Genome wide analysis for mouth ulcers identifies associations at immune regulatory loci

Tom Dudding¹,², Simon Haworth¹,², Penelope A. Lind³, J. Fah Sathirapongsasuti⁴, the 23andMe Research Team⁵, Joyce Y. Tung⁴, Ruth Mitchell¹, Lucía Colodro-Conde³, Sarah E. Medland³, Scott Gordon⁵, Benjamin Elsworth¹, Lavinia Paternoster¹, Paul W. Franks⁶,⁷,⁸, Steven J. Thomas², Nicholas G. Martin⁵ & Nicholas J. Timpson¹

Mouth ulcers are the most common ulcerative condition and encompass several clinical diagnoses, including recurrent aphthous stomatitis (RAS). Despite previous evidence for heritability, it is not clear which specific genetic loci are implicated in RAS. In this genome-wide association study (n = 461,106) heritability is estimated at 8.2% (95% CI: 6.4%, 9.9%). This study finds 97 variants which alter the odds of developing non-specific mouth ulcers and replicate these in an independent cohort (n = 355,744) (lead variant after meta-analysis: rs76830965, near IL12A, OR 0.72 (95% CI: 0.71, 0.73); P = 4.4e−483). Additional effect estimates from three independent cohorts with more specific phenotyping and specific study characteristics support many of these findings. In silico functional analyses provide evidence for a role of T cell regulation in the aetiology of mouth ulcers. These results provide novel insight into the pathogenesis of a common, important condition.
Oral ulceration is the most common ulcerative condition in humans, affecting up to 25% of young adults and a higher proportion of children. Depending on context, ulcers in the mouth are described as mouth ulcers or canker sores, both of which are descriptive terms rather than clinical diagnoses. In this article, mouth ulcers is used as an umbrella term describing the spectrum of clinical presentation.

To date, many causes of mouth ulcers are recognised including mucosal trauma and a range of autoimmune and inflammatory conditions. For example, mouth ulcers are common in patients with ulcerative colitis and Crohn’s disease, systemic lupus erythematosus, and are considered a diagnostic feature of Behçet’s disease, an inflammatory disorder of blood vessels that causes ulceration of the mouth, eyes and genitals.

Many people experience recurrent mouth ulcers which cannot be attributed to systemic disease or obvious oral trauma. Here, the clinical diagnosis of recurrent aphthous stomatitis (RAS) is used, referring to a group of closely related conditions of uncertain aetiology, whose defining feature is the presence of clinically characteristic oral ulcers, which are painful and associated with impaired quality of life. First line management strategies for RAS reduce the severity of ulceration without preventing recurrence. Second line management strategies involve topical or systemic use of potent and non-specific immunomodulatory drugs including thalidomide and dapsone, exposing patients to a range of side effects. Thus, there is unmet need for a wider range of therapeutic options in the management of RAS.

In part, the lack of satisfactory management for RAS reflects uncertainty in the exact aetiology. A number of predisposing factors have been reported including vitamin or haematinic deficiency, chemicals such as sodium laurel sulphate, mechanical trauma, stress and anxiety or infection with bacteria or viruses. Regardless of the initial trigger, it is believed that immune regulation plays a pivotal role in mediating tissue damage and the clinical presentation of RAS. Susceptible individuals experience focal infiltration of the oral mucosa by monocytes and T lymphocytes deep to the basal membrane, followed by loss of superficial mucosa and a proliferative healing phase. Family-based studies support a role of genetic susceptibility in the aetiology of RAS, yet to date the genetic basis of this susceptibility remains poorly characterised. Previous candidate gene association studies have investigated variation in the region of genes encoding key cytokines (TNF-α, IL-1α, IL-1β, IL-6, IL-10, IL-12), with varied results. Genome-wide association studies (GWAS) have the potential to identify genetic variants associated with both susceptibility to initial triggers and the immune reaction that leads to the tissue damage and ulcer formation.

A recent study used genome-wide data to look for associations between gene pathways and specific sub-phenotypic features (including mouth ulcers) in a case series of patients with systemic lupus erythematosus. The study found some evidence for an association between the vascular endothelial growth factor (VEGF) pathway and the oral ulcer sub-phenotype. It is not clear how these findings relate to the general population and there are no previous conventional genome-wide association studies for RAS or mouth ulcers.

There is therefore a need to undertake genome-wide analysis for RAS, but specific measures are unavailable in large cohorts. The most appropriate strategy given available data, is to use large collections to identify and replicate associated genetic variants in a well-powered GWAS of self-reported non-specific mouth ulcers and then validate the effects of these variants in smaller collections with more clinically relevant RAS-specific measures. Given that RAS is nested within mouth ulcers, the inclusion of other causes of ulcers would introduce, at worst, noise into the analysis and, at best, enhance it by highlighting mechanisms which are relevant to oral mucosal breakdown irrespective of trigger.

This genome-wide association study identifies 97 variants which alter the odds of developing non-specific mouth ulcers and replicates them in an independent cohort. In silico functional analyses provide evidence for a role of T cell regulation in the aetiology of mouth ulcers. These results provide novel insight into the pathogenesis of a common, important condition.

**Results**

**Contributing studies.** In UK Biobank (UKBB) participants with data on mouth ulcers, the mean age at questionnaire completion was 56.7 years (range = 38.0, 73.0), 54.2% of participants were female and 10.2% of participants reported having mouth ulcers within the last year. In research participants from the personal genetics company 23andMe, Inc. with data on canker sores, 67.8% were over 45 years old, 59.2% of participants were female and 72.4% of participants reported ever having canker sores. In the QIMR Berghofer Medical Research Institute (QIMR) Over 50s (Aged) study, which examined ageing and age-related disease in twin pairs from the Australian Twin Registry, the mean age was 61.2 years (range = 50.2, 85.6). A large proportion of participants were female (74.1%) and the proportion reporting a history of mouth ulcers was 18.7%. Two other studies exclusively or partially included younger participants, the Avon Longitudinal Study of Parents and Children (ALSPAC), a UK population-based birth cohort (mean age = 23.9 years, range = 22.8, 25.3; 65.2% females), and QIMR Melanocytic Naevi in Adolescent Twins (TW) study, which primarily examined melanotic naevas in twin pairs (mean age 24.2 years, range = 10.1, 62.3 (a combination of adolescents and their parents); 53.8% females). In these studies, the proportion reporting ulcers was much higher (ALSPAC: ulcer cases = 74%, TW: percentage reporting having ulcers at least rarely = 86.8%) (Table 1).

**Genome-wide discovery analysis for mouth ulcers.** The primary genome-wide analysis was undertaken in UK Biobank. At an aggregate, genome-wide level there was evidence for a genetic contribution to mouth ulcers, with heritability estimated at 8.2% (95% CI: 6.4%, 9.9%) under an infinitesimal model implemented in linkage disequilibrium score regression (LDSR). Under a non-infinitesimal model implemented in Heritability Estimator from Summary Statistics (HIES), heritability was estimated at 8.7% (95% CI: 8.2%, 9.2%). Genomic inflation factor (lambda GC) was estimated at 1.20. The intercept term from univariate LDSR was 1.03 (95% CI: 1.01, 1.05) suggesting that most inflation in lambda GC was attributable to polygenicity rather than bias.

After final quality control (QC), 9,851,866 genetic variants were included in GWAS. Evidence for genome-wide association ($P < 5e^{-8}$) with mouth ulcers was seen at 7127 single variants (Fig. 1, Supplementary Figure 1). After clumping, using a LD threshold ($r^2 = 0.1$) in PLINK, these formed 97 approximately independent lead variants.

**Signal validation.** All 97 of these independent variants showed directional consistency in 23andMe, with comparable effect sizes in both collections (Supplementary Figure 2). Summary statistics for the 10 strongest associated variants after meta-analysis of UK Biobank and 23andMe are shown in Table 2. All 97 variants are shown in Supplementary Data 1.

**Assessment of effect sizes in additional populations.** Three smaller samples were used to assess the effect sizes of the 97 lead
Table 1 Demographics of samples included in analysis

| Study (short name) | N (genotype and phenotype data) | Sex (female (%)) | Severity level n (%) | Mean age (range) | Proportion female (%) |
|--------------------|---------------------------------|------------------|----------------------|-----------------|-----------------------|
| UKBB               | 461,106                         | 47,079 (10.2)    | Never 51.1 [36.0, 66.0] | 56.7 [38.0, 73.0] | 54.2 |
| 23andMe            | 355,744                         | 98,298 (72.4)    | Rarely 23.9 [22.8, 25.3] | 51.1 [36.0, 66.0] | 59.2 |
| ALSPAC             | 2976                            | 2201 (74.0)      | Sometimes 61.2 [50.2, 85.6] | 23.9 [22.8, 25.3] | 65.2 |
| AG                 | 1120                            | 209 (18.7)       | Frequently 24.2 [10.1, 62.3] | 23.9 [22.8, 25.3] | 74.1 |
| TW                 | 2442                            | 115 (13.2)       |                      | 24.2 [10.1, 62.3] | 74.1 |

UKBB UK Biobank, ALSPAC Avon Longitudinal Study of Parents and Children, AG QIMR Berghofer Medical Research Institute’s (Aged) study, TW QIMR Berghofer Medical Research Institute Melanocytic Naevi in Adolescent Twins study

*For 23andMe the interquartile range is given instead of the range

variants with more RAS-specific phenotypes. In the AG and TW studies, clinical photographs were used to help participants identify RAS as opposed to traumatic ulcers. In the TW study, measures of ulcer severity were collected and the adolescents within the study (1572 of 2442 participants) were near the age of highest RAS prevalence. In the ALSPAC study all participants were in this high RAS prevalence age range. Twenty-four of the 97 lead variants showed consistent effect direction across all phenotypes from the three independent collections (Supplementary Data 1).

Description of lead novel associations. The strongest evidence for association, after meta-analysis of UK Biobank and 23andMe, was seen at rs76830965, a common variant lying within ILAS1-AS1 ~69 kb 5’ of IL12A on chromosome 3. This variant conferred large effects on the odds of mouth ulcers (odds ratio (OR) 0.72 per C allele, 95% CI: 0.71, 0.73; effect allele frequency (EAF) 0.89; χ² test P = 4.4e−483). Complementary evidence for a protective effect of the C allele was seen in all three lookup cohorts (ALSPAC: OR 0.67, 95% CI: 0.56, 0.79; P = 4.0e−06, AG: OR 0.75, 95% CI: 0.53, 1.07; P = 0.11 and TW: OR 0.86, 95% CI: 0.77, 0.96; P = 7.8e−03) (Table 2, Fig. 2a). After clumping, other variants in the same 3q25 locus as rs76830965 showed very strong evidence for association and showed consistent effects in all but the AG cohort. For example, rs7645203 2.0 kb 5’ of IL12A (OR 1.08 per C allele, 95% CI: 1.08, 1.09; EAF 0.90; P = 9.6e−80), rs11928736 within SCHIP1 (OR 1.07 per G allele, 95% CI: 1.06, 1.08; EAF 0.56; P = 2.62e−60), and rs55667203 within RPI1-43286.3 (OR 0.91 per C allele, 95% CI: 0.90, 0.92; EAF 0.84; P = 1.5e−64) (Table 2).

rs1800871, a variant within 1 kb 5’ of IL10, showed the second strongest evidence for association after meta-analysis, conferring

Table 2 Ten lead single variant results from GWAS in UK Biobank, replication in 23andMe and after Meta-analysis

| Variant   | Chr | Position | Effect allele | Other allele | UK Biobank (N = 461,106) | 23andMe (N = 355,744) | Meta-analysis (N = 816,850) |
|-----------|-----|----------|---------------|--------------|--------------------------|-----------------------|-----------------------------|
| rs1800871 | 1   | 206946634| A             | G            | EAF 0.92 (95% CI: 0.89, 0.93) | 0.94 (0.90, 0.95)     | 0.90 (0.89, 0.92)           |
| rs4683205 | 3   | 46334670 | G             | A            | 0.87 (1.05, 1.08)          | 0.83 (1.04, 1.08)     | 0.80 (1.02, 1.06)           |
| rs34390431| 3   | 46464940 | C             | T            | 0.81 (1.03, 1.06)          | 0.80 (1.04, 1.07)     | 0.78 (1.02, 1.06)           |
| rs76830965| 6   | 137540370| A             | G            | 0.88 (1.07, 1.09)          | 0.86 (1.07, 1.09)     | 0.84 (1.05, 1.09)           |
| rs7645203 | 3   | 15983069 | C             | T            | 0.81 (1.03, 1.06)          | 0.80 (1.05, 1.08)     | 0.78 (1.03, 1.07)           |
| rs55667203| 3   | 159950290| C             | T            | 0.81 (1.03, 1.06)          | 0.80 (1.05, 1.08)     | 0.78 (1.03, 1.07)           |
| rs7985203 | 6   | 137540370| A             | G            | 0.82 (1.05, 1.08)          | 0.80 (1.05, 1.08)     | 0.78 (1.03, 1.07)           |
| rs3764613 | 19  | 46896217 | A             | G            | 0.83 (1.07, 1.09)          | 0.81 (1.05, 1.08)     | 0.79 (1.03, 1.07)           |

All 97 variants reaching genome-wide significance given in Supplementary Data 1. Base pair positions are given with reference to build 37 of human genome reference consortium (human chromosome, CI confidence interval, EAF effect allele frequency in UK Biobank)
a large effect on mouth ulcers (OR 1.18 per A allele, 95% CI: 1.17, 1.19; EAF 0.22; P = 6.0e–236) (Table 2, Fig. 2b).

In the 3p21 locus, rs4683205 near CCR3 showed strong evidence for a moderate effect on mouth ulcers (OR 1.05; 95% CI: 1.03, 1.07; EAF 0.47; P = 4.9e–164) (Table 2, Fig. 2c). Additional variants in this region (e.g. rs493469 near CCR3 and rs34390431 near CCRL2) also showed strong evidence for an association with mouth ulcers after clumping (Supplementary Data 1).

Other associated common variants showed more modest effects on mouth ulcers. For example, rs7749390 an intronic variant in the IFNGR1 gene (Table 2, Fig. 2d) and rs3764613 within PPP5C (Table 2, Fig. 2e), showed very strong evidence for associations with mouth ulcers (OR 1.09, 95% CI: 1.06, 1.11; EAF 0.67; P = 2.7e–26). Additional variants in this region (rs9743469 near CCR3 and rs34390431 near CCRL2) also showed strong evidence for an association with mouth ulcers after clumping (Supplementary Data 1).

Within the HLA (chromosome 6p21) there was very strong evidence for association between rs2523589 (2.4 kb 5’ of HLA-B) and the odds of developing mouth ulcers (OR 0.94 per G allele, 95% CI: 0.89, 0.99; EAF 0.50; P = 7.4e–73, respectively).

Within the HLA (chromosome 6p21) there was very strong evidence for association between rs2523589 (2.4 kb 5’ of HLA-B) and the odds of developing mouth ulcers (OR 0.94 per G allele, 95% CI: 0.89, 0.99; EAF 0.50; P = 7.4e–73, respectively).

Additional haplotype analyses were undertaken to characterise this association, as described under sensitivity analyses, below.

**Gene prioritisation and enrichment analyses.** Gene prioritisation analysis in DEPICT was performed to nominate plausible biologically causal genes by identifying genes in different statistically associated loci (P < 5e–8 after clumping) that have similar predicted functions more often than expected by chance. This analysis suggested genes encoding chemokines or chemokine receptors were plausible candidates at many associated loci, for example suggesting IL12A at the lead associated chr3:15843176–160796695 locus and IL12B at the chr5:158741791–158757895 locus. In loci containing multiple genes with related functions, several potential candidates were identified; for example, association at chr3:45864808–46621589 produced a number of candidates including CCR1, CCR2, CCRL2, CCR3, and CCR5 among others (Supplementary Data 2).

DEPICT assesses whether any of 14,461 pre-computed gene sets are enriched for genes in the associated loci more than would be expected by a randomly distributed phenotype. This analysis identified enrichment in 895 sets with a false discovery rate < 0.01. The strongest statistical evidence for enrichment was seen for MP:0008560 (increased tumour necrosis factor secretion; Z-test P = 9.7e–12). Strong evidence for enrichment in several T cell regulatory gene sets was observed, for example, GO:0046632 (alpha-beta T cell differentiation; P = 2.0e–08), GO:0046634 (Cellular response to interleukin-1 beta; P = 2.4e–09).
(regulation of alpha-beta T cell activation; \( P = 2.70e-08 \), GO:200514 (regulation of CD4\(^+\), alpha-beta T cells; \( P = 2.88e-08 \), GO:0002286 (T cell activation involved in immune response; \( P = 3.75e-08 \), GO:0042098 (T cell proliferation; \( P = 3.87e-08 \) and GO:0045580 (regulation of T cell differentiation; \( P = 4.43e-08 \)) amongst others (Supplementary Data 3).

Furthermore, DEPICT assesses whether genes in associated loci are highly expressed in any of 209 tissue/cell type annotations. This revealed enrichment in 36 tissues with false discovery rate < 0.01, with the strongest evidence for enrichment in haemic and immune cell lines. The most robust single finding was evidence for enrichment in gene expression in leucocytes (\( P = 2.69e-10 \)) (Supplementary Data 4).

**Enrichment in regulatory motifs.** Non-parametric enrichment analysis in GARFIELD\(^{25} \) that accounts for LD, minor allele frequency, matched genotyping variants and local gene density, identified enrichment in genomic annotations and tissue-specific annotations that are present more than would be expected by chance. Single variants associated with mouth ulcers were enriched for five prime untranslated region variants by \( 18\)-fold compared to permuted matched controls, suggesting the variants identified in this study may predominantly regulate transcription rather than altering protein structure. Associated variants had \( 14\text{--}18\text{-fold} \) enrichment in DNase1 hypersensitive sites in a number of T cell lineages, including CD8\(^+\) primary cells, CD4\(^+\) primary cells, T helper (Th) 1 cells and Th2 cells. These results suggest active gene expression occurs near associated variants in a tissue-specific manner (Supplementary Figures 3--10).

**Imputed gene transcription levels.** Tests for association between ulcers and predicted gene expression were performed using S-PrediXcan\(^{26} \), which uses predictive models to impute transcript expression levels trained in 48 gene–tissue expression project (GTEx)\(^{27} \) tissues and then uses full GWAS summary statistics to test for associations between these predicted expression levels and phenotype. Results were then assimilated using the S-MultiXcn\(^{28} \) method, which integrates information from multiple tissue-specific predictions to improve statistical power. In total, 25,839 gene transcripts were tested for association, of which 244 transcripts passed a Bonferroni-corrected multiple testing threshold (\( P < 1.9e-06 \)). The strongest evidence for association with mouth ulcers across all tissues was at IL102A mirroring the single variant results, with an increase in expression predicted to increase the odds of mouth ulcers (\( Z\)-test \( P = 2.23e-103 \)). Other single variant results were mirrored in the results of this analysis with increased expression of SCHIPI increasing the odds of mouth ulcers and IL10 decreasing the odds of mouth ulcers (\( P = 8.99e-70 \) and \( 5.60e-55 \), respectively). The results for all genes are shown in Fig. 3.

**Tests for genetic correlation.** To assess whether the genetic variants contributing to the heritability of mouth ulcers also influenced other traits, genetic correlations were calculated against publicly available GWAS summary statistics\(^{29} \). Genetic correlation results were available for 222 traits, for which two passed a Bonferroni corrected \( P \)-value threshold of \( P < 2.3e-04 \). These were neuroticism (\( r_g = 0.23 \), \( P = 1.80e-08 \)) and depressive symptoms (\( r_g = 0.24 \), \( Z\)-test \( P = 5.73e-07 \) (Fig. 4). A full list of results is included in Supplementary Data 5. For neuroticism and depressive symptoms, the genetic correlation was further examined using the rho-HESS approach which estimates local genetic correlation between mouth ulcers and these traits. At an aggregate level (i.e. incorporating all common genetic variation in the genome), the \( r_g \) estimates from rho-HESS gave consistent interpretation with those from LDSR (neuroticism: \( r_g = 0.18 \), \( P = 8.43e-107 \); depressive symptoms: \( r_g = 0.33 \), \( P = 4.16e-21 \)). Additionally, it shows that genetic correlation between mouth ulcers and these traits is evenly distributed across the genome, without peaks in genetic correlation corresponding to peaks in local heritability of either mouth ulcers or these two traits (Supplementary Data 6 and 7).

**HLA haplotype analysis.** To characterise the association signal seen in single variant results near HLA-B, analysis of imputed haplotypes was performed within UK Biobank (\( n = 336,038 \)). This identified 24 haplotypes which were associated with mouth ulcers at a Bonferroni-corrected \( P \)-value threshold of 0.05. The most robust finding was DRB1*0103, an uncommon haplotype (frequency in controls = 0.017, frequency in cases = 0.022) which was associated with markedly increased odds of mouth ulcers in a fully adjusted logistic regression model (OR = 1.33, 95% CI: 1.26, 1.41; \( \chi^2 P = 2.03e-24 \)) (Supplementary Table 1, Supplementary Figure 11).

**Polygenic risk score (PRS) analysis.** A PRS approach was used to examine the ability of a series of PRS for mouth ulcers to predict phenotypic variance in the two QIMR samples (TW and AG, combined adults and adolescents). Using Genome-wide Complex Trait Analysis\(^{30} \) to control for genetic relatedness in linear mixed models (LMMs) in the predictions, up to 0.37% of variation in mouth ulcer severity (\( n = 2442 \)) and 0.86% in mouth ulcer case status (\( n = 3562 \)) could be accounted for using scores trained in UK Biobank. While scores trained only from loci meeting genome-wide significance in UKB accounted for 0.73% of variation in mouth ulcer severity and 0.21% in mouth ulcer case status. (Supplementary Tables 2 and 3).

**Drug repurposing.** To assess whether associated loci might represent targets for repurposed drug interventions we examined the Open Targets database for pharmacological interventions which might recapitulate the effects of naturally occurring genetic variation. Of the 244 gene transcripts that passed Bonferroni correction in S-PrediXcan, 27 were not recognised by the platform. As the platform limits the number of genes to 200, the 17 with the weakest evidence for association in S-PrediXcan were not included in the model. Fifty-two drugs were identified as potential targets (Supplementary Table 4). Fourteen of these are in phase IV trials including Ustekinumab, an antibody against the IL12 protein, encoded by the IL12A gene.

**Discussion** This large-scale genome-wide association study used a non-specific measure of mouth ulcers finding that, in common with studies of specific ulcer types, mouth ulcers are partly heritable\(^{2,16,17} \). Although the estimate of heritability from this study is likely an under-estimate and only provides the lower bound, it is substantially less than the heritability previously estimated in twin studies\(^{16} \). Some of this heritability is attributable to genetic variants with large effects on the odds of developing mouth ulcers, such as the lead variant identified here (rs76830965). Effect estimates in samples with RAS-specific phenotypes provide further evidence that these variants are associated with this specific type of oral ulceration. These variants with large effects are closely related to biological mechanisms thought to be relevant to the formation of mouth ulcers. The remainder of the heritability is likely driven by indirect effects of large numbers of variants with modest effects on the odds of developing mouth ulcers. Given these genetic variants, which act indirectly, likely also contribute to a wide range of distal...
health-related phenotypes, it is perhaps unsurprising that neu-roticism and depressive symptoms show detectable genetic cor-
relation with mouth ulcers in both the LDSR and rho-HESS
analyses. The view that this correlation is driven by non-speci-
fic associations across the genome, rather than effects at lead-
associated variants, is supported by the location specific genetic
correlation performed in rho-HESS.

The variants with larger effects on the odds of mouth ulcers are
likely to be clinically informative. Many are located in or near
genes or are enriched in pathways relating to T cell immunity,
and tend to impose a Th1-type immune response, a biologically
plausible mechanism that supports previous literature and reso-
nates with clinical practice. Complementary evidence of these
variants having true causal associations with RAS is provided by
the consistent effect directions seen for many of them in the three
independent look-up cohorts with more speci-
cfic phenotypes.

PRSs, generated using variants selected with a range of sig-
nificance thresholds, explained only a small amount of variance in
the ulcer phenotypes. At present, these PRSs are unlikely to be a
clinically useful tool in predicting which patients are at risk of
developing mouth ulcers or in predicting severity of mouth
ulcers.

Association in the HLA region at the single-SNP level was
recapitulated by haplotype analysis, which identified multiple
HLA haplotypes which are associated with odds of mouth ulcers.
The most striking finding was an association between the
DRB1*0103 haplotype and increased odds of mouth ulcers.
DRB1*0103 is an uncommon haplotype of
HLA-DRB1 which
encodes the beta chain of the HLA-DR heterodimer, forming a
ligand for the T cell receptor.

There are traits that commonly present with similar clinical
symptoms to mouth ulcers, and the loci identified in this study
show commonality with previous GWAS of these traits. Behçet’s
disease is thought to relate to inappropriate T-cell-mediated
inflammatory response and presents clinically with mouth ulcers
among other features31. Genetic variants at IL12A, IL10, STAT4,
RIPK2, IRF8, and CEBPB-PTPN1 which have been reported in
previous GWAS for Behçet’s disease32–34, are in high LD and
have consistent effect direction with those reported here. This
raises the possibility of a similar mechanism leading to clinical
presentation with mouth ulcers in both conditions. Elsewhere in
the literature there is striking overlap with the genetics of coeliac
disease, where T cell-mediated responses are also believed to be
important35. In particular, rs17810546 (a variant in close LD with

Fig. 3 Effect of predicted increased transcription of all genes on mouth ulcers. Each dot represented the effect of increased transcription (averaged across
all tissue-specific predictions using MultiXcan) on mouth ulcers. The size of the dot indicates the largest effect size in any tissue. The standard deviation
(SD) of tissue-specific Z-score is an indicator of tissue specificity with high values (yellow) indicating higher tissue specificity.
rs76830965, the top associated variant in our analysis, \( r^2 = 0.98 \) was reported as early as 2008 to have large effects on coeliac disease\(^{36} \), a finding which has since been validated\(^{37} \), whilst CCR3 has been reported more recently\(^{38} \). The HLA-DRB1*0103 haplotype has previously been reported as associated with ulcerative colitis\(^{39,40} \) and both Crohn’s and ulcerative colitis\(^{41} \) in candidate gene association studies. It is possible that the overlap in lead variants between these diseases is driven by specific tag variants flagging particular biological events affecting auto-immune traits of the digestive system.

Immune regulatory loci identified here may influence the susceptibility of infective or non-infective risk factors for mouth ulcers. A range of viral\(^{42-44} \), bacterial\(^{45} \) and other changes in the oral microbiome\(^{46} \) have been suggested as acquired risk factors for mouth ulcers. It is possible that the genetic loci identified in this study relate to mouth ulcers through regulation of the host microbiome; a similar view has recently been proposed for both Behçet’s and coeliac disease\(^{47,48} \). Hematinic status, especially deficiencies of folate, vitamin B12, ferritin or haemoglobin are thought to be risk factors for mouth ulcers\(^{49} \) and these deficiencies might occur secondary to genetically determined inflammatory or immune states, such as pernicious anaemia or coeliac disease.

Loci identified in the present study may also have an effect at a tissue response level, where the cellular response to minor oral trauma is either proportionate and leads to resolution or disproportionate, leading to the clinical experience of mouth ulcers. One current view is that dysregulation of local cell-mediated response leads to an inappropriate focal accumulation of CD8\(^+ \) T cell populations within the oral mucosa following minor triggers, leading to tissue damage and clinical manifestation as oral ulceration\(^{44,50} \). These upstream and downstream processes are not mutually exclusive and future research may wish to examine the effect of specific genetic loci across strata of potential risk factors to gain further insight into the aetiology.

A limitation of this study is that the presence or absence of mouth ulcers was inferred from questionnaire data rather than clinical examination. This is a necessary limitation as clinical oral examination data were not available and because the short duration and intermittent nature of mouth ulcers means they are often not visible on clinical examination even for affected individuals. In common with any questionnaire-derived data this may lead to some misclassification. For example, the phenotypes used in this study do not distinguish between RAS (which is in any case a clinical diagnosis) and other causes of mouth ulcers. For the most part we believe non-RAS ulceration (such as traumatic

---

**Fig. 4** Genetic correlation between mouth ulcers and 222 traits. Each dot represented the Rg between mouth ulcers and an individual trait. The red line represents the Bonferroni-corrected multiple testing threshold at the 5% level. Error bars showing 95% confidence intervals and labels are included for traits that pass the multiple testing threshold.
across the genome are anticipated to have small effects and are arthritis (In licensed for immune-related diseases, such as rheumatoid Some of these agents are tumour necrosis factor targets and are intervals for effect estimates in each of the three smaller lookup speci was to assess the effects of the lead genetic variants on more RAS-comparison to the studies with older participants (UK Biobank and AG). A major motivation for the use of the look-up collections was to assess the effects of the lead genetic variants on more RAS-specific phenotypes and in studies with younger participants. However, a limitation of this strategy is the wide confidence intervals for effect estimates in each of the three smaller lookup cohorts, meaning that only variants with larger effects are expected to exclude the null and the heterogeneity across the three studies which precludes the use of meta-analysis. The Open Targets analysis identified 52 drugs which might recapitulate the effects of naturally occurring genetic variation. Some of these agents are tumour necrosis factor targets and are licenced for immune-related diseases, such as rheumatoid arthritis (Infliximab51) while others such as Ustekinumab target IL12A and are licenced for several diseases (psoriasis52, psoriatic arthritis53 and Crohn’s disease54), and are being repurposed for other immune diseases such as systemic lupus erythematosus55. This vignette may help illustrate how the availability of GWAS results for mouth ulcers could facilitate repurposing of existing drug interventions or the development of novel, specific inter-ventions for mouth ulcers. In conclusion, this GWAS of mouth ulcers identified multiple associated loci including a common variant near IL12A with large effects on risk of mouth ulcers. Follow-up analyses provide insight into the aetiology of this common ulcerative condition and prioritise topics for future basic and applied research.

Methods
Overview. A GWAS for mouth ulcers was performed within UK Biobank. Variants passing a conventional threshold for genome-wide significance (P = 5.0e–08) were replicated in 23andMe. All variants which showed directional consistency across the two cohorts were further explored in three independent resources with more specific mouth ulcer phenotypes and genetic data; namely the ALSpac, the QIMR Over 50s (Aged) study (AG) and the QIMR Melanocytic Naevi in Adolescent Twins (TW) study.

Participants and phenotypes. UK Biobank is a population-based health research resource consisting of ~500,000 people, aged between 38 and 73 years, who were recruited between the years 2006 and 2010 from across the UK56. Participants provided a range of information pertinent to adult and later life health outcomes via questionnaires, interviews, physical measurement and donating biological samples (data showcase available at www.ukbiobank.ac.uk)57. In the baseline questionnaire participants were asked to supply information about their oral health. The question stem was: Do you have any of the following? (You can select more than one answer). The possible answers included mouth ulcers and participants were prompted to answer this question thinking about the past year, if they pressed the help button. Participants who selected this answer were coded as cases. Participants who did not select this answer were coded as controls. Participants who chose the option for prefer not to answer, or did not complete the questionnaire session were coded as missing and not included in further analysis. UK Biobank received ethical approval from the North West Multi-centre Research Ethics Committee (REC reference for UK Biobank is 11/NW/0382).

23andMe Inc. is a personal genomics company that provides genotype and health-related information to customers58. Individuals included in the analyses provided informed consent and are covered by the 23andMe human subjects protocol, which was reviewed and approved by Ethical and Independent Review Services, a private institutional review board (http://www. eandireview.com). Mouth ulcer cases were defined as those who answered yes to the question: Have you ever had a canker sore (an open sore on the soft tissue inside your mouth)? Those who were not included in the analysis. The ALSPAC longitudinal birth cohort recruited pregnant women living near Bristol, UK with an estimated delivery date between 1991 and 1992. There were 15,779 preterm births resulting in 10,873 live births59. Follow-up clinical assessment and questionnaires and is ongoing. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time. Study members were asked to complete questionnaires about oral health at age 23.9 years. Participants were asked if they had ever had mouth ulcers (no/yes, but only once or twice/yes, on several occasions). An ever ulcers phenotype was generated with any answer of Yes, used to define case status and No to define controls. Those who did not answer the question were set to missing. Study data were collected and managed using REDCap electronic data capture tools60.

The QIMR Over 50s (AG) study was initiated in 1992–1993 to understand the role of genetics in healthy aging and age-related disease. AG recruited twin pairs from the Australian Twin Register who were over 50 years of age. This study was approved by the QIMR Berghofer Human Research Ethics Committee (HREC reference number P1204). Participants were asked whether they had mouth ulcers now/previouly/both or never. This was used to derive case control status for ever or never having ulcers. The QIMR melanocytic naevi in adolescent twins (TW) study was established in 1992 to investigate melanotic naevoi. Twin pairs aged around 12 years of age were recruited from schools in Brisbane and the surrounding area in Queensland, Australia. This study was approved by the QIMR Berghofer Human Research Ethics Committee (HREC reference number P193). All twins and most parents donated blood for DNA extraction and completed a questionnaire. Children and mothers of nuclear families were asked independently about mouth ulcers. Mothers also answered a questionnaire on mouth ulcers on behalf of the entire family, with each family member given a score for ulcers which was scaled to allow comparison with the studies with binary ulcer phenotypes (never = 0, rarely = 1/3, sometimes = 2/3, frequently = 1). Pictures of RAS were used to help participants identify whether they had this specific type of oral ulceration. These ulcer questions were treated as a severity score preserving all four possible responses. Severity scores from both parents and children were combined into a mixed cohort.

Oral health phenotypic data collection in the ALSpac study was conducted between November 2015 and September 2016, the AG study was conducted between 1992 and 1993 (n = 1120). Mouth ulcer data in the TW study were collected from 1992 until 2016 (n = 2442).

Genotypes. The UK Biobank genotype data (July 2017 release) contains 488,377 successfully genotyped samples. 49,979 individuals were genotyped using the UK BiLEVE array and 438,398 using the UK Biobank array. Pre-imputation QC, phasing and imputation were completed64. In brief, prior to phasing, multiallelic variants or those with minor allele frequency ≤ 1% were removed. Phasing of genotype data was performed using a modified version of the SHAPEIT2 algorithm65. Genotype imputation was performed to a combined UK10K haplotype66 and Haplofix reference consortium (HRC) reference panels using IMPUTE2 algorithm67. A further QC protocol was then applied at the Wellcome Trust Centre for Human Genetics prior to release (http://biobank.ctsu.ox.ac.uk/crystal/docs/genotyping_qc.pdf). The analyses presented here were restricted to autosomal variants using a graded filtering so that rarer genetic var-iants are required to have a higher imputation INFO score (info > 0.3 for MAF > 1% and > 0.6 for MAF 1–3% and > 0.7 for MAF 0.1–0.5%). Monomorphic and uncommon variants with MAF < 0.1% were removed. In addition, all variants not included in the HRC site list were removed. Individuals with sex-mismatch (derived by comparing genetic sex and reported sex) or individuals with sex-chromosome aneuploidy were excluded from the analysis (n = 814). A k-means cluster analysis was performed with four clusters using the first four principal components provided by UK Biobank in the statistical software environment R68. The current analysis includes the largest cluster from this analysis (n = 464,708), of whom 461,106 individuals (47,100 with mouth ulcers) had non-missing phenotype data and were included in GWAS69. Sensitivity analyses of the HLA region used imputed HLA haplotypes provided by UK Biobank. As this analysis was performed using logistic regression rather than mixed models, analysis was restricted to individuals of white British ancestry. In addition, one individual from each pair of closely related (third degree or closer) individuals was removed until no related pairs remained. Following these exclusions criteria, had generated including which 336,038 had non-missing phenotype data and were included in logistic regression models.

8 NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-08923-6 | www.nature.com/naturecommunications
In 23andMe, DNA extraction and genotyping were performed on saliva samples by CLIA-certified and CAP-accredited clinical laboratories of Laboratory Corporation of America, Inc. Imputation was performed by 23andMe. Briefly, samples were genotyped on a 23andMe custom genotyping array platform (Illumina HumanHap550+ Bead chip V1 V2, OmniExpress+ Bead chip V3, custom array V4) with a minimum call rate of 98.5%. Missing participant genotype date was imputed using the UK10K and 1000 genome combined reference panel. Analysis was conducted on a maximal set of unrelated individuals.

Genetic data for the ALSPAC participants has been collected (n > 10,000). Genotyping was conducted on the Illumina HumanHap550 quad chip for children and Illumina Quad array for mothers. Prior to imputation, samples with >33% missingness, indeterminate heterozygosity, extreme autosomal heterozygosity or which clustered outside the CEU HapMap2 population using multidimensional scaling were removed. In addition, variants with minor allele frequency <1%, call rate of <95% or violations of Hardy–Weinberg equilibrium (P < 5e−7) were removed. Samples and variants passing these QC measures were carried forward to a joint phasing stage, prior to imputation to the HRC reference panel using the Michigan imputation server (r1.1, 2017 release).

Genetic data for the participants of AG and TW with mouth ulcer data have been collected as part of a larger project by QIMR that comprises multiple waves of genotyping. In total, 3562 participants had phenotype and genotype data available (n = 1120 for AG, n = 2442 for TW). Most participants were genotyped on the Illumina Human610-Quad v1-B (n = 1942) or HumanCoreExome-121v1-0-C (n = 1019) arrays; genotyping on a small number of additional individuals was conducted on the Illumina HumanOmni25M-8v1-1_B (n = 144), 317 K (n = 247), HumanCNV370 (n = 91), Human600W-Quad v1-C (n = 8), HumanOmniExpress-12v1-A (n = 21) and PsychArray-B (n = 92) platforms. Genotype data from all assays was jointly imputed using HRC reference panel (r1.1)^46-70.

### Statistical methods.

A GWAS was performed using a LMM implemented in BOLT-LMM (v2.3)^72 using an in-house GWAS pipeline. A subset of 143,006 high-quality variants were used to estimate and account for genetic relatedness and ancestry, allowing for the inclusion of closely and distantly related individuals in genetic analyses. Age, sex and genotype arrays were included as covariates in association testing. The Bayesian model was not predicted to provide a substantially better fit than a conventional LMM, so results are presented for the standard bolt_lmm_inf model. BOLT-LMM association statistics are on the linear scale. Analyses for adult twins and siblings (including standard LMM) were transformed to log odds ratios and their corresponding 95% confidence intervals on the liability scale using a Taylor transformation expansion series. These methods for transforming betas to odds ratios that take into account allele frequency have been suggested. Both methods for deriving transformed OR showed excellent concordance (Supplementary Figure 12) so only the values from Taylor transformation expansion series are reported. Genome-wide significance was defined at P < 5e−08. Associated loci that passed the genome-wide association threshold were clumped based on LD values using PLINK (version 1.9)^78 to identify approximately independent associated variants (options --clump-kb 10000 –clump-p2 1 --clump-r2 0.1), using an independent sample of HRC imputed genetic data to estimate LD. These variants were selected for replication in 23andMe.

In 23andMe, associations between the UK Biobank genome-wide associated variants and mouth ulcer phenotype were assessed using linear regression analysis, assuming an additive model and using a 23andMe internally developed pipeline. Replication was considered successful if directionality was consistent across the two studies. Where replication was successful the estimates from UK Biobank and 23andMe were meta-analysed using the meta package in R^79.

Association between the replicated variants and the mouth ulcer phenotype were assessed using linear regression analysis, assuming an additive model and using a 23andMe internally developed pipeline. Replication was considered successful if directionality was consistent across the two studies. Where replication was successful the estimates from UK Biobank and 23andMe were meta-analysed using the meta package in R^79.

### Tests for polygenicity bias and heritability.

Univariate LD score regression was performed using the LDSC software package (v1.0.0, April 2017 release)^25 to quantify the relative contribution of polygenicity and bias to inflation in single variant test statistics from UKBB. We used LDSC to estimate heritability under the multi-locus model assumption and obtained contrasting estimates from the HESS method which may be more robust under certain genetic architectures^28.

### HLA haplotype analysis.

Analysis of the HLA region in UK Biobank used 362 imputed haplotypes (provided by UK Biobank). Unadjusted analysis was performed using logistic regression and an additive model for each haplotype which used haplotype dosage to account for uncertainty in haplotype imputation. Seventeen haplotypes were not observed within the final sample, whilst 345 were present in at least one individual and produced test statistics. Of these 345 haplotypes, 24 passed the Bonferroni-corrected P-value threshold (P < 1.5e−04) and were examined further in a fully adjusted model which included adjustment for age, sex, 40 genetic principal components and genotyping array.

### In-silico functional analysis.

Tests for association between ulcer and predicted gene expression were performed using S-PrediXcan^26, which uses predictive models to impute transcript expression levels in specific tissues and then uses full GWAS summary statistics to test for associations between these predicted expression levels and phenotype. A threshold was applied on predicted performance (a measure of ability to accurately infer transcription levels at a given locus in a specific tissue) with a cut-off at 0.1. After filtering, all remaining predicted expression levels were tested for association with mouth ulcers. S-MultiXcan^25 was then applied to these tissue-specific estimates, which harnesses information from the sharing of expression quantitative trait loci (eQTL) across multiple tissues, to increase power to detect the effects of gene expression on the phenotype of interest. A Bonferroni correction for multiple testing was applied which takes into the correlation between tissues and equals to 0.05 divided by the number of genes with a prediction model in at least one tissue (P < 1.9e−06). Analysis was performed using the MultiXcan standalone package, which includes the S-PrediXcan and S-MultiXcan methods.

Gene prioritisation, gene set enrichment and tissue enrichment analysis were performed using the DEPICT software package^27. DEPICT anticipates that association signals for a biologically causal gene will co-localise with genes encoding other members of a co-regulated gene network elsewhere in the genome. By using publicly available gene expression data to define co-regulated gene networks, DEPICT can take account of association patterns elsewhere in the genome to nominate biologically plausible candidate genes at a statistically associated locus. DEPICT analysis was performed using the MultiXcan standalone package which passed the gene set threshold (P < 5e−8) after clamping and default settings in the standalone java software.

Enrichment in regulatory motifs was assessed using the non-parametric enrichment analysis package GARFIELD^25. GARFIELD performs greedy LD pruning of GWAS summary statistics using reference data from the UK10K project. LD-tagged genomic regions are then annotated with information on 1005 regulatory features, including ENCODE, GENCODE and Roadmap epigenomics projects. These features include genetic annotations, chromatin states, histone modifications, DNase hypersensitive sites and transcription factor-binding sites in a range of cell lines. GARFIELD tests whether regulatory information is present in trait-associated loci more frequently than expected by chance by performing adaptive permutations at various significance thresholds (p = 10−2, 10−4, 10−6, 10−8,..., 10−10) but requires full GWAS data. Trait-associated loci are matched with appropriate null loci using many features, such as LD information from the ENCODE and ROADMAP projects. Analysis was conducted using default settings in standalone GARFIELD (v1.0) software.

### Tests for genetic correlation.

Genetic correlation between mouth ulcer other and tests was assessed using bivariate LD score regression implemented in the LD Hub platform^28. For traits that passed the Bonferroni-corrected threshold for association the rho-HESS approach was applied to estimate local correlation between mouth ulcers and these phenotypes.

### PRS analysis.

PRS were generated in plink v1.9 at a range of P-value thresholds between 1 and 5e−08. PRS were standardised using a mean of 0 and a standard deviation of 1. Association between PRS and mouth ulcers was assessed using LMMs in GCTA, including a genetic relatedness matrix to account for family structure and relatedness in QIMR.

### Drug repurposing.

The Open Targets database for pharmacological interventions (https://www.targetvalidation.org) was used to assess whether any of the ulcer-associated genes might represent targets for drug repurposing. The S-PrediXcan analysis was used to select gene transcripts.

### Data availability.

All phenotypes and genotypes used in UK Biobank are available to bona fide researchers through a process agreed at ukbiobank.ac.uk. The ALSPAC study website contains details of all the data that are available through a fully searchable data dictionary (www.bristol.ac.uk/alspac/researchers/our-data). The data access procedures for ALSPAC data are described in full online (http://www.bristol.ac.uk/alspac/researchers/access/). Researchers interested in using the data should contact Nick Martin (Nick.Martin@qimrberghofer.edu.au). All software packages referred to in the methods are available online. The summary statistics of the UK Biobank GWAS have been deposited at data.brs [https://doi.org/10.5523/brs.459eyulzf2ys5hensf550y4] and the summary statistics for the 97 lead variants from 23andMe, Inc. and the meta-analysis of these with the UK Biobank variants can be found in Supplementary Data 1.
Received: 12 October 2018 Accepted: 5 February 2019
Published online: 05 March 2019

References
1. Slebioda, Z., Szponar, E. & Kowalska, A. Etiopathogenesis of recurrent aphthous stomatitis and the role of immunologic aspects: literature review. Adv. Immunol. Ther. Exp. 62, 205–215 (2014).
2. Miller, M. F., Garfunkel, A. A., Ram, C. A. & Ship, I. The inheritance of recurrent aphthous stomatitis - observations on susceptibility. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. Endod. 49, 409–412 (1980).
3. Katsanos, K. H. et al. Review article: non-malignant oral manifestations in inflammatory bowel diseases. Aliment. Pharmacol. Ther. 42, 40–60 (2015).
4. Chiewchongkol, D. et al. Mucocutaneous manifestations in a UK national cohort of juvenile-onset systemic lupus erythematosus patients. Rheumatology 53, 1504–1512 (2014).
5. Khatibi, M., Shakoorpour, A. H., Jahromi, Z. M. & Ahmadzadeh, A. The prevalence of oral mucosal lesions and related factors in 188 patients with systemic lupus erythematosus. Lupus 21, 1312–1315 (2012).
6. Al-Omiri, M. K. et al. Recurrent aphthous stomatitis (RAS): a preliminary genome-wide complex trait analysis. J. Hum. Genet. 88, 78–82 (2011).
7. Xie, M. et al. Behcet's disease: a review. Autoimmun. Highlights 7, 4 (2016).
8. Sousa, I. et al. Association of CCR1, KLRC4, IL12A-AS1, STAT4, and ERAPI with Behcet's disease in Iranians. Arthritis Rheumatol. 67, 2742–2748 (2015).
9. Kappen, J. H. et al. Genome-Wide Association Study in an admixed case series reveals IL12A as a new candidate in Behcet disease. PLoS ONE 10, e0119085 (2015).
10. Takeuchi, M. et al. Dense genotyping of immune-related loci implicates host responses to microbial exposure in Behcet's disease susceptibility. Nat. Genet. 49, 438–443 (2017).
11. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 common traits from summary association data. Am. J. Hum. Genet. 99, 139–153 (2016).
12. Puzanowska, B., Prokopowicz, D., Ziarko, S., Radziwon, P. & Lapinski, T. W. The incidence of HLA DRB1*0103 in ulcerative colitis patients in North-Eastern Poland. Hepatogastroenterology 50, 1436–1438 (2003).
13. Ippolite, M. S. et al. Population- and family-based study of Canadian families reveals association of HLA DRB1*0103 with colonic involvement in inflammatory bowel disease. Inflamm. Bowel Dis. 9, 1–9 (2003).
14. Sun, A. et al. Human cytomegalovirus as a potential etiologic agent in recurrent aphthous ulcers and Behcet's disease: J. Oral Pathol. Med. 25, 212–218 (1996).
15. Sun, A. et al. Preliminary evidence for an association of Epstein–Barr virus with pre-ulcerative oral lesions in patients with recurrent aphthous ulcers or Behcet's disease. J. Oral Pathol. Med. 27, 168–175 (1998).
16. Lin, S. et al. Study of the viral infections and cytokines associated with recurrent aphthous ulcerations. J. Oral Microbiol. 10, 1118–1124 (2008).
17. Silverberg, M. S. et al. A population- and family-based study of Canadian families reveals association of HLA DRB1*0103 with colonic involvement in inflammatory bowel disease. Inflamm. Bowel Dis. 9, 1–9 (2003).
18. Boer, G. J. et al. The oral microbiota of patients with recurrent aphthous stomatitis. J. Oral Microbiol. 6, 25739 (2014).
19. Morton, L. T., Situnayake, D. & Wallace, G. R. Genetics of Behcet's disease. Curr. Opin. Rheumatol. 28, 39–44 (2016).
20. Verdu, E., Galipeau, H. J. et al. Behcet's disease. Nat. Rev. Gastroenterol. Hepatol. 12, 497–506 (2015).
21. Chen, H., Sui, Q., Chen, Y., Ge, L. & Lin, M. Impact of haematologic deficiencies on recurrent aphthous ulceration: a meta-analysis. Br. Dent. J. 218, 6 (2015).
22. Akinowe, S. O. & Greenberg, M. S. Recurrent aphthous stomatitis. Dent. Clin. North Am. 58, 281–297 (2014).
23. Fleischmann, R. et al. Infliniximab efficacy in rheumatoid arthritis after an inadequate response to etanercept or adalimumab: results of a target-driven active switch study. Curr. Med. Res. Opin. 30, 2139–2149 (2014).
24. Papp, K. A. et al. Efficacy and safety of ustekinumab, a human interleukin-12/23 monoclonal antibody, in patients with psoriasis: 52-week results from a randomised, double-blind, placebo-controlled trial (PHOENIX 2). Lancet 371, 647–654 (2013).
25. van Vollenhoven, R. F. et al. Efficacy and safety of ustekinumab, an IL-12 and IL-23 inhibitor, in patients with active systemic lupus erythematosus: results of a phase 3, multicentre, double-blind, placebo-controlled PSMITM1 trial. Lancet 382, 780–789 (2013).
26. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. 9, 1825 (2018).
27. Lonsdale, J. et al. The Genotype-Tissue Expression (GTex) project. Nat. Genet. 45, 580–585 (2013).
28. Barbeira, A. N. et al. Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genetics 15, e1007889 (2019).
29. Zhou, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33, 272–279 (2017).
30. Yang, J. A., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88, 76–82 (2011).
31. Tezcan, M. et al. Behcet’s disease pathophysiology: a contemporary review. Autoimmun. Highlights 7, 4 (2016).
32. Sousa, I. et al. Association of CCR1, KLRC4, IL12A-AS1, STAT4, and ERAPI with Behcet's disease in Iranians. Arthritis Rheumatol. 67, 2742–2748 (2015).
33. Kappen, J. H. et al. Genome-Wide Association Study in an admixed case series reveals IL12A as a new candidate in Behcet disease. PLoS ONE 10, e0119085 (2015).
34. Takeuchi, M. et al. Dense genotyping of immune-related loci implicates host responses to microbial exposure in Behcet's disease susceptibility. Nat. Genet. 49, 438–443 (2017).
35. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the genetic architecture of 30 common traits from summary association data. Am. J. Hum. Genet. 99, 139–153 (2016).
of a multicentre, double-blind, phase 2, randomised, controlled study. *Lancet* 392, 1330–1340 (2019).

56. Allen, N. E., Sudlow, C., Peakman, T. & Collins, R. UK Biobank data: come and get it. *Sci. Transl. Med.* 6, 224ed4 (2014).

57. Collins, R. What makes UK Biobank special? *Lancet* 379, 1173–1174 (2012).

58. Pickrell, J. K. et al. Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* 48, 709–717 (2016).

59. Boyd, A. C. et al. Cohort profile: The children of the 90s—the index offspring of the Avon Longitudinal Study of Parents and Children. *Int. J. Epidemiol.* 42, 111–123 (2013).

60. Harris, P. A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* 42, 377–381 (2009).

61. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomics data. *Nature* 562, 203–209 (2018).

62. O’Connell, J. et al. Haplotype estimation for biobank-scale data sets. *Nat. Genet.* 48, 817–820 (2016).

63. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* 6, 8111 (2015).

64. Hoveve, R., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 Genes Genomes Genet.* 1, 457–470 (2011).

65. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018).

66. Mitchell, R., Hemani, G., Dudding, T. & Paternoster, L. UK Biobank Genetic Dose: Quality Control, Version 1. https://doi.org/10.5523/bris.3074kr6bb2r9zh2b0303wq (2017).

67. Sanchez-Roige, S. et al. Genome-wide association study of discounting in 23,217 adult research participants of European ancestry. *Nat. Neurosci.* 21, 16–20 (2018).

68. McGregor, B. et al. Genetic and environmental contributions to size, color, shape, and other characteristics of melanocytic naevi in a sample of adolescent twins. *Genet. Epidemiol.* 16, 40–53 (1999).

69. Zhu, G. et al. A major quantitative trait locus for mole density is linked to the familial melanoma gene CDKN2A: a maximum-likelihood combined linkage and association analysis in twins and their sibs. *Am. J. Hum. Genet.* 65, 482–493 (1999).

70. Zhu, G. et al. A genome-wide scan for naevus count: linkage to CDKN2A and to other chromosome regions. *Eur. J. Hum. Genet.* 15, 94–102 (2007).

71. Loh, P.-R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290 (2015).

72. Loh, P.-R. R., Kichaev, G., Gatzal, S., Schoeck, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. *Nat. Genet.* 50, 906–908 (2018).

73. Elsworth, B. et al. MRC IEU UK Biobank GWAS Pipeline Version 1. https://doi.org/10.5523/bris.2fahpksont1zi26xs0yamqo8rr (2017).

74. Zaitlen, N. et al. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* 9, e1003520 (2013).

75. Cortes, A. et al. Identification of multiple risk variant related to ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat. Genet.* 45, 730–738 (2013).

76. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7 (2015).

77. Schwarz, G. meta: an R package for meta-analysis. *R News* 7, 40–45 (2007).

78. Zhan, X., Hu, Y., Li, R., Albecais, G. & Liu, D. J. RVTESTX: an efficient and comprehensive tool for rare variant association analysis using sequence data. *Bioinformatics* 32, 1423–1426 (2016).

79. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* 6, 9 (2015).

80. Shi, H., Mancuso, N., Spendlove, S. & Pasaniuc, B. Local genetic correlation gives insights into the shared genetic architecture of complex traits. *Am. J. Hum. Genet.* 101, 737–751 (2017).

Acknowledgements

S.H. and T.D. conducted the ALSPAC data collection, UK Biobank and ALSPAC analysis, conducted the meta-analysis and in-silico functional analyses, interpreted the results and wrote the manuscript. N.T. supervised the analysis, interpreted the results and edited the manuscript. R.M., B.E. and L.P. created the infrastructure to enable the UK Biobank analysis within the MRC IEU and provided comments on the manuscript. P.A.L., L.C.-C., S.E.M. and S.G. analysed the TW and AG datasets and provided comments on the manuscript. S.J.T. and N.G.M. collected the TW and AG data and edited the manuscript. P.W.F. interpreted the results and provided comments on the manuscript. J.F.S. contributed replication in the 23andMe dataset and was supervised by J.Y.T. The 23andMe Research Team created the infrastructure to enable data collection and analysis.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-08923-6.

Competing interests: J.F.S., J.Y.T. and members of the 23andMe Research Team are employees of 23andMe, Inc. and hold stock or stock options in 23andMe. The remaining authors declare no competing interests.

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

Journal peer review information: *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019
23andMe Research Team
Michelle Agee4, Babak Alipanahi4, Adam Auton4, Robert K. Bell4, Katarzyna Bryc4, Sarah L. Elson4,
Pierre Fontanillas4, Nicholas A. Furlotte4, Barry Hicks4, David A. Hinds4, Karen E. Huber4, Ethan M. Jewett4,
Yunxuan Jiang4, Aaron Kleinman4, Keng-Han Lin4, Nadia K. Litterman4, Jennifer C. McCeight4,
Matthew H. McIntyre4, Kimberly F. McManus4, Joanna L. Mountain4, Elizabeth S. Noblin4,
Carrie A.M. Northover4, Steven J. Pitts4, G. David Poznik4, Janie F. Shelton4, Suyash Shringarpure4, Chao Tian4,
Vladimir Vacic4, Xin Wang4 & Catherine H. Wilson4