European Journal of Human Genetics (2016) 24, 895–902

www.nature.com/ejhg

ARTICLE

Prediction of male-pattern baldness from genotypes

Fan Liu[1,2], Merel A Hamer[3], Stefanie Heilmann[4,5], Christine Herold[6], Susanne Moebus[7], Albert Hofman[8], André G Uitterlinden[9,8], Markus M Nöthen[4,5], Cornelia M van Duijn[8], Tamar EC Nijsten[3] and Manfred Kayser*[,1]

The global demand for products that effectively prevent the development of male-pattern baldness (MPB) has drastically increased. However, there is currently no established genetic model for the estimation of MPB risk. We conducted a prediction analysis using single-nucleotide polymorphisms (SNPs) identified from previous GWASs of MPB in a total of 2725 German and Dutch males. A logistic regression model considering the genotypes of 25 SNPs from 12 genomic loci demonstrates that early-onset MPB risk is predictable at an accuracy level of 0.74 when 14 SNPs were included in the model, and measured using the area under the receiver-operating characteristic curves (AUC). Considering age as an additional predictor, the model can predict normal MPB status in middle-aged and elderly individuals at a slightly lower accuracy (AUC 0.69–0.71) when 6–11 SNPs were used. A variance partitioning analysis suggests that 55.8% of early-onset MPB genetic liability can be explained by common autosomal SNPs and 23.3% by X-chromosome SNPs. For normal MPB status in elderly individuals, the proportion of explainable variance is lower (42.4% for autosomal and 9.8% for X-chromosome SNPs). The gap between GWAS findings and the variance partitioning results could be explained by a large body of common DNA variants with small effects that will likely be identified in GWAS of increased sample sizes. Although the accuracy obtained here has not reached a clinically desired level, our model was highly informative for up to 19% of Europeans, thus may assist decision making on early MPB intervention actions and in forensic investigations.
European Journal of Human Genetics (2016) **24,** 895–902; doi:10.1038/ejhg.2015.220; published online 28 October 2015

INTRODUCTION

Male-pattern baldness (MPB) or androgenic alopecia is the most common type of hair loss in men, with a prevalence of around 20% at age 20–30, and the incidence growing at 10% per decade.[1,2] MPB is a chronic problem commonly seen by dermatologists,[3] with severe psychosocial consequences[4] and limited effective therapeutic options.[5] The effectiveness of most treatments (eg, minoxidil or finasteride) relies on how early they are applied. Therefore, the ability to predict the early-onset or normal MPB status using DNA variants may have important implications for treatment strategies. Furthermore, owing to its widespread prevalence and the fact that most criminals are men, MPB in principle could help identify unknown perpetrators via the concept of forensic DNA phenotyping,[6,7] especially in light of the current progress in predicting chronological age from DNA data.[8,9] However, so far there is no established genetic model for predicting MPB from genetic data providing motivation for the present study.

MPB is a highly heritable visible trait, with estimated heritability of about 80% in young[2] and elderly males.[10] A locus on chromosome Xq12 harboring the androgen receptor gene (*AR*) and its neighboring ectodysplasin A2 receptor gene (*EDA2R*) is known as the major locus for MPB.[11,12] In addition, two genetic loci on chromosome 20p11 (*PAX1*/*FOXA2*) and 7p21.1 (*HDAC9*) were identified to be involved in MPB.[13–15] A meta-analysis of seven GWASs for early-onset MPB involving ~13 000 individuals of European origin conducted by the International MAAN Consortium[16] replicated these loci and highlighted five additional loci showing genome-wide significant association with MPB; these included 1p36.22, 2q37.3, 7q11.22, 17q21.31, and 18q12.3. Individuals in the highest-risk quartile of a genotype score had an approximately sixfold increased risk of early-onset MPB. In addition, the most recent study by Heilmann *et al*[17] comprising 2759 cases and 2661 controls successfully identified four additional autosomal loci showing genome-wide significant association with MPB; these included 2q35, 3q25.1, 5q33.3, and 12p12.1 (identifying a total of 12 genetic loci so far). These findings on one hand highlight a relatively strong X-linked major gene locus effect (odds ratio per risk allele ~2.5 at the *AR* locus) and on the other hand suggest a highly polygenetic autosomal component (odds ratios at individual loci <1.5). Here we estimate to what degree MPB is predictable using the currently known DNA markers discovered from GWAS so far. This study includes a total of 2725 German and Dutch males, with included early-onset MPB patients and controls as well as middle-aged and elderly cases and controls.

MATERIALS AND METHODS

Ethics statement

All studies were approved by the institutional ethics review committees at the relevant organizations, and written informed consent was provided by all participating individuals.

[1]Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands; [2]Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China; [3]Department of Dermatology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands; [4]Department of Genomics, Life and Brain Center, University of Bonn, Bonn, Germany; [5]Institute of Human Genetics, University of Bonn, Bonn, Germany; [6]German Center for Neurodegenerative Disease (DZNE), Bonn, Germany; [7]Institute of Medical Informatics, Biometry, and Epidemiology, University Hospital of Essen, University Duisburg-Essen, Essen, Germany; [8]Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands; [9]Department of Internal Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands
*Correspondence: Professor M Kayser, Department of Genetic Identification, Erasmus MC University Medical Center Rotterdam, PO Box 2060, 3000 CB Rotterdam, The Netherlands. Tel: +31 10 7038073; Fax: +31 10 7044575; E-mail: m.kayser@erasmusmc.nl
Received 17 April 2015; revised 27 August 2015; accepted 1 September 2015; published online 28 October 2015

Rotterdam Study

The Rotterdam Study (RS) is a population-based prospective study of Dutch elderly subjects (>45 years of age) consisting of an initial cohort and two extensions.[18] MPB status was assessed by trained physicians according to the Norwood–Hamilton grading scale[1,19] with grades 1–12. Cases were defined as grade IV–VIII and otherwise controls. The current study included 1161 male RS subjects. RS samples have not been used for MPB GWAS before and are not part of MAAN; hence, RS sample are completely independent from previous MPB SNP discoveries. Anonymized individual-level phenotype and genotype data used for the prediction analysis from the Rotterdam Study participants are available in Supplementary Table S1.

Erasmus Rucphen Family Study

Erasmus Rucphen Family (ERF) is a family-based study that includes inhabitants of a genetically isolated community in the south–west of the Netherlands, studied as part of the Genetic Research in Isolated Population program.[20] Study population includes ~3000 individuals who are living descendants of 22 couples who had at least six children baptized in the community church. All data were collected between 2002 and 2005. The population shows minimal immigration and high inbreeding. Cases were defined as Norwood–Hamilton grading scale[1,19] IV–VIII at any age or grade II–III between 50 and 60 years of age or grade I before 50 years of age, and controls otherwise. The current study included 567 male ERF subjects. ERF samples have not been used for MPB GWAS before and are not part of MAAN; hence, ERF samples are completely independent from previous MPB SNP discoveries. The ERF Study data are archived in European Genome–Phenome Database (EGA) with the accession code EGAS00001001134.

BONN Study

The hair status of each participant was assessed by a dermatologist according to the Hamilton/Norwood (HN) classification.[1,3] Affected men were aged <30 years with HN grades IV–VII, or <40 years with HN grades V–VII, and were thus representative of the most severely affected 10% for the respective age classes ($N=581$). The control sample comprises 270 men aged ⩾60 years with no signs of AGA (20% least affected individuals in the population) and 146 male controls (HN⩽V) that were recruited as part of the Heinz Nixdorf Recall cohort (risk factors, evaluation of coronary calcium and lifestyle) at the University of Essen. All the cases and controls were of German descent.[13] Note that the 581 cases and the 146 controls from the BONN Study used here were part of the initial MAAN study,[16] and all BONN subjects were also part of the previous study of Heilmann *et al*,[17] based on which the SNPs used here for MPB prediction were initially discovered. There is thus a potential risk in overestimating the prediction accuracy in BONN but at most by only a small degree as BONN was only a small component of the MAAN study[16] (ie, 14.9% of cases and 1.6% of controls in MAAN and 21.1% of cases and 15.6% of controls in the Heilmann study). The RS and ERF subjects are completely independent from previous studies. The BONN Study data are archived in European Genome–Phenome Database (EGA) with the accession code EGAS00001001354.

Genotyping and quality control

The genotyping platforms, quality controls, and imputation methods used in participant studies have been described in detail previously. In brief, extensive quality control thresholds were applied to include common SNPs (minor allele frequency >1%) with a high call rate (95%) for genotyped SNPs; SNPs demonstrating deviation from Hardy–Weinberg equilibrium ($P<10^{-6}$) were excluded. SNP genotypes from all cohorts were imputed using the 1000-Genome Project as the reference panel with high-quality metrics (variance ratio 0.3 for MACH and proper info statistic 0.4 for IMPUTE).[21,22] The X-chromosome in ERF was imputed using HapMap-CEU samples as the reference. After all quality controls, this study includes 2 444 603 autosomal SNPs and 239 421 X-chromosome SNPs in RS; 2 266 959 autosomal SNPs and 89 191 X-chromosome SNPs in BONN; and 6 099 730 autosomal SNPs and 16 708 X-chromosome SNPs in ERF. All SNPs are described using dbSNP ID according to human reference assembly GRCh37.p13.

Statistical analyses

We initially selected 20 SNPs from 12 genomic loci for prediction analysis from the Table 2 and the Supplementary Table S2 of Li *et al*[16] as well as Table 1 of Heilmann *et al*.[17] A candidate SNP analysis was conducted using logistic regression in all three cohorts for the 20 SNPs assuming additive allele effect adjusted for age at examination when appropriate. As the *AR* locus has a relatively large effect, we additionally selected four SNPs from the *AR* locus showing some residual effect based on a conditional logistic regression analysis of all SNPs within 66.5–67.9 Mbp of the *AR* locus in a stepwise manner, that is, until the newly included SNP is not significant anymore at $P<0.05$ level in a multivariate analysis of all SNPs accumulated in previous steps. A prediction analysis including all 24 selected SNPs was conducted separately in all three cohorts; age at examination was included as a predictor in RS and ERF but not in BONN consisting of early-onset cases and screened elderly controls. Then the final selection and ranking of the SNP predictors was based on stepwise analysis of the Akaike information criterion[23] using R function '*step*'. The prediction models were trained separately in the three cohorts based on multivariate logistic regression. Because our sample size is not large and an explicit validation set was not available, we used the leave-one-out cross-validation method to prevent overfitting, that is, the prediction models were trained in all subjects except one, who was used for validation by applying the trained model on this subject. We repeated this procedure iteratively over the whole cohort by leaving each subject out, and obtained the predicted probability of baldness status for all subjects. The predicted probabilities are compared with observed baldness status, where the AUCs[24] were derived as an overall measurement of prediction accuracy. An AUC value ranges from 0.5 representing random prediction to 1.0 representing perfect prediction. Binary prediction of baldness status for each subject was defined if the predicted probability is >0.5 otherwise non-bald. The predicted and observed baldness status was compared using a confusion table, where sensitivity and specificity values are derived, both ranging from 0 to 1. Sensitivity and specificity values are a pair of inseparable accuracy parameters for a binary classifier; a perfect classifier would be described as 100% sensitive and 100% specific. All candidate SNP analysis and prediction analysis were conducted in R version 3.2.0 (http://www.r-project.org/).

Estimates of the proportion of variance explained were calculated using the Genome-wide Complex Trait Analysis (GCTA) tool v1.24 (http://gump.qimr.edu.au/gcta/).[25] Genetic relationships were estimated using all autosomal SNPs (–make-grm, –maf 0.03). The top 10 eigenvectors from PCA analysis (–pca) were then used as covariates in a restricted maximum likelihood analysis (–reml) to estimate the proportion of the variance explained by SNPs (V_G/V_P or narrow-sense heritability h^2), repeated for all autosomes as a whole, each autosome separately, and the X chromosome.

A GWAS for MPB status was separately conducted in all cohorts using logistic regression considering age at examination as a covariate (except in BONN, which used early-onset cases) using PLINK1.9 beta.[26] Family relationship was adjusted using first four principal components from EIGENSTRAT[27] analysis. A meta-analysis of GWAS results was conducted using inverse variance fixed-effect analysis. *P*-values equal to or smaller than 5×10^{-8} were considered as genome-wide significant.

RESULTS

The characteristics of study subjects (all male) are summarized in Supplementary Table S2 (BONN: 581 early-onset cases and 416 controls; RS: 619 cases and 542 controls; and ERF: 252 cases and 315 controls). The BONN data have been described in a previous GWAS,[13] whereas RS and ERF data have not been used before for the genetic investigation of male-pattern baldness. Note that the BONN Study included only early-onset MPB cases (<40 years of age) and screened elderly controls while the RS (mainly elderly individuals: mean age 67.79 years, min 51.54, max 96.73) and ERF (mainly middle-aged: mean age 49.05 years, min 18.07, max 79.03) are population-based studies without sample selection based on MPB status. The difference in MPB prevalence observed between RS

(53.3%) and ERF (44.4%) is fairly consistent with a ~10% increase per decade of MPB[1] (Supplementary Table S2).

Candidate SNP analysis

We focused on 20 candidate SNPs from 11 autosomal loci and the *AR/EDA2R* locus on X chromosome previously reported by Li *et al*[16] and Heilmann *et al*[17] (Table 1). Part of the BONN subjects was used in Li *et al* and Heilmann *et al*, so as expected, the allelic effects in BONN for all tested SNPs was consistent with previous reports and most of the tested SNPs showed significant association with MPB ($P<0.05$). In RS and ERF, the risk alleles previously reported in other populations also showed higher frequencies in cases than in controls for all SNPs tested, that is, OR > 1.0, and seven SNPs at five genetic loci were nominally significantly associated with MPB (chr1p36: rs12565727 $P=2.71\times10^{-4}$; chr7q11: rs6945541 $P=0.039$; chr17q21: rs17762954 $P=0.014$ and rs12373124 $P=5.18\times10^{-3}$; chr18q12: rs10502861 $P=0.036$; chrXq12: rs1511061 $P=2.16\times10^{-11}$; and rs2497938 $P=1.38\times10^{-12}$; Table 1). In ERF, the allelic effect for all tested SNPs was consistent with previous studies and five SNPs at four genetic loci showed significant association with MPB (chr1p36: rs12565727 $P=1.37\times10^{-3}$; chr7q11: rs2073963 $P=9.63\times10^{-4}$; chr7q11: rs6945541 $P=0.037$; chrXq12: rs1511061 $P=4.98\times10^{-5}$; and rs2497938 $P=6.38\times10^{-5}$). The genetic associations at 2q37 and 20p11 were not statistically significant associated in both RS and ERF, likely due to the smaller sample sizes compared with the previous GWASs. In the present meta-analysis of the three cohorts, the only locus that did not show significant association was 3q25.1, but the allelic effect was in the same direction as Li *et al.* Overall, the observed allele effect sizes were similar between RS and ERF, and similar to previous estimates in other populations.[14–16] Noticeable exceptions were: (1) the chromosome 20p11 SNPs showed much smaller effects in RS and ERF (eg, rs6047844 OR = 1.09 in RS and 1.18 in ERF, Table 1) than previous estimations (OR = 1.60 in Li *et al* and 1.82 in BONN); and (2) the SNPs at the *EDA2R/AR* locus showed much larger effect (rs1511061 OR = 9.07) in BONN than in RS and ERF (OR 2.68–2.75). These differences are likely explained by the extreme design in BONN, as BONN included only early-onset MPB cases (< 40 years of age) and a set of super controls (⩾ 60 years, no signs of MPB) while no MPB pre-selection was made in RS and ERF, in which the majority of the subjects were elderly people.

Prediction analysis

A prediction analysis included all 20 SNPs listed in Table 1 as predictors. We also added four extra X chromosome SNPs within a 66.5–67.9-Mbp region of the *AR* locus (rs113043121, rs471205, rs141476270, and rs182829063), which showed significant residual effect after conditioning on the top-associated X SNPs based on a stepwise conditional analysis (Table 2). For predicting MPB in RS and ERF we also added age as additional predictor (which in BONN was left out because of an extreme case–control design, that is, early-onset MBP patients and elderly screened controls). The overall prediction accuracy was derived using receiver-operating characteristic curves generated separately in the three studies (Figure 1a). In BONN, the AUC was estimated at 0.74 (sensitivity = 0.84, specificity = 0.50; Table 2) using 14 DNA markers and adding additional markers did not improve the prediction accuracy. In RS, the maximal AUC value was estimated at 0.71 (sensitivity = 0.73, specificity = 0.58) when age and 11 DNA markers were included (Table 2). In ERF, AUC was estimated at 0.69 (sensitivity = 0.60, specificity = 0.69) when age and 6 DNA markers were included (Table 2). Note that in ERF, some SNPs on chr2q35, chr17q21, and the X chromosome were not available, which could explain a lower accuracy in ERF than in RS. The extra SNPs were not available in ERF due to differences in the imputation procedures on the X chromosome (see Materials and Method section). The slightly lower prediction accuracies in RS/ERF *versus* BONN are likely explained by the fact that extreme early-onset MPB cases were selected in BONN, while RS and ERF are population-based cohorts without early-onset MPB selection. As may be expected,

Table 1 SNPs from previous Li *et al*[16] and Heilmann *et al*[17] associated with male-pattern baldness in BONN, RS, and ERF

Chr	*SNP*	*HGVS*	*Previous studies*				*BONN*		*RS*		*ERF*		*Meta*	
			EA/NEA	*OR*	*P*	*Ref*	*OR*	*P*	*OR*	*P*	*OR*	*P*	*OR*	*P*
1p36.22	rs12565727	g.11033082A>G	A/G	1.33	9.07E-11	Li *et al*[16]	1.40	**3.55E-03**	1.35	**2.71E-04**	1.64	**1.37E-03**	1.42	**3.26E-07**
2q35	rs10193725	g.218861775T>C	C/T	1.25	1.46E-10	Heilmann *et al*[17]	1.30	**1.99E-02**	1.01	7.74E-01	—	—	1.13	1.05E-01
	rs7349332	g.219756383C>T	T/C	1.34	3.55E-15	Heilmann *et al*[17]	1.46	**4.18E-03**	1.02	7.08E-01	—	—	1.20	**3.73E-02**
2q37.3	rs9287638	g.239694631C>A	A/C	1.31	1.01E-12	Li *et al*[16]	1.31	**7.05E-03**	1.13	2.53E-01	—	—	1.21	**4.59E-03**
	rs9751918	g.239735812A>G	G/A	1.27	1.71E-08	Li *et al*[16]	1.16	1.81E-01	1.19	4.60E-01	1.13	3.90E-01	1.16	**2.51E-02**
3q25.1	rs7648585	g.151639765A>G	G/A	1.18	1.20E-09	Heilmann *et al*[17]	1.08	4.33E-01	1.01	6.54E-01	1.11	4.35E-01	1.01	8.53E-01
	rs4679955	g.151653368A>T	T/A	1.19	1.79E-10	Heilmann *et al*[17]	1.10	2.88E-01	1.01	7.42E-01	1.05	6.82E-01	1.03	5.83E-01
5q33.3	rs929626	g.158310631A>G	A/G	1.19	2.12E-11	Heilmann *et al*[17]	1.39	**4.39E-04**	1.16	1.73E-01	1.26	6.66E-02	1.26	**4.57E-05**
	rs1081073	g.158381512T>A	T/A	1.17	8.52E-09	Heilmann *et al*[17]	1.25	**2.00E-02**	1.19	8.32E-02	1.23	9.98E-02	1.20	**8.66E-03**
7p21.1	rs2073963	g.18877874T>G	G/T	1.29	1.08E-12	Li *et al*[16]	1.43	**1.87E-04**	1.15	7.03E-01	1.50	**9.63E-04**	1.32	**1.31E-06**
7q11.22	rs6945541	g.68611960C>T	C/T	1.27	1.71E-09	Li *et al*[16]	1.52	**4.38E-06**	1.15	**3.97E-02**	1.30	**3.75E-02**	1.31	**1.34E-06**
12p12.1	rs9668810	g.26426420T>C	T/C	1.21	1.09E-10	Heilmann *et al*[17]	1.25	**3.16E-02**	1.02	4.23E-01	1.28	8.73E-02	1.15	**2.86E-02**
	rs7975017	g.26428793T>C	T/C	1.21	4.03E-10	Heilmann *et al*[17]	1.28	**3.16E-02**	1.02	5.35E-01	1.25	1.49E-01	1.09	3.25E-01
17q21.31	rs17762954	g.43899786C>T	C/T	1.32	4.87E-08	Li *et al*[16]	1.17	1.58E-01	1.18	**1.40E-02**	—	—	1.18	**3.74E-02**
	rs12373124	g.43924219T>C	T/C	1.33	5.07E-10	Li *et al*[16]	—	—	1.22	**5.18E-03**	—	—	—	—
18q12.3	rs10502861	g.42800148C>T	C/T	1.28	2.62E-09	Li *et al*[16]	1.42	**7.96E-04**	1.12	**3.68E-02**	1.10	5.20E-01	1.17	**1.12E-02**
20p11.22	rs6047844	g.22037575T>C	T/C	1.60	1.71E-39	Li *et al*[16]	—	—	1.09	2.77E-01	1.18	1.77E-01	1.12	1.25E-01
	rs804520	g.22119141A>G	G/A	1.56	6.82E-35	Li *et al*[16]	1.83	**2.78E-10**	1.11	7.59E-02	1.12	4.17E-01	1.29	**1.85E-05**
Xq12	rs1511061	g.66316124C>T	T/C	2.38	6.82E-90	Li *et al*[16]	9.07	**4.52E-18**	2.68	**2.16E-11**	2.85	**4.98E-05**	3.57	**2.56E-22**
	rs2497938	g.66563018T>C	T/C	2.20	2.40E-91	Li *et al*[16]	7.12	**1.64E-17**	2.73	**1.38E-12**	2.35	**6.38E-05**	3.52	**6.27E-25**

Abbreviations: EA/NEA, effect allele and non-effect allele; HGVS, Human Genome Variation Society, according to human reference assembly GRCh37.p13; OR, odds ratio per effect allele.
P-values smaller than 0.05 in BONN, RS, or ERF are indicated in bold.

Table 2 Multivariate analysis and leave-one-out cross-validated prediction of male-pattern baldness based on 25 SNPs from 12 genomic loci in BONN, RS, and ERF

Rank	*Predictor*	*CHR*	*EA*	*Multivariate analysis*		*Accumulative*		
				Beta	*P*	*AUC*	*SENS*	*SPEC*
BONN								
1	rs1511061	Xq12	C	−1.303	**2.52E-10**	—	0.971	0.230
2	rs804520	20p11.22	A	−0.575	**1.49E-07**	0.543	0.868	0.410
3	rs2073963	7p21.1	G	0.379	**5.90E-04**	0.651	0.894	0.385
4	rs6945541	7q11.22	C	0.355	**6.76E-04**	0.696	0.872	0.425
5	rs7349332	2q35	T	0.480	**1.69E-03**	0.705	0.897	0.423
6	rs10502861	18q12.3	T	−0.369	**2.22E-03**	0.719	0.865	0.471
7	rs9287638	2q37.3	A	0.341	**3.54E-03**	0.726	0.848	0.456
8	rs929626	5q33.3	G	−0.661	**7.66E-03**	0.731	0.830	0.491
9	rs12565727	1p36.22	G	−0.332	**1.15E-02**	0.731	0.848	0.466
10	rs17762954	17q21.31	T	−0.324	**1.19E-02**	0.733	0.836	0.491
11	rs113043121	Xq12	A	−0.247	**2.87E-02**	0.739	0.829	0.509
12	rs471205	Xq12	T	0.277	**2.89E-02**	0.741	0.845	0.501
13	rs9668810	12p12.1	T	0.229	6.33E-02	0.741	0.836	0.516
14	rs1081073	5q33.3	T	−0.410	9.70E-02	0.741	0.838	0.501
RS								
1	Age (year)			0.094	**4.14E-24**	0.671	0.728	0.514
2	rs2497938	Xq12	C	−0.695	**2.30E-04**	0.695	0.725	0.554
3	rs12565727	1p36.22	G	−0.329	**2.31E-03**	0.700	0.718	0.574
4	rs141476270	Xq12	T	0.581	**2.36E-03**	0.704	0.718	0.574
5	rs1081073	5q33.3	A	−0.228	**1.05E-02**	0.704	0.714	0.573
6	rs182829063	Xq12	G	0.496	**2.41E-02**	0.706	0.712	0.569
7	rs113043121	Xq12	A	0.266	**3.34E-02**	0.708	0.722	0.567
8	rs6945541	7q11.22	C	0.186	**4.42E-02**	0.708	0.718	0.559
9	rs12373124	17q21.31	C	−0.194	6.62E-02	0.709	0.722	0.584
10	rs1511061	Xq12	C	−0.280	1.07E-01	0.710	0.725	0.573
11	rs9287638	2q37.3	A	0.158	1.08E-01	0.710	0.727	0.578
12	rs2073963	7p21.1	G	0.147	1.12E-01	0.711	0.727	0.582
ERF								
1	Age (year)			0.041	**1.27E-07**	0.615	0.479	0.691
2	rs1511061	Xq12	G	−0.513	**1.01E-04**	0.654	0.545	0.649
3	rs2073963	7p21.1	G	0.384	**4.13E-03**	0.673	0.574	0.677
4	rs12565727	1p36.22	G	−0.461	**5.80E-03**	0.677	0.591	0.681
5	rs804520	20p11.22	A	0.470	**1.22E-02**	0.678	0.574	0.699
6	rs6047844	20p11.22	C	−0.348	**3.99E-02**	0.679	0.574	0.695
7	rs929626	5q33.3	G	−0.250	6.24E-02	0.685	0.599	0.691

Beta and *P*-values are from multivariate logistic regression including all predictors.
AUC/SENS/SPEC: cumulative AUC/sensitivity/specificity as more markers are included.
In ERF, chr2q35, chr17q21, and extra X SNPs were not available.
AUC —: predicted probability too discrete.
Adding these SNPs did not additionally contribute to the prediction accuracy:
2q35 rs10193725; 3q25.1 rs7648585 and rs4679955; and 12p12.1 rs7975017. $P<0.05$ are indicated in bold.

chronological age alone was the strongest predictor of MPB in both population-based sample sets (RS AUC = 0.67, ERF AUC = 0.62). Excluding the extra chrX-SNPs reduced the prediction accuracy in both RS and BONN, but only marginally (BONN AUC = 0.73, RS AUC = 0.71). Although the AUC values are lower than a clinically desired level (0.85) for diagnosis, our prediction model provided very accurate prediction results for a good proportion of individuals, that is, individuals with predicted probabilities of baldness <0.2 or >0.8 (BONN 8.2% <0.2 and 10.9% >0.8, Figure 1b; RS 5.2% <0.2 and 8.3% >0.8, Figure 1c; ERF 7.2% <0.2 and 0.9% >0.8, Figure 1d; also see Supplementary Table S3). In practice, our model may provide a highly informative test for these individuals (19% in BONN, 14% in RS, and 8% in ERF) but less informative for the rest.

GCTA analysis

A GCTA analysis was performed to estimate the proportion of MPB variance explained by all common SNPs available in BONN and RS (Table 3). ERF was excluded from this analysis as it is a family-based study and it has a limited number of available X-chromosome SNPs. In BONN, the variance in liability to early-onset MPB was partitioned 55.8% to all autosomal common variants and 23.3% to X-chromosome variants. In RS, the variance was partitioned 42.4% to all

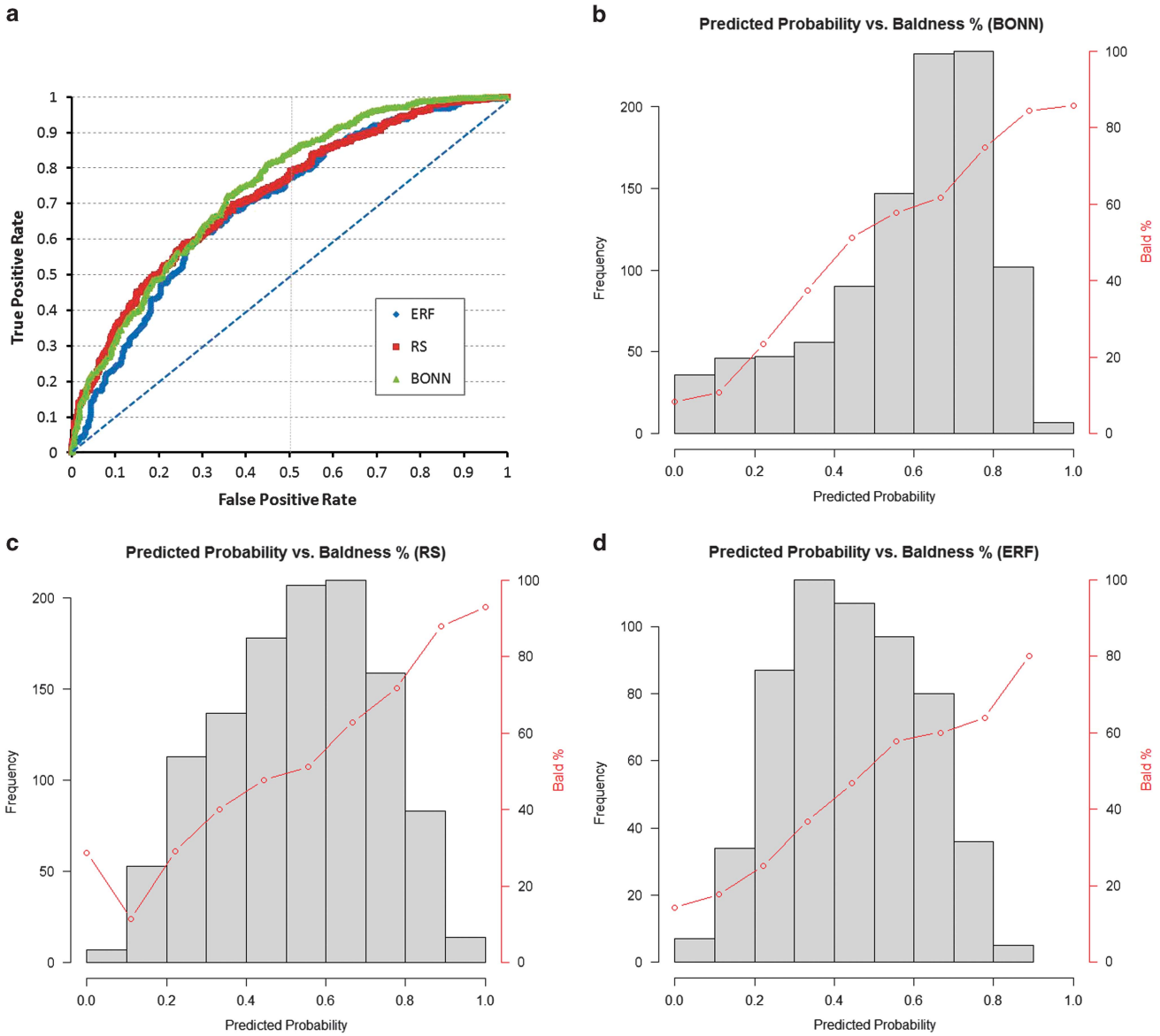

Figure 1 Prediction results of male-pattern baldness in the three study populations. (**a**) Receiver-operating characteristic curves for predicting male-pattern baldness in three European population samples (BONN, RS, and ERF). In BONN Study the prediction model for early-onset MPB included 14 SNPs as predictors (Table 2); the model for predicting MPB in elderly RS people included 11 SNPs and age as predictors (Table 2); and the prediction model in ERF included 6 SNPs and age as predictors (Table 2). (**b**) Histogram of predicted probability overplayed with percentage of baldness in each probability bin (BONN Study). (**c**) Histogram of predicted probability overplayed with percentage of baldness in each probability bin (Rotterdam Study). (**d**) Histogram of predicted probability overplayed with percentage of baldness in each probability bin (ERF Study).

autosomes and 9.8% to the X chromosome. Not surprisingly, early-onset MPB in BONN demonstrated an overall higher heritable component explainable by common DNA variants than normal MPB status in elderly individuals in RS. We further partitioned the proportion of variance explained by variants to each chromosome (Table 3,Supplementary Figure S1). In BONN, chromosomes 1 ($h^2 = 13.9\%$), 6 (12.5%), 17 (9.6%), 20 (24.3%), and X (27.2%) all explained a significant portion of early-onset MPB variance (Supplementary Figure S1A); the variance partitioned to other chromosomes was inconclusive due to large standard errors as the 95% confidence intervals included 0. In RS, only chromosome X showed a significant effect ($h^2 = 9.8\%$, 95% interval: 0.2–19.4%, Supplementary Figure S1B) in explaining MPB status in elderly individuals.

Meta-analysis of GWAS

GWASs for MPB were conducted in BONN, RS, and ERF. The QQ and Manhattan plot in ERF looks very different from RS and BONN (Supplementary Figure S2), that is, in ERF no genome-wide significant associations was detected including the *EDA2R/AR* locus (rs1511061 $P = 7.2 \times 10^{-5}$). This underlies the importance of large sample sizes in GWAS even for detecting major gene effects. A meta-analysis of the GWAS results in the three cohorts identified 946 SNPs in four genetic loci that showed genome-wide significant association with MPB ($P < 5 \times 10^{-8}$, Supplementary Table S4, Supplementary Figure S2). These include 739 SNPs on chromosome Xq12 between the *EDA2R* and *AR* genes, 205 SNPs on chromosome 20p11 (min $P = 1.76 \times 10^{-15}$ observed for rs6075850), one SNP on chromosome

Table 3 Proportion of variance in male-pattern baldness liability explained by common SNPs

Chr	*Rotterdam Study*		*BONN*	
	VG/VP	*SE*	*VG/VP*	*SE*
All autosomes	0.424	0.234	0.558	0.216
X	0.098	0.049	0.233	0.011
1	0.104	0.075	0.139	0.067
2	0.000	0.069	0.000	0.061
3	0.051	0.067	0.000	0.060
4	0.064	0.065	0.000	0.053
5	0.095	0.068	0.006	0.050
6	0.000	0.060	0.125	0.063
7	0.000	0.061	0.102	0.057
8	0.072	0.062	0.006	0.053
9	0.056	0.060	0.000	0.050
10	0.056	0.060	0.000	0.049
11	0.000	0.083	0.000	0.051
12	0.029	0.089	0.000	0.050
13	0.000	0.077	0.018	0.043
14	0.075	0.077	0.039	0.046
15	0.078	0.077	0.047	0.038
16	0.039	0.078	0.000	0.039
17	0.063	0.077	0.096	0.044
18	0.027	0.070	0.000	0.042
19	0.025	0.059	0.031	0.034
20	0.000	0.072	0.243	0.051
21	0.013	0.057	0.092	0.058
22	0.106	0.062	0.017	0.028

7p21.1 (rs756853 $P=3.3\times10^{-8}$), and one SNP on chromosome 6p25.1 (rs4959410 $P=3.35\times10^{-8}$). Except 6p25.1, all loci have been previously associated with MPB.[14–16] The associated SNP at 6p25.1 is a single genome-wide significant finding not supported by additional SNPs in the region and thus warrants further validation.

DISCUSSION

Candidate DNA variants in 12 previously discovered loci showed robust association with early-onset, as well as normal, MPB in German and Dutch males. Although one locus (3q25.1) was not statistically significantly associated, overall genetic effect sizes from all loci were similar to previous estimates. A major gene effect on Xq12 was confirmed with an allelic OR of up to 9.07 for early-onset MPB and 2.7–2.8 for normal MPB in middle-aged and elderly men. The combined predictive power of all candidate DNA variants was also higher for predicting early-onset MPB cases (AUC 0.74) than predicting MPB in middle-aged and elderly men from population-based studies (AUC 0.68–0.71), for whom age served as the strongest predictor (age-alone AUC 0.62–0.67). Although the overall predictive accuracy has not reached a practically desired level (>0.85), our risk model may prove useful in assisting decision making on early preventive actions for MPB and in forensic investigations for about 8–19% European individuals. Contrasting the observation that the 12 known DNA loci together explained rather limited variation of MPB liability, a variance partitioning analysis demonstrated that a large proportion of the phenotypic variance can be explained by all genotyped common SNPs available in the microarray data sets. This gap is likely due to many common variants with small effect sizes, which will likely be identified in future larger studies. Finally, our meta-analysis of the GWAS results identified one new locus at 6p25.1, which demonstrated significant association with MPB.

It has long been known that DNA variants in or near the *AR* gene increase the risk of patterned hair loss in both men and women, but the exact identity of the causal DNA variant(s) remains unclear. A common synonymous coding variant rs6152 G>A (StuI restriction site) in the exon 1 of the *AR* gene has been associated with MPB in previous studies.[11,12] For example, Ellis *et al*[11] found that in an Australian cohort the G allele was present in 98.1% of young bald men, 92.3% of older bald men, and only 76.6% of non-bald men. The variant rs6152 is available in RS and was also highly significantly associated with MPB ($P=5.6\times10^{-8}$) but much less so than the top-associated X-chromosomal SNP rs1511061 ($P=2.6\times10^{-11}$ in RS), and it became nonsignificant ($P>0.05$) when conditioning on the genotypes of rs1511061. The SNP rs1511061 is an intergenic noncoding SNP located 236 kbp upstream of the *AR* coding region. These results suggest noncoding sequences upstream of *AR* containing functional variants. These could have regulatory effects as we recently established for noncoding variants in another human appearance trait – pigmentation.[28,29] The T allele (or A allele on the reverse strand of the genome) of rs1511061 is the major allele in our sample with a pronounced frequency in MPB cases (see Table 1). Assuming MPB is a monogenic phenotype caused by a single variant (ie, rs1511061), it would suggest baldness is the default phenotype in advanced ages, that is, the majority of males (wild-type allele) will eventually develop baldness while the minor allele provides a protective effect. The T allele of rs1511061 is also the major allele in HapMap-CEU ($n=226$, $f=0.885$), fixed in HapMap-HCB ($n=90$, $f=1.0$) and JPT ($n=88$, $f=1.0$), but reversely fixed in HapMap-YRI ($n=120$, $f=0.0$). This contradicts rs1511061 (and its high LD SNPs such as rs2497938) being causal because such a pattern of allele frequency distribution can hardly be correlated to the prevalence pattern at a global level as Asians and African–Americans have lower MPB prevalence and less severe MPB than Europeans.[30,31] Functional analyses of the noncoding sequence in this region, such as those we previously carried out for pigmentation gene regions,[28,29] are required to reveal the exact identity of the causal MPB genetic variant(s).

It has been suggested that there might be different genetic influences on balding in young men (ie, early-onset MPB) and on non-balding in elderly men, based on the observation of an increased frequency of non-balding in the fathers of non-bald elderly men.[32] However, twin studies have shown that the heritability of MPB estimated in young men[2] was similar to that estimated in elderly men[10] (both around 80%). In our study, the genetic component of early-onset MPB (ie, BONN) showed somewhat different signature than that of normal MPB in elderly people (ie, RS), that is, the variance partitioning analysis suggests that a higher percentage of variance in early-onset MPB can be addressed by all common SNPs compared with that in elderly men. This is likely due to the extreme case–control design used in BONN Study, where early-onset cases and screened elderly controls were contrasted to enhance the genetic contrast. Furthermore, it is uncertain whether this discrepancy is due to phenotyping errors, as classifying MPB status in elderly people is much more error prone than defining early-onset cases. The variance partitioning analysis might be more sensitive to measurement errors than in twin studies, whereas for the latter even the effect of some differential misclassifications can be canceled out between monozygotic and dizygotic twins, that is, if the classification for one twin is differentially biased due to an unobserved factor, the same bias likely occurs to the other twin at a similar degree. Nevertheless, it is at least clear that the allelic effects at Xq12 and 20p11 are pronounced in early-onset MPB cases. For

example, in our data the risk allele at Xq12 was presented in 96.5% of the early-onset cases, in 88% of the middle-aged and elderly cases, and in 75% of all controls.

The genetic architecture of human complex traits differs substantially even between the ones with similar heritability. For example, eye color and adult body height both have heritability estimates of about 80%. The genetic architecture of eye color, however, is relatively simple with a very strong major gene effect provided by a noncoding SNP at the *HERC2* gene regulating transcription of the neighboring *OCA2* pigmentation gene,[29] and several minor-effect SNPs, which together predict the phenotype at very high accuracy (AUC > 0.9 for blue/brown, explaining over 50% phenotypic variance).[33,34] On the other hand, adult body height has long served as a model trait of extreme genetic complexity, that is, without any major gene effect, 180 SNPs together could predict height at AUC of 0.75 and explain ~12% of the phenotypic variance,[35] and with recent progress[36] the AUC for predicting height is expected to be larger. In light of the current and previous studies, MPB appears something in the middle. It does have a major gene effect, that is, on the X chromosome (although much weaker than that of eye color) and likely involves many common variants with small effects (probably more than eye color). Improving the prediction accuracy for MPB will rely on the identification of more trait-associated DNA variants in GWAS of increased sizes, which will be a continuous and accumulative effort but certainly achievable in the future, as scientists already have achieved in studying human height (accumulative sample size over 250 000, several thousand SNPs together explain about 30% of the phenotypic variance).[36] Therefore, we may expect that the prediction accuracy of MPB will eventually surpass that of body height given it has a known major gene effect, which is absent for height.

Our meta-analysis identified one new locus on chromosome 6p25.1 showing significant association with MPB. The associated SNP rs4959410 is surrounded by a pseudogene (*BTF3P7*, basic transcription factor 3 pseudogene 7) and several uncharacterized genes. The closest known genes are *LY86* (lymphocyte antigen 86) and *RREB1* (ras responsive element-binding protein 1). No existing evidence supports the involvement of any of these genes in hair loss. Therefore, this finding still needs to be confirmed in independent samples.

In conclusion, by taking all available SNPs previously found to be associated with MPB at a genome-wide significant level and testing their predictive value in 2725 German and Dutch males with early-onset MPB patients as well as from middle-aged to elderly men, we achieved prevalence-adjusted prediction accuracies expressed as AUC values of around 0.7 (where 0.5 means random prediction and 1.0 means completely accurate prediction). Although the prediction accuracy has not reached a level useful in practice, our preliminary genetic model may already assist decision making on early MPB preventive actions and in forensic investigations. Furthermore, our results imply that with more genome-wide significantly associated SNPs identified in the future and included in the prediction model together with the DNA markers presented here, male-pattern baldness will likely become predictable from DNA with high enough accuracy to allow routine practical applications such as in medicine and forensics.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Dr David Gunn for his valuable discussions and useful comments on the manuscript. We thank Pascal Arp; Mila Jhamai; Marijn Verkerk; Lizbeth Herrera; Marjolein Peters, MSc; Carolina Medina-Gomez, MSc; and Fernando Rivadeneira, MD PhD, for their help in creating the GWAS database, and Karol Estrada, PhD; Yurii Aulchenko, PhD; and Carolina Medina-Gomez, MSc, for the creation and analysis of imputed data. We thank Sophie Flohil, Emmilia Dowlatshahi, Robert van der Leest, Leonie Jacobs, Joris Verkouteren, Ella van der Voort, and Shmaila Talib for collecting the phenotype data in the RS. This work was supported in part by the Erasmus MC University Medical Center Rotterdam and funds from the Netherlands Genomics Initiative/Netherlands Organization of Scientific Research (NWO) within the framework of the Netherlands Consortium of Healthy Ageing (NCHA).

The generation and management of GWAS genotype data for the Rotterdam Study (RS I, RS II, RS III) was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. The GWAS data sets are supported by the Netherlands Organisation of Scientific Research NWO Investments (no. 175.010.2005.011, 911-03-012), the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO), and the Netherlands Consortium for Healthy Aging (NCHA), project no. 050-060-810. FL is supported by Chinese Thousand Talent Program for Distinguished Young Scholars and MAH is supported by Unilever. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam; Netherlands Organization for the Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly (RIDE); the Ministry of Education, Culture and Science; the Ministry for Health, Welfare and Sports; the European Commission (DG XII); and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

ERF Study as a part of EUROSPAN (European Special Populations Research Network) was supported by the European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007–2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the programme 'Quality of Life and Management of the Living Resources' of 5th Framework Programme (no. QLG2-CT-2002-01254) as well as the FP7 project EUROHEADPAIN (no. 602633). High-throughput analysis of the ERF data was supported by joint grant from Netherlands Organization for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043).

The BONN Study is supported by Heinz Nixdorf Foundation, the German Ministry of Education and Science and the German Research Council (D Glass; Project SI 236/8-1, SI236/9-1, ER 155/6-1); German Research Council (D Glass; FOR 423); and the Life and Brain GmbH (Bonn, Germany; project grant). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1 Norwood OT: Male pattern baldness: classification and incidence. *South Med J* 1975; **68**: 1359–1365.
2 Nyholt DR, Gillespie NA, Heath AC, Martin NG: Genetic basis of male pattern baldness. *J Invest Dermatol* 2003; **121**: 1561–1564.
3 Hamilton JB: Patterned loss of hair in man; types and incidence. *Ann N Y Acad Sci* 1951; **53**: 708–728.
4 Cash TF: The psychological effects of androgenetic alopecia in men. *J Am Acad Dermatol* 1992; **26**: 926–931.
5 Sinclair R: Male pattern androgenetic alopecia. *BMJ* 1998; **317**: 865–869.
6 Kayser M, de Knijff P: Improving human forensics through advances in genetics, genomics and molecular biology. *Nat Rev Genet* 2011; **12**: 179–192.
7 Kayser M: Forensic DNA phenotyping: predicting human appearance from crime scene material for investigative purposes. *Forensic Sci Int Genet* 2015; **18**: 33–48.
8 Zubakov D, Liu F, van Zelm MC *et al*: Estimating human age from T-cell DNA rearrangements. *Curr Biol* 2010; **20**: R970–R971.
9 Zbiec-Piekarska R, Spolnicka M, Kupiec T *et al*: Examination of DNA methylation status of the ELOVL2 marker may be useful for human age prediction in forensic science. *Forensic Sci Int Genet* 2015; **14**: 161–167.

10 Rexbye H, Petersen I, Iachina M *et al*: Hair loss among elderly men: etiology and impact on perceived age. *J Gerontol A Biol Sci Med Sci* 2005; **60**: 1077–1082.
11 Ellis JA, Stebbing M, Harrap SB: Polymorphism of the androgen receptor gene is associated with male pattern baldness. *J Invest Dermatol* 2001; **116**: 452–455.
12 Prodi DA, Pirastu N, Maninchedda G *et al*: EDA2R is associated with androgenetic alopecia. *J Invest Dermatol* 2008; **128**: 2268–2270.
13 Brockschmidt FF, Heilmann S, Ellis JA *et al*: Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness. *Br J Dermatol* 2011; **165**: 1293–1302.
14 Hillmer AM, Brockschmidt FF, Hanneken S *et al*: Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat Genet* 2008; **40**: 1279–1281.
15 Richards JB, Yuan X, Geller F *et al*: Male-pattern baldness susceptibility locus at 20p11. *Nat Genet* 2008; **40**: 1282–1284.
16 Li R, Brockschmidt FF, Kiefer AK *et al*: Six novel susceptibility loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet* 2012; **8**: e1002746.
17 Heilmann S, Kiefer AK, Fricker N *et al*: Androgenetic alopecia: identification of four genetic risk loci and evidence for the contribution of WNT signaling to its etiology. *J Invest Dermatol* 2013; **133**: 1489–1496.
18 Hofman A, Darwish Murad S, van Duijn CM *et al*: The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* 2013; **28**: 889–926.
19 Taylor R, Matassa J, Leavy JE, Fritschi L: Validity of self reported male balding patterns in epidemiological studies. *BMC Public Health* 2004; **4**: 60.
20 Pardo LM, MacKay I, Oostra B, van Duijn CM, Aulchenko YS: The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* 2005; **69**: 288–295.
21 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
22 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* 2007; **39**: 906–913.
23 Akaike H: Citation classic—a new look at the statistical-model identification. *CC/Eng Technol Appl Sci* 1981; **51**: 22–22.
24 Zweig MH, Campbell G: Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; **39**: 561–577.
25 Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011; **88**: 76–82.
26 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ: Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**: 7.
27 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
28 Visser M, Palstra RJ, Kayser M: Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. *Hum Mol Genet* 2014; **23**: 5750–5762.
29 Visser M, Kayser M, Palstra RJ: HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Res* 2012; **22**: 446–455.
30 Lee WS, Lee HJ: Characteristics of androgenetic alopecia in asian. *Ann Dermatol* 2012; **24**: 243–252.
31 Hoffmann R: Male androgenetic alopecia. *Clin Exp Dermatol* 2002; **27**: 373–382.
32 Birch MP, Messenger AG: Genetic factors predispose to balding and non-balding in men. *Eur J Dermatol* 2001; **11**: 309–314.
33 Liu F, Wollstein A, Hysi PG *et al*: Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet* 2010; **6**: e1000934.
34 Liu F, van Duijn K, Vingerling JR *et al*: Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol* 2009; **19**: R192–R193.
35 Liu F, Hendriks AEJ, Ralf A *et al*: Common DNA variants predict tall stature in Europeans. *Hum Genet* 2013; **133**: 587–597.
36 Wood AR, Esko T, Yang J *et al*: Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 2014; **46**: 1173–1186.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)