs201617046	152129480	G145S	A/G	Gain of glycosylation
rs149308960	152163757	G160S	G/T	Gain of glycosylation
rs146924427	152420000	T563A	A/G	Loss of glycosylation
rs201118302	152420040	S576L	C/T	Gain of glycosylation

Table 5. SNPs in UTR region predicted to have significance by UTRscan or MicroSNiPER.

SNP ID	Nucleotide change	Genomic position	Site change
rs9340771	C/T	152128555	No pattern → GY-BOX
rs148368610	G/T	152422731	IRES → no pattern
rs33986155	C/G/T	152420685	hsa-miR-4271 (G) → no pattern
rs3020385	A/G/T	152420848	hsa-miR-367-5p, hsa-miR-5691 (A) → no pattern
rs2982901	A/C	152421320	No pattern → hsa-miR-6077, hsa-miR-5197-3p, hsa-miR-3185, hsa-miR-3192
rs114409231	C/T	152420121	No pattern → hsa-miR-761, hsa-miR-299-3p, hsa-miR-4296
rs139705407	A/G	152420122	hsa-miR-4539, hsa-miR-299-3p, hsa-miR-892a → no pattern
rs189550638	A/T	152420350	No pattern → hsa-miR-494
rs187602901	A/G	152420428	No pattern → hsa-miR-105-5p
rs3798757	C/T	152424243	No pattern → hsa-miR-586, hsa-miR-548al, hsa-miR-491-3p, hsa-miR-222-5p, hsa-miR-4711-5p

Discussion

Understanding the functions of SNPs will greatly help in understanding the genetics of the human phenotype variation, especially the genetic basis of human complex diseases. However, to identify functional SNPs from a pool containing both functional and neutral SNPs is challenging.

A number of studies have been reported on associations between *ESR1* gene polymorphisms and many diseases, such as osteoporosis (Ioannidis *et al.* 2004; Harsløf *et al.* 2010; Luo *et al.* 2014), osteoarthritis (Tawonsawatruk *et al.* 2009; Riancho *et al.* 2010), breast cancer (Tapper *et al.* 2008; Ding *et al.* 2010), endometrial cancer (Sliwinski *et al.* 2010), some brain disorders as depressive disorders (Tsai *et al.* 2003), Alzheimer's disease (Lin *et al.* 2003; Monastero *et al.* 2006; Ma *et al.* 2009) and coronary artery disease (Wu *et al.* 2010). However, most of the SNPs of the *ESR1* gene have not been studied yet. In this study, we surveyed and prioritized 362 SNPs of the *ESR1* gene. Using different computational algorithm tools, we found that 18 nsSNPs were evolutionary conserved by SIFT, 18 nsSNPs might alter protein structure by Polyphen-2, nine nsSNPs were nonneutral by SNAP and eight nsSNPs were deleterious by PROVEAN. An agreement between the four programs used was observed in three nsSNPs (rs121913044, rs121913043 and rs182943916). This result suggests that these variants may be considered to be most likely damaging or deleterious SNPs. The analysis of protein stability change allows confirmation of this finding. In fact, using the INPS program, 22 nsSNPs were found to induce an increase in energy in comparison with the native structure and the most destabilizing SNPs were rs121913044 (V364E) and rs121913043 (C447A).

Using ESEfinder, our results showed that 17 nsSNPs might change functional pattern of putative splicing site. Among the functional nsSNPs predicted, a coding nonsense SNP (rs104893956) was found. This variation caused by a nucleotide change from C to T and leads to a codon stop gain. This result suggests that this SNP may have a very high level of risk to be involved in some diseases as it can truncate and even inactivate the ER protein.

Equally, our results showed that, four nsSNPs (rs200924028, rs121913044, rs141662120 and rs138891155) might alter protein structure and function as well as splicing phenomen. This result suggests that these markers might have a high potential to be candidate SNPs in association studies. Further, four nsSNPs (rs201617046, rs149308960, rs146924427 and rs201118302) were found to change putative sites of glycosylation among which three nsSNPs created new site of glycosylation. Protein glycosylation is an important post-translational modification that confers both structural and functional properties to the molecules. However, many studies have shown that up to 1.4% of known disease-causing missense mutations are predicted to give rise to gains of glycosylation and for some of these mutations, the novel glycans have been shown to be both necessary and sufficient to account for the deleterious impact of the mutation (Schulte am Esch *et al.* 2005; Vogt *et al.* 2007). Thereby, it may be suggested that these nsSNPs can be added to the list of candidate SNPs in association studies to determine their potential role in diseases.

Analysing SNPs in regulatory regions, we found that one SNP might change functional binding motif at 5′UTR, nine SNPs might change pattern of miRNA binding site and 28 SNPs might modulate gene regulation.

Table 6. SNPs in regulatory region predicted to have significance by TFsearch.

SNP ID	Nucleotide change	Position*	Site change
rs79543702[a]	C/T	−9930	No pattern → CdxA
rs851999[a]	A/C	−9673	USF → no pattern
rs75691244[a]	A/C	−9448	GATA1, GATA2 → no pattern
rs75453684[a]	G/T	−8373	No pattern → HFH-2
rs11969897[a]	A/G	−7451	No pattern → AP-1
rs2982574[a]	C/G	−6427	No pattern → GATA-1
rs851995[a]	A/G	−6097	Evi-1 → no pattern
rs851993[a]	C/T	−5620	No pattern → SRY
rs9383939[a]	A/G	−5453	IRF-2 → no pattern
rs851988[a]	A/G	−3815	No pattern → GATA-1, GATA-2
rs11963534[a]	A/G	−3670	GATA-2 → no pattern
rs2941740[a]	C/T	−1993	No pattern → Nkx-2
rs189179070[a]	C/T	−1629	No pattern → CdxA
rs17840349[a]	A/T	−1260	No pattern → CdxA
rs368772753[a]	A/G	−1226	C/EBPb → no pattern
rs2982572[a]	A/G	−1070	CdxA → no pattern
rs113692904[a]	C/T	−912	No pattern → HFH-1, 2, HNF-3b, CdxA
rs34620075[a]	−/T	−904	No pattern → CdxA
rs191313267[a]	A/G	−823	SRY → no pattern
rs375192774[a]	C/T	−815	No pattern → CdxA
rs75027116[a]	A/C	−781	CdxA → no pattern
rs147158208[a]	C/T	−391	AP-1 → no pattern
rs2881766[b]	G/T	−9694/−9336/−7688[&]	No pattern → CdxA
rs9479118[b]	C/T	−9689/−9329/−7683[&]	No pattern → Nkx-2
rs6914569[b]	C/G	−9196/−8836/−7190[&]	No pattern → Nkx-2
rs538098[b]	C/T	−7640/−7584/−5938[&]	c-Ets, Elk-1 → no pattern
rs11964281[b]	C/T	−7371/−7012/−5365[&]	No pattern → HSF2
rs4329125[b]	C/T	−7230/−6870/−5224[&]	No pattern → Oct-1
rs79994281[b]	A/C	−5221/−48861/−3215[&]	No pattern → SRY, HNF-3b
rs57977903[b]	A/G	−5141/−4781/−3135[&]	No pattern → C/EBP, HNF-3b
rs77480311[b]	A/G	−5004/−4644/−2998[&]	No pattern → GATA-2, GATA-1
rs523736[b]	A/G	−4918/−4558/−2912[&]	USF, delta-E → no pattern
rs6903180[b]	A/G	−3582/−3222/−1576[&]	Nkx-2 → no pattern
rs9478245[b]	C/T	−2640/−2866/−1220[&]	No pattern → sox-5
rs79646490[b]	G/T	−2202/−1842/−196[&]	Sp-1 → no pattern
rs28462265[c]	G/T	−1708/ − 1348[§]	HSF2 → no pattern
rs73780864[c]	C/T	−1378/ − 1018[§]	No pattern → CdxA
rs9371556[c]	A/G	−1154/ − 789[§]	CdxA → no pattern

[a]SNP in regulatory region of the variant NM_001122742.1, [b]SNP in regulatory region of NM_000125.3, NM_001122740.1 and NM_001122741.1 variants, [c]SNP in regulatory region of the variants NM_000125.3 and NM_001122740.1, *relative to the transcription start site, [&]position of the variants NM_000125.3/NM_001122740.1/NM_001122741.1, [§]position of the variants NM_000125.3/NM_001122740.1.

Our predictions are in good agreement with previous reports, especially those which have demonstrated that the variation rs121913044 (V364E) which has a single amino acid substitution in hormone-binding domain of the ERα, allows the receptor to act as a strong dominant negative inhibitor of oestrogen action (McInerney *et al.* 1996) and the rs121913043 (C447A) which causes a decoupling of hormone binding and transcriptional activation functions of the receptor (Reese and Katzenellenbogen 1991). We also identified rs397509428 (Q375H) as a key SNP, a prediction supported by the fact that it corresponds to a substitution in ligand-binding domain of ERα and causes a complete oestrogen insensitivity and puberty delay in women (Quaynor *et al.* 2013). Another functional SNP is rs104893956 which is characterized by cytosine to thymine transition at codon 157 and results in a premature stop codon and oestrogen resistance (Smith *et al.* 1994).

Reviewing the literature, 89 SNPs were selected as functional SNPs where most of them were located in noncoding regions. This result is consistent with the view of Frazer *et al.* (2009) who suggested that disease risk associated SNP map predominantly to noncoding regions of the human genome. In fact, among the most important polymorphic sites of the *ESR1* gene are rs2234693 and rs9340799, located in the first intron and are separated by only 46 bp. The rs2234693 (T397C) is caused by a T/C transition in intron 1, whereas the rs9340799 (G351A) polymorphism is caused by a G/A transition located 50 bp downstream of the rs2234693

(Shearman *et al.* 2003; Pollak *et al.* 2004). These two polymorphisms of the *ESR1* gene have been described and studied for possible association with several clinical outcomes including cardiovascular risk (Herrington *et al.* 2002; Alevizaki *et al.* 2007), multiple sclerosis (Niino *et al.* 2000; Kikuchi *et al.* 2002), osteoporosis (Harsløf *et al.* 2010; Kim *et al.* 2010), uterine leiomyomas (Al-Hendy and Salama 2006), cancer (Chattopadhyay *et al.* 2014) as well as type 2 diabetes, obesity (Speer *et al.* 2001), bone mineral density (Yamada *et al.* 2002), azoospermia or severe oligozoospermia (Kukuvitis *et al.* 2002; Suzuki *et al.* 2002; Lazaros *et al.* 2010) and systemic lupus erythematosus (Wang *et al.* 2010).

By combining the results of SNP prioritization and association studies, we come with six functional SNPs among which only SNP rs2077647 is located in the coding region and was already reported to be associated with numerous diseases such as coronary artery disease (Peter *et al.* 2005, 2009), cancers (Anghel *et al.* 2010; Sonoda *et al.* 2010), Alzheimer disease (Ma *et al.* 2009) as well as to response to drugs administration (Zhang *et al.* 2010). The correlation between our prediction results and data from association studies supports the results of this study and suggests that SNP-based pathogenicity detection tools can appropriately reflect the role of a disease associated SNP. Since association studies and SNP prioritization are two nonredundant source of knowledge, we think that a good correlation between them can support the use of computational tools for the selection of SNPs to be investigated by association studies.

Conclusion

Genetic screening of the *ESR1* gene locus has revealed the existence of thousands of polymorphic sites, some of them alter the function of the receptor and were associated to phenotypic traits and diseases risk. However, giving the high number of SNPs in this gene, association studies should be carried on genetic variants that have functional significance. The correlation between our results and data from association studies suggests that application of computational tools might provide an alternative approach to select functional SNPs in association studies. Since association studies and SNP prioritization are two nonredundant source of knowledge, we think that a good correlation between them can support the use of computational tools for the selection of SNPs to be investigated by association studies.

References

Adzhubei I. A., Schmidt S., Peshkin L., Ramensky V. E., Gerasimova A., Bork P. *et al.* 2010 A method and server for predicting damaging missense mitations. *Nat. Methods* **7**, 248–249.

Albagha O. M., McGuigan F. E., Reid D. M. and Ralston S. H. 2001 Estrogen receptor alpha gene polymorphisms and bone mineral density: haplotype analysis in women from the United Kingdom. *J. Bone Miner. Res.* **16**, 128–134.

Alevizaki M., Saltiki K., Cimponeriu A., Kanakakis I., Xita N., Alevizaki C. C. *et al.* 2007 Severity of cardiovascular disease in postmenopausal women: associations with common estrogen receptor alpha polymorphic variants. *Eur. J. Endocrinol.* **156**, 489–496.

Al-Hendy A. and Salama S. A. 2006 Ethnic distribution of estrogen receptor-alpha polymorphism is associated with a higher prevalence of uterine leiomyomas in black Americans. *Fertil. Steril.* **86**, 686–693.

Andrawiss M. 2005 First phase of HapMap project already helping drug discovery. *Nat. Rev. Drug Discov.* **4**, 947.

Anghel A., Narita D., Seclaman E., Popovici E., Anghel M. and Tamas L. 2010 Estrogen receptor alpha polymorphisms and the risk of malignancies. *Pathol. Oncol. Res.* **16**, 485–496.

Barenboim M., Zoltick B. J., Guo Y. and Weinberger D. R. 2010 MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. *Hum. Mutat.* **31**, 1223–1232.

Boffa M. B., Maret D., Hamill J. D., Bastajian N., Crainich P., Jenny N. S. *et al.* 2008 Effect of single nucleotide polymorphisms on expression of the gene encoding thrombin-activatable fibrinolysis inhibitor: a functional analysis. *Blood* **111**, 183–189.

Bromberg Y. and Rost B. 2007 SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835.

Cartegni L., Wang J., Zhu Z., Zhang M. Q. and Krainer A. R. 2003 ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acid Res.* **31**, 3568–3571.

Chattopadhyay S., Siddiqui S., Akhtar M. S., Najm M. Z., Deo S. V., Shukla N. K. and Husain S. A. 2014 Genetic polymorphisms of ESR1, ESR2, CYP17A1, and CYP19A1 and the risk of breast cancer: a case control study from North India. *Tumour Biol.* **35**, 4517–4527.

Choi Y., Sims G. E., Murphy S., Miller J. R. and Chan A. P. 2012 Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688.

Ding S. L., Yu J. C., Chen S. T., Hsu G. C., Hsu H. M., Ho J. Y. *et al.* 2010 Diverse associations between ESR1 polymorphism and breast cancer development and progression. *Clin. Cancer Res.* **16**, 3473–3484.

Fariselli P., Martelli P. L., Savojardo C. and Casadio R. 2015 INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics* **31**, 2816–2821.

Frazer K., Murray S., Schork N. and Topol E. 2009 Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251.

Grillo G., Turi A., Licciulli F., Mignone F., Liuni S., Banfi S. *et al.* 2010 UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regylatory motifs of the untranslated region of eukaryotic mRNAs. *Nucleic Acids Res.* (Database issue), D75–D80.

Hampe J., Franke A., Rosenstiel P., Till A., Teuber M., Huse K. *et al.* 2007 A genome wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crhon disease in ATG16L1. *Nat. Genet.* **39**, 207–211.

Harsløf T., Husted L. B., Carstens M., Stenkjaer L. and Langdahl B. L. 2010 Genotypes and haplotypes of the estrogen receptor genes, but not the retinoblastoma-interacting zinc finger protein 1 gene, are associated with osteoporosis. *Calcif. Tissue Int.* **87**, 25–35.

Heinemeyer T., Wingender E., Reuter I., Hermjakob H., Kel A. E., Kel O. V. *et al.* 1998 Databases on transcriptional regulation: TRANSFAC, TRRD, and COMPEL. *Nucleic Acids Res.* **26**, 364–370.

Herrington D. M., Howard T. D., Hawkins G. A., Reboussin D. M., Xu J., Zheng S. L. *et al.* 2002 Estrogenreceptor polymorphisms and effects of estrogen replacement on high-density lipoprotein cholesterol in women with coronary disease. *N. Engl. J. Med.* **346**, 967–974.

Ilhan I. and Tezel G. 2013 How to select tag SNPs in genetic association studies? The CLONTagger method with parameter optimization. *OMICS* **17**, 368–383.

Ioannidis J. P., Ralston S. H., Bennett S. T., Brandi M. L., Grinberg D., Karassa F. B. *et al.* 2004 Differential genetic effects of *ESR1* gene polymorphisms on osteoporosis outcomes. *JAMA* **292**, 2105–2114.

Kikuchi S., Fukazawa T., Niino M., Yabe I., Miyagishi R., Hamada T. and Tashiro K. 2002 Estrogen receptor gene polymorphism and multiple sclerosis in Japanese patients: interaction with HLA-DRB1*1501 and disease modulation. *J. Neuroimmunol.* **128**, 77–81.

Kim S. Y., Kim H. H., Nam C. M., Kim H. C., Suh I. and Kang B. Y. 2010 Association of estrogen receptor-alpha gene polymorphism with pathogenesis of osteoporosis in Korean vegetarian men. *Med. Princ. Pract.* **19**, 200–205.

Kim S., Pyun J. A., Kang H., Kim J., Cha D. H. and Kwack K. 2011 Epistasis between CYP19A1 and ESR1 polymorphisms is associated with premature ovarian failure. *Fertil. Steril.* **95**, 353–356.

Kimura-Kataoka K., Yasuda T., Fujihara J., Toga T., Ono R., Otsuka Y. *et al.* 2012 Genetic and expression analysis of SNPs in the human deoxyribonuclease II: SNPs in the promoter region reduce its *in vivo* activity through decreased promoter activity. *Electrophoresis* **33**, 2852–2858.

Knobil E. and Neill J. 1994 *The physiology of reproduction*, 2nd edition. Raven, New York, USA.

Kos M., Reid G., Denger S. and Gannon F. 2001 Minireview: genomic organization of the human ERα gene promoter region. *Mol. Endocrinol.* **15**, 2057–2063.

Kukuvitis A., Georgiou I., Bouba I., Tsirka A., Giannouli C. H., Yapijakis C. *et al.* 2002 Association of oestrogen receptor alpha polymorphisms and androgen receptor CAG trinucleotide repeats with male infertility: a study in 109 Greek infertile men. *Int. J. Androl.* **25**, 149–152.

Lazaros L. A., Xita N. V., Kaponis A. I., Zikopoulos K. A., Plachouras N. I. and Georgiou I. A. 2010 Estrogen receptor alpha and beta polymorphisms are associated with semen quality. *J. Androl.* **31**, 291–298.

Lee J. E., Choi J. H., Lee J. H. and Lee M. G. 2005 Gene SNPs and mutations in clinical genetic testing: haplotype-based testing and analysis. *Mutat. Res.* **573**, 195–204.

Lin G. F., Ma Q. W., Zhang D. S., Zha Y. L., Lou K. J. and Shen J. H. 2003 Polymorphism of alpha-estrogen receptor and aryl hydrocarbon receptor genes in dementia patients in Shanghai suburb. *Acta Pharmacol. Sin.* **24**, 651–656.

Luo L., Xia W., Nie M., Sun Y., Jiang Y., Zhao J., He S. and Xu L. 2014 Association of ESR1 and C6orf97 gene polymorphism with osteoporosis in postmenopausal women. *Mol. Biol. Rep.* **41**, 3235–3243.

Ma S. L., Tang N. L., Tam C. W., Lui V. W., Lau E. S., Zhang Y. P. *et al.* 2009 Polymorphisms of the estrogen receptor alpha (ESR1) gene and the risk of Alzheimer's disease in a southern Chinese community. *Int. Psychogeriatr.* **21**, 977–986.

Maruyama H., Toji H., Harrington C. R., Sasaki K., Izumi Y., Ohnuma T. *et al.* 2000 Lack of an association of estrogen receptor alpha gene polymorphisms and transcriptional activity with Alzheimer disease. *Arch. Neurol.* **57**, 236–240.

McInerney E. M., Ince B. A., Shapiro D. J. and Katzenellenbogen B. S. 1996 A transcriptionally active estrogen receptor mutant is a novel type of dominant negative inhibitor of estrogen action. *Mol. Endocrinol.* **10**, 1519–1526.

Milanese M., Segat L. and Crovella S. 2007 Transcriptional effect of DEFB1 gene 5′ untranslated region polymorphisms. *Cancer Res.* **67**, 5997.

Monastero R., Cefalù A. B., Camarda C., Noto D., Camarda L. K., Caldarella R. *et al.* 2006 Association of estrogen receptor alpha gene with Alzheimer's disease: a case control study. *J. Alzheimers Dis.* **9**, 273–278.

Niino M., Kikuchi S., Fukazawa T., Yabe I. and Tashiro K. 2000 Estrogen receptor gene polymorphism in Japanese patients with multiple sclerosis. *J. Neurol. Sci.* **179**, 70–75.

Paskulin D. D., Cunha-Filho J. S., Paskulin L. D., Souza C. A. and Ashton-Prolla P. 2013 ESR1 rs9340799 is associated with endometriosis-related infertility and *in vitro* fertilization. *Dis. Markers* **35**, 907–913.

Patnala R., Clements J. and Batra J. 2013 Candidate gene association studies: a comprehensive guide *in silico* tools. *BMC* **14**, 39.

Pesole G. and Liuni S. 1999 Internet resources for the functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs. *Trends Genet.* **15**, 378.

Peter I., Shearman A. M., Zucker D. R., Schmid C. H., Demissie S., Cupples L. A. *et al.* 2005 Variation in estrogen-related genes and cross-sectional and longitudinal blood pressure in the Framingham heart study. *J. Hypertens.* **23**, 2193–2200.

Peter I., Kelley-Hedgepeth A., Huggins G. S., Housman D. E., Mendelsohn M. E., Vita J. A. *et al.* 2009 Association between arterial stiffness and variations in oestrogen-related genes. *J. Hum. Hypertens.* **23**, 636–644.

Pollak A., Rokach A., Blumenfeld A., Rosen L. J., Resnik L. and Dresner Pollak R. 2004 Association of oestrogen receptor alpha gene polymorphism with the angiographic extent of coronary artery disease. *Eur. Heart J.* **25**, 240–245.

Quaynor S. D., Stradtman E. W., Kim H. G., Shen Y., Chorich L. P., Schreihofer D. A. and Layman L. C. 2013 Delayed puberty and estrogen resistance in a woman with estrogen receptor α variant. *N. Engl. J. Med.* **369**, 164–171.

Reese J. C. and Katzenellenbogen B. S. 1991 Mutagenesis of cysteines in the hormone binding domain of the human estrogen receptor. Alterations in binding and transcriptional activation by covalently and reversibly attaching ligands. *J. Biol. Chem.* **266**, 10880–10887.

Riancho J. A., García-Ibarbia C., Gravani A., Raine E. V., Rodríguez-Fontenla C., Soto-Hermida A. *et al.* 2010 Common variations in estrogen-related genes are associated with severe large-joint osteoarthritis: a multicenter genetic and functional study. *Osteoarthr. Cartil.* **18**, 927–933.

Schulte am Esch J. 2nd, Robson S. C., Knoefel W. T., Eisenberger C. F., Peiper M. and Rogiers X. 2005 Impact of O-linked glycosylation of the VWF-A1-domain flanking regions on platelet interaction. *Br. J. Haematol.* **128**, 82–90.

Shearman A. M., Cupples L. A., Demissie S., Peter I., Schmid C. H., Karas R. H. *et al.* 2003 Association between estrogen receptor alpha gene variation and cardiovascular disease. *JAMA* **290**, 2263–2270.

Sim N. L., Kumar P., Hu J., Henikoff S., Schneider G. and Ng P. C. 2012 SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40** (web server issue), W452–W457.

Sliwinski T., Sitarek P., Stetkiewicz T., Sobczuk A. and Blasiak J. 2010 Polymorphism of the ERalpha and CYP1B1 genes in endometrial cancer in a Polish subpopulation. *J. Obstet. Gynaecol. Res.* **36**, 311–317.

Smith E. P., Boyd J., Frank G. R., Takahashi H., Cohen R. M., Specker B. *et al.* 1994 Estrogen resistance caused by a mutation in the estrogen-receptor gene in a man. *N. Engl. J. Med.* **331**, 1056–1061.

Sonoda T., Suzuki H., Mori M., Tsukamoto T., Yokomizo A., Naito S. *et al.* 2010 Polymorphisms in estrogen related genes may modify the protective effect of isoflavones against prostate cancer risk in Japanese men. *Eur. J. Cancer. Prev.* **19**, 131–137.

Speer G., Cseh K., Winkler G., Vargha P., Braun E., Takács I. and Lakatos P. 2001 Vitamin D and estrogen receptor gene

polymorphisms in type 2 diabetes mellitus and in android type obesity. *Eur. J. Endocrinol.* **144**, 385–389.

Steentoft C., Vakhrushev S. Y., Joshi H. J., Kong Y., Vester-Christensen M. B., Schjoldager K. T. *et al.* 2013 Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488.

Stenson P. D., Mort M., Ball E. V., Howells K., Phillips A. D., Thomas N. S. and Cooper D. N. 2009 The human gene mutation database: 2008 update. *Genome Med.* **1**, 13.

Suzuki Y., Sasagawa I., Itoh K., Ashida J., Muroya K. and Ogata T. 2002 Estrogen receptor alpha gene polymorphism is associated with idiopathic azoospermia. *Fertil. Steril.* **78**, 1341–1343.

Takahashi Y., Mimori K. and Mori M. 2012 Significance of genome-wide association study in cancer. *Nihon Geka Gakkai Zasshi* **113**, 210–214.

Tapper W., Hammond V., Gerty S., Ennis S., Simmonds P. and Collins A. 2008 The influence of genetic variation in 30 selected genes on the clinical characteristics of early onset breast cancer. *Breast Cancer Res.* **10**, R108.

Tawonsawatruk T., Trachoo O., Channoom T., Sura T., Eu-Ahsunthornwattana J., Woratanarat P. and Wajanavisit W. 2009 Association of estrogen receptor-alpha single-nucleotide polymorphism (codon 594 G–>A) and Thai patients affected by knee osteoarthritis. *J. Med. Assoc. Thai.* **92** suppl 6, S45–S50.

Tsai S. J., Wang Y. C., Hong C. J. and Chiu H. J. 2003 Association study of oestrogen receptor alpha gene polymorphism and suicidal behaviours in major depressive disorder. *Psychiatr. Genet.* **13**, 19–22.

Vogt G., Vogt B., Chuzhanova N., Julenius K., Cooper D. N. and Casanova J. L. 2007 Gain-of-glycosylation mutations. *Curr. Opin. Genet. Dev.* **17**, 245–251.

Wang J., Nuite M. and McAlindon T. E. 2010 Association of estrogen and aromatase gene polymorphisms with systemic lupus erythematosus. *Lupus* **19**, 734–740.

Wu M. M., Hsieh Y. C., Lien L. M., Chen W. H., Bai C. H., Chiu H. C. *et al.* 2010 Association of estrogen receptor {alpha} genotypes/haplotypes with carotid intima-media thickness in Taiwanese women. *Angiology* **61**, 275–282.

Yamada Y., Ando F., Niino N., Ohta S. and Shimokata H. 2002 Association of polymorphisms of the estrogen receptor alpha gene with bone mineral density of the femoral neck in elderly Japanese women. *J. Mol. Med.* **80**, 452–460.

Zhang T. T., Zhang Z. H., Gao Y. F., Zhang Y. F., Yang D. L. and Li X. 2010 T29C genotype polymorphism of estrogen receptor alpha is associated with initial response to interferon-alpha therapy in chronic hepatitis B patients. *Hepatobiliary Pancreat. Dis. Int.* **9**, 275–279.

Zhernakova A., van Diemen C. C. and Wijmenga C. 2009 Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55.

Received 16 April 2015, in final revised form 8 March 2016; accepted 17 March 2016
Unedited version published online: 21 March 2016
Final version published online: 18 November 2016

Corresponding editor: RAJIVA RAMAN