nature genetics

Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome

Yongyong Shi[1–3,37*], Han Zhao[4–7,37], Yuhua Shi[4–7,37], Yunxia Cao[8,37], Dongzi Yang[9,37], Zhiqiang Li[1–3,37], Bo Zhang[10], Xiaoyan Liang[11], Tao Li[4], Jianhua Chen[1,12], Jiawei Shen[1], Junzhao Zhao[13], Li You[4–7], Xuan Gao[4–7], Dongyi Zhu[14], Xiaoming Zhao[15], Ying Yan[16], Yingying Qin[4–7], Wenjin Li[1], Junhao Yan[4–7], Qingzhong Wang[1], Junli Zhao[17], Ling Geng[4–7], Jinlong Ma[4–7], Yueran Zhao[4–7], Guang He[1], Aiping Zhang[1], Shuhua Zou[18], Aijun Yang[19], Jiayin Liu[20], Weidong Li[1], Baojie Li[1], Chunling Wan[1], Ying Qin[21], Juanzi Shi[22], Jing Yang[23], Hong Jiang[24], Jin-e Xu[25], Xiujuan Qi[25], Yun Sun[15], Yajie Zhang[26], Cuifang Hao[27], Xiuqing Ju[28], Dongni Zhao[29], Chun-e Ren[30], Xiuqing Li[31], Wei Zhang[32], Yiwen Zhang[33], Jiangtao Zhang[4,5], Di Wu[4,5], Changming Zhang[4,5], Lin He[1,2,34,35] & Zi-Jiang Chen[4–7,15,36]

Following a previous genome-wide association study (GWAS 1) including 744 cases and 895 controls, we analyzed genome-wide association data from a new cohort of Han Chinese (GWAS 2) with 1,510 polycystic ovary syndrome (PCOS) cases and 2,016 controls. We followed up significantly associated signals identified in the combined results of GWAS 1 and 2 in a total of 8,226 cases and 7,578 controls. In addition to confirming the three loci we previously reported, we identify eight new PCOS association signals at $P < 5 \times 10^{-8}$: 9q22.32, 11q22.1, 12q13.2, 12q14.3, 16q12.1, 19p13.3, 20q13.2 and a second independent signal at 2p16.3 (the *FSHR* gene). These PCOS association signals show evidence of enrichment for candidate genes related to insulin signaling, sexual hormone function and type 2 diabetes (T2D). Other candidate genes were related to calcium signaling and endocytosis. Our findings provide new insight and direction for discovering the biological mechanisms of PCOS.

Polycystic ovary syndrome is a complex endocrinopathy in women of reproductive age, with a prevalence worldwide of 6–8% (refs. 1,2). The syndrome is defined by hyperandrogenism, menstrual irregularity and polycystic ovarian morphology. PCOS also manifests clinically as obesity, infertility, impaired glucose tolerance (IGT), insulin resistance and an increased risk of endometrial cancer, metabolic syndrome, T2D and cardiovascular disease[3–5].

Over the past decades, many genes involved in gonadotropin secretion, steroid hormone synthesis, insulin signaling and chronic inflammation have been interrogated as susceptibility genes for PCOS. Our previous GWAS (here called GWAS 1) identified three susceptibility loci—2p16.3, 2p21 and 9q33.3—associated with PCOS in the Han Chinese population[6]. However, these genetic signals are not sufficient to account for the considerable genetic susceptibility for this endocrine-metabolic disorder; therefore, additional genetic risk factors remain to be discovered.

To explore additional risk regions for PCOS, we performed a further GWAS (GWAS 2) of an additional 1,510 individuals with PCOS and 2,106 controls of northern Han Chinese ancestry using the Affymetrix Axiom array. To identify potential population stratification, principal-component analysis (PCA) was performed for the study samples and HapMap references (see URLs; **Supplementary Fig. 1**). To minimize the effect of potential population stratification, logistic regression was applied to the test association for each SNP, adjusting by the significant principal components (Online Methods). Little evidence was observed for population stratification ($\lambda = 1.07$, $\lambda_{1000} = 1.04$). To maximize statistical power, we combined the new data with our previous GWAS 1 data (744 cases and 895 controls of northern Han Chinese ancestry analyzed with the Affymetrix SNP6.0 chip). We then performed a meta-analysis (GWAS-meta) of both genotyped and imputed SNPs (**Fig. 1** and Online Methods). In total, 2,254 PCOS cases and 3,001 normal controls were used for GWAS-meta (GWAS 1 and 2). Replication for GWAS-meta included two independent sample sets (replication 1: 1,908 cases and 1,913 controls; replication 2: 6,318 cases and 5,665 controls) (Online Methods).

We confirmed the three previously identified loci in the GWAS-meta stage (**Fig. 1**, **Supplementary Fig. 2** and **Supplementary Table 1**): 2p16.3 (rs13405728, $P_{\text{GWAS-meta}} = 3.77 \times 10^{-9}$), 2p21 (rs13429458, $P_{\text{GWAS-meta}} = 4.17 \times 10^{-13}$) and 9q33.3 (rs2479106, $P_{\text{GWAS-meta}} = 5.14 \times 10^{-10}$). In addition, SNPs in 19 new regions showed association at $P_{\text{GWAS-meta}} < 1 \times 10^{-5}$ with PCOS susceptibility in the GWAS-meta stage (**Supplementary Table 2**). Additionally, variants in the *FSHR* gene, which are located at 2p16.3 but for which associations were not directly supported in the previous GWAS 1, also associated at $P_{\text{GWAS-meta}} < 1 \times 10^{-5}$. We selected the most significantly associated SNPs from all 20 regions for validation (replication 1; **Supplementary Table 2**).

*A full list of author affilations appears at the end of the paper.

Received 17 January; accepted 19 July; published online 12 August 2012; doi:10.1038/ng.2384

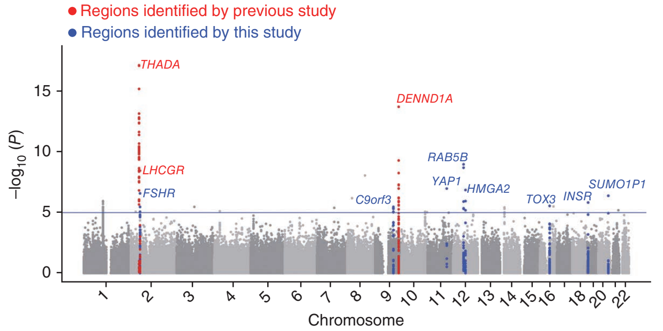

Figure 1 Genome-wide Manhattan plot for the GWAS meta-analysis. Shown are $-\log_{10} P$ values for SNPs that passed quality control. The solid horizontal line indicates $P < 1 \times 10^{-5}$. Markers within 50 kb of a SNP associated with PCOS are marked in red for those identified in a previous GWAS[10] and replicated here and in blue for those first identified in the current study. Associations at *THADA*, *LHCGR* and *DENND1A* were also reported in a previous GWAS[10].

Of these 20 regions, 7 were validated in the replication 1 stage ($P < 0.05$ with the same allelic odds ratio (OR) direction), and another 3 regions had SNPs with association P values of $<5 \times 10^{-6}$ in the meta-analysis of GWAS and replication 1 data (Online Methods and **Supplementary Table 2**). SNPs from these ten regions were genotyped again in a second independent sample set (replication 2). Overall, common variants in eight regions—9q22.32, 11q22.1, 12q13.2, 12q14.3, 16q12.1, 19p13.3, 20q13.2 and the *FSHR* gene (2p16.3)—showed combined evidence of association at $P < 5 \times 10^{-8}$ in meta-analysis of all stages under a fixed-effects model (GWAS-rep-meta analysis) (**Table 1**).

Most but not all of the candidate genes at the associated loci were related to hormones, insulin resistance and organ growth. At 9q22.32, the most significantly associated SNP was rs3802457 ($P_{\text{GWAS-rep-meta}} = 5.28 \times 10^{-14}$; $\text{OR}_{\text{GWAS-rep-meta}} = 0.77$), which is located in the intron of the *C9orf3* gene (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs3802457, rs4385527 ($P_{\text{GWAS-rep-meta}} = 5.87 \times 10^{-9}$; $\text{OR}_{\text{GWAS-rep-meta}} = 0.84$) showed independent association in conditional logistic regression analysis (**Supplementary Table 4**) and is also located in *C9orf3*. The encoded C9orf3 protein is a member of the M1 zinc aminopeptidase family. The rs3802458 SNP within *C9orf3* has been reported to be associated with the development of erectile dysfunction in African-American men following radiotherapy for prostate cancer[7]. Erectile dysfunction in men and PCOS in women occur when individuals have either inadequate or excessive amounts of sexual hormones. Notably, the *FSHR* gene (rs2268363) has been identified as the locus most significantly associated with erectile dsyfunction[7], and we also found strong association evidence for *FSHR* with PCOS.

At 11q22.1, rs1894116 ($P_{\text{GWAS-rep-meta}} = 1.08 \times 10^{-22}$; $\text{OR}_{\text{GWAS-rep-meta}} = 1.27$) is located in the intron of *YAP1* (MIM 606608) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs1894116, conditional logistic regression analysis identified no additional association signal (**Supplementary Table 4**). The YAP1 protein, also verified to be associated with PCOS in our previous study[8], is a transcriptional regulator that can act both as a coactivator and a corepressor and is the critical downstream regulatory target in the Hippo signaling pathway. It also has a pivotal role in organ size control and tumor suppression by restricting proliferation and promoting apoptosis. YAP overexpression alters the expression of genes associated with cell proliferation, apoptosis, migration, adhesion and the epithelial-to-mesenchymal transition[9]. However, the function of YAP in the ovary needs further study.

At 12q13.2, the most significantly associated SNP was rs705702 ($P_{\text{GWAS-rep-meta}} = 8.64 \times 10^{-26}$; $\text{OR}_{\text{GWAS-rep-meta}} = 1.27$), which is located in the intergenic region between *RAB5B* (MIM 179514) and *SUOX* (MIM 606887) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs705702, conditional logistic regression analysis showed that there is no additional association signal (**Supplementary Table 4**). Notably, this region has been reported to be a type 1 diabetes (T1D) susceptibility locus in several studies[10–15]. Of the SNPs showing evidence of association with PCOS risk (**Supplementary Table 2**), rs2292239 and rs11171739 are

Table 1 GWAS, replication studies and meta-analysis results for the most significant SNPs

					GWAS-meta		Replication 1 (1,908 cases and 1,913 controls)		Replication 2 (6,318 cases and 5,665 controls)		GWAS-rep-meta		
SNP	Chr.	Nearby gene(s)	Alleles	MAF	OR	P	OR (95% CI)	P	OR (95% CI)	P	OR	P	P_{het}
rs2268361	2p16.3	*FSHR*	T/C	0.496	0.84	8.82×10^{-6}	0.91 (0.83-0.99)	3.31E-02	0.87 (0.83-0.92)	1.05×10^{-7}	0.87	9.89×10^{-13}	5.71×10^{-1}
rs2349415	2p16.3	*FSHR*	T/C	0.181	1.33	2.56×10^{-7}	1.32 (1.18-1.47)	2.51E-06	1.11 (1.05-1.19)	6.93×10^{-4}	1.19	2.35×10^{-12}	3.29×10^{-3}
rs4385527	9q22.32	*C9orf3*	A/G	0.219	0.78	9.62×10^{-6}	0.90 (0.82-0.99)	2.32E-02	0.86 (0.81-0.93)	4.64×10^{-5}	0.84	5.87×10^{-9}	1.74×10^{-1}
rs3802457	9q22.32	*C9orf3*	A/G	0.096	0.69	3.81×10^{-6}	0.88 (0.75-1.03)	1.17E-01	0.76 (0.69-0.83)	1.07×10^{-9}	0.77	5.28×10^{-14}	1.47×10^{-1}
rs1894116	11q22.1	*YAP1*	G/A	0.194	1.30	1.11×10^{-7}	1.21 (1.09-1.35)	6.29E-04	1.27 (1.20-1.36)	4.45×10^{-14}	1.27	1.08×10^{-22}	3.58×10^{-1}
rs705702	12q13.2	*RAB5B*, *SUOX*	G/A	0.245	1.32	1.09×10^{-9}	1.21 (1.09-1.34)	2.27E-04	1.26 (1.19-1.34)	7.44×10^{-15}	1.27	8.64×10^{-26}	4.73×10^{-1}
rs2272046	12q14.3	*HMGA2*	C/A	0.093	0.67	1.43×10^{-7}	0.80 (0.68-0.95)	1.07E-02	0.69 (0.63-0.76)	9.10×10^{-15}	0.70	1.95×10^{-21}	4.52×10^{-1}
rs4784165	16q12.1	*TOX3*	G/T	0.325	1.26	2.82×10^{-6}	1.09 (0.99-1.20)	7.24×10^{-2}	1.14 (1.08-1.21)	1.57×10^{-6}	1.15	3.64×10^{-11}	1.20×10^{-1}
rs2059807	19p13.3	*INSR*	G/A	0.301	1.24	1.58×10^{-6}	1.16 (1.05-1.28)	4.40×10^{-3}	1.09 (1.02-1.15)	6.61×10^{-3}	1.14	1.09×10^{-8}	1.83×10^{-2}
rs6022786	20q13.2	*SUMO1P1*	A/G	0.339	1.24	4.05×10^{-7}	1.11 (1.01-1.22)	3.77×10^{-2}	1.10 (1.04-1.16)	4.82×10^{-4}	1.13	1.83×10^{-9}	6.58×10^{-2}

Chr., chromosome; alleles, minor allele/major allele; MAF, minor allele frequency; P_{het}, P value of the heterogeneity test between study stages. MAFs for the controls are shown. OR is for the minor allele.

also T1D susceptibility SNPs. The direction of association is consistent with that seen in T1D studies (**Supplementary Table 5**). rs2292239 is located in intron 7 of *ERBB3*. The ERBB3 protein has a critical role in determining the function of antigen-presenting cells[16]. rs11171739 is also associated with the expression of *RPS26* and *SUOX*.

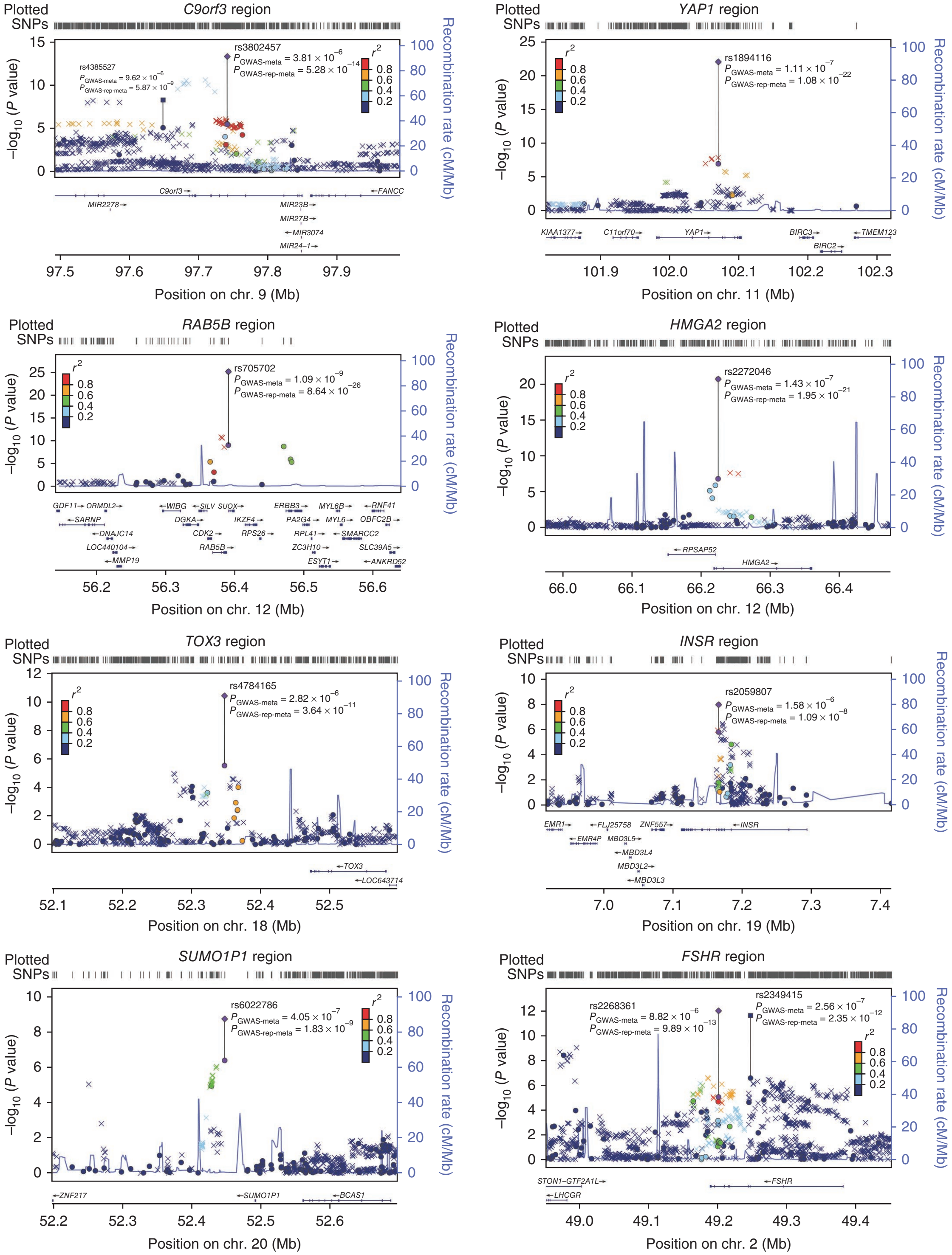

Figure 2 Regional plots of the eight newly discovered PCOS loci. Genotyped and imputed SNPs passing quality control are plotted with their meta-analysis *P* values as a function of genomic position (NCBI Build 37). At each locus, genotyped SNPs are plotted as circles, and imputed SNPs are shown as crosses. The index association SNP is represented in purple. $P_{GWAS\text{-}meta}$ indicates the combined results of the initial data sets, and $P_{GWAS\text{-}rep\text{-}meta}$ indicates the combined results of the initial and follow-up data sets, represented by the diamond (for the index SNP) or a square (for another independent SNP in this region). Estimated recombination rates (from 1000 Genome Asian populations (ASI)) are plotted to reflect the local LD structure. The color of the SNPs indicates LD with the index SNP according to pairwise r^2 values from 1000 Genome ASI individuals. Gene annotations were taken from the UCSC Genome Browser.

At 12q14.3, rs2272046 ($P_{\text{GWAS-reo-meta}} = 1.95 \times 10^{-21}$; $\text{OR}_{\text{GWAS-reo-meta}} = 0.70$) is localized in an intronic region of *HMGA2* (MIM 600698), which encodes a protein with structural DNA-binding domains that acts as a transcription-regulating factor (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs2272046, there were no additional association signals in this region (**Supplementary Table 4**). *HMGA2* has previously been shown to be associated with adult stature[17], vascular tumors, including angiomyxomas and pulmonary hamartomas[18], and T2D[19]. Women with PCOS have increased risk of T2D[5]. We investigated linkage between susceptibility variants and found that rs2272046 and rs1531343 (the strongest T2D susceptibility variant in a previous study)[19] showed little linkage disequilibrium (LD; $r^2 = 0.002$ and 0.003 in HapMap Han Chinese in Beijing, China (CHB) and Utah residents of Northern and Western European ancestry (CEU) populations) (**Supplementary Table 5**). Disruption of both *Hmga2* alleles results in the pygmy mouse, which has significantly lower body weight, lower amounts of fat tissue and infertility in both sexes relative to wild-type mice[20], all suggesting a vital role for the encoded protein in growth and reproduction.

At 19p13.3, rs2059807 ($P_{\text{GWAS-rep-meta}} = 1.09 \times 10^{-8}$; $\text{OR}_{\text{GWAS-rep-meta}} = 1.14$) is located in the intron of the *INSR* gene (MIM 147670) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs2059807, conditional logistic regression analysis identified no additional association signals (**Supplementary Table 4**). *INSR* has an important role in insulin metabolism, consistent with a very common explanation for the pathogenesis of PCOS—insulin resistance. Mutations affecting the tyrosine kinase domain of the insulin receptor are known to cause severe hyperinsulinemia and insulin resistance[21–23]. In previous studies, common SNPs in the *INSR* gene have been reported to be associated with PCOS in both Han Chinese and individuals of European ancestry[24,25]. *Insr*-null mice grow slowly and die by 7 days of age with ketoacidosis, high serum insulin and triglyceride concentrations, low glycogen stores and fatty livers[26].

The 2p16.3 locus was also reported in our previous GWAS to be associated with PCOS[6]. In that study, genome-wide significant signals in this region only mapped to the *LHCGR* gene (MIM 152790). The top signal was not directly linked with the *FSHR* gene (MIM 136435). In the current study, SNPs in the *FSHR* gene met the selection criteria for validation in the initial stage, reaching genome-wide significance in the combined analysis (top signal: rs2268361, $P_{\text{GWAS-rep-meta}} = 9.89 \times 10^{-13}$; $\text{OR}_{\text{GWAS-rep-meta}} = 0.87$) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Conditional logistic regression analysis supports the notion that the association of variants at *FSHR* is independent from the previously identified association signals in *LHCGR* (**Supplementary Table 4**). *FSHR* has long been considered one of the most compelling candidate genes for PCOS[27]. *FSHR*-null females are sterile, with small ovaries, blocked follicular development, atrophic uterus and imperforate vagina, and null males are fertile, despite lower testis weight, oligozoospermia and lower testosterone levels[28].

In addition to identifying candidate genes as the newly associated loci related to hormones, insulin resistance and growth, as may have been expected, we identified other candidate genes for which a connection to the PCOS pathogenesis mechanism is less clear. At 16q12.1, the most significantly associated SNP was rs4784165 ($P_{\text{GWAS-rep-meta}} = 3.64 \times 10^{-11}$; $\text{OR}_{\text{GWAS-rep-meta}} = 1.15$) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs4784165, conditional logistic regression analysis identified no additional association signal (**Supplementary Table 4**). *TOX3* (MIM 611416) is the nearest gene to this signal. The TOX3 protein belongs to the large and diverse family of high-mobility-group (HMG)-box proteins that function as architectural factors in the modification of chromatin structure by bending and unwinding DNA[29].

At 20q13.2, the top signal was rs6022786 ($P_{\text{GWAS-rep-meta}} = 1.83 \times 10^{-9}$; $\text{OR}_{\text{GWAS-rep-meta}} = 1.13$), located in an intergenic region between *SUMO1P1* and *ZNF217* (MIM 602967) (**Fig. 2**, **Table 1** and **Supplementary Table 3**). Controlling for rs6022786, conditional logistic regression analysis identified no additional association signals (**Supplementary Table 4**). *SUMO1P1* is a pseudogene of SUMO1 (pseudogene 1). *ZNF217 (*encoding zinc finger protein 217) attenuates apoptotic signals resulting from telomere dysfunction as well as from doxorubicin-induced DNA damage, potentially promoting neoplastic transformation by increasing cell survival during telomeric crisis, and may promote later stages of malignancy by increasing cell survival during chemotherapy[30]. These observations could be related to organ growth.

Notably, strong evidence for an intersection between the genetic basis of PCOS and diabetes mellitus was observed in the current study. Gonadotropin secretion and insulin signaling genes (for example, *FSHR* and *INSR*)[24,31] were already known to be relevant for both T2D and PCOS, but there has been insufficient evidence to support a relationship between other pathways and PCOS. Among the candidate genes identified within the associated loci (**Supplementary Table 3**), *HMGA2* and *THADA* (one of our previous GWAS findings) were also identified as candidate genes for T2D in recent GWAS[19,32]. The 12q13.2 locus also confers risk to T1D[10,11,13–15,33]. By investigating the potential functional relationships among these genes (**Supplementary Table 3**) using the PANTHER[34,35], Reactome[36,37] and KEGG[38] databases, we found that the T2D candidate genes *INSR* and *ERBB3* are implicated in the processes of female gamete generation, *LHCGR* and *FSHR* both encode hormone-hormone–binding receptors, *FBP1* (situated 76 kb downstream of rs4744370) and *INSR* are related through the insulin signaling pathway, *YAP1* and *ERBB3* are both involved in the ERBB signaling pathway[36,37], *LHCGR* and *ERBB3* are involved in calcium signaling[38], and *RAB5B* and *ERBB3* have roles in endocytosis[38]. New PCOS susceptibility loci may provide clues to PCOS etiology and identify gene networks of functional importance.

We also compared gene expression of these candidate genes in PCOS cases and healthy control subjects (**Supplementary Table 6**) and investigated the relationship between the PCOS susceptibility alleles and expression of related genes (**Supplementary Table 7**; see the **Supplementary Note** for discussion on the pathophysiology of PCOS).

In summary, our GWAS and meta-analyses confirmed three previously reported loci and identified eight new loci associated with PCOS susceptibility. Our study identified a number of candidate genes at these associated loci related to hormones and organ growth that are shared with T2D, and other candidate genes were related to cytokinesis and cell division. These findings may provide new directions for genetic and functional research of PCOS. In the future, we need to overcome challenges in discovering the biological mechanisms of disease behind these common variant associations.

URLs. R, http://www.r-project.org/; LocusZoom, http://csg.sph.umich.edu/locuszoom/; PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/; The International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/; SHEsis, http://analysis.bio-x.cn/myAnalysis.php; MACH, http://www.sph.umich.edu/csg/abecasis/MACH/index.html; IMPUTEv2, https://mathgen.stats.ox.ac.uk/impute/impute_v2.html; The 1000 Genome Project, http://www.1000genomes.org/; Gene Expression Omnibus (GEO), http://www.ncbi.nlm.nih.gov/geo/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all participants involved in this study. This study was supported by grants from the National Basic Research Program of China (973 Program: 2012CB944700, 2010CB529600, 2011CB944502 and 2010CB945002), the 863 Program (2012AA02A515), the National Natural Science Foundation of China (81130022, 30973170, 81121001, 31000553, 81000238, 81070461 and 81000236), the National Key Technology Research and Development Program (2011BAI17B00), the Science Research Foundation Item of No-earnings Health Vocation (201002013), the Program for Changjiang Scholars and the Innovative Research Team in University (IRT1025), the Foundation for the Author of National Excellent Doctoral Dissertation of China (201026) and the Program for New Century Excellent Talents in University (NCET-09-0550).

AUTHOR CONTRIBUTIONS

Z.-J.C., L.H. and Yongyong Shi designed the whole study. Yongyong Shi supervised the experiments and data analysis. Z.-J.C. supervised patient diagnosis and sample recruitment. Yongyong Shi, Z.L., H.Z., T.L. and J. Shen conducted data analyses and drafted the manuscript. H.Z., Yuhua Shi, L.G., J.M., Yingying Qin and J. Yan recruited samples. Y.C., D.Y., B.Z., X. Liang, Junzhao Zhao, D. Zhu, X.Z., Y.Y., Junli Zhao, S.Z., A.Y., J.L., J. Shi, J. Yang, H.J., J.X., X.Q., Y. Sun, Yajie Zhang, C.H., X.J., D. Zhao, C.R., X. Li, W.Z. and Yiwen Zhang coordinated and provided samples from different hospitals. J.C., Wenjin Li, Q.W., G.H., A.Z., Weidong Li, C.W., B.L. and Ying Qin performed or contributed to the main experiments. L.Y., Y. Zhao, D.W. and C.Z. performed DNA extraction. X.G and J.Z. performed endocrine biochemical examination. All authors critically reviewed the manuscript and approved the final version.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2384.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ehrmann, D.A., Barnes, R.B., Rosenfield, R.L., Cavaghan, M.K. & Imperial, J. Prevalence of impaired glucose tolerance and diabetes in women with polycystic ovary syndrome. *Diabetes Care* **22**, 141–146 (1999).
2. Goodarzi, M.O. & Azziz, R. Diagnosis, epidemiology, and genetics of the polycystic ovary syndrome. *Best Pract. Res. Clin. Endocrinol. Metab.* **20**, 193–205 (2006).
3. Carmina, E. Cardiovascular risk and events in polycystic ovary syndrome. *Climacteric* **12**, 22–25 (2009).
4. Kandaraki, E., Christakou, C. & Diamanti-Kandarakis, E. Metabolic syndrome and polycystic ovary syndrome and vice versa. *Arq. Bras. Endocrinol. Metabol.* **53**, 227–237 (2009).
5. Wild, S., Pierpoint, T., Jacobs, H. & McKeigue, P. Long-term consequences of polycystic ovary syndrome: results of a 31 year follow-up study. *Hum. Fertil. (Camb)* **3**, 101–105 (2000).
6. Chen, Z.J. *et al.* Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat. Genet.* **43**, 55–59 (2011).
7. Kerns, S.L. *et al.* Genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with the development of erectile dysfunction in African-American men after radiotherapy for prostate cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **78**, 1292–1300 (2010).
8. Li, T. *et al.* Identification of *YAP1* as a novel susceptibility gene for polycystic ovary syndrome. *J. Med. Genet.* **49**, 254–257 (2012).
9. Hao, Y., Chun, A., Cheung, K., Rashidi, B. & Yang, X. Tumor suppressor LATS1 is a negative regulator of oncogene *YAP. J. Biol. Chem.* **283**, 5496–5509 (2008).
10. Barrett, J.C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41**, 703–707 (2009).
11. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
12. Cooper, J.D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **40**, 1399–1401 (2008).
13. Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7**, e1002216 (2011).
14. Todd, J.A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat. Genet.* **39**, 857–864 (2007).
15. Hakonarson, H. *et al.* A novel susceptibility locus for type 1 diabetes on Chr12q13 identified by a genome-wide association study. *Diabetes* **57**, 1143–1146 (2008).
16. Wang, H. *et al.* Genetically dependent ERBB3 expression modulates antigen presenting cell function and type 1 diabetes risk. *PLoS ONE* **5**, e11789 (2010).
17. Weedon, M.N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nat. Genet.* **39**, 1245–1250 (2007).
18. Kazmierczak, B. *et al.* Cloning and molecular characterization of part of a new gene fused to HMGIC in mesenchymal tumors. *Am. J. Pathol.* **152**, 431–435 (1998).
19. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
20. Ashar, H.R. *et al.* Disruption of the architectural factor HMGI-C: DNA-binding AT hook motifs fused in lipomas to distinct transcriptional regulatory domains. *Cell* **82**, 57–65 (1995).
21. Moller, D.E. & Flier, J.S. Detection of an alteration in the insulin-receptor gene in a patient with insulin resistance, acanthosis nigricans, and the polycystic ovary syndrome (type A insulin resistance). *N. Engl. J. Med.* **319**, 1526–1529 (1988).
22. Moller, D.E., Yokota, A., White, M.F., Pazianos, A.G. & Flier, J.S. A naturally occurring mutation of insulin receptor alanine 1134 impairs tyrosine kinase function and is associated with dominantly inherited insulin resistance. *J. Biol. Chem.* **265**, 14979–14985 (1990).
23. Taylor, S.I. *et al.* Mutations in insulin-receptor gene in insulin-resistant patients. *Diabetes Care* **13**, 257–279 (1990).
24. Chen, Z.J. *et al.* Correlation between single nucleotide polymorphism of insulin receptor gene with polycystic ovary syndrome. *Zhonghua Fu Chan Ke Za Zhi* **39**, 582–585 (2004).
25. Siegel, S. *et al.* AC/T single nucleotide polymorphism at the tyrosine kinase domain of the insulin receptor gene is associated with polycystic ovary syndrome. *Fertil. Steril.* **78**, 1240–1243 (2002).
26. Accili, D. *et al.* Early neonatal death in mice homozygous for a null allele of the insulin receptor gene. *Nat. Genet.* **12**, 106–109 (1996).
27. Simoni, M., Tempfer, C.B., Destenaves, B. & Fauser, B. Functional genetic polymorphisms and female reproductive disorders: Part I: polycystic ovary syndrome and ovarian response. *Hum. Reprod. Update* **14**, 459–484 (2008).
28. Sun, L. *et al.* FSH directly regulates bone mass. *Cell* **125**, 247–260 (2006).
29. O'Flaherty, E. & Kaye, J. TOX defines a conserved subfamily of HMG-box proteins. *BMC Genomics* **4**, 13 (2003).
30. Huang, G. *et al.* ZNF217 suppresses cell death associated with chemotherapy and telomere dysfunction. *Hum. Mol. Genet.* **14**, 3219–3225 (2005).
31. Sudo, S. *et al.* Genetic and functional analyses of polymorphisms in the human FSH receptor gene. *Mol. Hum. Reprod.* **8**, 893–899 (2002).
32. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40**, 638–645 (2008).
33. Cooper, J.D. *et al.* Analysis of 55 autoimmune disease and type II diabetes loci: further confirmation of chromosomes 4q27, 12q13.2 and 12q24.13 as type I diabetes loci, and support for a new locus, 12q13.3-q14.1. *Genes Immun.* **10**, S95–S120 (2009).
34. Mi, H. *et al.* PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* **38**, D204–D210 (2010).
35. Thomas, P.D. *et al.* PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res.* **31**, 334–341 (2003).
36. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
37. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37**, D619–D622 (2009).
38. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).

[1]Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Bio-X Institutes, Ministry of Education, Shanghai Jiao Tong University, Shanghai, China. [2]Shanghai genomePilot Institutes for Genomics and Human Health, Shanghai, China. [3]Changning Mental Health Center, Shanghai, China. [4]Center for Reproductive Medicine, Shandong Provincial Hospital, Shandong University, Jinan, China. [5]National Research Center for Assisted Reproductive Technology and Reproductive Genetics, Jinan, China. [6]The Key Laboratory for Reproductive Endocrinology, Ministry of Education of the People's Republic of China, Jinan, China. [7]Shandong Provincial Key Laboratory of Reproductive Medicine, Jinan, China. [8]Reproductive Medicine Center, The First Affiliated Hospital, Anhui Medical University, Hefei, China. [9]Department of Obstetrics and Gynecology, Sun Yat-sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, China. [10]Reproductive Medicine Center, The Maternal and Child Health Hospital of Guangxi Zhuang Autonomous Region, Nanning, China. [11]Reproductive Medicine Center, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China. [12]Shanghai Institute of Mental Health, Shanghai, China. [13]Reproductive Medicine Unit, The First Affiliated Hospital of Wenzhou Medical College, Wenzhou, China. [14]Reproductive Medicine Center, Linyi People's Hospital, Linyi, China. [15]Center for Reproductive Medicine, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [16]National Institute of Hospital Administration, Ministry of Health of the People's Republic of China, Beijing, China. [17]Reproductive Medicine Center, Affiliated Hospital of Ningxia Medical University, Yinchuan, China.

[18]Center for Reproductive Medicine, Qingdao Women & Children Medical Healthcare Center, Qingdao, China. [19]Reproductive Medicine Center, Affiliated Hospital of Jining Medical College, Jining, China. [20]Department of Obstetrics and Gynecology, First Affiliated Hospital of Nanjing Medical University, Nanjing, China. [21]Oncology Translational Medicine, GlaxoSmithKline R&D China, Shanghai, China. [22]Assisted Reproduction Center, Maternal and Child Health Care Hospital of Shaanxi Province, Xi'an, China. [23]Reproductive Medical Center, Renmin Hospital of Wuhan University, Wuhan, China. [24]Department of Obstetrics and Gynecology, 105th Hospital of People's Liberation Army, Hefei, China. [25]Department of Obstetrics and Gynecology, The Affiliated Hospital of Medical College Qingdao University, Qingdao, China. [26]Department of Obstetrics and Gynecology, Jinan Health Institute of Maternity and Infant, Jinan, China. [27]Reproductive Medical Center, Yantai Yuhuangding Hospital, Yantai, China. [28]Department of Obstetrics and Gynecology, Tengzhou People's Hospital, Tengzhou, China. [29]Reproductive Medical Center, Shengjing Hospital of China Medical University, Shenyang, China. [30]Department of Obstetrics and Gynecology, Affiliated Hospital of Weifang Medical College, Weifang, China. [31]Department of Obstetrics and Gynecology, Anqiu People's Hospital, Anqiu, China. [32]Department of Gynecology, Obstetrics & Gynecology Hospital of Fudan University, Shanghai, China. [33]Reproductive Medicine Unit, Weihai Women and Children Hospital, Weihai, China. [34]Institutes of Biomedical Sciences, Fudan University, Shanghai, China. [35]Institute for Nutritional Sciences, Shanghai Institute of Biological Sciences, Chinese Academy of Sciences, Shanghai, China. [36]Bio-X Institutes, Shanghai Jiao Tong University, Shanghai, China. [37]These authors contributed equally to this work. Correspondence should be addressed to Z.-J.C. (chenzijiang@hotmail.com) or L.H. (helinhelin@gmail.com).

ONLINE METHODS

Subjects. All Han Chinese samples in this study were obtained in multiple collaborating hospitals from China. The discovery sets (GWAS 1 and 2) of 2,254 Han Chinese PCOS samples and 3,001 controls were recruited mainly from northern China. Subsequent replication samples (replication 1 and 2) were collected from 29 provinces (Shandong, Heilongjiang, Jilin, Liaoning, Inner Mongolia, Hebei, Henan, Tianjin, Beijing, Shanxi, Shaanxi, Gansu, Ningxia, Jiangsu, Anhui, Shanghai, Guangdong, Guangxi, Fujian, Zhejiang, Hubei, Hunan, Jiangxi, Sichuan, Chongqing, Xinjiang, Yunnan, Guizhou and Hainan) throughout China. Individuals were diagnosed with PCOS according to the Rotterdam Consensus proposed from 2003 (ref. 39), requiring the presence of any two of the following three criteria: oligoovulation and/or anovulation; clinical and/or biochemical signs of hyperandrogenism; and polycystic ovary morphology. PCOS cases in GWAS 1 were diagnosed strictly and satisfied all three features. Oligoovulation and/or anovulation were defined as menstrual cycles of more than 35 days in length or a history of ≤8 menstrual cycles in a year. A finding of polycystic ovarian morphology was determined when ≥12 follicles measuring 2–9 mm in diameter were scanned in either ovary or the ovarian volume was greater than 10 ml. Hyperandrogenism was confirmed if there was evidence for hyperandrogenemia and/or hirsutism. Affected individuals with other causes of oligomenorrhea or hyperandrogenism (for example, non-classical 21-hydroxylase deficiency, Cushing syndrome, hypothyroidism or elevated prolactin) were excluded. Subject information is summarized in **Supplementary Table 8**. Among the subjects with PCOS, 37.12% had BMI of ≥25, 91.35% had irregular menstrual cycle, 45.14% had increased serum testosterone levels, 11.06% had clinical hyperandrogenism with hirsutism (F-G score ≥ 6), and 96.22% had polycystic ovarian morphology as evaluated by ultrasound examination.

All controls were healthy women recruited through routine physical examination or tubal factor infertility and matched PCOS cases from each institute. Endocrine and biochemical parameters were also measured to exclude hyperandrogenism, and ultrasound imaging was performed to exclude ovarian morphology indicative of PCOS. All participants provided written informed consent. The study was approved by the Institutional Ethical Committee of each hospital and was conducted according to Declaration of Helsinki principles.

DNA extraction. EDTA anti-coagulated venous blood samples were collected from all participants. Genomic DNA was extracted from peripheral blood lymphocytes by standard procedures using Flexi Gene DNA kits (Qiagen) and was diluted to working concentrations of 50 ng/µl for genome-wide genotyping and 15–20 ng/µl for the validation study.

GWAS genotyping and quality control. Affymetrix Genome-Wide Arrays were used for the discovery phase: GWAS data set 1 was genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0, and samples from GWAS data set 2 were genotyped using Axiom Genome-Wide Arrays. Quality control filtering of the GWAS data was performed as follows. For the SNP 6.0 arrays, samples with contrast quality control of 0.4 or greater were left out of further data analysis. For the Axiom arrays, samples with dish quality control (DQC) of 0.82 or better were considered to have passed. Genotype data were generated using the birdseed algorithm for SNP 6.0 and the Axiom GT1 algorithm for Axiom arrays. For sample filtering, arrays for which genotypes were generated for fewer than 95% of loci were excluded. For SNP filtering (after sample filtering), SNPs with call rates of <95% in either case or control samples were removed. SNPs whose MAF was <1% or deviated significantly from Hardy-Weinberg Equilibrium (HWE, $P \leq 1 \times 10^{-5}$) in controls were excluded.

Imputation analysis of ungenotyped SNPs. To conduct meta-analysis across array types, imputation was carried out separately for both GWAS data sets using MACH[40,41] (see URLs). Phased haplotypes for 90 CHB and Japanese in Tokyo, Japan (JPT) subjects (180 haplotypes) were used as the reference for imputing genotypes. Any SNP imputed with information content $r^2 < 0.3$ was excluded from association analysis because of lack of power. In addition, we performed a second imputation step using IMPUTEv2 (refs. 42,43) (see URLs) for the eight newly identified regions (0.5 MB on either side of any SNP that achieved $P_{\text{GWAS-meta}} < 1 \times 10^{-5}$), using the 1000 Genomes haplotypes Phase 1 interim release (Jun2011; see URLs) as reference. SNPs imputed with proper info of <0.4 were treated as poor quality. The criteria for SNP quality control filtering were the same as for the genotyped ones.

Analysis of population substructure. Population substructure was evaluated using PCA, as implemented in the EIGENSTRAT software[44]. Twenty principal components were generated for each subject. PCA was conducted on the study samples combined with the HapMap samples. The first two principal components were plotted (**Supplementary Fig. 1**). Those samples that deviated from the main body of test samples (**Supplementary Fig. 1**, gray crosses; 43 cases and 9 controls) were excluded.

Association analysis. Logistic regression was used to determine whether there was a significant difference in principal component scores between cases and controls; significant principal components were used as covariates in the association analysis to correct for population stratification. After adjustment, little stratification was observed ($\lambda = 1.07$, $\lambda_{1000} = 1.04$, standardized to a sample size of 1000 (ref. 45)).

Meta-analysis of GWAS. The GWAS data sets were combined using meta-analysis. The meta-analysis was conducted using PLINK[46] (see URLs). Heterogeneity across the three stages was evaluated using Q-statistic P values. The Mantel-Haenszel method was used to calculate the fixed-effect estimate.

SNP selection and replication. The following criteria were used for the selection of SNPs for validation. SNPs with strong, significant association ($P_{\text{GWAS-meta}} \leq 1 \times 10^{-5}$) in the GWAS meta-analysis were selected for replication 1. Generally, those SNPs that showed nominal significance ($P < 0.05$) in replication 1 or were not significant in replication 1 but had a GWAS and replication 1 meta-analysis P value of less than 5×10^{-6} were kept for replication 2. The Sequenom MassArray system was used for most of the replication studies, except for rs2059807, which was genotyped using TaqMan assays (Life Technologies).

Statistical analysis. Genome-wide association analysis at the single-marker level and the HWE analysis in the case-control samples were performed using PLINK[46], and the R package was used to generate the genome-wide P-value plot (see URLs). Regional plots were generated using LocusZoom[47] (see URLs). In the replication studies, allelic association analysis was conducted using SHEsis[48] (see URLs). The GWAS and replication data were also combined using meta-analysis performed with PLINK[46]. Conditional logistic regression was used to test for independent effects of an individual SNP[6,49]. The association analysis of subphenotype groups of PCOS versus controls was conducted using PLINK (**Supplementary Table 9**).

Gene expression analysis and eQTL analysis. Gene expression profile and eQTL data sets were downloaded from NCBI GEO (see URLs). Because all our samples are female, only female samples of the data sets were kept for further analysis. For Affymetrix gene expression arrays, the raw data were normalized by MAS5.0 and robust multi-array average (RMA), and gene expression analysis was carried out by t test after normalization. The eQTL analysis was performed as described in a previous study[50].

39. Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Fertil. Steril.* **81**, 19–25 (2004).
40. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
41. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
42. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
43. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
44. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

doi:10.1038/ng.2384

45. Lindgren, C.M. *et al.* Genome-wide association scan meta-analysis identifies three loci influencing adiposity and fat distribution. *PLoS Genet.* **5**, e1000508 (2009).
46. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
47. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
48. Shi, Y.Y. & He, L. SHEsis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci. *Cell Res.* **15**, 97–98 (2005).
49. Petukhova, L. *et al.* Genome-wide association study in alopecia areata implicates both innate and adaptive immunity. *Nature* **466**, 113–117 (2010).
50. Shi, Y. *et al.* Common variants on 8p12 and 1q24. 2 confer risk of schizophrenia. *Nat. Genet.* **43**, 1224–1227 (2011).

doi:10.1038/ng.2384