A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features

Kaustubh Adhikari1, Tania Fontanil2, Santiago Cal2, Javier Mendoza-Revilla1,3, Macarena Fuentes-Guajardo1,4, Juan-Camilo Chacón-Duque1, Farah Al-Saadi1, Jeanette A. Johansson5, Mirsha Quinto-Sanchez6, Víctor Acuña-Alonzo1,7, Claudia Jaramillo8, William Arias8, Rodrigo Barquera Lozano7,9, Gastón Macín Pérez7,9, Jorge Gómez-Valdés10, Hugo Villamil-Ramírez9, Tábita Hunemeier11, Virginia Ramallo6,11, Caio C. Silva de Cerqueira6,11, Malena Hurtado3, Valeria Villegas3, Vanessa Granja3, Carla Gallo3, Giovanni Poletti3, Lavinia Schuler-Faccini11, Francisco M. Salzano11, Maria-Cátira Bortolini11, Samuel Canizales-Quinteros9, Francisco Rothhammer12, Gabriel Bedoya8, Rolando Gonzalez-José6, Denis Headon5, Carlos López-Otín2, Desmond J. Tobin13, David Balding1,14 & Andrés Ruiz-Linares1

We report a genome-wide association scan in over 6,000 Latin Americans for features of scalp hair (shape, colour, greying, balding) and facial hair (beard thickness, monobrow, eyebrow thickness). We found 18 signals of association reaching genome-wide significance ($P$ values $5 \times 10^{-8}$ to $3 \times 10^{-119}$), including 10 novel associations. These include novel loci for scalp hair shape and balding, and the first reported loci for hair greying, monobrow, eyebrow and beard thickness. A newly identified locus influencing hair shape includes a Q30R substitution in the Protease Serine S1 family member S3 (PRSS53). We demonstrate that this enzyme is highly expressed in the hair follicle, especially the inner root sheath, and that the Q30R substitution affects enzyme processing and secretion. The genome regions associated with hair features are enriched for signals of selection, consistent with proposals regarding the evolution of human hair.
**Results**

**Study population and phenotypes.** Our study sample consists of 6,630 volunteers from the CANDELA cohort recruited in five Latin American countries (Brazil, Colombia, Chile, México and Perú; Supplementary Table 1)\(^{17,18}\). In these individuals, we performed a categorical assessment (in men and women) of: scalp hair shape (curliness), colour, balding and greying as well as (in men) of beard thickness (that is, density), monobrow and eyebrow thickness (Fig. 1 and Supplementary Fig. 1). These individuals were genotyped on Illumina’s Omni Express BeadChip. After applying quality control filters, 669,462 single-nucleotide polymorphisms (SNPs) and 6,357 individuals were retained for further analyses (including 2,922 males). Average autosomal admixture proportions for this sample were estimated as: 48% European, 46% Native American and 6% African, but with substantial inter-individual variation (Supplementary Fig. 2). Several of the traits examined show low to moderate, but significant, correlations between them and with basic covariates (using a 10\(^{-5}\) Bonferroni-adjusted threshold; Supplementary Table 2). The highest significant trait correlations occur between: beard and eyebrow density and monobrow (r = 0.14 to r = 0.24); balding and beard density (r = 0.15); hair greying and balding (r = 0.13) and hair greying and beard density (r = 0.13). The highest trait-covariate correlations detected were: age with hair greying (r = 0.56), balding (r = 0.15), beard density (r = 0.28) and eyebrow thickness (r = 0.17); balding and hair colour with sex (r = −0.35 and r = −0.10); European ancestry correlates most strongly with hair colour (r = −0.32) and beard density (r = 0.3). Based on a kinship matrix derived from the SNP data\(^{19}\), we estimated narrow-sense heritability using GCTA\(^{20}\). We found significant values for all the traits examined (Supplementary Table 3), with the highest heritability being estimated for hair colour (1) and the lowest for hair greying (0.27). These estimates of heritability are similar to other available estimates based on family data\(^{21,10}\).

**GWAS for hair features.** We performed genome-wide association tests on 9,143,600 chip genotyped and imputed SNP data using multivariate linear regression, as implemented in PLINK\(^{21}\), using an additive genetic model adjusting for: age, sex and the first five SNP principal components (Supplementary Fig. 3). All the traits scored showed genome-wide significant association (P values < 5 × 10\(^{-8}\)) with SNPs in at least one genomic region (Fig. 1 and Table 1). We re-examined the association signals for each index SNP in every country sample separately (by performing independent association tests) and combined results as a meta-analysis using METAL\(^{22}\). For each SNP, significant effects were in the same direction in all countries, the variability of effect size across countries reflecting sample size (Fig. 2, Supplementary Fig. 4 and Supplementary Table 4). To probe the extent of phenotypic variation captured by the genetic data, we constructed a phenotype prediction model combining the index SNPs (Table 1) and a BLUP (Best Linear Unbiased Predictor) random effects component calculated from the genome-wide kinship matrix\(^{19}\), in addition to covariates (Supplementary Table 5). The BLUP term is sensitive to the effect of SNPs associated below the threshold for genome-wide significance. The prediction accuracy of this model was broadly consistent with the heritability estimates (Supplementary Table 3). Highest prediction accuracy was observed for hair colour, with 41% of the total phenotypic variance for this trait being captured by the model, including 20% explained by the index SNPs and the BLUP component. The lowest prediction accuracy was observed for monobrow, for which 16.2% of the total phenotypic variance was captured by the model. The difference between the heritability estimates and the prediction accuracy partly reflects limitations of the model, including that the BLUP component is likely to capture imperfectly the polygenicity of the traits examined (Supplementary Fig. 5).

**Candidate genes in genome regions associated with hair traits.** Several of the regions showing genome-wide significant association include strong candidate genes. Scalp hair shape and beard thickness are strongly associated with SNPs in 2q12 (Fig 3a,b) and two other traits (eyebrow thickness and monobrow) show genome-wide suggestive P values for SNPs in this same region (Supplementary Fig. 6). Associated SNPs overlap the EDAR (ectodysplasin A receptor) gene. EDAR acts as part of the EDA-EDAR-EDARADD signalling pathway\(^{23}\) during prenatal development to specify the location, size and shape of ectodermal appendages, such as hair follicles, teeth and glands\(^{24}\). Strongest association with hair shape was observed for SNP

There is great diversity in primates with regard to hair appearance, including its distribution, shape and colour\(^{1}\). Hair plays a range of important functions, including thermal regulation, camouflage, sensory and social signalling, and the evolution of hair has been proposed to be influenced by both natural and sexual selection\(^{1}\). Humans differ from other primates in having lost most terminal body hair (via hair follicle miniaturization), possibly in connection to the adaptive development of more efficient sweating, linked to bipedalism\(^{2,3}\). However, considerable head hair has been retained in modern humans and its appearance shows extensive variation between individuals\(^{4–6}\). Head hair appearance is highly heritable\(^{7–10}\) and certain traits show high differentiation between continental native populations. For instance, variation in hair colour is essentially absent from sub-Saharan Africa\(^{4}\). It has been proposed that variation in head hair appearance has been influenced by selection during the evolution of modern humans\(^{3,11}\). Although genetic association studies for human hair traits are few to date, loci influencing male-pattern baldness, scalp hair colour and shape (curliness) have been identified in samples of European and East Asian ancestry\(^{6}\). Consistent with the key role of androgens in balding, the androgen receptor (AR) gene region has been highlighted as a major determinant among several loci associated with male-pattern baldness\(^{12}\). The genes associated with hair colour, involved in various aspects of melanocyte biology and melanin synthesis, also often impact on skin and eye pigmentation\(^{13}\). Interestingly, different genes have been associated with straight hair in Europeans and East Asians, suggesting that this trait evolved independently at least twice. The most robust associations for straight hair have implicated Trichohyalin (TCHH, a structural hair protein) in Europeans\(^{14,15}\), and EDAR (a cell signalling receptor) in East Asians\(^{16}\), illustrating the range of cellular mechanisms that can impact on hair shape.

To further our understanding of the genetic basis of variation in human hair, we performed a genome-wide association study (GWAS) in individuals of mixed European, Native American and African ancestry, who show high genetic diversity and extensive variation in head hair appearance. We report ten novel genetic associations, including the first reported loci for hair greying and for facial hair distribution/density. A newly identified locus influencing scalp hair shape encodes a serine protease (PRSS53) that we show is expressed in the hair follicle, with strongest association seen for a Q30R substitution that we show affects enzyme processing. Similar to what has been observed in the EDAR gene region, we find evidence of recent selection in East Asians at PRSS53.
rs3827760 (Fig. 3a and Table 1), coding for a V370A substitution in EDAR. This variant has been robustly associated with hair shape in East Asians. In a previous study of the CANDELA sample, which examined 30 ancestry informative markers, we found association between hair shape and a SNP in EDAR that is in strong linkage disequilibrium (LD) with rs3827760 in the 1000 genomes data. The EDAR SNP showing strongest association with beard density is not the coding SNP rs3827760 associated with hair shape but rather rs365060 located 62 kb upstream of rs3827760 in the first intron of EDAR (Fig. 3b). Several SNPs in this intron have smaller association $P$ values than rs3827760 and analyses conditioning on rs3827760 suggest that the association signal of SNPs in this intron is independent from that observed at rs3827760 (Supplementary Fig. 7). These intronic SNPs are located in a different LD block and there is a recombination hotspot between them and rs3827760 (Fig 3a,b and Supplementary Fig. 7A). Interestingly, the first intron of EDAR is rich in regulatory elements and SNPs in this intron show evidence of recent selection in Europeans (discussed further below). Hypohidrotic ectodermal dysplasia is a Mendelian disorder caused by mutations in the EDA-EDAR-EDARADD pathway and is characterized by sparse scalp hair, eyebrows and eye lashes. Transgenic mice with increased Edar function have been shown to have thickened and straightened hair fibres. We therefore examined chin hair follicle density in wild-type mice and in an Edar gain-of-function transgenic strain.
(EdarTg951/Tg951)26. Consistent with the effect of EDAR on beard thickness, we found that the EdarTg951/Tg951 strain has significantly lower chin hair follicle density compared with wild-type mice (Fig. 4a,b).

Apart from 2q12, scalp hair shape is associated with SNPs in three other genomic regions (in 1q21, 10p14 and 16p11). Genome-wide significant association in the 16p11 region was strongest for rs11150606, the derived allele of which codes for a Q30R substitution in the Protease Serine S1 family member 53 (PRSS53; Fig. 3c and Table 1). Proteases and protease inhibitors are known to be important for epidermal keratinization and regulate hair growth and cycling28. A spontaneous mouse mutant (frizzy, fr), characterized by curly whiskers, carries an amino-acid substitution in another serine protease (Prss8)29 and a conditional knock-out with no expression of Prss8 in the epidermis shows hair abnormalities in newborns and defects in corneocyte morphogenesis, epidermal lipid composition, profilaggrin processing and tight junction assembly30.

Apart from 2q12, scalp hair shape is associated with SNPs in three other genomic regions (in 1q21, 10p14 and 16p11). Genome-wide significant association in the 16p11 region was strongest for rs11150606, the derived allele of which codes for a Q30R substitution in the Protease Serine S1 family member 53 (PRSS53; Fig. 3c and Table 1). Proteases and protease inhibitors are known to be important for epidermal keratinization and regulate hair growth and cycling28. A spontaneous mouse mutant (frizzy, fr), characterized by curly whiskers, carries an amino-acid substitution in another serine protease (Prss8)29 and a conditional knock-out with no expression of Prss8 in the epidermis shows hair abnormalities in newborns and defects in corneocyte morphogenesis, epidermal lipid composition, profilaggrin processing and tight junction assembly30.

**Table 1 | Features of index SNPs showing strongest genome-wide significant association (P value < 5 × 10⁻⁸) to the scalp and facial hair features examined in the CANDELA sample.**

| Trait       | Region | SNP               | Closest gene | P value | Alleles ancestral > derived |
|-------------|--------|-------------------|--------------|---------|-----------------------------|
| Hair shape  | 1q21   | rs11803731        | TCHH         | 3 × 10⁻¹² | A > T                      |
| Hair shape  | 2q12   | rs3827760         | EDAR         | 3 × 10⁻¹¹ | A > G                      |
| Hair shape  | 10p14  | rs17143387        | GATA3        | 4 × 10⁻⁸  | T > G                      |
| Hair shape  | 16p11  | rs11150606        | PRSS53       | 7 × 10⁻⁹  | T > C                      |
| Beard thickness | 2q12 | rs365060          | EDAR         | 1 × 10⁻⁵  | G > C                      |
| Beard thickness | 4q12 | rs4864809         | LNX1         | 1 × 10⁻⁸  | A > G                      |
| Beard thickness | 6q21 | rs6901317         | PREP         | 4 × 10⁻⁸  | G > T                      |
| Beard thickness | 7q31 | rs117717824       | FOXP2        | 1 × 10⁻⁸  | G > T                      |
| Eyebrow thickness | 3q22 | rs112458845       | FOXL2        | 4 × 10⁻¹⁰ | G > A                      |
| Monobrow | 2q36   | rs2218065         | FOX5         | 5 × 10⁻⁸  | A > G                      |
| Hair greying | 6p25   | rs12203592        | IRF4         | 3 × 10⁻¹³ | C > T                      |
| Hair colour | 5p13   | rs183671          | SLC4A5       | 2 × 10⁻⁵  | G > A                      |
| Hair colour | 6p25   | rs12203592        | IRF4         | 1 × 10⁻¹³ | C > T                      |
| Hair colour | 1q14   | rs598952          | TYR          | 2 × 10⁻⁸  | T > A                      |
| Hair colour | 15q13  | rs12913832        | HERC2/OC2A2  | 5 × 10⁻¹⁰ | G > A                      |
| Hair colour | 15q21  | rs1426654         | SLC24A5      | 1 × 10⁻¹⁸ | G > A                      |
| Bolding | 10q22  | rs2814331         | GRID1        | 4 × 10⁻⁹  | C > T                      |
| Bolding | 10q12  | rs4258142         | AR/EDA2R     | 2 × 10⁻⁸  | C > T                      |

SNPs, single-nucleotide polymorphism. Regions with intragenic index SNPs are shown in bold (the gene name is underlined if this index SNP is coding). SNPs in regions reaching only genome-wide suggestive association, P value < 10⁻⁵, are listed in Supplementary Table 6.

Apart from 2q12, scalp hair shape is associated with SNPs in three other genomic regions (in 1q21, 10p14 and 16p11). Genome-wide significant association in the 16p11 region was strongest for rs11150606, the derived allele of which codes for a Q30R substitution in the Protease Serine S1 family member 53 (PRSS53; Fig. 3c and Table 1). Proteases and protease inhibitors are known to be important for epidermal keratinization and regulate hair growth and cycling28. A spontaneous mouse mutant (frizzy, fr), characterized by curly whiskers, carries an amino-acid substitution in another serine protease (Prss8)29 and a conditional knock-out with no expression of Prss8 in the epidermis shows hair abnormalities in newborns and defects in corneocyte morphogenesis, epidermal lipid composition, profilaggrin processing and tight junction assembly30.

The 1q21.3 region overlaps the trichohyalin gene (TCHH), which has been previously associated with scalp hair curliness in Europeans14,15. Strongest association was seen for SNP rs11803731, encoding a M790L substitution in TCHH (Supplementary Fig. 8), consistent with previous GWASs14,15. TCHH is expressed in cornifying keratinocytes of epithelia, particularly in the inner root sheath (IRS) and the hair fibre medulla of hair follicles where it is involved in the cross-linking of the cornified envelope with cellular keratin filaments31. SNPs in 10p14 overlap LINC00708, ~150 kb downstream of the GATA-binding protein 3 gene (GATA3), an interesting candidate in this region (Supplementary Fig. 8). Gata3 is expressed in the hair follicle IRS of mice and a Gata3 null mutant shows abnormal hair growth and shape32. Interestingly, Gata3 mutant hair follicles have greatly reduced expression of trichohyalin33.

Eyebrow thickness shows genome-wide significant association to SNPs on 3q23 overlapping the forkhead box L2 (FOXL2) gene (Fig. 3d). Rare mutations in the FOXL2 gene region (including coding variants as well as upstream and downstream intergenic rearrangements) cause blepharophimosis syndrome (BPES)34, an autosomal dominant eyelid malformation often accompanied by thick eyebrows. Mouse experiments have shown that Foxl2 is expressed around the eyes up to the time that hair is formed (E13.5)35 and a mutant with altered Foxl2 expression (and BPES features) typically shows hair loss around the eyes36.

Monobrow shows genome-wide significant association with SNPs in 2q36 with strongest association being observed for marker rs2395845 located ~70 kb downstream of the paired box gene 3 (PAX3) gene, an interesting candidate in the region (Fig. 3e). Rare mutations of PAX3 have been shown to cause Waardenburg syndrome type 1 (WS1). WS is a clinically and genetically heterogeneous Mendelian disorder of neural crest derivatives whose manifestations include deafness, a range of pigmentation abnormalities, broad nasal bridge and monobrow (seen in ~85% of WS1 patients37). Intrinsic SNPs within PAX3 have been implicated in recent GWASs of facial morphology, particularly in relation with nasion position (the point just above the nasal bridge)38,39. PAX3 is a key transcription factor during embryogenesis and analysis of mouse mutants have confirmed that it is essential to guide normal development of neural crest derivatives40.

Hair greying shows genome-wide significant association to SNP rs12203592 in intron 4 of the interferon regulatory factor 4 gene (IRF4; Fig. 3f). This SNP also shows association with hair colour in our sample (Table 1) and in previous studies this SNP has been associated with skin, hair and eye pigmentation13. Recent in-vitro studies have shown that rs12203592 impacts on the function of an enhancer element regulating IRF4 expression and the induction of tyrosinase (TYR), a key enzyme in melanin synthesis41. In addition to IRF4, hair colour (but not hair greying) shows genome-wide significant association to four other well-established pigmentation gene regions (SLC4A5A, TYR, OCA2/HERC2 and SLC24A5; Table 1 and Supplementary Fig. 8)13. Finally, for balding, we replicate association to the well-established Androgen Receptor/Ectodysplasin A2 Receptor (AR/EDA2R) locus on 2q12 (ref. 12; Supplementary Fig. 8).

The other four genomic regions showing genome-wide association include no strong candidate genes with established roles in hair biology (Table 1 and Supplementary Fig. 8). Beard thickness is associated with SNPs in 7q31, 4q12 and 6q21 where the nearest genes are, respectively: forkhead box P2 (FOXP2), ligand of numb-protein X 1 (LNX1) and prolyl endopeptidase...
None of these genes have known functions specifically related to hair. Similarly, balding is associated with SNPs in the second intron of the Glutamate receptor delta-1 subunit gene (GRID1) on 10q22 but this gene has no documented role in hair biology.

PRSS53 and hair shape. Among the SNPs in regions showing genome-wide significant association, a functional role potentially underlying the observed association is most suggestive for rs11150606 encoding the R30Q substitution in PRSS53, associated with scalp hair shape. Bioinformatic analysis indicates that this amino-acid change could introduce a subtilisin/kexin-like proprotein convertase site in PRSS53 (Supplementary Fig. 9), and thus could have the potential of affecting processing of the enzyme. To probe into the role of PRSS53 in hair development, we performed immunohistochemistry of human scalp hair follicles during active hair growth (anagen phase; Fig. 4c–j). Expression of PRSS53 was mainly detected in the developing IRS and pre-cortex of the hair fibre and in some...
bulbar melanocytes (Fig. 4c). PRSS53 expression was increased in subpopulations of IRS keratinocytes corresponding to early and late stages of hair fibre keratinization and IRS cornification (Fig. 4d,e). This pattern of expression is consistent with the hair-shape association we observe, as the cornifying IRS is thought to impart significant hair-shaping influence on the hair fibre. Expression of PRSS53 appears to be modulated as a function of hair fibre differentiation, as evidenced by high PRSS53 expression at the level of the hair follicle where the hair fibre undergoes dissolution of nuclear DNA (Fig. 4e,f). Double immunofluorescence for PRSS53 and TCHH revealed co-expression of these proteins in specific elements of the IRS (Fig. 4g–j), suggesting that PRSS53 may also be associated with the maturation of this hair follicle sheath. The only other component of the hair follicle that expresses TCHH is the medulla of the hair fibre and this was also found to express PRSS53 (Fig. 4i). The medulla is a non-obligate component of the hair fibre, but when present it impacts on the mechanical properties of hair. It is not yet known how the ‘slippage’ of the inner versus outer hair follicle components occurs during outward hair fibre growth, but it is likely that protease activity is involved. Expression of PRSS53 changes in the upper IRS (Fig. 4f), at the level of the sebaceous gland, where the IRS undergoes a controlled dissolution, required for exiting of the hair fibre to the skin surface.

To evaluate the cellular impact of the Q30R substitution in PRSS53, we expressed both forms of the enzyme in 293-EBNA cells, a human cell line showing pro-protein convertase activity. Western blot analysis of transfected cell extracts revealed a different signal peptide processing in the two forms of the enzyme and that PRSS53 Q30R has a slightly faster electrophoretic mobility, consistent with an extra proteolytic cleavage (Fig. 4a and Supplementary Fig. 9). These differences were not detected when the cell cultures were incubated with an inhibitor of pro-protein convertases (decanoyl-RVKR-CMK, DECA). Consistent with an altered processing of the enzyme affecting its secretion, the PRSS53 Q30R variant was less abundant in the cell culture media, except when cells were incubated in the presence of DECA (Fig. 5b).

Figure 3 | Association plots for six regions with SNPs showing genome-wide significant association to hair traits. (a) 2q12 hair shape, (b) 2q12 beard thickness, (c) 16p11 hair shape, (d) 3q22 eyebrow thickness, (e) 2q36 monobrow, (f) 6q25 hair greying. The index SNP in each region (Table 1) is shown as a purple diamond. At the top of the figure are shown the association results (on a -log10 P scale; left y axis) for all genotyped and imputed SNPs. The dot colour indicates the strength of LD (r2) between the index SNP and each SNP (based on the 1000genomes AMR data set). Recombination rate across the region, in the AMR data, is shown as a continuous blue line (scale on the right y axis). Genes in the region are shown in the middle. These plots were produced using LocusZoom. Below each LocusZoom plot we show an LD heatmap (using r2; red indicating r2 = 1 and white indicating r2 = 0) produced using Haploview. Coordinates used are from human genome sequence build 37. Plots for regions not shown here are presented in Supplementary Fig. 8.
indicates a greater accumulation of this variant in the ER-transGolgi network (Fig. 5c). Altogether, these in-vitro analyses confirm that the Q30R substitution in PRSS53 can affect processing and secretion of the enzyme.

**Signatures of selection at associated gene regions.** Some of the strongest signals of selection in the human genome detected in recent genome-wide searches involve pigmentation genes in Europeans and EDAR in East Asians. To assess whether selection could have contributed broadly to shape variation at gene regions impacting on human hair features, we examined whether there is enrichment for signatures of selection at the regions showing evidence of association with hair traits in the CANDELA sample. For this purpose, we used the Composite of Multiple Signals (CMS) statistic calculated in the three main reference East Asian, European and African populations from the 1000 Genomes Project data (ASN, CEU and YRI, respectively). We contrasted the distribution of CMS scores across the genome. We found significantly higher CMS scores in the hair-associated gene regions compared with...
Figure 5 | Processing of PRSS53 and signals of selection in the PRSS53 gene region. Comparison of PRSS53 and PRSS53 (Q30R) from cell extracts (a) and media (b), after expression in 293-EBNA cells cultured in the absence (−) or presence (+) of DECA (decanoyl-RVKR-CMK, a pro-protein convertase inhibitor).

(a) The top two bands seen in lanes loaded with PRSS53 cell extracts, result from partial processing of the signal peptide. In the absence of DECA, there appears to be an accumulation of PRSS53 (Q30R) without signal peptide. PRSS53 (Q30R) also has a slightly faster mobility compared with PRSS53 (a difference of eight amino acids based on the location of the pro-protein convertase site Supplementary Fig. 9). In the presence of DECA, both forms of the enzyme appear identically processed and there is no difference in mobility of the proteins. (b) The medium from cell cultures grown in the absence of DECA shows a reduced amount of PRSS53 (Q30R), compared with PRSS53. In the presence of DECA, the amount of protein in the media is similar for the two forms of the enzyme. Recombinant proteins were detected using an anti-FLAG antibody after migration on 13% SDS–polyacrylamide gel electrophoresis gels. Molecular weight markers (kDa) are indicated on the left. Beta-actin was used as a loading control (C = cells transfected with an empty vector). Full immunoblots for a and b are shown in Supplementary Fig. 11. (c) Immunostaining of 293-EBNA cells expressing PRSS53 and PRSS53 (Q30R). The enzyme was detected using an anti-FLAG antibody and the endoplasmic reticulum (ER) stained with an anti-calreticulin antibody. There is more abundant co-localization of the enzyme with the ER (white arrows) for PRSS53 (Q30R) compared with PRSS53, consistent with the intracellular accumulation of PRSS53 (Q30R). Negative controls, lacking primary antibodies for FLAG and calreticulin, are also shown. The scale bar indicates 10 μm. Magnification was the same for all photographs. (d) The top panel shows CMS scores (red dots) in the 1000 genomes ASN (JPT + CHB) data for SNPs in the 16p11 region associated with hair shape. The dotted line indicates the empirical significance threshold (1%). The SNP with the highest CMS score is rs11150606, coding for the Q30R substitution in PRSS53 (highlighted in purple) associated with hair curliness (Table 1 and Fig. 3c). Genes in the region are shown in the panel below (introns in magenta, exons in green). Coordinates are from human genome sequence Build 37.
subject of recent selection in humans 2,3. These observations, and the fact that SNPs with empirically significant CMS scores (that is, SNPs in the top 1% of the distribution; one-sided Mann–Whitney U-test P values of $5 \times 10^{-5}$, $1 \times 10^{-4}$ and $4 \times 10^{-4}$, in ASN, CEU and YRI, respectively). Noticeably, significant CMS scores were observed in ASN for SNPs in the PRSS53 region associated with hair shape, the highest CMS score (12.96, empirical P value $3 \times 10^{-4}$) being observed for SNP rs11150606, encoding the Q30R substitution in PRSS53 (Fig. 5d). In the EDAR region, we observe the previously documented strong signal of selection in ASN for variants around SNP rs3827760 (refs 27,46, associated with hair shape (Table 1). We also observe a significant signal of selection in CEU for the intrinsic EDAR variants associated here with beard thickness, independently of rs3827760, discussed above (Fig. 3b and Supplementary Fig. 7b).

Discussion

The analyses presented here have enabled us to expand substantially the set of gene regions known to impact on variation in human head hair appearance. This task has been facilitated by the extensive phenotypic and genetic diversity of the CANDELA sample, a result of Latin American history involving admixture between Africans, Europeans and Native Americans3,7. The predominant European and Native American ancestry of the CANDELA sample is expected to provide especially high power for the detection of genetic effects at loci with differentiated allele frequencies between those two continental populations. Consistently, allele frequencies at most index SNPs identified here show large differences between Europeans and East Asians/Americans (Supplementary Table 7). This strong differentiation in allele frequencies could relate partly to selection acting on the associated gene regions, as proposed for the evolution of human hair appearance25,31. The enrichment we observe for significant CMS scores at gene regions associated with hair features is consistent with this scenario. Interestingly, the pattern of variation at PRSS53 is similar to that seen for EDAR, with significant CMS scores in East Asians and a functional amino acid changing variant reaching high frequency only in East Asia (Supplementary Table 7). These observations, and the fact that other genes are associated with straight hair in Europeans, are in line with the proposal that scalp hair shape has been the subject of recent selection in humans25.

There is increasing interest in elucidating the mechanisms influencing the shape of the growing hair fibre. Based on mouse studies it has been proposed that EDAR signalling is involved in determining hair shape through regulating expression of the key hair growth signal Shh (Sonic hedgehog)32. Higher Edar function increases Shh expression and causes it to become symmetrically expressed in the hair bulb, leading to straight hair growth presumably through promotion of symmetric cell proliferation36. Structurally, a key player in shaping hair is TCHH, involved in crosslinking the cornified envelope and keratins of IRS cells3. TCHH is one of the earliest differentiation proteins of the growing (anagen) hair follicle bulb, and is enriched in the IRS (representing ~30% of its protein content). TCHH confers significant mechanical strength to the IRS via its enrichment in arginines and glutamines, which represent over 40% of TCHH amino acids. Many of the arginine residues undergo citrullination/deimination, whereas the glutamines are involved in intra- and interprotein chain crosslinks during the cornification/hardening of the IRS3. The progressive hardening of the IRS is thus thought to contribute to the moulding of the still-pliable hair fibre during hair growth. It is interesting to note that the TCHH SNP with smallest P value observed here (rs11803731) has also been strongly implicated by previous analyses in Europeans38,48, suggesting that this variant might be directly affecting hair shape. The M790L substitution in TCHH encoded by rs11803731 is not predicted to result in major structural alterations of TCHH, but it has been proposed that it could affect the post-translational processing of TCHH, with regulatory implications38. It will be important to assess whether TCHH is a substrate for PRSS53, given our observation that this enzyme is enriched in all layers of the IRS during the cornification/hardening stages of the hair fibre, as well as in the medulla (a component influencing the physico-mechanical characteristics of the hair fibre). Proteases and protease inhibitors are important for epidermal keratinization and can regulate hair growth and cycling28. For example, optimal desquamation of the IRS (and overall hair growth cycling) is perturbed in lysosomal cysteine protease cathepsin L knock-out mice49. Moreover, given the similarity of PRSS53 with the kallikreins (the largest family of secreted serine protease endopeptidases) it is possible that PRSS53 lyases substrates in cornified tissues that ultimately desquamates50. Interestingly, it has been suggested that GATA3 (in the 10p14 region associated with hair shape) inhibits the expression of serine protease inhibitor Kazal type-5 (SPINK5), mutations of which cause Netherton syndrome, an autosomal recessive congenital ichthyosis characterized by so-called Bamboo hair, epidermal hyperplasia and an impaired epidermal barrier function51.

The colour of hair results from melanin pigments transferred to hair fibre keratinocytes from hair follicle melanocytes. These melanocytes differentiate from melanoblasts that migrate from the neural crest into hair follicles early in development6. Some hair follicle melanoblasts remain undifferentiated and serve as stem cells for the periodic replenishment of mature melanocytes, melanogenesis occurring only in the anagen phase of the hair growth cycle. Among the several hundred gene products known to participate in melanogenesis, recent association studies have identified a handful that influence hair colour variation in Europeans13. Most of these associations have been replicated here (Table 1), including that of the derived T allele at SNP rs12203592 in IRF4 with lighter hair colour. It has been shown experimentally that IRF4 interacts with the microphthalmia-associated transcription factor (MITF, a key regulator of the expression of many pigment enzymes and differentiation factors), to activate the expression of TYR (a rate-limiting essential enzyme in melanin synthesis). The derived T allele at rs12203592 leads to reduced TYR expression and melanin synthesis, consistent with the association of this allele with lighter hair colour41. In line with the geographic distribution of light hair colour, the T allele at rs12203592 is essentially absent outside Europe (Supplementary Table 7). Interestingly, we find that the T allele at SNP rs12203592 is also associated with increased hair greying. Experimental evidence suggests that the mechanism of hair greying involves incomplete maintenance of melanocyte stem cells in the hair follicle6. Importantly, MITF is known to affect melanocyte survival via its regulation of anti-apoptotic Bcl2 expression, a key factor in protection of the hair follicle against oxidative stress5. To probe the mechanism by which IRF4 might impact on hair greying, it will therefore be important to evaluate whether the T allele at rs12203592 influences MITF in terms of melanoblast stem cell maintenance and survival or via melanocyte loss post differentiation.

The skin on different parts of the body has different hair characteristics, with the final hair distribution dependent upon the spacing pattern laid down during development, the extent of skin growth that occurs subsequent to pattern establishment, and to hormonal and ageing effects. The development of skin at
different body sites is known to be controlled by an underlying transcription factor code to which FOXL2, FOXP2 and PAX3 represent a step in that direction.

Methods

Study subjects. A total of 6,630 volunteers from five Latin American countries (Brazil, Chile, Colombia, Mexico and Peru), part of the CANDELA consortium sample (http://www.ucl.ac.uk/silva/candela) were included in this study (Supplementary Table 1). Ethics approval was obtained from: the Universidad Nacional Autónoma de México (Mexico), the Universidad de Antioquia (Colombia), the Universidad Peruana Cayetano Heredia (Peru), the Universidad de Tarapacá (Chile), the Universidad Federal do Rio Grande do Sul (Brazil) and the University College London (UK). All participants provided written informed consent. Blood samples were collected by a certified phlebotomist and DNA extracted following standard laboratory procedures.

Hair phenotyping. Scalp hair features were recorded by physical examination of the volunteers. Natural hair colour was scored in four categories (1-red/darkred, 2-blonde, 3-dark blond/light brown or 4-brown/black). Greying was scored on a five-point scale: 1-for no greying, 2-for predominant non-greying, 3-for ~50% greying, 4-for predominant greying and 5-for totally white hair. Hair curliness was scored as 1-straight, 2-wavy, 3-curly or 4-frizzy. Baling was scored on a three-point scale (none, medium, high) in both women and men. Although frequency of balding in women is low, their inclusion more than doubles sample size, thus adding considerable power to the analyses (Supplementary Fig. 10). Facial hair traits were scored using photographs of the faces of the individuals. Beard density was scored in men using a three-point scale (low, medium or high), separately for shaven and unshaven individuals and scores for these two groups subsequently merged. As an interview of the volunteers indicated that most women do not have beard, eyebrows, monobrow and eyebrow thickness were only scored in men. Both these traits were scored on three-point scales: eyebrow thickness as low, medium or high, and monobrow as none, medium or high. The frequency distribution of the traits in the CANDELA sample analysed here is shown in Supplementary Fig. 1.

DNA genotyping and quality control. DNA samples from participants were genotyped on the Illumina HumanOmniExpress chip containing 730,558 SNPs. PLINK v1.9 (ref. 57) was used to exclude SNPs and individuals with more than 5% missing data, markers with minor allele frequency <1%, related individuals and those who failed the X-chromosome sex concordance check (sex estimated from X-chromosome heterozygosity not matching recorded sex information). After quality control filters, 669,462 individuals and 6,357 individuals (3,730 males and 3,623 females) were retained for further analysis. Because of the admixed nature of the study sample (Supplementary Fig. 2), there is an inflation in Hardy–Weinberg P values. We therefore did not exclude markers based on Hardy–Weinberg deviation.

Statistical genetic analyses. P values for Pearson correlation coefficients were obtained by permutation. Narrow-sense heritability (computed as the additive genetic variance explained by a kinship matrix computed from the chip genotypes) was estimated using GCTA (as fitted by an additive linear model with a random effect term whose variance is given by the kinship matrix, with age and sex as covariates. The kinship matrix was obtained using the LDAP approach, which accounts for LD between SNPs, African, European and Native American ancestry was estimated from a set of 9,728 autosomal SNPs (LD-pruned and full chip data) via supervised runs of ADMIXTURE. Reference putative parental population data included in the ADMIXTURE analyses for Africans and Europeans were chosen from HAPMAP and for Native Americans from selected Amerindian populations as described in Ruiz-Linares et al. The chip genotype data were phased using SHAPEIT2 (ref. 59). IMPUTE2 (ref. 60) was then used to impute genotypes at untyped SNPs using variant positions from the 1000 Genomes Phase I data. The 1000 Genomes data reference set haplotype information for 1,092 individuals for 36,820,992 variant positions. Positions that are monomorphic in 1000 Genomes Latin American populations were removed from the 1000 Genome Latin American Panel (CLM, MM, PUR) and IMPUTE2 on-examples were excluded, leaving 2,053,002 SNPs being imputed in our data set. Of these, 22,737 SNPs had imputation quality scores <0.3 and were excluded. The IMPUTE2 genotype probabilities at each locus were converted into best-guess genotypes using PLINK. SNPs with uncalled genotypes in >5% of samples or minor allele frequency <1% were excluded. The final imputed data set used in the GWAS analyses included genotypes for 9,143,600 SNPs.

PLINK 1.9 (ref. 57) was used to perform genome-wide association tests for each phenotype using multiple linear regression with an additive genetic model incorporating age, sex and five genetic PCs (principal components) as covariates. Results for these tests were obtained using the LD-pruned genotype data using PLINK 1.9. These PCs were selected by inspecting the proportion of variance explained and checking scatter and screen plots (Supplementary Fig. 3A). Individual outliers were removed and PCs recalculated after each removal. Using these PCs, the Q–Q plots for all association tests showed no sign of inflation, the genomic control factor λ being in all cases <1, indicating no departures from HWE. We previously showed that using five PCs in GWAS of the CANDELA sample completely removes the inflation, which is observed when PC adjustment is not used, and increasing the number of PCs included in the regression from five to ten does not provide additional gain. Association analysis on the imputed data set were performed using the best-guess imputed genotypes in PLINK and using the IMPUTE2 genotype probabilities obtained in SNPTEST v2.5 (ref. 62). Association results from both approaches were consistent with each other and with the results from the chip genotype data. For analysis of the X chromosome, an inactivation model was used (male genotypes encoded as 0/2 and female genotypes as 0/1). The main effect and the effect size heterogeneity across country samples with significant heterogeneity, a random effects model was used for meta-analysis. A prediction model was constructed for each hair trait examined, including the associated index SNPs (Table 1) as fixed effects and a random effect term obtained via BLUP (Best Linear Unbiased Predictor) using the genome-wide kinship matrix obtained from the chip data. BLUP components is thus sensitive to the effect of SNPs below the threshold for genome-wide significance. Prediction results were obtained through tenfold cross-validation and the set of samples were randomly split.
into ten chunks of 10% and used as the test data set, whereas the remaining 90% was used as training data set to fit a prediction model. For each run, R² estimates were calculated (conditional on the covariates) to obtain the proportion of phenotypic variance explained by the model. Average prediction scores across the tenfold runs were calculated. Genetic PCs, age and sex (except for facial hair traits which were only male data sets used as covariates).

To evaluate an enrichment of signal at gene regions associated with hair traits, we examined the CMS scores of selection calculated for the three main 1000 Genomes Project populations46: ASN (IPT + CHB), CEU and YRI. We obtained empirical significance cutoffs (1%) separately for each population based on CMS scores for 3,071,032, 3,179,944 and 3,312,050 SNPs (in ASN, CEU and YRI, respectively). Each population contained the regions of CMS scores for SNPs in a ±2 kb region around each gene, based on the UCSC ReSeq annotation (excluding gene regions with less than four SNPs). We obtained the distribution of mean CMS scores for genes including SNPs associated with the hair traits examined here and compared it with the distribution for all other genes in the genome (e.g. 23, 20,093 and 20,366 genes in ASN, CEU and YRI, respectively). We used a suggestive significance level cutoff for inclusion in the set of hair-associated genes (that is, SNP association P value < 10⁻³; Table 1 and Supplementary Table 6), resulting in 53, 53 and 55 gene regions being included for ASN, CEU and YRI, respectively. If associated SNPs were in intergenic regions, we included the gene most closely located to this SNP. We contrasted the distribution of mean CMS scores at these gene regions with the distribution for all other gene regions in the genome using a one-sided Mann–Whitney U-test.

Assessment of mouse chin hair follicle density. We examined chin hair placode (follicle primordium) density in E14.5 day mouse embryos. We focused on embryonic day 14.5 (E14.5) as at this embryonic stage the hair pattern is laid out, so that the sites of future hair follicle growth are readily quantifiable by detecting focal expression of the marker gene Sostdc1 (ref. 65). We performed in situ hybridization staining embryos for Sostdc1 to visualize the placodes present on the lower jaw. At this stage, placodes are visible as rings or filled rings of Sostdc1 expression. Embryos were positioned to view the lower jaw, imaged from a frontal view and placodes were counted within a rectangular area on the lower jaw. Placode density for each embryo was determined using Image-Pro software (Media Cybernetics).

Immunohistochemistry of PRSS53. Unshaven, full-thickness human adult scalp with terminal hair was used snap frozen in liquid nitrogen in cubes of 2 cm³. Cryosections of 6–8 µm were cut using a cryostat onto adhesive glass slides and incubated with anti-PRSS53 antibody (NBPI, #90678 from Novus) and anti-trichohyalin antibody (A15, # I0373 from Immusquest) using standard double immunofluorescence protocols. In vitro transfection with a FLAG epitope was used at the same concentration as the smallest primary antibody dilution. Co-localization and co-localization of both antigens in the hair follicle were determined by merging of the PRSS53- and trichohyalin-positive channels.

In vitro analysis of PRSS53. Mutagenesis. To introduce the Q30R substitution in the PRSS53 sequence, the QuickChange II XL Site-Directed Mutagenesis Kit (Agilent) was employed following the instructions by the manufacturer. Primers designed for mutagenesis were: poli3MUT-FOR 5’-AGGCAACGCTG-3’ and poli3MUT-REV 5’-GGCGGGCCAGGCGGCACCGTG-3’, and poli3MUT-REV 5’-GGCGGGCCAGGCGGCACCGTG-3’, and we employ as template a previously published plasmid (plasmid (empty vector) was used as a negative control for all these assays. Ponceau staining was used as training data set to fit a prediction model. For each run, R² estimates were calculated (conditional on the covariates) to obtain the proportion of phenotypic variance explained by the model. Average prediction scores across the tenfold runs were calculated. Genetic PCs, age and sex (except for facial hair traits which were only male data sets used as covariates).
41. Matsui, A., Gomes, A. C. & Cavaco-Paulo, A. Hair coloration by gene

61. Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).

Acknowledgements

We are grateful to the volunteers for their enthusiastic support for this research. We thank Alvaro Alvarado, Monica Ballestero Romero, Ricardo Cebreiros, Miguel Angel Cordero Siqueiro, Francisco de Avila Becerril, Joyce De la Piedra, Maria Teresa Del Solar, Paola Everardo Martinez, William Flores, Martha Granados Riveros, Ilich Jafet Moreno, Jodie Lampert, Paola Leon-Mimila, Francisco Queispealya, Diana Rogel Diaz, Ruth Rojas, Norman Russell, Vanessa Sarabia, Rosilene Palmit, Ricardo Gunski, Sergeant Joao Felisberto Menezes Cavalheiro and Mayor Eugenio Correa de Souza Junior for assistance with volunteer recruitment, sample processing and data entry. We also thank Richard Baker for technical assistance with the human skin immunohistochemistry, Pardis Sabeti and Matteo Furnagali for advice on the analysis of CMS scores, Elfride De Baere for information on clinical features of BPES cases, Doug Speed for advice on the prediction analysis, Barbara Kremeyer for comments on the manuscript and Emiliano Bellini for the face illustrations in Fig. 1. The following institutions kindly provided facilities for the assessment of volunteers: Escuela Nacional de Antropología e Historia and Universidad Nacional Autónoma de México (Mexico); Pontificia Universidad Católica del Perú, Universidad de Lima and Universidad Nacional Mayor de San Marcos (Perú); Universidad Federal do Rio Grande do Sul (Brazil); 13º Companhia de Comunicações Mecanizadas do Exército Brasileiro (Brazil). This work was funded by grants from: the Leverhulme Trust (F/07134/D;F to ARL), BBSRC (BB/B021213/1 to ARL and Institute Strategic Programme grant to The Roslin Institute); Universidad de Antioquia, Colombia (CODI sostenibilidad de grupos 2013–2014 and MASO 2013–2014); Ministerio de Economía y Competitividad and Instituto de Salud Carlos III (RTICC). Spain. C.L.O. is an Investigator of the Botin Foundation supported by the Banco Santander through its Santander Universities Global Division.

Author contributions

K.A., M.Q.-S., R.G.-J., D.H., D.B., A.R.-L conceived and designed the study. K.A., M.Q.-S., R.G.-J., D.H., D.B., A.R.-L. provided the resources. K.A., M.Q.-S., R.G.-J., D.B., A.R.-L. performed the experiments. K.A., M.Q.-S., R.G.-J., D.B., A.R.-L. analyzed the data. V.A.-A., J.A.J., M.Q.-S., J.M.R., R.G.-J., D.H., D.B., A.R-L. contributed reagents/ materials. K.A., M.Q.-S., R.G.-J., D.B., A.R.-L. wrote the manuscript. Critical revision of the manuscript was done by C.G., G.P., S.C.-Q., R.G.-J., D.H., D.T.J. and D.B. and A.R.-L. supervised the research (PJ). K.A. and A.R.-L. wrote the manuscript, incorporating input from other authors.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/ reprintsandpermissions/

How to cite this article: Adhikari, K. et al. A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat. Commun. 7:10815 doi: 10.1038/ncomms10815 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/