

Genome-wide association study analysis of disease severity in Acne reveals novel biological insights.

Authors: Zhaohui Du¹, Tejaswi Iyyanki¹, Samuel Lessard¹, Michael Chao¹, FinnGen, Christian Asbrand², Dany Nassar³, Katherine Klinger⁴, Emanuele de Rinaldis¹, Shameer Khader¹, Clément Chatelain¹

1. Precision Medicine & Computational Biology, Sanofi, Cambridge, MA, USA
2. Type 1/17 Inflammation Cluster, Sanofi, Frankfurt, Germany
3. Early Immunology & Inflammation Development Clinical Team, Sanofi, Chilly-Mazarin, France
4. Genetics Research, Sanofi, Cambridge, MA, USA

Corresponding author

Name	Clément Chatelain
Address	350 Water Street
	Cambridge, MA, USA 02141
E-mail	Clement.Chatelain@sanofi.com

Abstract

Acne vulgaris is a common skin disease that affects >85% of teenage young adults among which >8% develop severe lesions that leaves permanent scars. Genetic heritability studies of acne in twin cohorts have estimated that the heritability for acne is 80%. Previous genome-wide association studies (GWAS) have identified 50 genetic *loci* associated with increased risk of developing acne when compared to healthy individuals. However only a few studies have investigated genetic association with disease severity. GWAS of disease progression may provide a more effective approach to unveil potential disease modifying therapeutic targets.

Here, we performed a multi-ethnic GWAS analysis to capture disease severity in acne patients by using individuals with normal acne as a control. Our cohort consists of a total of 2,956 participants, including 290 severe acne cases and 930 normal acne controls from FinnGen, and 522 cases and 1,214 controls from BioVU. We also performed mendelian randomization (MR), colocalization analyses and transcriptome-wide association study (TWAS) to identify putative causal genes. Lastly, we performed gene-set enrichment analysis using MAGMA to implicate biological pathways that drive disease severity in Acne.

We identified two new loci associated with acne severity at the genome-wide significance level, six novel associated genes by MR, colocalization and TWAS analyses, including genes *CDC7*, *SLC7A1*, *ADAM23*, *TLL10*, *CDK20* and *DNAJA4*, and 5 novel pathways by MAGMA analyses. Our study suggests that the etiologies of acne susceptibility and severity have limited overlap, with only 26% of known acne risk loci presenting nominal association with acne severity and none of the novel severity associated genes reported as associated with acne risk in previous GWAS.

Introduction

Acne vulgaris is a chronic inflammatory dermatosis that is characterized by comedones and inflammatory lesions including papules, pustules, and nodules¹. It is the most common skin disorder that affects >85% of teenage young adults where >8% develop severe lesions that leaves permanent scars². Severe acne may have a significant long-term negative impact on emotional and mental health including low self-esteem and social isolation³. The pathogenesis of acne involves a complex interplay between sebum production, follicular keratinisation, inflammation, and colonization of pilosebaceous follicles by *Propionibacterium acnes*¹.

Genetic heritability studies of acne in twin cohorts have shown that the heritability for acne is 80%⁴. Previous genome-wide association studies (GWAS) have identified 50 genetic *loci* associated with increased risk of developing acne when compared to healthy individuals among Caucasian or Asian populations⁵⁻⁹. However only a few studies have investigated genetic associations with disease severity. Whether biological mechanisms driving disease onset and progression are the same is still an open question in human genetics, as for many complex diseases. GWAS that focus on disease severity or progression may provide a more effective approach to unveil potential disease modifying therapeutic targets. Such studies have recently become possible with the rise of large-scale biobanks, such as FinnGen and UK Biobank, combining genetic and electronic health records.

Here, we performed a multi-ethnic GWAS analysis to capture disease severity in acne patients by using individuals with normal acne as a control. Our cohort consists of a total of 2,956 participants, including 290 severe acne cases and 930 normal acne controls from FinnGen release 10, and 522 cases and 1,214 controls from the BioVU biorepository extracted by Nashville Biosciences. We also performed mendelian randomization (MR), colocalization analyses and transcriptome-wide association study (TWAS) to identify putative causal genes. Lastly, we performed gene-set enrichment analysis using MAGMA to implicate biological pathways that drive disease severity in acne.

Results

GWAS analyses

We discovered one novel locus at 7q22.3 significantly associated with acne severity at the genome-wide significance level ($P\text{-value} < 5 \times 10^{-8}$) among participants with African ancestry (AA) in the BioVU cohort (Figure 1, Supplementary figure 1). The lead variant, rs115325598 (OR = 8.22; 95% CI: 3.72, 18.13; $P\text{-value} = 3.15 \times 10^{-8}$), is located at the intron region of gene *ATXN7L1*. The effect allele (C) of rs115325598 has a frequency of 6% in controls and 16% in severe acne cases among AA population in BioVU. There is one additional genome-wide significant locus at 12p12.2 among participants with European ancestry (EA) in BioVU cohort (supplementary figure 2). The lead variant, rs116779562 (OR = 11.04; 95% CI: 4.61, 26.39; $P\text{-value} = 3.25 \times 10^{-8}$), located at an intergenic region between genes *LINC02398* and *PDE3A*. However, this association was not replicated among FinnGen population (OR = 1.49; 95% CI: 0.47, 4.77; $P\text{-value} = 0.50$). The effect allele (T) of this variant is rare in controls (0.3%) and has low frequency among severe acne cases (2%). No genome-wide signal was detected in any other GWAS analyses (Supplementary figure 1).

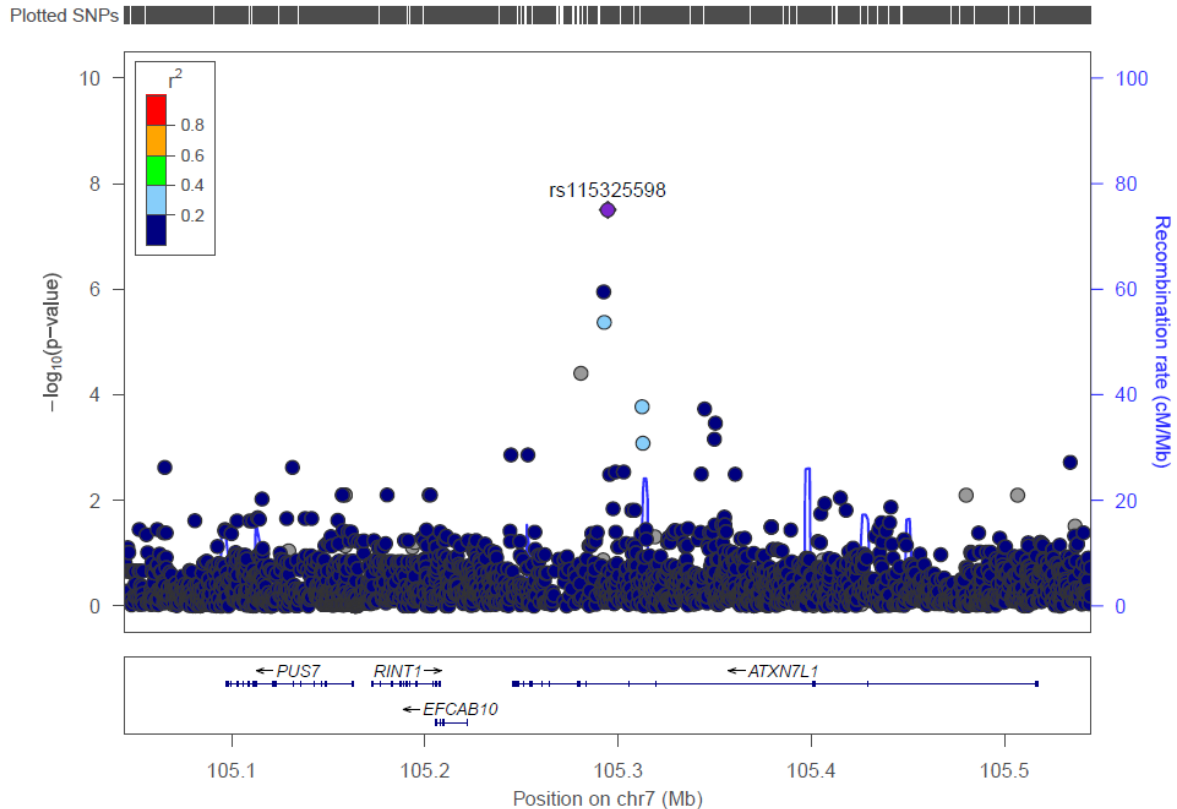


Figure 1: Regional association plots of the 7q22.3 risk region for acne severity among African ancestry (AA) population in BioVU cohort. Single-nucleotide polymorphisms (SNPs) are plotted by position (x-axis) and $-\log_{10}$ p-value (y-axis). r^2 was estimated from AA individuals in phase III 1000 Genomes Project (1KGP) data. The most statistically significant associated SNP (purple diamond) at 7q22.3 is rs115325598. The surrounding SNPs are colored to indicate pairwise correlation with the index SNP.

Variant annotation

We annotated the impact of all suggestive variants with $P < 1 \times 10^{-6}$ including coding variants and predicted gain or loss of function variants (GoLoF), and also mapped these variants to genes using activity-by-contact (ABC) enhancer-promoter interactions¹⁰.

Among the 5 single and meta-analyses GWASs, a total of 38 unique variants passed suggestive significance level. Two variants, rs3742185 and rs186795319 are annotated as coding variants of genes *ANKRD10* and *UMODL1*, with predicted functional consequences being missense and stop gained, respectively (Supplementary table 1).

Mendelian randomization and Colocalization analyses

In MR analyses, 36 genes were identified as significantly associated with acne severity at q-value < 0.05 . We then conducted a colocalization analysis to further assess the probability that severe acne associated SNPs and eQTLs share the same causal

genetic variant. Six genes (*ADAM23*, *CDC7*, *CDK20*, *DNAJA4*, *SLC7A1*, *TLL10*) were found to have a colocalization probability ≥ 0.8 in at least one tissue/cell type. (Table 1, Supplementary table 2).

Table 1. Genes that are associated with severe acne risk in the MR (Q-value < 0.05) and Colocalization analyses (PP.H4 ≥ 0.8).

Gene	GWAS	Region	MR beta*	MR q-value*	No. tissues/ cell types with Coloc PP.H4 ≥ 0.8	Max. Coloc PP.H4*	Tissue/cell type with Max. Coloc PP.H4*
<i>ADAM23</i>	BioVU EA	chr2: 207062772-207562772	1.29	5.92×10^{-3}	2	0.97	Sigmoid colon
<i>CDC7</i>	BioVU EA	chr1: 91716406-92216406	1.29	1.26×10^{-3}	13	1.00	Monocyte
<i>CDK20</i>	BioVU EA	chr9: 90214701-90714701	-8.15	1.61×10^{-3}	1	0.93	Brain
<i>DNAJA4</i>	BioVU EA + FinnGen	chr15: 78315188-78815188	0.32	9.86×10^{-3}	2	1.00	Monocyte
<i>SLC7A1</i>	BioVU EA	chr13: 30017359-30517359	4.88	8.06×10^{-4}	4	0.97	Blood
<i>SLC7A1</i>	BioVU EA + FinnGen	chr13: 30018119-30518119	3.50	8.89×10^{-3}	3	0.96	Blood
<i>SLC7A1</i>	BioVU EA + AA + FinnGen	chr13: 30017359-30517359	3.05	0.02	1	0.97	Blood
<i>TLL10</i>	BioVU EA	chr1: 816097-1316097	-1.34	0.04	1	0.87	Stomach

*Displayed results in tissues with MR Q-value < 0.05 and with the highest colocalization PP.H4.

Gene-set analyses using MAGMA identified pathways associated with severe acne risk.

In analyses using MAGMA, no associated gene was identified in the gene level analyses for severe acne risk. Subsequent gene-set analysis demonstrated significant enrichment in four pathways/gene sets at FDR < 0.05 (Table 2). The top pathway is “response to fluoride”, which includes five genes, including *COL1A1*, *PLCB1*, *ATG5*, *PON1* and *ASPN*, that collectively regulate the change in state or activity of a cell or an organism (in terms of movement, secretion, enzyme production, gene expression, etc.) as a result of a fluoride stimulus.

Table 2. Pathways statistically significantly associated with acne severity in MAGMA analysis.

Pathway Name	Discovery GWAS	P-value	Number of genes
GOBP: response to fluoride	BioVU EA + FinnGen	1.72×10^{-7}	5
GOCC: ISWI type complex	BioVU EA + FinnGen	1.12×10^{-6}	11
GOCC: intercellular bridge	BioVU EA + AA + FinnGen	1.59×10^{-6}	67
GOBP: sperm DNA condensation	BioVU EA	1.23×10^{-6}	10

Potential pathogenic gene prioritization

After getting the list of candidate pathogenic genes associated with severe acne risk using GWAS (the nearest gene of the genome-wide significant risk variant), MR and colocalization analyses, we further gathered additional evidences that may support the

pathogenicity of those genes through evaluating genetic associations with other diseases in a phenome-wide search using an in-house gene prioritization scheme, differential expressions in other diseases across multiple tissues/cell types, bulk and single cell gene expression levels by tissues and cell types. The credentialing results for each gene are summarized below (Table 3, Supplementary table 3-5, Supplementary figures 3-4).

Table 3: Summary of genetic-/transcriptomic- associated phenotypes and the highest expression tissues/cell types of candidate pathogenic genes.

Gene	Discovery GWAS	Phenotype with genetic association evidence (MR direction)	Phenotype with the most significant DEG (tissue: cell type)	Baseline expression	Highest expressed single cell (percent of cells expressing the gene, %)	Gene function
<i>ADAM23</i>	BioVU EA	Hypertension (-), Chronic disease of tonsils and adenoids (+), other diseases of upper respiratory tract (+), anxiety disorders (+), lower back pain, Siatica+ with lumbago	psoriasis vulgaris (Skin: epidermal keratinocyte)	Highly expressed in brain.	Highly expressed in neuron.	May play a role in cell-cell and cell-matrix interactions. This is a non-catalytic metalloprotease-like protein.
<i>ATXN7L1</i>	BioVU AA	Type 2 diabetes	atopic eczema (skin)	Ubiquitous expression, low tissue specificity	extratelencephalic-projecting glutamatergic cortical neuron (87.3)	One of the ATXN7 gene family members. An important paralog of this gene is <i>ATXN7</i> ,
<i>CDC7</i>	BioVU EA	Hemorrhage of rectum and anus, other diseases of anus and rectum	psoriasis vulgaris (Skin: epidermal keratinocyte)	Ubiquitous expression, enhanced in testis	oocyte (88)	Kinase involved in initiation of DNA replication. Phosphorylates critical substrates that regulate the G1/S phase transition and initiation of DNA replication, such as MCM proteins and CLASPIN. Required for high-level Shh responses in the developing neural tube. Together with TBC1D32, controls the structure of the primary cilium by coordinating assembly of the ciliary membrane and axoneme, allowing GLI2 to be properly activated in response to SHH signaling (By similarity). Involved in cell growth. Activates CDK2, a kinase involved in the control of the cell cycle, by phosphorylating residue 'Thr-160'.
<i>CDK20</i>	BioVU EA		crohn's disease (ileum)	Ubiquitous expression, low tissue specificity	oocyte (46)	Enables chaperone binding activity and unfolded protein binding activity. Involved in several processes, including negative regulation of endothelial cell migration; negative regulation of inclusion body assembly; and protein refolding.
<i>DNAJA4</i>	BioVU EA + FinnGen	Pneumonia (-)	psoriasis vulgaris (Skin: epidermal keratinocyte)	Ubiquitous expression, enhanced in brain and tongue	sperm (84)	High-affinity, low-capacity permease involved in the transport of the cationic amino acids (arginine, lysine and ornithine) in non-hepatic tissues. Predicted to enable protein-glycine ligase activity, elongating. Predicted to inactive polyglycylase.
<i>SLC7A1</i>	BioVU EA	Hypertension (+)	juvenile dermatomyositis (skeletal muscle)	Ubiquitous expression, enhanced in esophagus	distal lung epithelial progenitor cell (71)	
<i>TLL10</i>	BioVU EA	Asthma (Mixed)	rheumatoid arthritis (synovial tissue)	Enhanced expression in testis	sperm (74)	

***ADAM23* (ADAM Metallopeptidase Domain 23)**

In genetic prioritization ranking with other potential diseases, *ADAM23* showed evidence of association with a variety of diseases including hypertension, upper respiratory tract diseases (Supplementary table 3). In differential gene expression analysis, *ADAM23* is shown to be mostly upregulated in epidermal keratinocyte cell in skin tissue of psoriasis vulgaris (log₂-fold change = 2.8, adjusted P=9.4x10⁻¹⁸) and downregulated in CD14+ monocyte cell in skin tissue of clinically isolated syndrome (log₂-fold change = -3.7, adjusted P=1.4x10⁻¹³) (Supplementary figure 3, Supplementary table 4). In baseline gene expression data, *ADAM23* is mostly expressed in brain and heart, followed by CD4-positive, alpha-beta thymocyte cells (Supplementary figure 4). In single cell data, *ADAM23* is most expressed in neural cells, with 99.1% of extratelencephalic-projecting glutamatergic cortical neuron cells expressing the gene, followed by corticothalamic-projecting glutamatergic cortical neuron (97.7%) (Supplementary table 5).

***ATXN7L1* (Ataxin 7 Like 1)**

In genetic prioritization ranking with other potential related diseases, *ATXN7L1* showed evidence of association with type 2 diabetes, with the instrument variant being a lead coding variant in this gene (Supplementary table 3). In differential gene expression analysis, *ATXN7L1* is shown to be mostly upregulated in neutrophil cell in blood of septic arthritis (log₂-fold change = 1.3, adjusted P=6.6x10⁻⁸) and downregulated in epidermal keratinocyte cell in skin tissue of psoriasis vulgaris (log₂-fold change = -1.6, adjusted P=7.3x10⁻¹⁵) (Supplementary figure 3, Supplementary table 4). *ATXN7L1* is ubiquitously expressed across multiple tissues with low tissue specificity (Supplementary figure 4). In single cell data, *ATXN7L1* is most expressed in extratelencephalic-projecting glutamatergic cortical neuron, with 87.3% of cells expressing this gene (Supplementary table 5).

***CDC7* (Cell Division Cycle 7)**

In genetic prioritization ranking with other potential related diseases, *CDC7* showed evidence of association with hemorrhage of rectum and anus (Supplementary table 3).

In differential gene expression analysis, *CDC7* is shown to be mostly upregulated in epidermal keratinocyte cell of skin tissue in psoriasis vulgaris (log₂-fold change = 4.2, adjusted P=6.7x10⁻¹⁸) and downregulated in platelet of colorectal cancer (log₂-fold change = -2.5, adjusted P=1.0x10⁻⁴) (Supplementary figure 3, Supplementary table 4). *CDC7* is broadly expressed in testis, lymph node, bone marrows and multiple other tissues (Supplementary figure 4). In single cell data, *CDC7* is most expressed in germ cell, with 87.9% of oocyte cells expressing this gene (Supplementary table 5).

***CDK20* (Cyclin Dependent Kinase 20)**

CDK20 did not show genetic evidence of association with any disease using the genetic prioritization ranking scheme. In differential gene expression analysis, *CDK20* is shown to be mostly upregulated in skin tissue of atopic dermatitis (log₂-fold change = 1.1, adjusted P=4.0x10⁻²) and downregulated in PBMC of clinically isolated syndrome (log₂-fold change = -1.5, adjusted P=1.0x10⁻¹²) (Supplementary figure 3, Supplementary table 4). *CDK20* is ubiquitously expressed across multiple tissues with low tissue specificity (Supplementary figure 4). In single cell data, *CDK20* is most expressed in germ cell, with 45.6% of oocyte cells expressing this gene (Supplementary table 5).

***DNAJA4* (DnaJ Heat Shock Protein Family (Hsp40) Member A4)**

In genetic prioritization ranking with other potential related diseases, *DNAJA4* showed evidence of association with pneumonia, with a significant negative MR (MR beta = -0.03) association with plasma pQTL (Supplementary table 3). In differential gene expression analysis, *DNAJA4* is shown to be mostly upregulated in cytotoxic T cell in blood of asthma (log₂-fold change = 1.7, adjusted P=4.7x10⁻⁵) and downregulated in epidermal keratinocyte cell in skin tissue of psoriasis vulgaris (log₂-fold change = -3.7, adjusted P=2.4x10⁻¹⁸) (Supplementary figure 3, Supplementary table 4). *DNAJA4* is broadly expressed in bone marrow, brain, heart, skeletal muscle tissue and other tissues (Supplementary figure 4). In single cell data, *DNAJA4* is most expressed in germ cell, with 84.5% of sperm cells expressing this gene (Supplementary table 5).

***SLC7A1* (Solute Carrier Family 7 Member 1)**

In genetic prioritization ranking with other potential related diseases, *CDC7* showed evidence of association with hypertension, with a significant positive MR (MR beta =

0.11~0.28, q -value < 0.05) association and high colocalization (maximum PP.H4 = 0.99) in adrenal gland (Supplementary table 3). In differential gene expression analysis, *SLC7A1* is shown to be mostly upregulated in epidermal keratinocyte cell in skin tissue of psoriasis vulgaris (log₂-fold change = 1.6, adjusted $P=2.1 \times 10^{-17}$) and downregulated in the PBMC in relapsing-remitting MS (log₂-fold change = -1.9, adjusted $P=9.9 \times 10^{-21}$) (Supplementary figure 3, Supplementary table 4). *SLC7A1* is ubiquitously expressed in mucosa tissue in, transformed skin fibroblast, EBV-Transformed lymphocyte and other tissues (Supplementary figure 4). In single cell data, *SLC7A1* is most expressed in distal lung epithelial progenitor cell, with 71.2% of cells expressing this gene, followed by syncytiotrophoblast cell (65.8%) (Supplementary table 5).

***TLL10* (Tubulin Tyrosine Ligase Like 10)**

In genetic prioritization ranking with other potential related diseases, *TLL10* showed evidence of association with asthma, with the instrument variant of MR being a lead coding variant of this gene, a significant positive MR association and high colocalization (MR beta = 0.07, maximum PP.H4 = 0.99) in Brain Putamen basal ganglia cell, and a significant negative MR association and high colocalization in iPSC (MR beta = -0.07, maximum PP.H4 = 0.98, Supplementary table 3). It also has moderate evidence of association with hypothyroidism, with positive MR association (MR beta = 0.38) and high colocalization (PP.H4 = 0.89) in whole blood. In differential gene expression analysis, *TLL10* is shown to be mostly upregulated in skin tissue of psoriasis (log₂-fold change = 1.3, adjusted $P=4.4 \times 10^{-2}$) and downregulated in plasma blast cell of systemic lupus erythematosus (log₂-fold change = -1.8, adjusted $P=3.7 \times 10^{-3}$) (Supplementary figure 3, Supplementary table 4). In baseline gene expression, *TLL10* is mostly expressed in the testis tissue (Supplementary figure 4). In single cell data, *TLL10* is most expressed in sperm, with 74.5% of cells expressing this gene (Supplementary table 5).

Evaluation of known acne risk *loci* in severe versus normal acne comparisons

We extracted a total of 66 unique SNP associations from all 5 GWASs for the 50 reported acne risk *loci*. Among the 66 SNPs, 17 (26%) were nominally significant (P -value < 0.05) in any of the current GWASs, with the lowest P -value being 3.6×10^{-3} at

rs17692425 in FinnGen GWAS; however, no SNP remained significant after Bonferroni correction (Supplementary table 6, figure 2).

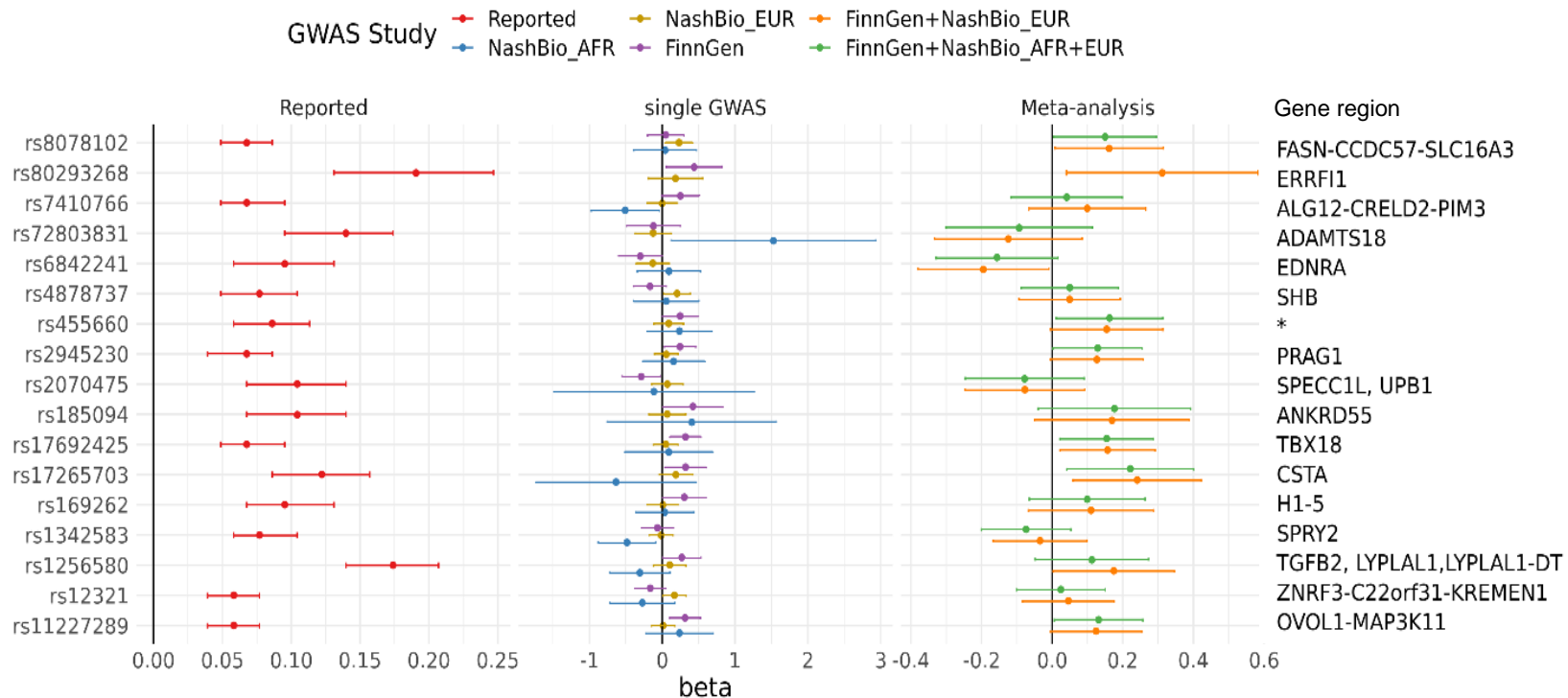


Figure 2. SNP associations that passed nominal significance level in any of the 5 severe acne GWASs at previously reported leading risk variant at acne risk *loci*.

Polygenic risk score analysis

We evaluated if a polygenic risk score (PRS) constructed with reported lead risk variants of acne at genome-wide significant *loci* and weights also associate with acne severity. The PRS score is statistically significantly associated with severe acne in FinnGen population (OR = 1.77, 95% CI: 1.37 - 2.29, P-value = 1.4×10^{-5}), but not associated in EA (OR = 1.22, 95% CI: 0.97 - 1.52, P-value = 0.08) or AA (P-value = 0.99) population in BioVU (Table 4, Figure 3).

Table 4: Associations between PRS and severe acne risk by population.

Study	Mean PRS in cases	Mean PRS in controls	OR (95%CI)	P-value
FinnGen	5.61	5.45	1.77 (1.37 - 2.29)	1.40×10^{-5}
BioVU EA	5.24	5.19	1.22 (0.97 - 1.52)	8.40×10^{-2}
BioVU AA	5.28	5.29	1.00 (0.52 - 1.91)	0.99
BioVU EA + FinnGen			1.46 (1.01 - 2.11)	4.37×10^{-2}
BioVU EA + AA + FinnGen			1.37 (1.00 - 1.88)	4.85×10^{-2}

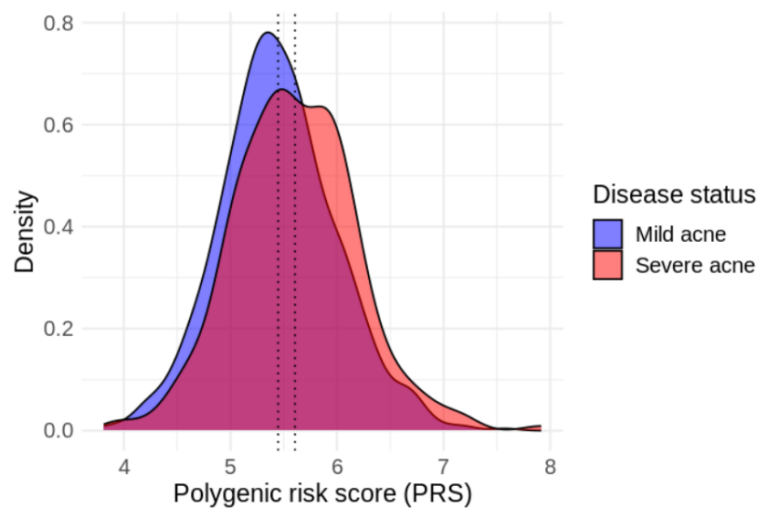


Figure 3. Distribution of acne PRS in FinnGen individuals with severe (red color) or mild (blue color) acne. Dotted vertical lines represent average values of PRS among each group.

Discussion

In this study, we identified one novel genome-wide significant locus associated with severe acne risk among African ancestry population in the BioVU cohort. Leveraging gene-level MR and colocalization analyses, we prioritized six novel genes that confer genetic risks for acne severity. We also identified four pathways that are enriched in severe acne genetic risk using gene-set analysis. Furthermore, we showed that the vast majority of the reported risk alleles for acne onset could not capture genetic risk for acne severity; however, we observed that a PRS constructed by aggregation of these risk variants has significant positive association with acne severity among European ancestry population but not in African ancestry population, possibly due to lower statistical power or lack of portability of PRS across populations^{11,12}.

The genome-wide significant SNP rs115325598 is located at the intron region of gene *ATXN7L1*. *ATXN7L1* is poorly characterized functionally. An important paralog of this gene is *ATXN7*, which acts as component of the STAGA transcription coactivator-HAT complex, mediates the interaction of STAGA complex with the CRX and is involved in CRX-dependent gene activation^{13,14}, and associates with microtubules and stabilizes the cytoskeletal network¹⁵. In our integrated differential gene expression analysis, this gene is demonstrated to be upregulated in skin tissue of atopic dermatitis patients and downregulated in epidermal keratinocyte cell in skin tissue of psoriasis vulgaris patients. Through phenome-wide search for associated diseases using our internal gene prioritization scheme, we found this gene has high evidence of association with type 2 diabetes. This is in consistent with previous studies which have shown that both acne vulgaris and type 2 diabetes mellitus are members of the family of mTORC1-driven metabolic diseases¹⁶. We identified another gene, *ADAM23*, associated with acne severity through our MR and colocalization analysis, and upregulated in keratinocytes in psoriasis patients. *ADAM23* is a non-catalytic metalloprotease-like protein involved in cell-cell and cell-matrix interaction. This finding highlights the role of skin barrier dysfunction in acne pathophysiology¹⁷, not only in term of disease onset but also as a factor driving disease severity and progression.

Our study only replicated a small proportion of the known acne susceptibility *loci* at the nominal significance level. The disparity may reflect differences in the biological mechanisms involved in acne onset and acne severity, but could also be influenced by the population-specific genetic determinants, case and control recruitment definitions, and lack of sufficient statistical power. Of note, the polygenic risk score which comprised of aggregated effects of all risk variants is demonstrated to be able to stratify severe versus mild acne, with each one unit increase in PRS is associated with a 1.77- and 1.22- fold elevated risk of developing severe acne. This observation is in concordance of the previous study which showed that the acne PRS has the greatest predictive ability in individuals with severe acne⁶, which implicates the potential application of using a common acne PRS to identify population with high risk of developing severe acne.

One limitation of our study is the small sample size, which makes it underpowered for genome-wide level discovery and replication of the known risk alleles, especially for less common variants with small effect size. Nevertheless, our analysis represents the first study to evaluate the association between genetic variants and acne severity among African ancestry populations and enabled a discovery of a novel genome-wide risk variant. The second limitation of this study is that we defined severe acne cases and controls based on EHR data including ICD codes, frequency of clinical visit, and medication records, rather than using a global acne grading system score (GAGS)¹⁸, which may cause some selection bias and may impact on the ability to detect true associations. Another limitation is the absence of replication in an independent cohort and functional validation. Future studies are needed to gather more resources to validate our findings.

In summary, our study shows that genetic risk *loci* driving acne onset could not effectively capture genetic risk of acne severity. Disease progression GWAS could contribute to unveiling novel biological mechanisms and may shed light on the discovery of potential therapeutic targets.

Methods

Study sample and GWAS

The study participants were from two independent biobanks: the BioVU biorepository extracted by Nashville Biosciences (NashBio), a subsidiary of the Vanderbilt University Medical Center (VUMC) and FinnGen (release 10).

The BioVU has been described previously¹⁹. Individuals included in this study were BioVU subjects with non-compromised DNA samples available and were previously genotyped on the Illumina Infinium Multiethnic Genotyping Array (MEGAEX) platform. We first selected all subjects with one or more instances of ICD-9/ICD-10 code for acne (L70.0, 706.1) in their electronic medical records (EHR). Among them, severe acne cases were defined as individuals who meet at least one of the following criteria: 1) subjects with a record of isotretinoin, 2) subjects with five or more instances of ICD-9/ICD-10 code of acne in their EHR, or 3) subjects with evidence of treatment with antibiotics for a duration of more than 60 days. Individuals with non-severe acne records and with similar ethnicities were selected as controls. After quality control (QC), a total of 522 severe acne cases, including 449 individuals of European ancestry (EA) and 73 individuals of African ancestry (AA), and 1,214 non-severe acne controls (970 EA individuals and 244 AA individuals) were included in the analysis. Per-allele odds ratios (ORs) and P-value of each variant were estimated using unconditional logistic regression, adjusted for age, sex, and the first 10 principal components (PC).

A detailed description of the FinnGen biobank has been published previously²⁰. Severe acne cases were defined as individuals who is elder than five years old, with at least five instances of acne diagnosis records (defined by 7061A, 70610, L70.0, L12_ACNE) in their EHR, or with at least five isotretinoin treatment records. Non-severe acne controls were defined as individuals who is elder than five years old, with one to four instances of acne diagnosis in their EHR and with at most four isotretinoin treatment records. After QC, a total of 290 severe acne cases and 908 non-severe acne controls were included in the analysis. Logistic regressions were first performed to determine associated genotyping batches with severe acne. GWAS analysis was then performed

using the REGENIE pipeline in FinnGen sandbox²¹, adjusting for age, sex, the first 10 PCs, and the identified associated genotyping batches.

Fixed effect meta-analyses with inverse variance weights were performed using the METAL software²² to combine GWAS results from BioVU and FinnGen. Meta-analyses were performed both for EA population only (BioVU EA with FinnGen) and multi-ethnic population (BioVU EA, AA with FinnGen).

Post-GWAS analyses

Variant annotation

We annotated the impact of all variants with $P < 1 \times 10^{-6}$ using the variant effect predictor (VEP v102)²³ with the following options: --everything --offline --check_existing. Coding variants were defined as those impacting protein coding transcript annotated as missense variant or predicted to have “high” impact. We also retrieved predicted gain or loss of function variants⁶. We also mapped variants to genes using activity-by-contact (ABC) enhancer-promoter interactions¹⁰. We used “ABCmax” to represent variant-gene pairs with the highest ABC score.

Mendelian randomization and colocalization

We conducted mendelian randomization (MR) using the TwoSampleMR R package²⁴. The following exposures were considered: protein quantitative trait loci (pQTL) from Sun et al²⁵, and expression quantitative trait loci from Blueprint²⁶, eQTLGen²⁷ and other datasets from the EBI eQTL catalogue^{26–41}. For each of those studies, variants with a MAF < 1% were excluded. Variants were clumped using PLINK2⁴² using the options –clump-p1 1 –clump-p2 1 –clump-r2 0.01 –clump-kb 10000. Only genes 250kb around significant *loci* were considered in this analysis. For each QTL, independent variants with $P < 1 \times 10^{-4}$ were used as instruments. When more than one instrument was present, the inverse variant weighted approach was used, otherwise the Wald Ratio approach was used. False Discovery Rate (FDR) corrections were applied to adjust *P-values* in multiple testing, and a threshold of FDR < 0.05 was used to determine significance. For genes with significant MR results, we performed colocalization analyses to detect shared causal SNPs between molecular-QTL and acne severity using the coloc R package with default priors⁴³, using a region of 250kb around the local lead GWAS

variant. We set a threshold of colocalization posterior probability at H4 ($PP.H4 \geq 0.8$), and genes with both significant MR and colocalization were considered as potential targeted molecular. Only autosomes were included in this analysis.

Gene-set analyses using MAGMA

We performed gene-set level analyses using MAGMA (v1.10)⁴⁴. SNP annotation to genes was done using the 1000 Genomes Phase 3 European panel and GRCh38 (downloaded from <http://ctglab.nl/software/magma>). Gene sets were downloaded from the Molecular Signatures Database (MSigDB)^{45,46}, including 7763 gene sets in Gene Ontology (GO), 186 in Kyoto Encyclopedia of Genes and Genomes (KEGG), 292 in BioCarta, 1,635 in Reactome and 4,872 in ImmuneSigDB.

Associations of identified genes with other phenotypes.

For each severe acne related gene identified from GWAS (nearest gene of the genome-wide significant risk variant), MR and Colocalization, we performed a phenome-wide search to identify other diseases that showed evidence of association with the focal severe acne-related gene. We employed an in-house gene prioritization scheme, which contained precalculated results across 6,918 GWAS from 4 sources, including UK Biobank, FinnGen release 10 (R10), EBI GWAS catalog, and meta-analyses of UK Biobank, FinnGen R10, and Estonian biobank [ref: GiTIPs paper]. Briefly, this gene-ranking scheme used a combinatory approach to prioritize disease putative genes based on MR, eQTL and pQTL colocalization, ABC enhancer-promoter interactions, and variant annotations. This approach mitigates the pleiotropic effect of eQTL while retaining important information about directionality and has been shown to be enriched for “gold standard” genes and drug targets with successful clinical trial [ref: GiTIPs paper].

Differential expression in other diseases

We further examined whether the identified acne severity related genes are differentially expressed in diseased compared to normal tissues of other diseases. Differential gene expressions (DGE) were extracted from integrated transcriptomics service from Qiagen Omicsoft database (QIAGEN OmicSoft DiseaseLand⁴⁷). We extracted both microarray and RNA-seq based transcriptomic (TxP) studies and disease comparisons/contrasts available for all studies in the DiseaseLand in the immunology and inflammation,

neurodegenerative and metabolic diseases⁴⁸. Common ontologies were applied to all samples, tissues, diseases, treatments, and comparison types to systematize database curation. Database raw data was derived from published repositories including gene expression omnibus (GEO) and ArrayExpress (<https://www.ebi.ac.uk/biostudies/arrayexpress>). In total, the integrated database consists of >250,000 samples with TxP profiles, 32,000 study comparison contrasts in ~700 disease areas. The contrasts that were of interest to us are “Disease.vs.Normal” which we subsequently used to perform meta-analysis.

Several studies have delved into the exploration of integrated transcriptomics data through various meta-analyses^{48,49}. We opted for the Fisher's exact test for its straightforward implementation, heightened sensitivity to small sample sizes, non-parametric nature, and platform independence across transcriptomics techniques (such as microarray and RNA-seq). Overall, meta-analysis examined whether a gene_i is a recurrent DGE enriched for a given disease, tissue, or sample (e.g.: Alzheimer's disease and brain tissue) when compared to recurrency of the gene in the entire database. We constructed a contingency table as outlined in the methodology suggested earlier by Wang et al.⁵⁰ where we used a cut-off of adjusted P values ($P_{adj} < 0.05$) & $\text{abs}(\log_2\text{Fold-change}) > 1.0$ to define a DGE, and at least 70% of studies having the same direction of log₂Fold-Change (positive or negative). A recurrent DGE enriched across different studies increases the confidence of the gene in playing a role in the disease mechanism of action (MoA).

Gene expression levels by cell types and tissues

We then assessed the expression levels of the identified acne severity related genes by tissues and cell types. Baseline gene expression data were retrieved from GTEx and Expression Atlas (<http://www.ebi.ac.uk/gxa>)⁵¹, and single-cell transcriptomics were obtained from the BioTuring Single-cell gene expression database⁵².

Association testing of known risk regions.

We examined whether the previously reported genetic risk *loci* of acne also contribute to severe acne risk. Directional consistency of effect was defined as ORs in the current GWASs that were in the same direction of effect (ie, >1) as those reported in the

discovery population. A nominal P value of .05 was used to determine statistical significance.

Polygenic risk score analysis.

The aggregate effect of known risk alleles was examined using a weighted polygenic risk score (PRS) for each individual: $PRS_i = \sum_{m=1}^C \beta_m g_{im}$, where g_{im} is the risk allele dosage for individual i at SNP m , C defines risk SNPs at 66 known acne susceptibility *loci*, and β_m is the weight for SNP m . The association between PRS and severe acne risk was examined separately in each dataset using logistic regression, adjusting for the first 10 PCs, age, and sex. The results were combined using a fixed-effect meta-analysis with inverse variance weights.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Patients and control subjects in FinnGen provided informed consent for biobank research, based on the Finnish Biobank Act. Alternatively, separate research cohorts, collected prior the Finnish Biobank Act came into effect (in September 2013) and start of FinnGen (August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by Fimea, the National Supervisory Authority for Welfare and Health. Recruitment protocols followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol Nr HUS/990/2017. The FinnGen study is approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020, and THL/2364/14.02/2020), Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3), the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020 and Statistics Finland (permit numbers: TK-53-1041-17 and TK-53-90-20). The Biobank Access Decisions for FinnGen samples and data utilized in FinnGen Data Freeze 6 include: THL Biobank BB2017_55, BB2017_111, BB2018_19, BB_2018_34, BB_2018_67, BB2018_71, BB2019_7, BB2019_8, BB2019_26, BB2020_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland_2017_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017, and Terveystalo Biobank STB 2018001. UK Biobank has received ethical approval from the NHS National Research Ethics Service North West (approval numbers 11/NW/0382 and 16/NW/0274). All participants provided written informed consent.

DATA AVAILABILITY

All data produced in the present study are available upon reasonable request to the authors.

ACKNOWLEDGEMENT

FinnGen

We want to acknowledge the participants and investigators of FinnGen study. The FinnGen project is funded by two grants from Business Finland (HUS 4685/31/2016 and UH 4386/31/2016) and the following industry partners: AbbVie Inc., AstraZeneca UK Ltd, Biogen MA Inc., Bristol Myers Squibb (and Celgene Corporation & Celgene International II Sàrl), Genentech Inc., Merck Sharp & Dohme LCC, Pfizer Inc., GlaxoSmithKline Intellectual Property Development Ltd., Sanofi US Services Inc., Maze Therapeutics Inc., Janssen Biotech Inc, Novartis Pharma AG, and Boehringer Ingelheim International GmbH. Following biobanks are acknowledged for delivering biobank samples to FinnGen: Auria Biobank (www.auria.fi/biopankki), THL Biobank (www.thl.fi/biobank), Helsinki Biobank (www.helsinginbiopankki.fi), Biobank Borealis of Northern Finland (<https://www.ppsHP.fi/Tutkimus-ja-opetus/Biopankki/Pages/Biobank-Borealis-briefly-in-English.aspx>), Finnish Clinical Biobank Tampere (www.tays.fi/en-US/Research_and_development/Finnish_Clinical_Biobank_Tampere), Biobank of Eastern Finland (www.ita-suomenbiopankki.fi/en), Central Finland Biobank (www.ksshP.fi/fi-FI/Potilaille/Biopankki), Finnish Red Cross Blood Service Biobank (www.veripalvelu.fi/verenluovutus/biopankkitoiminta), Terveystalo Biobank (www.terveystalo.com/fi/Yritystietoa/Terveystalo-Biopankki/Biopankki/) and Arctic Biobank (<https://www.oulu.fi/en/university/faculties-and-units/faculty-medicine/northern-finland-birth-cohorts-and-arctic-biobank>). All Finnish Biobanks are members of BBMRI.fi infrastructure (www.bbmri.fi). Finnish Biobank Cooperative -FINBB (<https://finbb.fi/>) is the coordinator of BBMRI-ERIC operations in Finland. The Finnish biobank data can be accessed through the Fingenious® services (<https://site.fingenious.fi/en/>) managed by FINBB.

BioVU

The datasets used for the analyses described were obtained from Vanderbilt University Medical Center's BioVU, which are supported by numerous sources: institutional funding, private agencies, and federal grants. These include NIH funded Shared Instrumentation Grant S10OD017985, S10RR025141, and S10OD025092; CTSA

grants UL1TR002243, UL1TR000445, and UL1RR024975. Genomic data are also supported by investigator-led projects that include U01HG004798, R01NS032830, RC2GM092618, P50GM115305, U01HG006378, U19HL065962, and R01HD074711.

UK Biobank

This research has been conducted using the UK Biobank Resource (www.ukbiobank.ac.uk), a large-scale biomedical database and research resource containing genetic, lifestyle and health information from 500,000 UK participants. UK Biobank is supported by its founding funders the Wellcome Trust and UK Medical Research Council, as well as the Department of Health, Scottish Government, the Northwest Regional Development Agency, British Heart Foundation and Cancer Research UK. The UK biobank pan-ancestry analysis was conducted under project ID 31063 (<https://pan.ukbb.broadinstitute.org>).

Estonian Biobank

Estonian Biobank research was supported by the European Union through Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008, by the Estonian Research Council grant PUT (PRG1291, PRG687 and PRG184) and by the European Union through the European Regional Development Fund project no. MOBERA21 (ERA-CVD project DETECT ARRHYTHMIAS, GA no JTC2018-009), Project No. 2014-2020.4.01.15-0012 and Project No. 2014-2020.4.01.16-0125.

COMPETING INTERESTS

All authors were employees of Sanofi US Services at the time of study and hold shares and/or stock options in the company. All authors declare no other competing interests.

FUNDING

This study was funded by Sanofi (Cambridge, MA, United States). The funder had the following involvement with the study: Sanofi reviewed the manuscript.

References

1. Williams, H. C., Dellavalle, R. P. & Garner, S. Acne vulgaris. *The Lancet* **379**, 361–372 (2012).
2. Katsambas, A. & Dessinioti, C. New and emerging treatments in dermatology: acne. *Dermatol. Ther.* **21**, 86–95 (2008).
3. Misery, L. Consequences of Psychological Distress in Adolescents with Acne. *J. Invest. Dermatol.* **131**, 290–292 (2011).
4. Bataille, V., Snieder, H., MacGregor, A. J., Sasieni, P. & Spector, T. D. The Influence of Genetics and Environmental Factors in the Pathogenesis of Acne: A Twin Study of Acne in Women. *J. Invest. Dermatol.* **119**, 1317–1322 (2002).
5. He, L. *et al.* Two new susceptibility loci 1q24.2 and 11p11.2 confer risk to severe acne. *Nat. Commun.* **5**, 2870 (2014).
6. Mitchell, B. L. *et al.* Genome-wide association meta-analysis identifies 29 new acne susceptibility loci. *Nat. Commun.* **13**, 702 (2022).
7. Navarini, A. A. *et al.* Genome-wide association study identifies three novel susceptibility loci for severe Acne vulgaris. *Nat. Commun.* **5**, 4020 (2014).
8. Petridis, C. *et al.* Genome-wide meta-analysis implicates mediators of hair follicle development and morphogenesis in risk for severe acne. *Nat. Commun.* **9**, 5075 (2018).
9. Teder-Laving, M. *et al.* Genome-wide meta-analysis identifies novel loci conferring risk of acne vulgaris. *Eur. J. Hum. Genet.* 1–8 (2023) doi:10.1038/s41431-023-01326-8.
10. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* **593**, 238–243 (2021).

11. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
12. Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat. Commun.* **11**, 3865 (2020).
13. Palhan, V. B. *et al.* Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8472–8477 (2005).
14. Zhao, Y. *et al.* A TFTC/STAGA Module Mediates Histone H2A and H2B Deubiquitination, Coactivates Nuclear Receptors, and Counteracts Heterochromatin Silencing. *Mol. Cell* **29**, 92–101 (2008).
15. Nakamura, Y. *et al.* Ataxin-7 associates with microtubules and stabilizes the cytoskeletal network. *Hum. Mol. Genet.* **21**, 1099–1110 (2012).
16. Melnik, B. C. Acne vulgaris: The metabolic syndrome of the pilosebaceous follicle. *Clin. Dermatol.* **36**, 29–40 (2018).
17. Thiboutot, D. & Del Rosso, J. Q. Acne Vulgaris and the Epidermal Barrier. *J. Clin. Aesthetic Dermatol.* **6**, 18–24 (2013).
18. Bernardis, E. *et al.* Development and Initial Validation of a Multidimensional Acne Global Grading System Integrating Primary Lesions and Secondary Changes. *JAMA Dermatol.* **156**, 296–302 (2020).
19. Roden, D. *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).
20. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).

21. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
22. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
23. Hunt, S. E. *et al.* Annotating and prioritizing genomic variants using the Ensembl Variant Effect Predictor—A tutorial. *Hum. Mutat.* **43**, 986–997 (2022).
24. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408.
25. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
26. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).
27. Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53**, 1300–1310 (2021).
28. Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat. Genet.* **53**, 1290–1299 (2021).
29. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
30. Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).

31. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).
32. Ng, B. *et al.* An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
33. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).
34. Bunt, M. van de *et al.* Transcript Expression Data from Human Islets Links Regulatory Signals from Genome-Wide Association Studies for Type 2 Diabetes and Glycemic Traits to Their Downstream Effectors. *PLOS Genet.* **11**, e1005694 (2015).
35. Taylor, D. L. *et al.* Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10883–10888 (2019).
36. Panopoulos, A. D. *et al.* iPSCORE: A Resource of 222 iPSC Lines Enabling Functional Characterization of Genetic Variation across a Variety of Cell Types. *Stem Cell Rep.* **8**, 1086–1100 (2017).
37. Pashos, E. E. *et al.* Large, Diverse Population Cohorts of hiPSCs and Derived Hepatocyte-like Cells Reveal Functional Genetic Variation at Blood Lipid-Associated Loci. *Cell Stem Cell* **20**, 558-570.e10 (2017).
38. Hoffman, G. E. *et al.* CommonMind Consortium provides transcriptomic and epigenomic data for Schizophrenia and Bipolar Disorder. *Sci. Data* **6**, 180 (2019).
39. Guelfi, S. *et al.* Regulatory sites for splicing in human basal ganglia are enriched for disease-relevant information. *Nat. Commun.* **11**, 1041 (2020).

40. Young, A. M. H. *et al.* A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
41. Schwartzentruber, J. *et al.* Molecular and functional variation in iPSC-derived sensory neurons. *Nat. Genet.* **50**, 54–61 (2018).
42. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
43. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).
44. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).
45. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
46. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
47. 'Omics Disease Databases | QIAGEN Digital Insights. *Bioinformatics Software | QIAGEN Digital Insights* <https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/content-exploration-and-databases/qiagen-diseaseland/>.
48. Wu, C. *et al.* Integrative Analysis of DiseaseLand Omics Database for Disease Signatures and Treatments: A Bipolar Case Study. *Front. Genet.* **10**, (2019).

49. Chang, L.-C., Lin, H.-M., Sibille, E. & Tseng, G. C. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics* **14**, 368 (2013).
50. Wang, Y.-Y. *et al.* CeDR Atlas: a knowledgebase of cellular drug response. *Nucleic Acids Res.* **50**, D1164–D1171 (2022).
51. Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **46**, D246–D251 (2018).
52. Luu, T. *et al.* BBrowser: Making single-cell data easily accessible.