Genetic Prediction of Male Pattern Baldness

Saskia P Hagenaars[1,2,3,*], W David Hill[1,2,*], Sarah E Harris[1,4], Stuart J Ritchie[1,2], Gail Davies[1,2], David C Liewald[1], Catharine R Gale[1,2,5], David J Porteous[1,4], Ian J Deary[1,2], Riccardo E Marioni[1,4,6]

1. Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, UK
2. Department of Psychology, University of Edinburgh, Edinburgh, UK
3 Division of Psychiatry, University of Edinburgh, Edinburgh, UK
4. Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
5. Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK
6. Queensland Brain Institute, The University of Queensland, Brisbane, QLD, Australia

* These authors contributed equally

Correspondence: Riccardo E. Marioni, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, Scotland, UK, Telephone: +44 131 651 8528 Email: riccardo.marioni@ed.ac.uk

Running Title: GWAS of Baldness

Author contributions:

SPH, WDH: Formal analysis, writing - review and editing

SEH, SJR: writing - review and editing

GD, DL: Data curation

DJP, CG, IJD: funding acquisition, writing - review and editing

REM: conceptualization, formal analysis, writing - original draft preparation, supervision

Abstract

Male pattern baldness can have substantial psychosocial effects, and it has been phenotypically linked to adverse health outcomes such as prostate cancer and cardiovascular disease. We explored the genetic architecture of the trait using data from over 52,000 male participants of UK Biobank, aged 40-69 years. We identified over 250 independent novel genetic loci associated with severe hair loss. By developing a prediction algorithm based entirely on common genetic variants, and applying it to an independent sample, we could discriminate accurately (AUC = 0.82) between those with no hair loss from those with severe hair loss. The results of this study might help identify those at the greatest risk of hair loss and also potential genetic targets for intervention.

Introduction

Male pattern baldness affects around 80% of men by the age of 80 years [1] and it can have substantial psychosocial impacts via changes in self-consciousness and social perceptions [2, 3]. In addition to alterations in physical appearance, some, but not all, studies have identified negative health outcomes associated with baldness including increased risk of prostate cancer [4-6] and cardiovascular disease [7-9]. Baldness is known to be substantially heritable [10]. Here, we use a large population-based dataset to identify many of the genes linked to variation in baldness, and build a genetic score to improve prediction of severe hair loss.

The total proportion of variance in male pattern baldness that can be attributed to genetic factors has been estimated in twin studies to be approximately 80% for both early- and late-onset hair loss [11, 12]. The remaining variance in these twin studies was attributable to non-shared environmental factors. Newer molecular-genetic methods have estimated the single-nucleotide polymorphism (SNP)-based, common-variant heritability of baldness at 13.7% [10]. Molecular methods also indicate some overlap between genetic variants linked to baldness and those linked to phenotypes such as height, waist-hip ratio, age at voice drop in males, age at menarche in females, and presence of a unibrow [13].

A number of studies have identified specific genetic variants linked to variations in baldness, usually with the *AR* gene showing the strongest association. The largest published genome-wide association study (GWAS) to date highlighted eight independent genetic loci that were linked to baldness; the top *AR* SNP yielded an odds ratio of 2.2 in a case-control meta-analysis of 12,806 individuals of European ancestry [14]. One of the autosomal hits identified

in that study was found to be in a gene linked to Parkinson's disease. More recently, a review paper highlighted fifteen loci from six studies that have been associated at genome-wide significance ($P<5x10^{-8}$) with baldness; two of the loci were located on the X chromosome [15].

Several attempts have been made to build predictors of male pattern baldness using polygenic risk scores. Heilman et al. found, using a case-control design with ~600 per arm, that a predictor based on 34,186 SNPs explained 4.5% of the variance on the liability scale [16]. Marcinska et al. used candidate genes to build 5-SNP and 20-SNP polygenic predictors, which performed quite well when considering prediction of early-onset male pattern baldness, but poorly when considering those with no baldness versus those with severe baldness across all ages [17]. Most recently, a 20-SNP predictor was assessed in three European studies [18]. It achieved a maximum Area Under the Curve (AUC) prediction of 0.74 in an early-onset cohort, but weaker estimates in the other two, late-onset cohorts (AUC = 0.69 and 0.71). These values correspond to poor-to-fair predictions of baldness. In that study, age was included in the predictor, explaining the bulk of the differences. A meta-analysis of the three cohorts' GWAS studies identified a novel locus on chromosome 6. The study also estimated the SNP-based heritability of early-onset (56% (SE 22%) from the autosomes, 23% (SE 1.1%) from the X chromosome) and late-onset baldness (42% (SE 23%) from the autosomes, 10% (SE 5%) from the X chromosome).

The Present Study

The UK Biobank study [19] (http://www.ukbiobank.ac.uk) is a large, population-based genetic epidemiology cohort. At its baseline assessment (2006-2010), around 500,000 individuals aged between 40 and 70 years and living in the UK completed health and lifestyle questionnaires and provided biological samples for research.

The present study reports a GWAS of male pattern baldness in the UK Biobank cohort. It is over four times the size of the previously-largest meta-analytic study [14]. We then split the cohort into a large 'discovery' sample and smaller 'test' sample to perform a prediction analysis, determining the accuracy of a polygenic profile score built from the GWAS findings in discriminating between those with severe hair loss and those with no hair loss.

Results

The mean age of the 52,874 men was 57.2 years (SD 8.0). 16,724 (31.6%) reported no hair loss, 12,135 (23.0%) had slight hair loss, 14,234 (26.9%) had moderate hair loss, and 9,781 (18.5%) had severe hair loss.

The genome-wide association study of the four-category self-reported baldness measure in 52,874 White British men from UK Biobank yielded 13,029 autosomal hits from the imputed data ($P<5x10^{-8}$) in addition to 117 hits (out of 14,350 genotyped SNPs) on the X chromosome (**Figure 1**). The QQ plot for the autosomal GWAS is shown in **Figure S1**. An LD clumping analysis indicated that these hits can be attributed to 247 independent autosomal regions. All of the 8 autosomal hits identified by Li et al. [14] and the additional hits noted in Heilmann-Heimbach et al. [15] replicated in the UK Biobank dataset with a maximum P value of

$6.4x10^{-14}$ (**Table 1**). The previously reported X chromosome variant from Li et al. [14] and the variant from Richards et al. [10] also replicated with P-values that were effectively zero (**Table 1**). The chromosome 6 hit from Liu et al. [18] failed to replicate (P=0.37). A list of the top 20 independent autosomal hits is presented in **Table 2**. The top 10 independent X chromosome hits are presented in **Supplementary Table 1**; rs140488081 and rs7061504 are intronic SNPs in the *OPHN1* gene.

The gene-based analysis identified 112 autosomal genes and 13 X chromosome genes that were associated with baldness after a Bonferroni correction (P<0.05/18,061 and P<0.05/567, respectively). The top gene-based hit was, as expected, the androgen receptor on the X chromosome ($P=2.0x10^{-269}$). A full list of the autosomal significant gene-based hits is provided in **Supplementary Table 2** and significant genes on the X chromosome are shown in **Supplementary Table 3**.

Using common genetic variants with a minor allele frequency of at least 1%, GCTA-GREML analysis found that 47.3% (SE 1.3%) of the variance in baldness can be explained by common autosomal genetic variants, 4.6% (SE 0.3%) can be explained by common X chromosome variants.

Genetic correlations were examined between male pattern baldness and numerous (n=22) cognitive, health, and anthropometric traits using LD Score regression. No significant associations were found and all estimates were close to the null (**Supplementary Table 4**).

The GWAS for self-reported baldness was re-run on a sub-sample of 40,000 individuals - retaining an equal proportion of each of the four baldness patterns as observed in the full cohort - to allow a polygenic prediction score to be built and applied to the remaining, independent sample of 12,874 individuals. The most powerful predictions from comparing the extreme phenotype groups were observed at the $P<1x10^{-5}$ threshold for the autosomal polygenic score and at P<1 for the X chromosome polygenic score (**Table 3**). The optimal autosomal polygenic score yielded an AUC of 0.75 for discriminating between those with no hair loss (n=4,123) and those with severe hair loss (n=2,456). The AUC values between those with no hair loss and those with slight and moderate hair loss were 0.586 and 0.652, respectively. The corresponding AUCs for the X chromosome polygenic score were 0.724 (severe loss), 0.592 (moderate loss), and 0.523 (slight loss, P=0.001), when examined against the reference group of no hair loss. All logistic regression P-values for the effect of the polygenic score were $<2x10^{-16}$ except where noted. An additive combination of the autosomal and X chromosome polygenic scores gave an AUC of 0.816 for severe baldness, 0.707 for moderate baldness, and 0.619 for slight baldness (**Figure 2**).

The results of the partitioned heritability analysis indicated that 27 of the functional annotations from the baseline model were statistically significant (**Supplementary Figure 2** and **Supplementary Table 5**). These significant annotations included a broad array of functional elements including histone marks, enhancer regions, conserved regions, and DNaseI hypersensitivity sites (DHS). The ten tissue types were then tested for significance after controlling for the baseline model. Following correction for multiple testing all ten of the tissue groups showed significant enrichment (**Supplementary Figure 3** and **Supplementary Table 5**).

Discussion

In this large GWAS study of male pattern baldness, we identified 287 independent genetic signals that are linked to differences in the trait, a substantial advance over the previous largest GWAS meta-analysis, which identified eight independent signals [14]. We showed, in line with a previous study [18], but with much greater precision, that a substantial proportion of individual differences in hair loss patterns can be explained by common genetic variants on the autosomes as well as the X chromosome. Finally, by splitting our cohort into a discovery and a prediction sample, we show a good predictive discrimination (AUC = 0.82) between those with no hair loss and those with severe hair loss.

Despite there being genetic overlap for SNP hits associated with baldness and Parkinson's Disease—first noted in Li et al. [14] and replicated here—we observed no genetic correlations between baldness and any of the health, cognitive, or anthropometric outcomes we studied. The point estimates for the genetic correlations were all near zero, suggesting true null associations as opposed to a lack of statistical power to detect modest-sized correlations.

As mentioned above, the GWAS identified 247 independent autosomal loci and 40 independent X chromosome loci. The top 20 hits from the autosomes were located in/near to genes that have been associated with, for example, hair growth/length in mice (*FGF5*) [20], grey hair (*IRF4*) [21], cancer (breast: *MEMO1* [22], bladder: *SLC14A2* [23]), histone acetylation (*HDAC9*), and frontotemporal dementia (*MAPT*) [24]. Two of the top 10 X

chromosome SNPs were located in *OPHN1,* a gene previously associated with X-linked mental retardation [25].

Of the top autosomal gene-based findings (maximum $P=3.1x10^{-15}$), *RSPO2* has been linked to hair growth in dogs. *PGDFA* has been linked to hair follicle development [26]; *EBF1* is expressed in dermal papillae in mature hair follicles [27]; *PRR23B* is proximal to a GWAS hit for eyebrow thickness [21]; and *WNT10A* has been linked to both straight hair [28] and dry hair [29].

The top X chromosome gene-based findings included the androgen receptor (*AR*), which has been well established as a baldness associated gene [30], along with its upstream (*EDA2R*) and downstream (*OPHN1*) genes. *EDA2R* plays a role in the maintenance of hair and teeth as part of the tumor necrosis factor receptor. Onset of male pattern baldness could by influenced by *EDA2R* via activation of nuclear proto-oncoprotein *c-Jun*, which is linked to transcription activation of *AR* [31]. Two other genes included in the gene-based findings, *OPHN1* and *ZC4H2*, have previously been associated with X-linked mental retardation [25, 32].

The optimal thresholds from the prediction results highlight the polygenicity of the baldness trait and suggest that there may be many opportunities to identify novel pathways and possible therapies linked to baldness. The release of the full set of UK Biobank genotypes at the beginning of the next calendar year (2017) will likely enhance the predictive power of the polygenic score, and will probably allow for the identification of even more novel genetic loci in future GWAS studies.

The main strength of this study is its large sample size and phenotypic homogeneity. Many meta-analytic studies of complex traits are weakened by different cohorts collecting data at different time-points, under different protocols, in different populations. The present study replicated all of the previously identified autosomal hits for baldness from Li et al.[14] and Heilmann-Heimbach et al., [15] suggesting a degree of robustness in phenotypic measurement. Whereas the genomic inflation factor from the GWAS was large (1.09, Q-Q plot in **Supplementary Figure 1**), this is likely to be a result of genuine polygenic effects. We have used identical analysis protocols for other traits with far lower SNP-based heritabilities in the same UK Biobank cohort and observed no evidence of inflation [33].

Conclusion

We identified hundreds of independent, novel genetic correlates of male pattern baldness - an order of magnitude greater than the list of previous genome-wide hits. Our top SNP and gene-based hits were in genes that have previously been associated with hair growth and development. We also generated a powerful polygenic predictor that discriminated well between those with no hair loss and those with severe hair loss.

Methods

Data

Data came from the first release of genetic data of the UK Biobank study and analyses were performed under the data application 10279. Ethical approval for UK Biobank was granted by the Research Ethics Committee (11/NW/0382).

Genotyping details including quality control steps have been reported previously [33]. Briefly, from the sample with genetic data available as of June 2015, 112,151 participants remained after the following exclusion criteria were applied: SNP missingness, relatedness, gender mismatch, non-British ancestry, and failed quality control for the UK BiLEVE study [33]. For the current analysis, an imputed dataset was used for the autosomes (reference set panel combination of the UK10K haplotype and 1000 Genomes Phase 3 panels: http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=157020). Imputed data were not available for the X chromosome; hence only genotyped variants were considered. X chromosome quality control steps included a minor allele frequency cut-off of 1% and a genotyping call rate cut-off of 98% [34]. For the imputed autosomal data, we restricted the analyses to variants with a minor allele frequency >0.1% and an imputation quality score >0.1.

From the sample of 112,151 unrelated White British participants with genetic data, we identified 52,874 men with a self-reported response to UK Biobank question 2395, which was adapted from the Hamilton-Norwood scale [35, 36]. These men were asked to choose, from four patterns (no loss; slight loss; moderate loss; severe loss), the one that matched their hair coverage most closely. **Figure 3** shows a screenshot of the four options.

Statistical Analysis

A genome-wide association study was conducted using baldness pattern residuals as the dependent variable. The residuals were obtained from a linear regression model of baldness

pattern on age, assessment centre, genotyping batch and array, and 10 principal components to correct for population stratification.

The GWAS for the imputed autosomal dataset was performed in SNPTest v2.5.1 [37] via an additive model, using genotype probability scores. The GWAS for the X chromosome was performed in PLINK [38, 39].

The number of independent signals from the GWAS was determined using LD-clumping [38, 39] based on the LD structure annotated in the 1000 genomes project [40]. SNPs were considered for the clumping analysis if they passed a genome-wide significance threshold of $P<5x10^{-8}$. Following this, SNPs that were within a 500kb region and in LD of at least 0.1 with the top SNPs were included in the clump; SNPs from the region were assigned to the clump if they had an association $P<1x10^{-5}$.

GWAS lookups were performed for the 17 top hits reported in Li et al 2012, Heilmann-Heimbach et al. 2016 and Liu et al. 2016 [14, 15, 18].

Gene-based analyses were performed using MAGMA [41]. SNPs were mapped to genes according to their position in the NCBI 37.3 build map. No additional boundary was added beyond the genes start and stop site. For the autosomal genes the summary statistics from the imputed GWAS were used to derive gene-based statistics using the 1000 genomes (phase 1, release 3) to model linkage disequilibrium. For genes on the X chromosome the genotype

data from UK Biobank was used and the gene-based statistic was derived using each participant's phenotype score.

For the prediction analysis, the GWAS was re-run on a randomly selected cohort of 40,000 individuals, leaving an independent cohort of 12,874 in which to test the polygenic predictor. The methods for the GWAS were identical to those reported for the full sample. The regression weights from the GWAS on the 40,000 cohort were used to construct polygenic scores in the target dataset at P value thresholds of $<1x10^{-20}$, $<5x10^{-8}$, $<1x10^{-5}$, <0.01, <0.05, <0.1, <0.2, <0.5, <1 using PRSice software [42]. PRSice creates polygenic scores by calculating the sum of alleles associated with male pattern baldness across many genetic loci, weighted by their effect sizes estimated from the male pattern baldness GWAS. Thereafter, each threshold was used to discriminate between those with no hair loss and those with severe hair loss via logistic regression with results being reported for the optimal predictor only. A predictor for both the autosomes and X chromosome were built and assessed independently and additively. Receiver operator characteristic (ROC) curves were plotted and areas under the curve (AUC) were calculated using the pROC package in R [43, 44].

SNP-based heritability of baldness was estimated using GCTA-GREML [45] after applying a relatedness cut-off of >0.025 in the generation of the autosomal (but not X chromosome) genetic relationship matrix. Linkage disequilibrium score (LDS) regression analyses [46] were used to generate genetic correlations between baldness and 22 cognitive, anthropometric, and health outcomes, where phenotypic correlations or evidence of shared genetic architecture have been found (**Supplementary Table 6**) . Due to the large effects in the *APOE* region, 500kb was removed from around each side of this region and the analysis

was repeated. The Alzheimer's data set without this region is referred to as Alzheimer's 500kb and makes a total of 23 tests performed and controlled for using False discovery rate (FDR) correction [47]. An overview of the GWAS summary data for the anthropometric and health outcomes is provided in **Appendix 1**.

Partitioned heritability analysis was performed using stratified linkage disequilibrium score (SLDS) regression [48]. SLDS regression examines groups of SNPs sharing the same functional properties and was used to estimate a heritability metric for each group. The goal of this analysis was to determine if a specific group of SNPs made a greater contribution to the total heritability of male pattern baldness than would be expected by the size of the SNP set. Firstly, a baseline model was derived using 52 overlapping, functional categories. Secondly, a cell-specific model was constructed by adding each of the 10 cell-specific functional groups to the baseline model one at a time to the baseline model. Multiple testing was controlled for using FDR correction [47] to the to the baseline model using 52 categories. For the cell-specific analysis the baseline model was first included and the level of enrichment for each cell specific category as derived. Here, 10 tests were controlled for using a FDR correction.

Acknowledgements

This research was conducted, using the UK Biobank Resource, in The University of Edinburgh Centre for Cognitive Ageing and Cognitive Epidemiology, part of the cross-council Lifelong Health and Wellbeing Initiative (MR/K026992/1). Funding from the Biotechnology and Biological Sciences Research Council (BBSRC) and Medical Research Council (MRC) is gratefully acknowledged. WDH is supported by a grant from Age UK (Disconnected Mind Project).

Conflicts of Interest

Ian Deary and David Porteous are participants in UK Biobank.

References

1. Hamilton JB. Patterned loss of hair in man: types and incidence. Annals of the New York Academy of Sciences. 1951;53(3):708-28. doi: 10.1111/j.1749-6632.1951.tb31971.x.

2. Alfonso M, Richter-Appelt H, Tosti A, Viera MS, García M. The psychosocial impact of hair loss among men: a multinational European study. Current Medical Research and Opinion. 2005;21(11):1829-36. doi: 10.1185/030079905X61820.

3. Cash. The psychosocial consequences of androgenetic alopecia: a review of the research literature. British Journal of Dermatology. 1999;141(3):398-405. doi: 10.1046/j.1365-2133.1999.03030.x.

4. Cremers RG, Aben KK, Vermeulen SH, den Heijer M, van Oort IM, Kiemeney LA. Androgenic alopecia is not useful as an indicator of men at high risk of prostate cancer. European Journal of Cancer. 2010;46(18):3294-9. doi: http://dx.doi.org/10.1016/j.ejca.2010.05.020.

5. Zhou CK, Levine PH, Cleary SD, Hoffman HJ, Graubard BI, Cook MB. Male Pattern Baldness in Relation to Prostate Cancer–Specific Mortality: A Prospective Analysis in the NHANES I Epidemiologic Follow-up Study. American Journal of Epidemiology. 2016;183(3):210-7. doi: 10.1093/aje/kwv190.

6. Zhou CK, Pfeiffer RM, Cleary SD, Hoffman HJ, Levine PH, Chu LW, et al. Relationship Between Male Pattern Baldness and the Risk of Aggressive Prostate Cancer: An Analysis of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. Journal of Clinical Oncology. 2014. doi: 10.1200/jco.2014.55.4279.

7. Christoffersen M, Frikke-Schmidt R, Schnohr P, Jensen GB, Nordestgaard BG, Tybjaerg-Hansen A. Visible Age-Related Signs and Risk of Ischemic Heart Disease in the

General Population: A Prospective Cohort Study. Circulation. 2013. doi: 10.1161/circulationaha.113.001696.

8. Shahar E, Heiss G, Rosamond WD, Szklo M. Baldness and Myocardial Infarction in Men: The Atherosclerosis Risk in Communities Study. American Journal of Epidemiology. 2008;167(6):676-83. doi: 10.1093/aje/kwm365.

9. Trieu N, Eslick GD. Alopecia and its association with coronary heart disease and cardiovascular risk factors: A meta-analysis. International Journal of Cardiology. 2014;176(3):687-95. doi: dx.doi.org/10.1016/j.ijcard.2014.07.079.

10. Richards JB, Yuan X, Geller F, Waterworth D, Bataille V, Glass D, et al. Male-pattern baldness susceptibility locus at 20p11. Nat Genet. 2008;40(11):1282-4.

11. Nyholt DR, Gillespie NA, Heath AC, Martin NG. Genetic Basis of Male Pattern Baldness. Journal of Investigative Dermatology. 2003;121(6):1561-4. doi: 10.1111/j.1523-1747.2003.12615.x.

12. Rexbye H, Petersen I, Iachina M, Mortensen J, McGue M, Vaupel JW, et al. Hair Loss Among Elderly Men: Etiology and Impact on Perceived Age. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences. 2005;60(8):1077-82. doi: 10.1093/gerona/60.8.1077.

13. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709-17. doi: 10.1038/ng.3570.

14. Li R, Brockschmidt FF, Kiefer AK, Stefansson H, Nyholt DR, Song K, et al. Six Novel Susceptibility Loci for Early-Onset Androgenetic Alopecia and Their Unexpected Association with Common Diseases. PLoS Genet. 2012;8(5):e1002746. doi: 10.1371/journal.pgen.1002746.

15. Heilmann-Heimbach S, Hochfeld LM, Paus R, Nöthen MM. Hunting the genes in male-pattern alopecia: how important are they, how close are we and what will they tell us? Experimental Dermatology. 2016;25(4):251-7. doi: 10.1111/exd.12965.

16. Heilmann S, Brockschmidt FF, Hillmer AM, Hanneken S, Eigelshoven S, Ludwig KU, et al. Evidence for a polygenic contribution to androgenetic alopecia. British Journal of Dermatology. 2013;169(4):927-30. doi: 10.1111/bjd.12443.

17. Marcińska M, Pośpiech E, Abidi S, Andersen JD, van den Berge M, Carracedo Á, et al. Evaluation of DNA Variants Associated with Androgenetic Alopecia and Their Potential to Predict Male Pattern Baldness. PLoS ONE. 2015;10(5):e0127852. doi: 10.1371/journal.pone.0127852.

18. Liu F, Hamer MA, Heilmann S, Herold C, Moebus S, Hofman A, et al. Prediction of male-pattern baldness from genotypes. Eur J Hum Genet. 2016;24(6):895-902. doi: 10.1038/ejhg.2015.220.

19. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015;12(3):e1001779. doi: 10.1371/journal.pmed.1001779.

20. Hébert JM, Rosenquist T, Götz J, Martin GR. FGF5 as a regulator of the hair growth cycle: Evidence from targeted and spontaneous mutations. Cell. 1994;78(6):1017-25. doi: http://dx.doi.org/10.1016/0092-8674(94)90276-3.

21. Adhikari K, Fontanil T, Cal S, Mendoza-Revilla J, Fuentes-Guajardo M, Chacon-Duque J-C, et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun. 2016;7. doi: 10.1038/ncomms10815.

22. Sorokin AV, Chen J. MEMO1, a new IRS1-interacting protein, induces epithelial-mesenchymal transition in mammary epithelial cells. Oncogene. 2013;32(26):3130-8. doi: 10.1038/onc.2012.327.

23. Rafnar T, Vermeulen SH, Sulem P, Thorleifsson G, Aben KK, Witjes JA, et al. European genome-wide association study identifies SLC14A1 as a new urinary bladder cancer susceptibility gene. Human Molecular Genetics. 2011. doi: 10.1093/hmg/ddr303.

24. Rademakers R, Cruts M, van Broeckhoven C. The role of tau (MAPT) in frontotemporal dementia and related tauopathies. Human Mutation. 2004;24(4):277-95. doi: 10.1002/humu.20086.

25. Billuart P, Bienvenu T, Ronce N, des Portes V, Vinet MC, Zemni R, et al. Oligophrenin-1 encodes a rhoGAP protein involved in X-linked mental retardation. Nature. 1998;392(6679):923-6.

26. Karlsson L, Bondjers C, Betsholtz C. Roles for PDGF-A and sonic hedgehog in development of mesenchymal components of the hair follicle. Development. 1999;126(12):2611-21.

27. Rendl M, Lewis L, Fuchs E. Molecular Dissection of Mesenchymal–Epithelial Interactions in the Hair Follicle. PLoS Biol. 2005;3(11):e331. doi: 10.1371/journal.pbio.0030331.

28. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, Zhu G, et al. Common Variants in the Trichohyalin Gene Are Associated with Straight Hair in Europeans. The American Journal of Human Genetics. 2009;85(5):750-5. doi: http://dx.doi.org/10.1016/j.ajhg.2009.10.009.

29. Adaimy L, Chouery E, Mégarbané H, Mroueh S, Delague V, Nicolas E, et al. Mutation in WNT10A Is Associated with an Autosomal Recessive Ectodermal Dysplasia: The Odonto-onycho-dermal Dysplasia. The American Journal of Human Genetics. 2007;81(4):821-8. doi: http://dx.doi.org/10.1086/520064.

30. Hillmer AM, Hanneken S, Ritzmann S, Becker T, Freudenberg J, Brockschmidt FF, et al. Genetic Variation in the Human Androgen Receptor Gene Is the Major Determinant of

Common Early-Onset Androgenetic Alopecia. The American Journal of Human Genetics. 2005;77(1):140-8. doi: http://dx.doi.org/10.1086/431425.

31. Prodi DA, Pirastu N, Maninchedda G, Sassu A, Picciau A, Palmas MA, et al. EDA2R Is Associated with Androgenetic Alopecia. Journal of Investigative Dermatology. 2008;128(9):2268-70. doi: http://dx.doi.org/10.1038/jid.2008.60.

32. Hirata H, Nanda I, van Riesen A, McMichael G, Hu H, Hambrock M, et al. ZC4H2 Mutations Are Associated with Arthrogryposis Multiplex Congenita and Intellectual Disability through Impairment of Central and Peripheral Synaptic Plasticity. The American Journal of Human Genetics. 2013;92(5):681-95. doi: http://dx.doi.org/10.1016/j.ajhg.2013.03.021.

33. Davies G, Marioni RE, Liewald DC, Hill WD, Hagenaars SP, Harris SE, et al. Genome-wide association study of cognitive functions and educational attainment in UK Biobank (N=112 151). Mol Psychiatry. 2016;21(6):758-67. doi: 10.1038/mp.2016.45.

34. König IR, Loley C, Erdmann J, Ziegler A. How to Include Chromosome X in Your Genome-Wide Association Study. Genetic Epidemiology. 2014;38(2):97-103. doi: 10.1002/gepi.21782.

35. Giles GG, Severi G, Sinclair R, English DR, McCredie MRE, Johnson W, et al. Androgenetic Alopecia and Prostate Cancer: Findings from an Australian Case-Control Study. Cancer Epidemiology Biomarkers & Prevention. 2002;11(6):549-53.

36. Norwood OT. Male pattern baldness: classification and incidence. South Med J. 1975;68(11):1359-65. PubMed PMID: 1188424.

37. 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(7):906-13.

38. Purcell S. PLINK v1.07.

39. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics. 2007;81(3):559-75. doi: http://dx.doi.org/10.1086/519795.

40. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

41. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015;11(4):e1004219. doi: 10.1371/journal.pcbi.1004219.

42. Euesden J, Lewis CM, O'Reilly PF. PRSice: Polygenic Risk Score software. Bioinformatics. 2014. doi: 10.1093/bioinformatics/btu848.

43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12(1):1-8. doi: 10.1186/1471-2105-12-77.

44. Team RC. R: A language and environment for statistical computing. 2013.

45. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010;42(7):565-9.

46. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics C, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47(3):291-5. doi: 10.1038/ng.3211

http://www.nature.com/ng/journal/v47/n3/abs/ng.3211.html#supplementary-information.

47. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological). 1995:289-300.

48. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet. 2015;47(11):1228-35. doi: 10.1038/ng.3404.

Table 1. Lookup of GWAS hits from Li et al. 2012 [14], Heilmann-Heimbach et al. 2016 [15] and Liu et al. 2016 [18].

Chr	Position	SNP ID	Effect allele	MAF	Beta	SE	P
1	11033082	rs12565727	G	0.22	-0.11	0.007	8.3E-53
2	219756383	rs7349332	T	0.15	0.07	0.009	2.1E-16
2	239694631	rs9287638	A	0.33	0.08	0.007	6.5E-37
3	151653368	rs4679955	T	0.41	0.05	0.006	6.4E-14
5	158310631	rs929626	G	0.47	-0.08	0.006	4.0E-40
6	6803438	rs4959410	T	0.37	0.01	0.006	3.7E-01
7	18877874	rs2073963	G	0.39	0.10	0.006	1.4E-51
7	18896011	rs2249817	G	0.43	0.10	0.006	3.8E-60
7	68611960	rs6945541	T	0.45	-0.08	0.006	1.3E-39
12	26426420	rs9668810	C	0.26	-0.06	0.007	5.6E-18
17	43924219	rs12373124	C	0.22	-0.08	0.007	1.2E-27
18	42800148	rs10502861	T	0.28	-0.07	0.007	1.7E-26
20	21853100	rs2180439	T	0.40	0.12	0.006	6.9E-87
20	22037575	rs6047844	C	0.44	-0.13	0.006	1.2E-103
20	22050503	rs1160312	G	0.46	-0.13	0.006	2.6E-98
X	66563018	rs2497938	C	0.18	-0.53	0.012	<1E-300
X	66510984	rs6625163	G	0.18	-0.53	0.012	<1E-300

Table 2. Top 20 independent autosomal GWAS hits

Chr	Position	SNP ID	Effect allele	MAF	Beta	SE	P	SNPs in Clump	SNPs with P<0.0001	Gene	Function
1	11040385	rs7542354	A	0.22	-0.12	0.007	5.74E-55	86	80	*C1orf127*	intronic
1	25467880	rs12745121	A	0.31	-0.07	0.007	5.51E-26	134	56	NA	NA
1	170361164	rs10919382	G	0.38	0.07	0.006	4.54E-25	514	463	NA	NA
2	32181424	rs13021718	A	0.14	-0.09	0.009	7.26E-26	541	396	*MEMO1/DPY30*	intronic
2	239695893	rs11684254	G	0.35	0.09	0.006	1.10E-40	140	134	AC144525.1	3downstream
3	139032333	rs7642536	C	0.14	0.11	0.009	1.28E-34	49	40	*MRPS22*	intronic
4	81197949	rs7680591	A	0.42	0.07	0.006	1.44E-26	107	103	*FGF5*	intronic
5	158320877	rs1422798	G	0.38	-0.09	0.006	2.84E-46	261	176	*EBF1*	intronic
6	396321	rs12203592	T	0.22	0.11	0.007	2.64E-49	8	8	*IRF4*	intronic/5upsteam
6	9327556	rs9357047	C	0.44	0.08	0.006	3.07E-35	229	194	NA	NA
7	18896988	rs71530654	G	0.40	0.11	0.006	4.68E-70	98	98	*HDAC9*	intronic
7	68587797	rs939963	C	0.45	-0.09	0.006	3.58E-48	104	94	NA	NA
8	109145555	rs79206101	T	0.01	0.31	0.028	4.91E-27	25	24	AP001331.1	5upstream
17	44066172	rs112385572	G	0.24	-0.08	0.007	8.45E-29	376	366	*MAPT*	intronic
17	44787313	rs538628	C	0.22	-0.08	0.007	3.60E-27	81	77	*NSF*	intronic
18	42814156	rs8085664	A	0.28	-0.07	0.007	3.47E-27	338	306	*SLC14A2*	intronic
20	21894764	rs6035986	T	0.43	0.12	0.006	1.20E-87	0	0	NA	NA
20	22033819	rs201593	A	0.43	0.13	0.006	1.06E-103	662	662	*LOC100270679/ RP11-125P18.1*	5upstream
20	22100070	rs7362397	T	0.30	-0.11	0.007	3.02E-65	0	0	NA	NA
20	22100072	rs7362398	T	0.30	-0.11	0.007	3.02E-65	0	0	NA	NA

Table 3. AUC results for severe hair loss versus no hair loss for all autosomal and X chromosome polygenic thresholds.

Genomic region	P value threshold	n SNPs	AUC
Autosomal	<1x10-20	37	0.648
	<5x10-8	224	0.725
	<1x10-5	546	0.748
	<0.01	10581	0.725
	<0.05	40047	0.701
	<0.10	74068	0.687
	<0.50	322885	0.663
	<1	622739	0.662
X chromosome	<1x10-20	52	0.611
	<5x10-8	109	0.630
	<1x10-5	159	0.625
	<0.01	479	0.649
	<0.05	1139	0.673
	<0.10	1914	0.691
	<0.50	7761	0.720
	<1	14653	0.724

List of Figures

Figure 1. Manhattan Plots of imputed autosomal GWAS (top panel) and genotyped X chromosome GWAS (bottom panel) of male pattern baldness.

Figure 2. Area under the curve plot for discriminating those with hair loss from those with no loss.

Figure 3. Screenshot of UK Biobank question 2395 on male-pattern baldness

Supplementary Table 1. List of top 10 genotyped male-pattern baldness GWAS hits for the X chromosome.

Chr	BP	SNP ID	Effect Allele	Beta	SE	P	SNPs in clump	SNPs with P<0.0001
X	65933285	rs73221556	A	-0.53	0.02	<5.1E-178	19	19
X	66481800	rs12558842	C	-0.54	0.03	<5.1E-178	8	8
X	67003584	rs5919427	C	-0.35	0.02	5.1E-178	11	11
X	67496002	rs140488081	T	-0.40	0.01	4.6E-61	9	4
X	65083247	rs147154263	T	-0.43	0.02	3.6E-58	2	2
X	67139063	rs148652266	A	-0.51	0.01	7.7E-45	1	1
X	65541956	rs145867342	T	-0.32	0.01	1.1E-44	2	2
X	67363801	rs7061504	G	0.19	0.04	1.7E-38	5	4
X	58005480	rs147829649	G	-0.28	0.01	1.3E-31	1	0
X	66337545	rs17216820	T	0.19	0.02	5.8E-26	2	2

Supplementary Table 2. Genome-wide significant autosomal gene-based hits (Bonferroni correction of $\alpha < 2.769e\text{-}06$) in the MAGMA gene-based analysis for male pattern baldness. NSNPS is the number of SNPs in the gene.

(See attached Excel spreadsheet)

Supplementary Table 3. Genome-wide significant gene-based hits (Bonferroni correction of $\alpha < 8.818e\text{-}05$) in the MAGMA gene-based analysis for male pattern baldness performed on the X chromosome. NSNPS is the number of SNPs in the gene.

(See attached Excel spreadsheet)

Supplementary Table 4. Genetic correlations between baldness and the 22 cognitive, health, psychiatric, and anthropometric variables. The heritability Z-score and the mean χ^2 indicate the level of power to detect association where a heritability Z-score of >4 and a mean χ^2 >1.02 being considered well powered [46]. None of the 23 tests performed survived FDR control for multiple comparisons.

Phenotypes	**Genetic correlation**	**Standard error**	***P*-value**	**Heritability Z-score**	**Mean χ^2**
Cognitive abilities					
Childhood intelligence	-0.039	0.064	0.544	5.830	1.076
Educational attainment	-0.046	0.038	0.221	13.250	1.226
Verbal-numerical reasoning	-0.034	0.048	0.482	10.526	1.168
Health and Metabolic					
Coronary artery disease	0.003	0.049	0.958	8.111	1.146
HbA1c	-0.033	0.054	0.541	5.231	1.060
HDL cholesterol	-0.018	0.042	0.673	5.600	1.153
HOMA B	0.095	0.062	0.125	6.273	1.053
HOMA IR	0.123	0.068	0.072	5.400	1.053
LDL cholesterol	-0.043	0.047	0.361	3.654	1.141
Obesity	0.018	0.032	0.567	17.100	1.125
Triglycerides	-0.023	0.034	0.510	5.846	1.154
Type 2 diabetes	-0.028	0.051	0.584	8.727	1.134
Waist to hip ratio	-0.016	0.04	0.692	13.286	1.358
Psychiatric					
ADHD	-0.083	0.105	0.434	2.417	1.016
Alzheimer's disease 500bp	0.058	0.078	0.462	5.545	1.106
Alzheimer's disease	0.064	0.096	0.506	1.909	1.115
Autism	-0.035	0.052	0.498	8.519	1.059
Bipolar disorder	-0.126	0.052	0.015	10.341	1.186
MDD	0.123	0.083	0.141	5.586	1.078
Schizophrenia	-0.055	0.03	0.068	22.208	1.814
Neuroticism	0.091	0.078	0.241	4.500	1.057
Anthropometric/Developmental					
Height	-0.061	0.028	0.030	17.400	2.982
Age at menarche	-0.096	0.032	0.003	20.625	1.401

ADHD, attention deficit hyperactivity disorder; MDD, major depressive disorder

Supplementary Table 5. Showing the full output of the partitioned heritability analysis for male pattern baldness. Prop._SNPs refers to the proportion of SNPs from the data set that were a part of the corresponding functional annotation. Statistical significance indicated in bold.

(See attached Excel spreadsheet)

Supplementary Table 6. The 22 health-related phenotypes included in the genetic correlation analysis with male pattern baldness. Verbal-numerical reasoning and childhood intelligence were examined as educational attainment (genetic association with baldness reported by Pickrell et al. 2016 [13]) can be used as a proxy phenotype for general cognitive ability. Metabolic traits were included as metabolic disease has been associated with baldness (references noted in the review paper by Heilmann-Heimbach et al. 2016 [15]). Psychiatric disorders were included due to the association between baldness and neurological conditions such as Parkinson's disease [14]. Genetic correlations have been observed between baldness and the listed anthropometric and developmental traits [13].

Cognitive abilities
Childhood intelligence
Educational attainment
Verbal-numerical reasoning (Intelligence)
Health and Metabolic
Coronary artery disease
Homeostatic model assessment beta-cell function (HOMA B)
Homeostatic model assessment insulin resistance (HOMA IR)
Glycated haemoglobin (HbA1c)
High density lipoprotein (HDL) cholesterol
Low density lipoprotein (LDL) cholesterol
Obesity
Triglycerides
Type 2 Diabetes
Waist to hip ratio
Psychiatric disease
Attention deficit hyperactivity disorder (ADHD)
Alzheimer's disease
Autism
Bipolar disorder
Major depressive disorder (MDD)
Schizophrenia

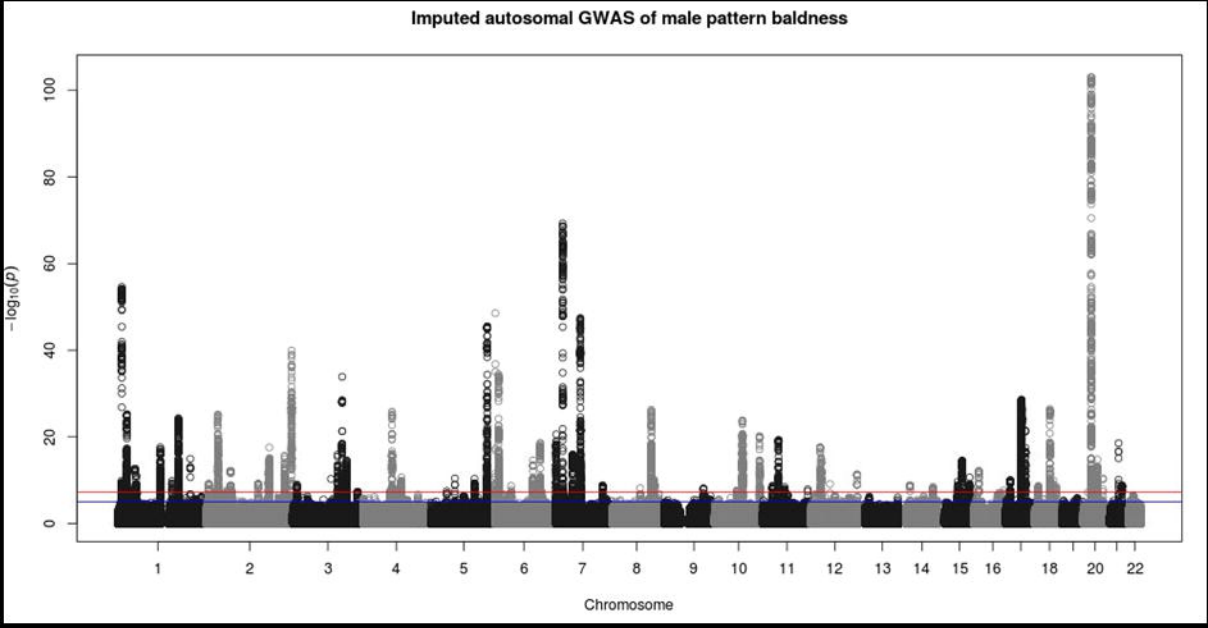
Neuroticism
Anthropometric and developmental traits
Height
Age at menarche

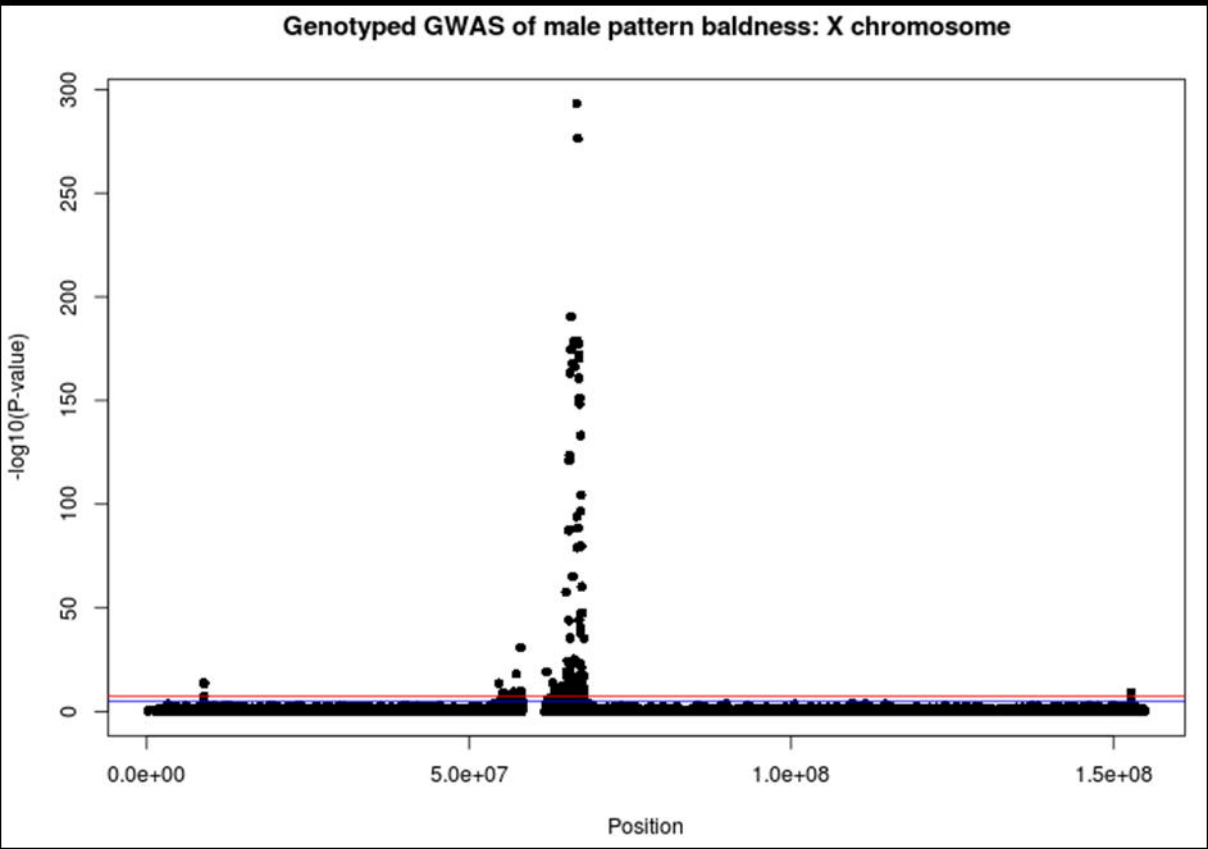
List of Supplementary Figures

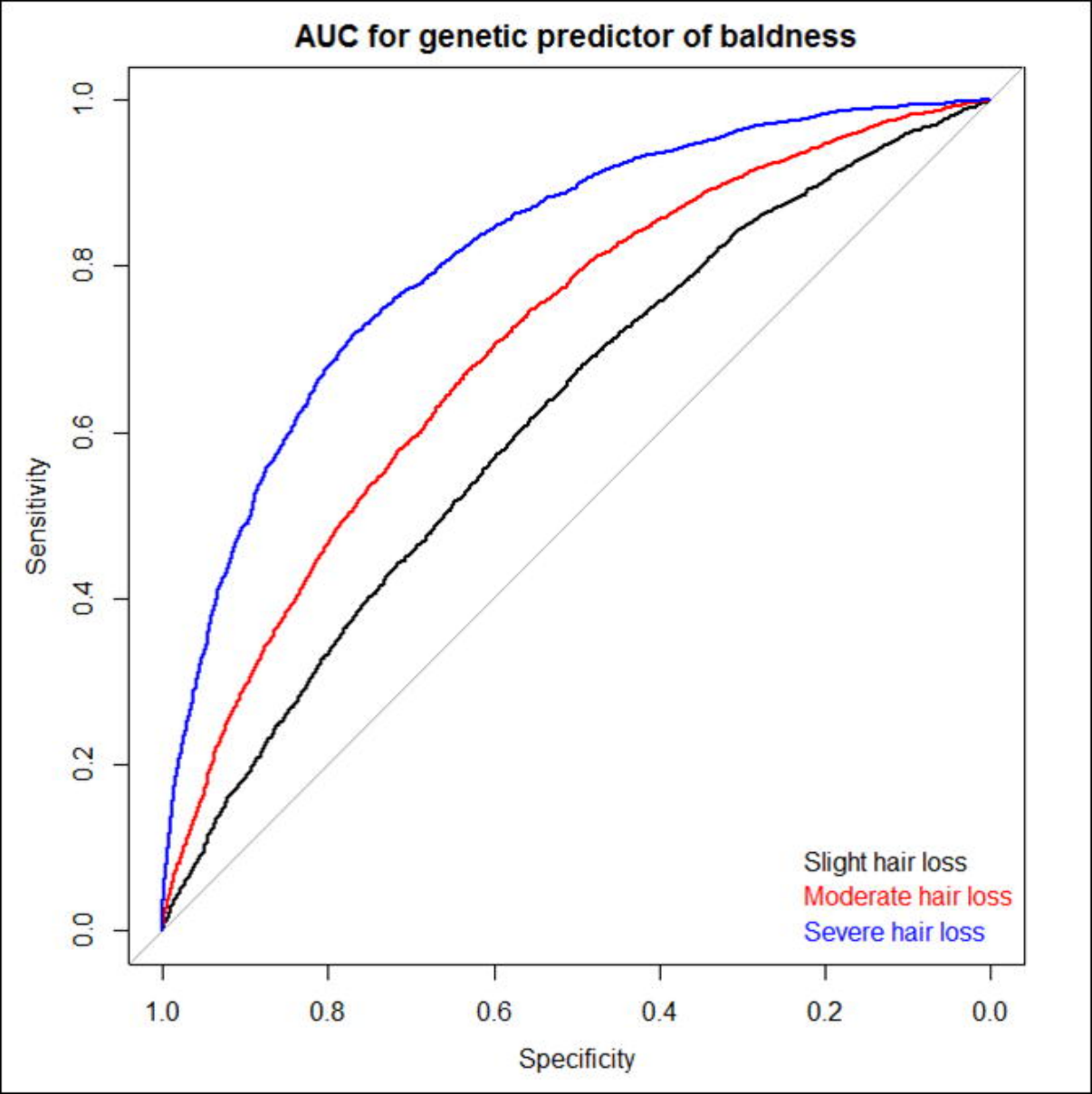
Supplementary Figure 1. Male-pattern baldness QQ Plot for imputed GWAS of autosomal variants (p-values truncated at $1x10^{-30}$)

Supplementary Figure 2. Enrichment analysis for male pattern baldness using the 52 functional categories in 52,874 individuals. The enrichment statistic is the proportion of heritability found in each functional group divided by the proportion of SNPs in each group ($Pr(h^2)/Pr(SNPs)$). Error bars are jackknife standard errors around the estimate of enrichment. The dashed line indicates no enrichment found when $Pr(h^2)/Pr(SNPs) = 1$. FDR correction indicated significance at P=0.011 indicated by asterisk.

Supplementary Figure 3. Enrichment analysis for male pattern baldness using the 10 cell specific functional The enrichment statistic is the proportion of heritability found in 52,874 individuals. in each functional group divided by the proportion of SNPs in each group ($Pr(h^2)/Pr(SNPs)$. Error bars are jackknife standard errors around the estimate of enrichment. The dashed line indicates no enrichment found when $Pr(h^2)/Pr(SNPs) = 1$. FDR correction indicated significance at P=0.037 indicated by asterisk.


Imputed autosomal GWAS of male pattern baldness
$-\log_{10}(p)$
0
20
40
60
80
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
18
20
22
Chromosome


Genotyped GWAS of male pattern baldness: X chromosome
-log10(P-value)
0
50
100
150
200
250
300
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position


AUC for genetic predictor of baldness
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Slight hair loss
Moderate hair loss
Severe hair loss

Which of the following best describes your hair/balding pattern?

Pattern 1

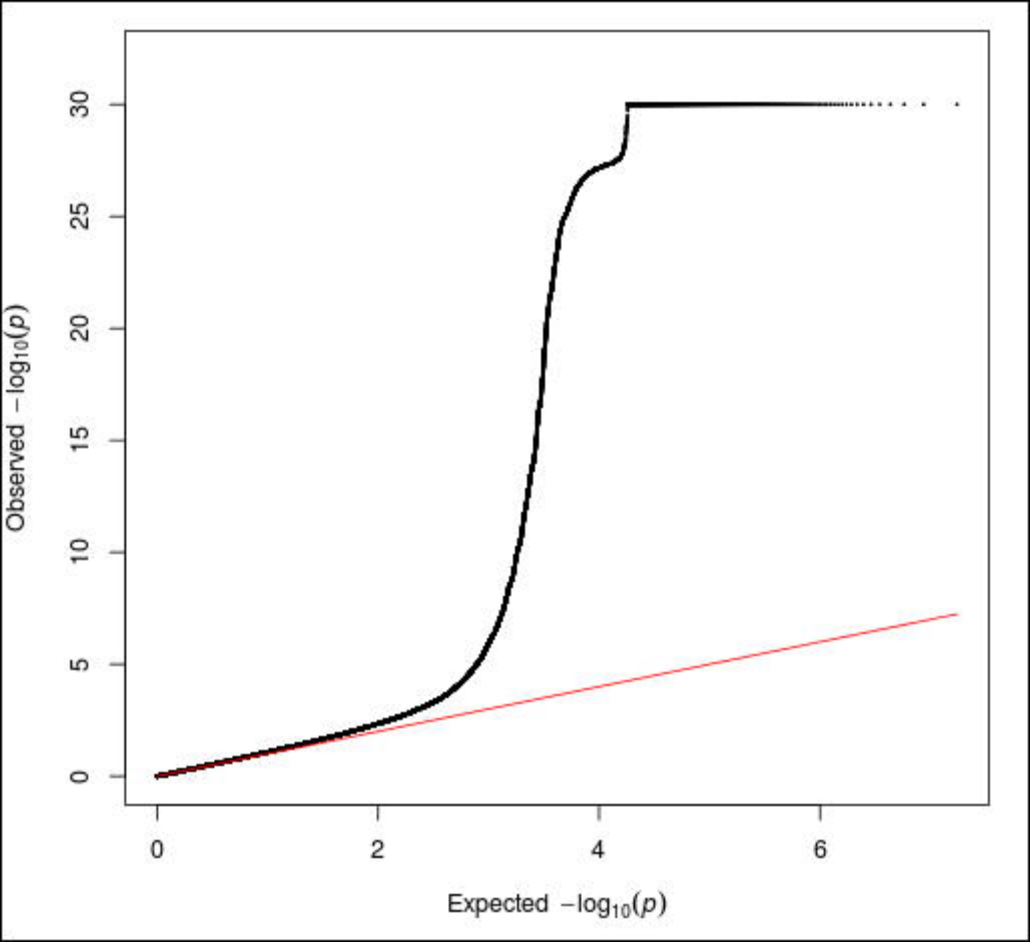
Pattern 2

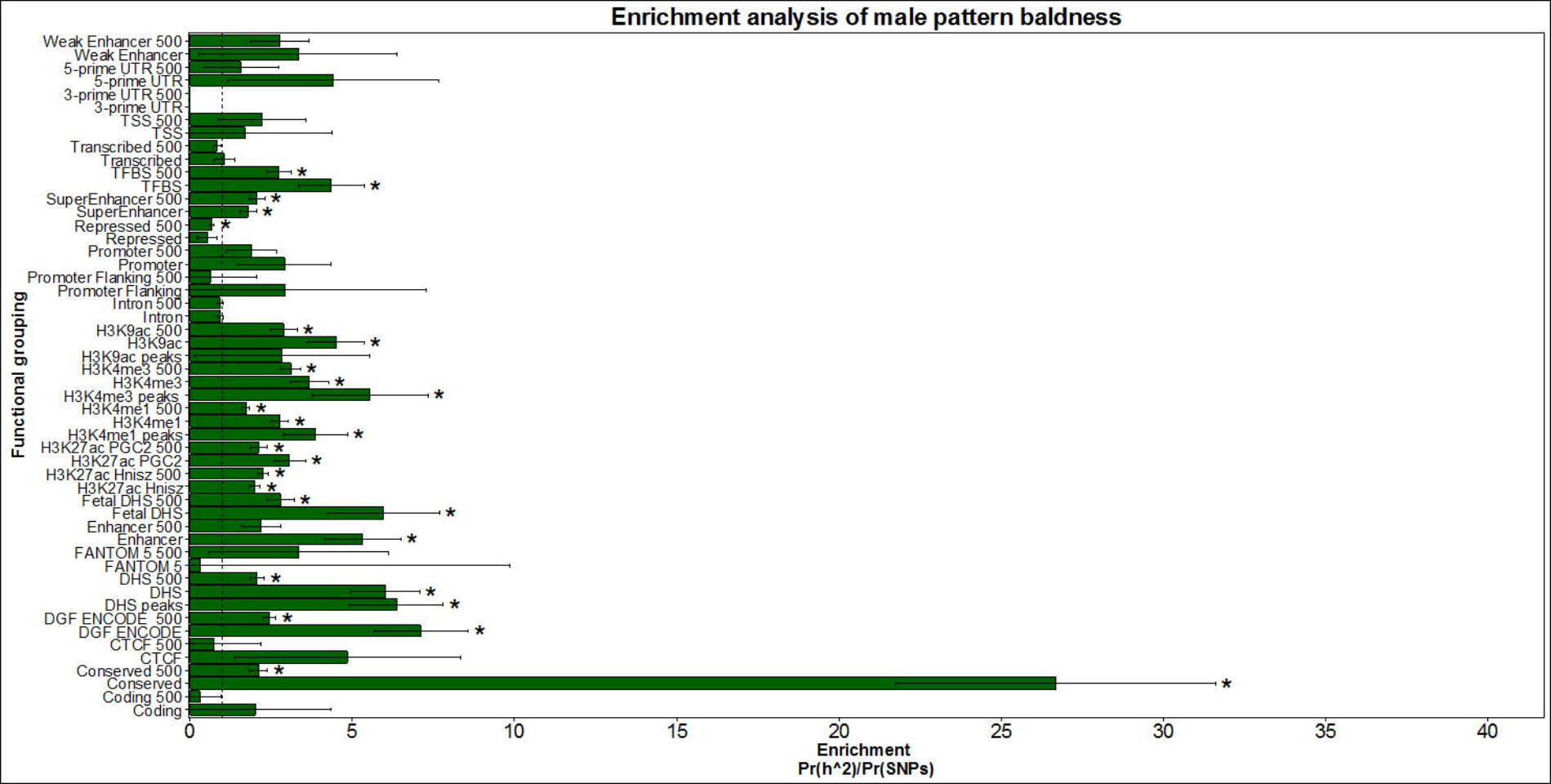
Pattern 3

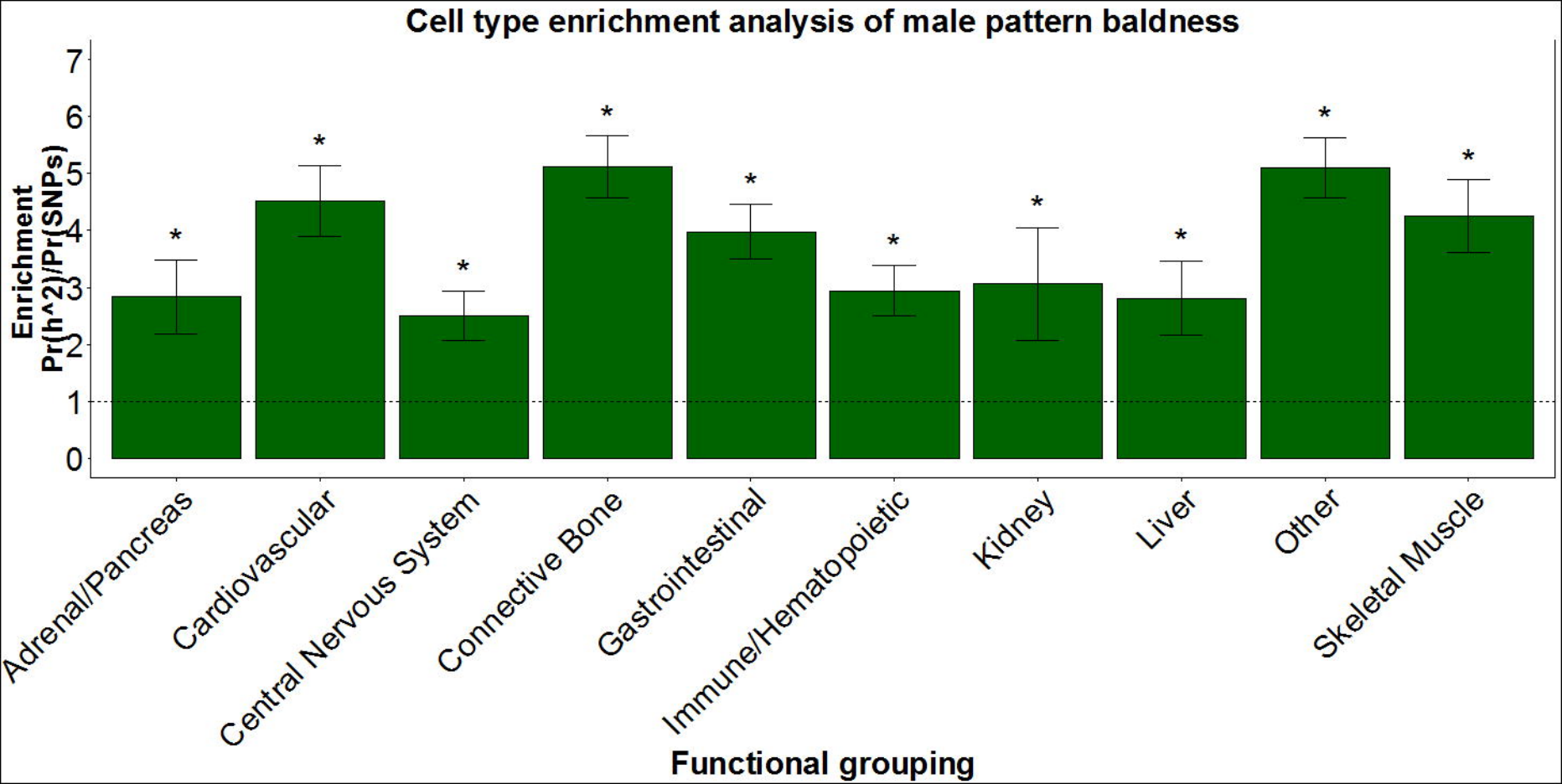
Pattern 4

Do not know

Prefer not to answer


Observed $-\log_{10}(p)$
Expected $-\log_{10}(p)$
0
5
10
15
20
25
30
0
2
4
6


Enrichment analysis of male pattern baldness
Functional grouping
Weak Enhancer 500
Weak Enhancer
5-prime UTR 500
5-prime UTR
3-prime UTR 500
3-prime UTR
TSS 500
TSS
Transcribed 500
Transcribed
TFBS 500
TFBS
SuperEnhancer 500
SuperEnhancer
Repressed 500
Repressed
Promoter 500
Promoter
Promoter Flanking 500
Promoter Flanking
Intron 500
Intron
H3K9ac 500
H3K9ac
H3K9ac peaks
H3K4me3 500
H3K4me3
H3K4me3 peaks
H3K4me1 500
H3K4me1
H3K4me1 peaks
H3K27ac PGC2 500
H3K27ac PGC2
H3K27ac Hnisz 500
H3K27ac Hnisz
Fetal DHS 500
Fetal DHS
Enhancer 500
Enhancer
FANTOM 5 500
FANTOM 5
DHS 500
DHS
DHS peaks
DGF ENCODE 500
DGF ENCODE
CTCF 500
CTCF
Conserved 500
Conserved
Coding 500
Coding
0
5
10
15
20
25
30
35
40
Enrichment
Pr(h^2)/Pr(SNPs)


Cell type enrichment analysis of male pattern baldness
Enrichment
Pr(h^2)/Pr(SNPs)
0
1
2
3
4
5
6
7
*
Adrenal/Pancreas
Cardiovascular
Central Nervous System
Connective Bone
Gastrointestinal
Immune/Hematopoietic
Kidney
Liver
Other
Skeletal Muscle
Functional grouping