Quraishi et al. Clinical Epigenetics (2015) 7:68
DOI 10.1186/s13148-015-0108-y

CLINICAL EPIGENETICS

RESEARCH Open Access

# Identifying CpG sites associated with eczema via random forest screening of epigenome-scale DNA methylation

B. M. Quraishi[1], H. Zhang[1*], T. M. Everson[2], M. Ray[1], G. A. Lockett[3], J. W. Holloway[3,4], S. R. Tetali[1], S. H. Arshad[4,5], A. Kaushal[1], F. I. Rezwan[6] and W. Karmaus[1]

## Abstract

**Background:** The prevalence of eczema is increasing in industrialized nations. Limited evidence has shown the association of DNA methylation (DNA-M) with eczema. We explored this association at the epigenome-scale to better understand the role of DNA-M.

Data from the first generation ($F_1$) of the Isle of Wight (IoW) birth cohort participants and the second generation ($F_2$) were examined in our study. Epigenome-scale DNA methylation of $F_1$ at age 18 years and $F_2$ in cord blood was measured using the Illumina Infinium HumanMethylation450 Beadchip. A total of 307,357 cytosine-phosphate-guanine sites (CpGs) in the $F_1$ generation were screened via recursive random forest (RF) for their potential association with eczema at age 18. Functional enrichment and pathway analysis of resulting genes were carried out using DAVID gene functional classification tool. Log-linear models were performed in $F_1$ to corroborate the identified CpGs. Findings in $F_1$ were further replicated in $F_2$.

**Results:** The recursive RF yielded 140 CpGs, 88 of which showed statistically significant associations with eczema at age 18, corroborated by log-linear models after controlling for false discovery rate (FDR) of 0.05. These CpGs were enriched among many biological pathways, including pathways related to creating transcriptional variety and pathways mechanistically linked to eczema such as cadherins, cell adhesion, gap junctions, tight junctions, melanogenesis, and apoptosis. In the $F_2$ generation, about half of the 83 CpGs identified in $F_1$ showed the same direction of association with eczema risk as in $F_1$, of which two CpGs were significantly associated with eczema risk, cg04850479 of the *PROZ* gene (risk ratio (RR) = 15.1 in $F_1$, 95 % confidence interval (CI) 1.71, 79.5; RR = 6.82 in $F_2$, 95 % CI 1.52, 30.62) and cg01427769 of the *NEU1* gene (RR = 0.13 in $F_1$, 95 % CI 0.03, 0.46; RR = 0.09 in $F_2$, 95 % CI 0.03, 0.36).

**Conclusions:** Via epigenome-scaled analyses using recursive RF followed by log-linear models, we identified 88 CpGs associated with eczema in $F_1$, of which 41 were replicated in $F_2$. Several identified CpGs are located within genes in biological pathways relating to skin barrier integrity, which is central to the pathogenesis of eczema. Novel genes associated with eczema risk were identified (e.g., the *PROZ* and *NEU1* genes).

**Keywords:** Eczema, Allergic disease, DNA methylation, Epigenome-scale, Epigenetics, Random forest, $F_1$ and $F_2$ generations, CpG

* Correspondence: hzhang6@memphis.edu
[1]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, 236A Robison Hall, Memphis, TN 38152, USA
Full list of author information is available at the end of the article

BioMed Central

© 2015 Quraishi et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

## Background

The increasing prevalence of allergic conditions including eczema is a major public health concern in industrialized nations [1]. The prevalence of eczema is reported to be 10–30 % in children and 1–3 % in adults of the developed world [2]. In addition to the physical discomfort to the affected individual and the social burden on their families, eczema has a huge economic impact on nations' health care budgets [3].

Eczema is a chronic condition involving a complex interplay of genetic, epigenetic, and environmental factors [4–6]. So far, DNA methylation (DNA-M) remains the most studied mechanism with potential answers to epigenetic regulation of gene function [7, 8]. The Illumina Infinium HumanMethylation450 Beadchip has the ability to measure DNA methylation at more than 450 K cytosine-phosphate-guanine sites (CpGs), which provides rich information for various epigenetic studies. Epigenome-scale studies offer an immense opportunity to understand disease pathophysiology, but there are also concerns about the challenges associated with this type of studies. A recent review published in 2014 by Paul et al. highlighted the potential challenges in the field of epigenomics [9] such as study design, methodologies of obtaining biologic samples, high dimensionality, and highly correlated data [9, 10].

Random forest (RF) is a machine learning algorithm used for classification and has the ability to efficiently handle high dimensionality and highly correlated data [11]. The R package was used in this study to screen CpG sites potentially associated with eczema. RF is composed of classification trees with each tree constructed using randomly selected bootstrap samples. Misclassification rates calculated based on testing samples can be used to estimate the accuracy of the forests.

In this study, we utilized a method built upon RF to screen specific CpGs potentially associated with eczema using data in the first generation ($F_1$) at age 18 years and functionally annotated the genes of the identified CpGs using DAVID [12] to understand the biological pathways. For the identified CpGs via the RF-based method, we further examined their statistical significance on their linear association with eczema risk at age 18 years using log-linear models and replicated the findings from the $F_1$ in the second generation ($F_2$).

## Results

Eczema frequencies in $F_1$ (18 years) and in $F_2$ (3, 6, and 12 months) generations stratified by sex indicated that females had higher eczema prevalence than males at 18 years of age in the $F_1$ generation, and the prevalence switched in the newborns of the $F_2$ generation (Table 1). This is consistent with the gender-reversal pattern of eczema reported in our earlier work [13].

**Table 1** Eczema status in male and female cohort participants in the $F_1$ and $F_2$ generations (chi-square tests)

| $F_1$ generation | | | | |
|---|---|---|---|---|
| Independent variables | | Females | Males | Chi-square |
| | | (*n* = 244) | (*n* = 122) | *P* value |
| Eczema status | Yes | 37 (15.2 %) | 9 (7.3 %) | 0.051 |
| | No | 207 (84.8 %) | 113 (92.6 %) | |
| $F_2$ generation | | | | |
| Independent variables | | Boys | Girls | Chi-square |
| | | (*n* = 60) | (*n* = 56) | *P* value |
| Age 3 months | Yes | 9 (15.0 %) | 2 (3.6 %) | 0.048 |
| Eczema status | No | 44 (73.3 %) | 53 (94.6 %) | |
| | Missing | 7 (11.7 %) | 1 (1.8 %) | |
| Age 6 months | Yes | 13 (21.7 %) | 6 (10.7 %) | 0.162 |
| Eczema status | No | 39 (65.0 %) | 43 (76.8 %) | |
| | Missing | 8 (13.3 %) | 7 (12.5 %) | |
| Age 12 months | Yes | 9 (15.0 %) | 5 (8.9 %) | 0.521 |
| Eczema status | No | 37 (61.7 %) | 36 (64.3 %) | |
| | Missing | 14 (23.3 %) | 15 (26.8 %) | |

In the screening process using recursive RF [14, 15], the parameters (*sampsize*, *mtry*, and *ntree*—details are in the "Statistical analysis" section) in the *randomForest*() R package were selected to achieve stabilized error rates. In total, pre-processed DNA methylation of 307,357 CpGs in the $F_1$ generation was included in the screening. The results of the recursive RF (Table 2, Fig. 1; details in the "Statistical Analysis" section) Indicated that a total of 140 CpGs (after excluding 8 CpGs located on the X chromosome) passed the screening showing potential association with eczema. The exclusion of the 8 CpGs were due to the potential bias measurement of DNA-M for different genders. Nevertheless, in the following analyses (log-linear models below), we assessed whether gender played a role in the association of DNA-M with eczema.

Further examination of these 140 CpGs from $F_1$ using log-linear models indicated that 88 out of 140 CpGs had a statistically significant linear association with eczema at age 18 (FDR-adjusted *P* value <0.05) (Additional file 1: Table S1). We also tested the statistical significance of the interaction between DNA-M and gender; none of the FDR-adjusted *P* values were <0.05.

We assessed the biological pathways enriched within the genes annotated to those 140 CpGs using DAVID (Table 3). The most significantly enriched pathways related to the creation of transcriptional variety through genetic (e.g., polymorphism) and regulatory (e.g., alternative splicing) mechanisms. The remainder of the significantly enriched pathways included several pathways mechanistically linked to epithelial barrier integrity and

**Table 2** The performance of recursive RF at each iteration

| Iteration | Number of CpGs | (OOB-ER) Overall misclassification | Eczema misclassification | Non-eczema misclassification |
|---|---|---|---|---|
| 1 | 307,357 | 18.6 % | 95.7 % | 7.5 % |
| 2 | 153,678 | 15.3 % | 82.6 % | 5.6 % |
| 3 | 76,838 | 18.6 % | 87.0 % | 8.8 % |
| 4 | 38,419 | 16.1 % | 65.2 % | 9.1 % |
| 5 | 19,208 | 17.8 % | 80.4 % | 8.8 % |
| 6 | 9604 | 14.2 % | 78.3 % | 5.0 % |
| 7 | 4802 | 12.3 % | 58.7 % | 5.6 % |
| 8 | 2401 | 10.7 % | 52.2 % | 4.7 % |
| 9 | 1200 | 7.9 % | 37.0 % | 3.8 % |
| 10 | 599 | 6.8 % | 26.1 % | 4.1 % |
| 11 | 298 | 6.6 % | 30.4 % | 3.1 % |
| 12[a] | 148 | 5.2 % | 17.4 % | 3.4 % |
| 13 | 74 | 6.3 % | 19.6 % | 4.4 % |
| 14 | 37 | 9.3 % | 21.7 % | 7.5 % |
| 15 | 18 | 8.5 % | 26.1 % | 5.9 % |
| 16 | 9 | 10.7 % | 19.6 % | 9.4 % |
| 17 | 3 | 16.9 % | 28.2 % | 15.3 % |

*OOB-ER* out of bag error rate
[a]The 12th iteration had the lowest misclassification error rate

cell adhesion, which are of key importance in eczema: examples include cadherins (protocadherin gamma, $P = 1.8 \times 10^{-16}$; cadherin 6 domain, $P = 1.5 \times 10^{-10}$; and cadherin N-terminal domain, $P = 3.6 \times 10^{-10}$), gap junctions ($P = 2.6 \times 10^{-6}$), cell adhesion (cell-cell adhesion, $P = 2.7 \times 10^{-10}$; cell adhesion, $P = 2.2 \times 10^{-5}$), tight junctions ($P = 1.6 \times 10^{-5}$), melanogenesis ($P = 7.1 \times 10^{-5}$), and apoptosis ($P = 7.3 \times 10^{-3}$).

### Replication results

We then replicated the findings from the $F_1$ generation in the $F_2$ generation. In total, 83 out of the 88 CpGs identified in the $F_1$ were also present in the $F_2$ dataset (the 5 CpG sites in the F2 were excluded after quality control). DNA methylation at 41 CpGs (out of 83) showed the same direction of changes with eczema in both the $F_1$ and $F_2$ generations (Table 4, Fig. 2). Of these 41 CpGs, two were statistically significantly associated with eczema risk in both generations (Table 4); cg04850479 in the *PROZ* gene showed adjusted risk ratio (RR) of 15.19 (95 % confidence interval (CI) 1.71 to 79.50) in the $F_1$ and 6.82 (95 % CI 1.52 to 30.62) in the $F_2$ and cg01427769 in the *NEU1* gene showed adjusted RR of 0.13 (95 % CI 0.03 to 0.46) in the $F_1$ and 0.09 (95 % CI 0.03 to 0.36) in the $F_2$. We further assessed the association of DNA methylation of these 2 CpGs with corresponding gene expressions in the $F_2$ generation. No statistically significant associations were identified. Among the remaining CpGs not replicated in the F2 generation, about 60 % CpGs ($n = 25$ CpGs) showed a statistically significant difference in DNA methylation between the two generations (based on two sample two sided $t$ tests) after adjusting for multiple testing. Since some of the F2 generation are offsprings of subjects in the F1 generation, the findings tend to be conservative. The above analyses were adjusted for estimated cell type proportions [16].

## Discussion

This is the first study to explore epigenome-scale DNA methylation patterns associated with eczema. Using data from two generations, our study based on data of the $F_1$ generation identified CpGs potentially associated with eczema status using the RF technique, which was further corroborated via log-linear models. In total, 140 CpGs were identified via RF, which were further assessed using log-linear models with 88 CpGs being statistically significantly associated with eczema risk after adjusting for cell type proportions and controlling for multiple testing. The remaining 52 CpGs were not corroborated in log-linear models. This is likely due to two reasons. Firstly, the 140 CpGs were identified based on their importance values in terms of minimizing misclassification errors other than statistical testing [11]. It is possible that the identified CpG sites did not have a statistically significant main effect on eczema risk. Secondly, among the 140 CpGs, complex non-linear interactions are likely to exist between multiple CpGs which may be difficult to parametrically identify using log-linear models. Using $F_2$ generation data, around 50 % (41 CpGs) of these 88 CpGs identified in the $F_1$ generation were further replicated. In particular, two CpGs showed statistically significant results in both $F_1$ and $F_2$: cg04850479 in the *PROZ* gene and cg01427769 in the *NEU1* gene. Although some studies have linked *NEU1* gene with asthma [17] via Th2-mediated airway inflammation [18, 17], and it is known that the Th2 pathway is also important for eczema [19, 20], based on our knowledge, no study has so far spotted its role in eczema. The insignificant findings on the association of DNA methylation of cg04850479 (in the *PROZ* gene) and cg01427769 (in the *NEU1* gene) is likely due to tissue-specific gene expression. That is, an early exposure has left a change in methylation in all tissues including blood but the gene is not expressed in blood but skin for eczema. It is also possible that the DNA methylation of these two CpGs is related to the production of dysfunctional transcripts.

Enrichment analysis of the CpG sites identified in the $F_1$ generation highlighted pathways related to the creation of transcriptional variation and several biological pathways related to the epidermal barrier and involved in eczema (Table 3).

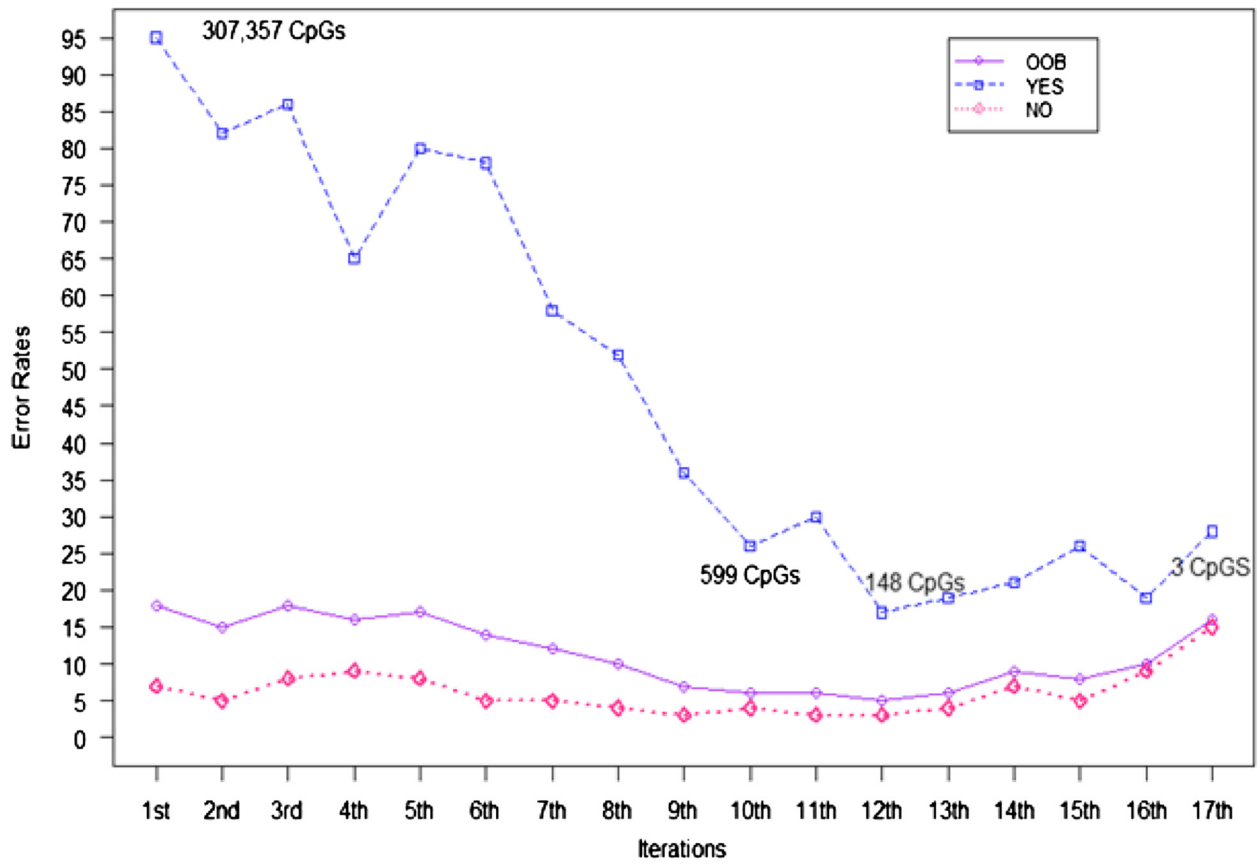


**Fig. 1** Misclassification error rates at each iteration of the recursive RF. *OOB* out of bag error rate (overall error), *YES* eczema, *No* non-eczema

The skin barrier is crucial in maintaining skin integrity, and disruption of the epidermal barrier is one of the important mechanisms in the pathogenesis of eczema [21, 22]. Several studies reported that skin barrier dysfunction is a result of the impairment of tight junction function in eczema patients [23–26]. Cadherins and protocadherins are transmembrane proteins important for cell-to-cell adhesion and epithelial integrity and have been associated with eczema and asthma in genetic studies [27]. Chronic eczema and several other dermatoses are also related to hyperpigmentation of the skin [28]. Our study detected differentially methylated CpGs within genes in pathways relating to epidermal barrier integrity and eczema pathogenesis, including cadherins, gap junction, cell adhesion, tight junction, melanogenesis, and apoptosis (Table 3). Their biological functions suggest these eczema-associated CpGs are of special interest, and they are potential epigenetic biomarkers for eczema. The detection of eczema-associated differential methylation within pathways already known to be associated with eczema is reasonable and suggests that epigenetic and genetic variation may work together to regulate eczema-associated gene expression in the genes identified here, as has already been observed in other eczema-associated genes [22].

Several limitations were identified in the process of our study. Although the 140 CpGs were chosen based on the least misclassification error rate, it is possible that some CpGs were incorrectly removed and vice versa. Also, cord blood contains a small amount of maternal cells [29], which may bias the measure of DNA methylation, but our cell type correction performed in this study was expected to reduce the bias. Findings from the $F_1$ generation were partially replicated in the $F_2$ generation. This could be due to age playing a role in the CpGs predicting eczema; adolescence transition has the potential to revise DNA methylation. This is supported by our comparison of DNA methylation between the F1 and F2 generations among the CpGs not replicated. Not all CpGs selected by random forests were involved in known eczema-associated biological pathways, which may be due to complex interactions between the CpGs hence requires further investigation. It is possible that some of the identified CpGs may be associated with the severity of eczema. Hence, there is a need to further examine potential associations of DNA methylation of those CpG sites with

**Table 3** Terms significantly enriched in functional annotation and pathway analysis and genes present in the pathways potentially associated with eczema (FDR-adjusted *P* value; FDR = 0.05)

| Term | FDR-adjusted *P* value | |
|---|---|---|
| Polymorphism | $4.7 \times 10^{-145}$ | |
| Sequence variant | $2.3 \times 10^{-111}$ | |
| Alternative splicing | $6.8 \times 10^{-74}$ | |
| Splice variant | $1.6 \times 10^{-46}$ | |
| Phosphoprotein | $6.2 \times 10^{-25}$ | |
| Protocadherin gamma | $1.8 \times 10^{-16}$ | *PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1,* |
| Disease mutation | $4.9 \times 10^{-16}$ | *PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5* |
| Domain: cadherin 6 | $1.5 \times 10^{-10}$ | *FAT1, PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5* |
| Cadherin, N-terminal | $3.6 \times 10^{-10}$ | *PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5* |
| Pathways in cancer | $8.5 \times 10^{-8}$ | |
| Membrane | $1.1 \times 10^{-7}$ | |
| Regulation of actin cytoskeleton | $1.8 \times 10^{-7}$ | |
| Long-term depression | $9.0 \times 10^{-7}$ | |
| Calcium ion binding | $1.1 \times 10^{-6}$ | |
| Plasma membrane | $2.2 \times 10^{-6}$ | |
| Glycoprotein | $2.4 \times 10^{-6}$ | |
| Gap junction[a] | $2.6 \times 10^{-6}$ | *GNAS, GNAI2, GNAI3, GUCY1A3, MAP2K1, PDGFA,PRKG1* |
| Cell-cell adhesion | $2.7 \times 10^{-6}$ | *CD164, CLDN5, CDSN,DAB1, FAT1, FGF6, PARD3, PTPRF, PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1,PCDHGB2,PCDHGB3, PCDHGB4, PCDHGB5* |
| Homophilic cell adhesion | $6.1 \times 10^{-6}$ | |
| Chemokine signaling pathway | $1.0 \times 10^{-5}$ | |
| Focal adhesion | $1.3 \times 10^{-5}$ | |
| Axon guidance | $1.3 \times 10^{-5}$ | *CLDN5, GNAI2, GNAI3, CSNK2B, MAGI2, MYL12B, PARD3* |
| Tight junction[a] | $1.6 \times 10^{-5}$ | |
| Biological adhesion | $1.7 \times 10^{-5}$ | |
| Cell adhesion | $2.2 \times 10^{-5}$ | *AEBP1, CD164, CD36, CLDN5, COL11A2, COL20A1, CDSN, DAB1, FAT1, FGF6, IGSF11, LAMA4, LAMC1, NELL2, NTM, PARD3, PTPRF, PPFIA1, PCDHGA1, PCDHGA2, PCDHGA3, PCDHGA4, PCDHGA5, PCDHGA6, PCDHGA7, PCDHGA8, PCDHGA9, PCDHGB1, PCDHGB2, PCDHGB3, PCDHGB4, PCDHGB5* |
| Coiled coil | $2.6 \times 10^{-5}$ | |
| Melanogenesis[a] | $7.1 \times 10^{-5}$ | *GNAS, CREB3, GNAI2, GNAI3, MAP2K1, WNT10B* |
| Vascular smooth muscle contraction | $1.1 \times 10^{-4}$ | |
| Chromosomal rearrangement | $2.6 \times 10^{-4}$ | |
| Cardiac muscle contraction | $2.7 \times 10^{-4}$ | |
| Intracellular signaling cascade | $4.5 \times 10^{-4}$ | |
| Cell membrane | $4.7 \times 10^{-4}$ | |
| Cell fraction | $4.8 \times 10^{-4}$ | |
| Prostate cancer | $6.2 \times 10^{-4}$ | |
| Ion binding | $6.8 \times 10^{-4}$ | |
| Acetylation | $7.6 \times 10^{-4}$ | |
| Signal | $8.3 \times 10^{-4}$ | |

**Table 3** Terms significantly enriched in functional annotation and pathway analysis and genes present in the pathways potentially associated with eczema (FDR-adjusted *P* value; FDR = 0.05) *(Continued)*

| | | |
|---|---|---|
| Transmembrane | $1.0 \times 10^{-3}$ | |
| Mutagenesis site | $1.1 \times 10^{-3}$ | |
| Cation binding | $1.2 \times 10^{-3}$ | |
| Lysine degradation | $1.4 \times 10^{-3}$ | |
| Leukocyte trans endothelial migration | $1.5 \times 10^{-3}$ | |
| Lysosome | $1.5 \times 10^{-3}$ | |
| Transcription factor binding | $3.9 \times 10^{-3}$ | |
| Melanoma[a] | $4.6 \times 10^{-3}$ | *E2F2, FGF6, MAP2K1, PDGFA* |
| Tumor suppressor | $5.0 \times 10^{-3}$ | |
| Nucleotide binding | $5.0 \times 10^{-3}$ | |
| Endocytosis | $7.0 \times 10^{-3}$ | |
| Apoptosis[a] | $7.3 \times 10^{-3}$ | *CHP2, NTRK1, PPP3CA, RIPK1* |
| Small cell lung cancer | $7.3 \times 10^{-3}$ | |
| Nucleus | $1.1 \times 10^{-2}$ | |
| Cell projection | $1.7 \times 10^{-2}$ | |
| Positive regulation of cellular biosynthetic process | $4.4 \times 10^{-2}$ | |
| Transcription co-activator activity | $4.9 \times 10^{-2}$ | |

[a]Represents pathways which are involved in eczema with their genes

eczema severity. For multiple CpG sites, DNA methylation was associated with eczema in the $F_1$ generation at age 18. These CpG sites could be risks or consequences of eczema. However, CpGs replicated in the $F_2$ generation were measured in cord blood before the onset of eczema and thus have the potential to predict eczema.

## Conclusions

This is the first epigenome-scale association study of eczema employing a classification technique (recursive RF), and we identified eczema-associated CpG sites. The findings added to the existing knowledge that recursive RF can be successfully employed in drawing actionable results from complex datasets. Genes annotated to eczema-associated CpGs were significantly enriched in pathways related to the creation of transcriptional variation and pathways relating to epidermal barrier function and eczema. Furthermore, the study identified for the first time that the *PROZ* and *NEU1* genes are potential predictors of eczema.

## Methods

### Isle of Wight birth cohort

The Isle of Wight (IoW) birth cohort was established to study the natural history of allergic diseases among children who were born between January 1, 1989 and February 28, 1990 on the Isle of Wight, UK. The study was approved by the local research ethics committee and written informed consent was obtained from the parents. After exclusion of adoptions, perinatal deaths, and refusal, 1456 children (95 %) were enrolled. Children were followed-up at ages 1 ($n = 1167$), 2 ($n = 1174$), 4 ($n = 1218$), 10 ($n = 1373$), and 18 years ($n = 1313$); detailed questionnaires were administered at each follow-up. Details of the birth cohort have been described elsewhere [4, 30, 31]. A total of 244 women and 122 men at age 18 years were randomly selected from the cohort for epigenome-scale DNA methylation studies. Ethics approvals were obtained from the Isle of Wight Local Research Ethics Committee (now named the National Research Ethics Service, NRES Committee South Central – Southampton B) at recruitment and for the subsequent follow-ups (06/Q1701/34).

### Outcome: eczema phenotype data collection

Eczema was defined as chronic or chronically relapsing itchy dermatitis lasting more than 6 weeks with characteristic morphology and distribution [32], following Hanifin and Rajka criteria [5].

### DNA methylation

DNA was extracted from whole blood and umbilical cord blood using a standard salting out procedure [33]. DNA concentration was determined by Qubit quantitation. One microgram of DNA was bisulfite-treated using the EZ 96-DNA methylation kit (Zymo Research, Irvine, CA, USA) following the manufacturer's standard protocol.

**Table 4** The 41 CpGs that had the same direction of effect with eczema in both $F_1$ and $F_2$ generations based on log-linear models

| CpGs | $F_1$-Risk Ratio | 95 % CI-$F_1$ | $F_2$-risk ratio | 95 % CI-$F_2$ | Gene |
|---|---|---|---|---|---|
| cg00193668 | 17.29 | 2.90, 102.87 | 4.86 | 0.89, 26.4 | *HINT2* |
| cg04850479[a] | 15.19 | 3.07, 75.17 | 6.82 | 1.52, 30.6 | *PROZ* |
| cg02641560 | 14.50 | 3.39, 62.65 | 1.33 | 0.13, 12.8 | *RCAN3* |
| cg05839818 | 13.02 | 2.34, 72.26 | 1.3 | 0.13, 12.1 | |
| cg05411056 | 9.73 | 2.64, 35.81 | 5.61 | 1.44, 21.85 | |
| cg02077766 | 9.60 | 2.14, 43.07 | 1.29 | 0.38, 4.33 | *PTCRA* |
| cg00667315 | 7.66 | 1.88, 31.21 | 1.25 | 0.19, 8.0 | |
| cg00900242 | 6.86 | 1.26, 37.20 | 6.04 | 0.75, 48.6 | |
| cg02583247 | 6.61 | 2.05, 21.33 | 1.27 | 0.26, 6.10 | *FGF6* |
| cg01802073 | 6.10 | 1.40, 26.43 | 1.43 | 0.24, 8.61 | *CGRRF1* |
| cg14839837 | 5.90 | 1.63, 21.39 | 2.94 | 0.73, 11.7 | *ARHGEF10* |
| cg00354884 | 5.77 | 1.95, 17.03 | 1.8 | 0.59, 6.03 | *ABR* |
| cg00158434 | 5.43 | 1.75, 16.78 | 2.47 | 0.52, 11.5 | *ALMS1P* |
| cg03049303 | 4.73 | 1.44, 15.57 | 4.61 | 0.77, 27.4 | *C10orf76* |
| cg24303123 | 4.68 | 1.73, 12.65 | 1.49 | 0.50, 4.46 | *RIPK1* |
| cg11570082 | 4.46 | 1.85, 10.71 | 2.56 | 0.58, 11.2 | |
| cg02237186 | 4.26 | 1.24, 14.63 | 2.89 | 0.16, 51.1 | *RRM2* |
| cg02654265 | 3.92 | 1.56, 9.87 | 0.29 | 0.05, 1.52 | |
| cg00369908 | 3.65 | 1.34, 9.92 | 4.05 | 0.75, 21.6 | *ING4* |
| cg00722180 | 3.64 | 1.22, 0.85 | 2.84 | 0.63, 12.7 | *RBM25* |
| cg02433979 | 2.91 | 1.35, 6.27 | 1.17 | 0.37, 3.68 | |
| cg00035220 | 2.62 | 1.19, 5.72 | 1.18 | 0.34, 4.03 | *PTPRN2* |
| cg00252472 | 2.62 | 1.27, 5.40 | 1.22 | 0.44, 3.38 | |
| cg00306063 | 2.59 | 1.12, 5.97 | 2.12 | 0.48, 9.19 | *LOC100129066* |
| cg00742851 | 2.23 | 1.16, 4.28 | 1.26 | 0.45, 3.48 | *LRRN1* |
| cg02203881 | 2.07 | 1.07, 4.00 | 1.67 | 0.46, 6.02 | *PLA2G4D* |
| cg00576402 | 0.57 | 0.35, 0.92 | 0.76 | 0.28, 2.07 | *PTPN12* |
| cg01560119 | 0.41 | 0.21, 0.80 | 0.79 | 0.37, 1.68 | *SETDB2* |
| cg01651499 | 0.37 | 0.16, 0.85 | 0.41 | 0.12, 1.34 | *GUCY1A3* |
| cg02098905 | 0.35 | 0.16, 0.76 | 0.41 | 0.14, 1.12 | |
| cg04797820 | 0.33 | 0.17, 0.64 | 0.93 | 0.31, 2.76 | *GLT1D1* |
| cg00247571 | 0.31 | 0.13, 0.75 | 0.89 | 0.26, 2.50 | |
| cg00071869 | 0.30 | 0.13, 0.70 | 0.77 | 0.12, 4.88 | *ATP1B3* |
| cg00797821 | 0.29 | 0.10, 0.82 | 0.36 | 0.06, 2.12 | |
| cg01158447 | 0.24 | 0.09, 0.60 | 0.35 | 0.11, 1.14 | *SLC40A1* |
| cg00077547 | 0.21 | 0.06, 0.70 | 0.91 | 0.25, 3.24 | *TMEM26* |
| cg04980849 | 0.21 | 0.07, 0.60 | 0.57 | 0.16, 1.96 | *LOC145663*; *GATM* |
| cg00050654 | 0.19 | 0.07, 0.51 | 0.71 | 0.20, 2.43 | |
| cg20077343 | 0.19 | 0.06, 0.58 | 0.25 | 0.03, 1.83 | *MUC6* |
| cg17602756 | 0.14 | 0.03, 0.64 | 0.26 | 0.05, 1.39 | *SQSTM1* |
| cg01427769[a] | 0.13 | 0.03, 0.46 | 0.09 | 0.02, 0.36 | *NEU1* |

[a]CpG sites significantly associated with eczema in both generations. For cg04850479, the *P* values are 0.0006 in the $F_1$ generation and 0.0121 in the $F_2$ generation, and for cg01427769, the *P* values are 0.0015 and 0.0007, respectively

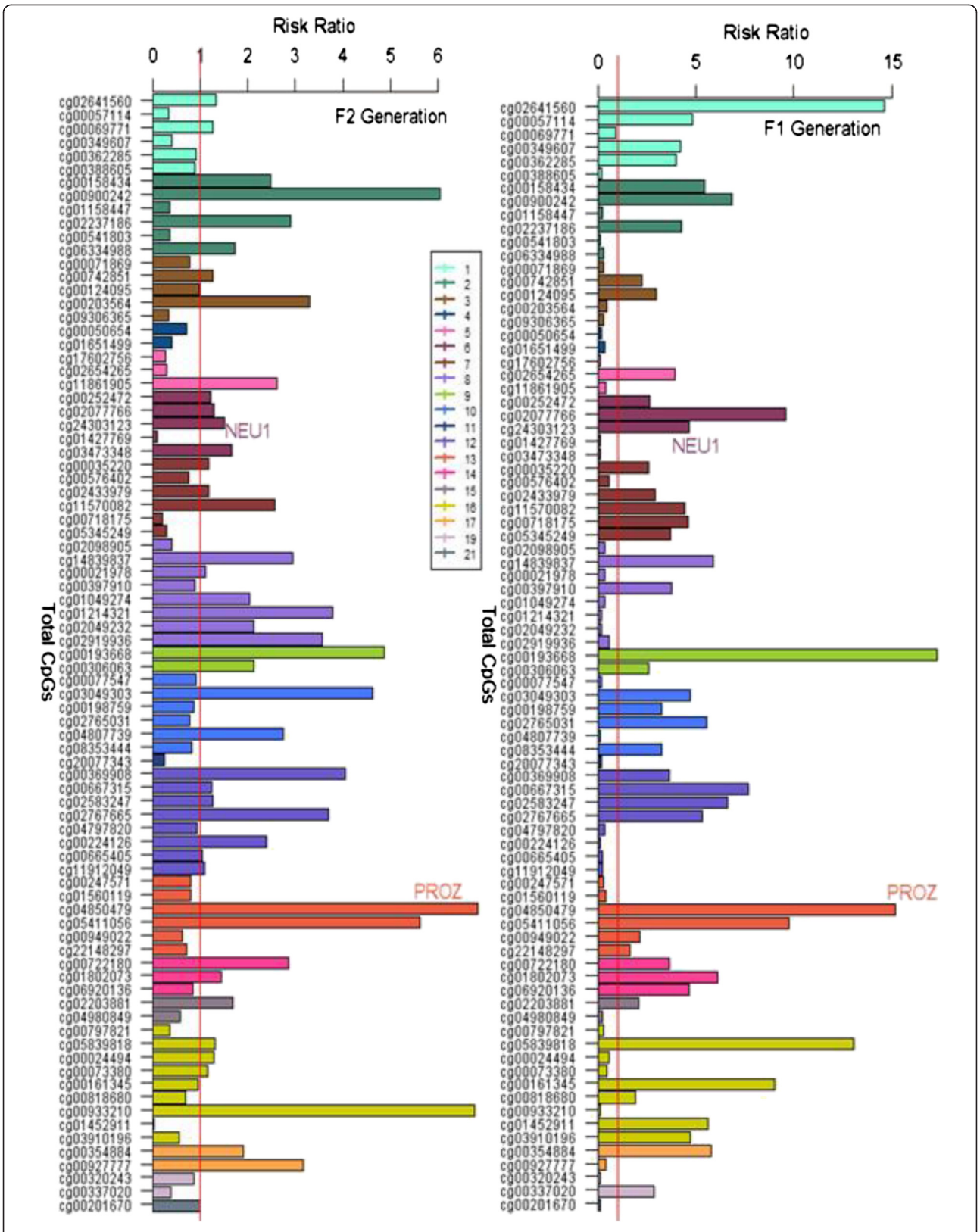


**Fig. 2** The risk ratios of 83 eczema-associated CpGs sorted by chromosome from 1 to 21. The *numbers* in the textbox are chromosome indices, which are represented by different colors in the bar graphs. The horizontal *red line* represents the risk ratio of one

Epigenome-scale DNA methylation was assessed using the Illumina Infinium HumanMethylation450 Beadchip (Illumina, Inc., San Diego, CA, USA), which interrogates >484,000 CpGs associated with approximately 24,000 genes. Arrays were processed using a standard protocol as described elsewhere [7], with multiple identical control samples assigned to each batch to assess assay variability, and samples were randomly distributed on microarrays to control against batch effects. The methylation level ($\beta$ value) for each CpG was determined using the Methylation module of GenomeStudio software (Illumina, Version 2011.1).

Methylation levels for each CpG site are recorded as beta ($\beta$) values, which represent the proportion of methylated ($M$) over methylated ($M$) plus unmethylated ($U$) probes ($\beta = M/[c + M + U]$, with constant $c$ introduced for the situation of too small $M + U$) and can be interpreted as percentage methylation. These values were utilized in the RF screening process described below; however, $\beta$ values close to 0 or 1 tend to suffer from severe heteroscedasticity; therefore, logit-transformed $\beta$ values ($M$ values, approximated by $\log_2(\beta / (1-\beta))$ [34] were used in other analyses including log-linear models.

### Pre-processing DNA methylation data

In our study, the detection $P$ value reported by GenomeStudio was used as a QC measure of probe performance. Probes whose detection $P$ values were >0.01 in >10 % of the samples were removed [35]. Methylation data were then pre-processed using the Bioconductor IMA (Illumina methylation analyzer) package and ComBat was used to perform peak correction and adjust for inter-array variation [36, 37]. To ensure that our findings were not biased by SNPs affecting measurement of methylation levels, we excluded all probes with a potential SNP in the probe sequence. After pre-processing, a total of 307,357 CpGs were retained in the DNA methylation dataset.

### Statistical analysis

Pearson's $\chi^2$ tests were used to determine if prevalence of eczema differed between the sexes. $P$ values were considered significant at a level of 0.05. To make sure that our findings are not a result of confounding due to cell types, we ran the analyses by adjusting for estimated proportions of $CD8^+$ T cells, $CD4^+$ T cells, natural killer cells, B cells, monocytes, and granulocytes. Cell type proportions were estimated as described previously [16].

The random forest package, *randomForest*(), in R was utilized to conduct the recursive RF analyses [38, 15, 14]. The parameter *sampsize* refers to the size of the sample of training data sets that is to be obtained for classification. The number of variables that are randomly sampled as predictors at each split is called *mtry*, whereas, *ntree* is a parameter referring to the total number of trees that are to be grown in the forest. In order to improve the prediction accuracy of the RF algorithm, these three parameters were repeatedly altered until the lowest misclassification rate was obtained. We decided whether to use a balanced *sampsize* of equal eczema and non-eczema cases such as 20 eczema and 20 non-eczema cases or 30/30 or 40/40. We also studied imbalanced RFs with *sampsize* such as 46/320 or 20/40 for the training sets by using the default values for *mtry and ntree*. We then tested the prediction accuracy of the RFs at different combinations of *mtry* ($\sqrt{p}$, $2^*\sqrt{p}$, $0.1p$, $0.15p$, $0.2p$, and $0.25p$) where $p$ is number of variables and *ntree* (200, 500, 1000, and 1500). Once the optimal parameter values were selected, the recursive RF algorithm was implemented. Mean Decrease Gini (MDG) served as a variable importance measure (VIM) for our study as it was shown to be more robust in previous research [39].

DNA methylation at 307,357 CpGs along with sex and eczema status in the $F_1$ generation served as input in *randomForest*(), and the CpGs were subjected to data reduction, repeatedly dropping 50 % of variables with the lowest VIMs until the misclassification rate showed a significant increase.

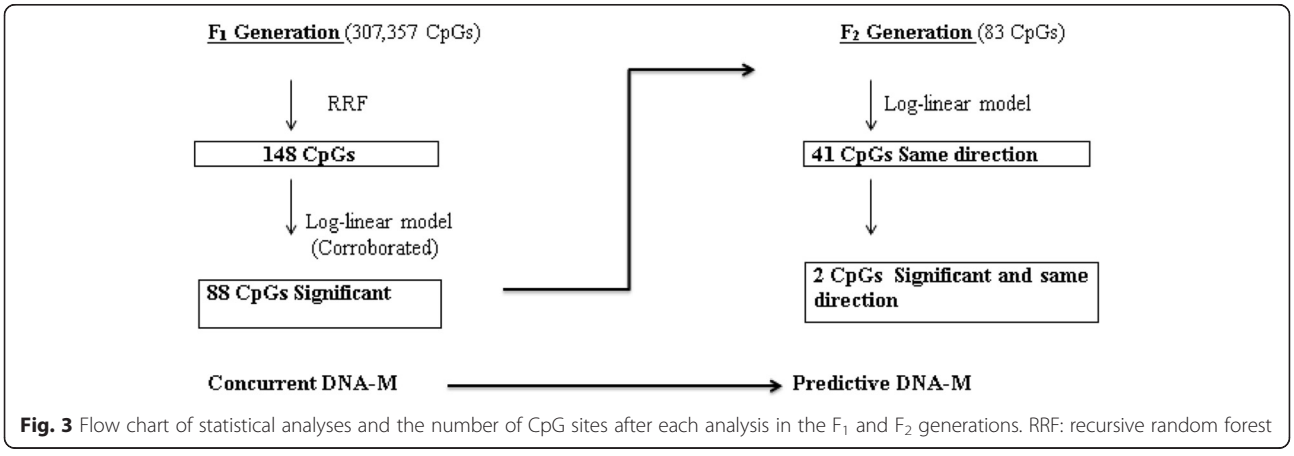


**Fig. 3** Flow chart of statistical analyses and the number of CpG sites after each analysis in the $F_1$ and $F_2$ generations. RRF: recursive random forest

After testing for *sampsize* (both equal and unequal) with different combinations (both with and without eczema), we set *sampsize* = (31, 31), *mtry* = 0.2*p* (where *p* is the available number of variables) and *ntree* = 500. We applied RF to pre-processed DNA methylation data containing 307,357 CpGs in the $F_1$ generation, ran a total of 17 iterations, and at each iteration, recorded the misclassification rate (Table 2, Fig. 1). The lowest overall misclassification error rate (of eczema and eczema-free) was 5.2 %, with a corresponding least misclassification rate of 17.4 % for eczema at the 12th iteration. The overall misclassification rate dropped from 18.6 % in the first iteration to 5.2 % in the 12th iteration, and the eczema misclassification error rate dropped to 17.4 % at the end of 12th iteration from 95.7 % in the first iteration.

The CpGs identified from the recursive RF [40] were assessed for enrichment of biological pathways using DAVID [12] bioinformatics tool and examined for their association with eczema at age of 18 years by use of log-linear models. Multiple testing was adjusted by controlling false discovery rate of 0.05 in the pathway analysis and log-linear models. Since differential cell types in the peripheral blood are known to have confounding effect on the final result [16], we adjusted for cell type correction. For genes of particular interest (e.g., showing statistical significance in both generations in log-linear models), robust regressions are applied to assess the association of DNA methylation and corresponding gene expressions in the $F_2$ generation. For this last test, multiple testing is adjusted within genes based on the number of CpG sites available of that gene.

### Replication cohort

The IoW $F_2$ generation cohort includes the offspring of the IoW 1989 birth cohort. In the $F_2$ generation, repeated measures of eczema at ages 3, 6, and 12 months were recorded in a sample of $n = 116$ children. DNA methylation was measured in umbilical cord blood. To replicate the findings from the $F_1$ generation, log-linear models with repeated measures of eczema were used in $F_2$ generation analyses. Figure 3 represents the summary of statistical analysis and sample size for each analysis conducted in this study.

## Additional file

**Additional file 1: Table S1.** Beta coefficients and *P* values after log-linear model of the 140 CpGs present in the 12th iteration of the RF algorithm output and their genetic details. Software necessary to view: Adobe Reader.

#### Abbreviations

CpG: cytosine-phosphate-guanine; DNA-M: DNA methylation; IoW: Isle of Wight; RF: random forest; RR: risk ratio.

#### Competing interests

The authors declare that they have no competing interests.

#### Authors' contributions

GAL, HZ, and BMQ cleaned the methylation data. BMQ conducted the statistical analyses, interpreted the data, and drafted the manuscript. HZ directed the analysis and guided interpretation of the results. FIR pre-processed the DNA methylation data. TME supervised the assessment of DNA methylation. MR, GAL, SRT, AK, and JWH helped in conducting analyses. SHA was responsible for cohort assessments, eczema phenotype data, and sample collection and provided critical revision. WK designed the study, reviewed the data quality, and helped with statistical analyses. All authors contributed to manuscript drafting and final proofreading. All authors read and approved the final manuscript.

#### Acknowledgements

The study conveyed in this publication was supported by the National Institute of Allergy and Infectious Diseases under award number R01 AI091905-01 (PI: Wilfried Karmaus). The 10-year follow-up of this study was funded by the National Asthma Campaign, UK (Grant No 364) and the 18-year follow-up by NIH/NHLBI R01 HL082925-01 (PI: S. Hasan Arshad). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors gratefully acknowledge the cooperation of the children and parents who participated in this study and appreciate the hard work of Mrs. Sharon Matthews and the Isle of Wight research team in collecting data and Nikki Graham for technical support. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532/Z/09/Z and MRC Hub grant G0900747 91070) for the generation of the methylation data.

#### Author details

[1]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, 236A Robison Hall, Memphis, TN 38152, USA. [2]Department of Epidemiology, and Biostatistics, Arnold School of Public Health, University of South Carolina, 800 Sumter Street, Columbia, SC 29208, USA. [3]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK. [4]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton SO16 6YD, UK. [5]The David Hide Asthma and Allergy Research Centre, St Mary's Hospital, Parkhurst Road, Newport, Isle of Wight PO30 5TG, UK. [6]Human Development and Health, Faculty of Medicine, University of Southampton, University Road, Southampton SO17 1BJ, UK.

Received: 14 April 2015 Accepted: 2 July 2015
Published online: 21 July 2015

#### References

1. Bieber T. Atopic dermatitis. Ann Dermatol. 2010;22(2):125–37.
2. Leung DYM, Boguniewicz M, Howell MD, Nomura I, Hamid QA. New insights into atopic dermatitis. J Clin Invest. 2004;113(5):651–7. doi:10.1172/JCI21060.
3. Gupta R, Sheikh A, Strachan D, Anderson H. Burden of allergic disease in the UK: secondary analyses of national databases. Clin Exp Allergy. 2004;34(4):520–6.
4. Arshad SH, Hide DW. Effect of environmental factors on the development of allergic disorders in infancy. J Allergy Clin Immunol. 1992;90(2):235–41.
5. Hanifin J, Rajka G. Diagnostic features of atopic eczema. Acta Dermatol Venereol (Stockh). 1980;92:44–7.
6. Orteu C, Rustin M, O'toole E, Sabin C, Salmon M, Poulter L, et al. The inhibition of cutaneous T cell apoptosis may prevent resolution of inflammation in atopic eczema. Clin Exp Immunol. 2000;122(2):150–6.
7. Bibikova M, Fan J-B. GoldenGate® Assay for DNA Methylation Profiling. DNA Methylation. Springer; 2009. p. 149–63.
8. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692–702.
9. Paul DS, Beck S. Advances in epigenome-wide association studies for common diseases. Trends Mol Med. 2014;20(10):541–3.

10. Xiong M, Ma L. An efficient sufficient dimension reduction method for identifying genetic variants of clinical significance. arXiv preprint arXiv:13013528. 2013.
11. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
12. Da Wei Huang BTS, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
13. Ziyab A, Raza A, Karmaus W, Tongue N, Zhang H, Matthews S, et al. Trends in eczema in the first 18 years of life: results from the Isle of Wight 1989 birth cohort study. Clin Exp Allergy. 2010;40(12):1776–84.
14. Hapfelmeier A, Ulm K. A new variable selection approach using random forests. Comput Stat Data Anal. 2013;60:50–69.
15. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of random forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. 2010;11(1):49.
16. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13(1):86.
17. Lu LQ, Liao W. Screening and functional pathway analysis of genes associated with pediatric allergic asthma using a DNA microarray. Mol Med Rep. 2015. doi:10.3892/mmr.2015.3277.
18. Katoh S, Maeda S, Fukuoka H, Wada T, Moriya S, Mori A, et al. A crucial role of sialidase Neu1 in hyaluronan receptor function of CD44 in T helper type 2-mediated airway inflammation of murine acute asthmatic model. Clin Exp Immunol. 2010;161(2):233–41. doi:10.1111/j.1365-2249.2010.04165.x.
19. Abrahamsson TR, Sandberg Abelius M, Forsberg A, Björkstén B, Jenmalm MC. A Th1/Th2-associated chemokine imbalance during infancy in children developing eczema, wheeze and sensitization. Clin Exp Allergy. 2011;41(12):1729–39. doi:10.1111/j.1365-2222.2011.03827.x.
20. Eyerich K, Novak N. Immunology of atopic eczema: overcoming the Th1/Th2 paradigm. Allergy. 2013;68(8):974–82. doi:10.1111/all.12184.
21. Cork MJ, Danby SG, Vasilopoulos Y, Hadgraft J, Lane ME, Moustafa M, et al. Epidermal barrier dysfunction in atopic dermatitis. J Investig Dermatol. 2009;129(8):1892–908.
22. Ziyab AH, Karmaus W, Holloway JW, Zhang H, Ewart S, Arshad SH. DNA methylation of the filaggrin gene adds to the risk of eczema associated with loss-of-function variants. J Eur Acad Dermatol Venereol. 2013;27(3):e420–3. doi:10.1111/jdv.12000.
23. Wapenaar MC, Monsuur A, van Bodegraven A, Weersma RK, Bevova M, Linskens R, et al. Associations with tight junction genes PARD3 and MAGI2 in Dutch patients point to a common barrier defect for celiac disease and ulcerative colitis. Gut. 2007;57(4):463–7.
24. De Benedetto A, Rafaels NM, McGirt LY, Ivanov AI, Georas SN, Cheadle C, et al. Tight junction defects in patients with atopic dermatitis. J Allergy Clin Immunol. 2011;127(3):773–86. e7.
25. Furuse M, Hata M, Furuse K, Yoshida Y, Haratake A, Sugitani Y, et al. Claudin-based tight junctions are crucial for the mammalian epidermal barrier a lesson from claudin-1–deficient mice. J Cell Biol. 2002;156(6):1099–111.
26. Niessen CM. Tight junctions/adherens junctions: basic structure and function. J Investig Dermatol. 2007;127(11):2525–32.
27. Faura Tellez G, Nawijn MC, Koppelman GH. Protocadherin-1: epithelial barrier dysfunction in asthma and eczema. Eur Respir J. 2014;43(3):671–4. doi:10.1183/09031936.00179713.
28. Harmon LE. Melanogenesis and pigmentary disturbances. J Natl Med Assoc. 1964;56(6):501.
29. Hall JM, Lingenfelter P, Adams S, Lasser D, Hansen J, Bean M. Detection of maternal cells in human umbilical cord blood using fluorescence in situ hybridization. Blood. 1995;86(7):2829–32.
30. Kurukulaaratchy R, Fenn M, Waterhouse L, Matthews S, Holgate S, Arshad S. Characterization of wheezing phenotypes in the first 10 years of life. Clin Exp Allergy. 2003;33(5):573–8.
31. Soto-Ramírez N, Ziyab AH, Karmaus W, Zhang H, Kurukulaaratchy RJ, Ewart S, et al. Epidemiologic methods of assessing asthma and wheezing episodes in longitudinal studies: measures of change and stability. J Epidemiol. 2013;23(6):399.
32. Arshad S, Karmaus W, Kurukulaaratchy R, Sadeghnejad A, Huebner M, Ewart S. Polymorphisms in the interleukin 13 and GATA binding protein 3 genes and the development of eczema during childhood. Br J Dermatol. 2008;158(6):1315–22.
33. Miller S, Dykes D, Polesky H. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16(3):1215.
34. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11(1):587.
35. Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, et al. Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. PLoS One. 2010;5(3):e9749.
36. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, et al. IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics. 2012;28(5):729–30.
37. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
38. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
39. Calle ML, Urrea V. Letter to the editor: stability of random forest importance measures. Brief Bioinform. 2011;12(1):86–9.
40. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, et al. Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol. 2005;28(2):171–82.



**Submit your next manuscript to BioMed Central and take full advantage of:**

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

**Supplementary Table 1. Beta coefficients and *P* values after log-linear model of the 140 CpGs present in the 12th iteration of the RF algorithm output and their Genetic details**

| CpGs | CHR[$] | UCSC* chromosomal location | Gene | UCSC* Gene Region | USCS* CpG Island Relation | β-coefficients | P-Value |
|---|---|---|---|---|---|---|---|
| **cg00057114** | 1 | 6344362 | *ACOT7* | Body | S_Shelf | 1.57 | 0.032 |
| **cg00069771** | 1 | 233089275 | *C1orf57* | Body | S_Shelf | -0.11 | 0.031 |
| **cg00349607** | 1 | 39546688 | *MACF1* | TSS1500 | N_Shore | 1.44 | 0.026 |
| **cg00362285** | 1 | 236444212 | *ERO1LB* | Body | N_Shore | 1.38 | 0.026 |
| **cg00388605** | 1 | 89150448 | *PKN2* | Body | Island | -1.55 | 0.028 |
| **cg02641560** | 1 | 24828811 | *RCAN3* | TSS1500 | N_Shore | 2.68 | 0.011 |
| **cg00158434** | 2 | 73871782 | *ALMS1P* | TSS1500 | | 1.69 | 0.017 |
| **cg00541803** | 2 | 120301835 | *PCDP1* | TSS1500 | Island | -2.02 | 0.031 |
| **cg00900242** | 2 | 200929967 | | | | 1.92 | 0.043 |
| **cg01158447** | 2 | 190445074 | *SLC40A1* | Body | N_Shore | -1.40 | 0.016 |
| **cg02237186** | 2 | 10263043 | *RRM2;RRM2* | Body | Island | 1.45 | 0.039 |
| **cg06334988** | 2 | 220095083 | *ANKZF1;ATG9A* | Body; TSS1500 | S_Shore | -1.12 | 0.009 |
| **cg00071869** | 3 | 141599902 | *ATP1B3* | Body | S_Shelf | -1.17 | 0.023 |
| **cg00124095** | 3 | 72305082 | | | | 1.10 | 0.016 |
| **cg00203564** | 3 | 176133845 | | | | -0.80 | 0.038 |
| **cg00742851** | 3 | 3843342 | *LRRN1* | 5'UTR | S_Shore | 0.80 | 0.034 |
| **cg09306365** | 3 | 61547944 | *PTPRG* | 1stExon; 5'UTR | Island | -1.17 | 0.047 |
| **cg00050654** | 4 | 4576493 | | | N_Shore | -1.64 | 0.014 |
| **cg01214900** | 4 | 66784395 | | | | 0.84 | 0.104 |
| **cg01651499** | 4 | 156646379 | *GUCY1A3* | Body | | -0.99 | 0.038 |
| **cg02654265** | 5 | 137086868 | | | N_Shelf | 1.36 | 0.017 |
| **cg11861905** | 5 | 176045983 | | | N_Shore | -0.89 | 0.038 |
| **cg17602756** | 5 | 179246001 | *SQSTM1* | 5'UTR | N_Shore | -1.92 | 0.028 |
| **cg00252472** | 6 | 150739173 | | | | 0.96 | 0.026 |
| **cg02077766** | 6 | 42891866 | *PTCRA* | Body | N_Shore | 2.26 | 0.017 |
| **cg03473348** | 6 | 33160976 | *COL11A2* | TSS1500 | Island | -1.84 | 0.017 |
| **cg24303123** | 6 | 3076314 | *RIPK1* | TSS1500 | | 1.54 | 0.016 |
| **cg00035220** | 7 | 158197980 | *PTPRN2* | Body | N_Shore | 0.96 | 0.034 |
| **cg00576402** | 7 | 77167212 | *PTPN12* | TSS200; Body | Island | -0.55 | 0.042 |
| **cg00718175** | 7 | 158807362 | *LOC154822* | Body | | 1.53 | 0.043 |
| **cg02433979** | 7 | 128824829 | | | N_Shelf | 1.07 | 0.024 |
| **cg05345249** | 7 | 80291832 | *CD36* | Body | | 1.30 | 0.016 |
| **cg11570082** | 7 | 116899446 | | | | 1.49 | 0.014 |
| **cg00021978** | 8 | 105476917 | *DPYS* | Body | N_Shore | -1.06 | 0.042 |
| **cg00397910** | 8 | 80522577 | *STMN2* | TSS1500 | N_Shore | 1.33 | 0.016 |
| **cg01049274** | 8 | 146228340 | *ZNF252;C8orf77* | TSS200; Body | Island | -1.02 | 0.046 |
| **cg01214321** | 8 | 26435429 | *DPYSL2* | 5'UTR; 1stExon | Island | -1.79 | 0.038 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **cg02049232** | 8 | 145556129 | *SCRT1* | 3'UTR | Island | -1.65 | 0.024 |
| **cg02098905** | 8 | 119759090 | | | | -1.02 | 0.026 |
| **cg02919936** | 8 | 70982285 | *PRDM14* | 5'UTR | Island | -0.57 | 0.038 |
| **cg14839837** | 8 | 1834174 | *ARHGEF10* | Body | S_Shelf | 1.77 | 0.025 |
| **cg00193668** | 9 | 35814283 | *HINT2* | Body | N_Shore | 2.85 | 0.016 |
| **cg00306063** | 9 | 92289643 | *LOC100129066* | Body | N_Shore | 0.95 | 0.043 |
| **cg00077547** | 10 | 63170372 | *TMEM26* | Body | | -1.55 | 0.03 |
| **cg00198759** | 10 | 81444321 | *LOC650623* | Body | Island | 1.18 | 0.047 |
| **cg02765031** | 10 | 85462980 | | | | 1.71 | 0.009 |
| **cg03049303** | 10 | 103816154 | *C10orf76* | TSS1500 | S_Shore | 1.55 | 0.028 |
| **cg04807739** | 10 | 88779914 | *FAM25A* | TSS200 | | -2.03 | 0.024 |
| **cg08353444** | 10 | 60506653 | *BICC1* | Body | | 1.18 | 0.026 |
| **cg20077343** | 11 | 1036307 | *MUC6* | Body | N_Shore | -1.62 | 0.017 |
| **cg00224126** | 12 | 62861053 | *MON2* | 1stExon | S_Shore | -2.18 | 0.024 |
| **cg00369908** | 12 | 6772357 | *ING4* | TSS200 | Island | 1.29 | 0.028 |
| **cg00665405** | 12 | 50899548 | *DIP2B* | Body | S_Shore | -1.45 | 0.022 |
| **cg00667315** | 12 | 295951 | | | N_Shelf | 2.03 | 0.021 |
| **cg02583247** | 12 | 4555311 | *FGF6* | TSS1500 | S_Shore | 1.88 | 0.016 |
| **cg02767665** | 12 | 129788385 | *TMEM132D* | Body | S_Shore | 1.67 | 0.016 |
| **cg04797820** | 12 | 129460559 | *GLT1D1* | Body | | -1.09 | 0.016 |
| **cg11912049** | 12 | 50361401 | | | Island | -1.38 | 0.016 |
| **cg00949022** | 13 | 73356224 | *PIBF1;DIS3* | TSS200; 1stExon; 5'UTR | Island | 0.77 | 0.038 |
| **cg01560119** | 13 | 50025780 | *SETDB2* | 5'UTR; 1stExon | | -0.87 | 0.026 |
| **cg04850479** | 13 | 113811569 | *PROZ* | TSS1500 | S_Shelf | 2.72 | 0.014 |
| **cg05411056** | 13 | 112852819 | | | S_Shelf | 2.27 | 0.014 |
| **cg22148297** | 13 | 99910587 | *GPR18;UBAC2* | 1stExon; Body; 5'UTR | | 0.49 | 0.037 |
| **cg00722180** | 14 | 73525139 | *RBM25* | TSS200 | Island | 1.29 | 0.038 |
| **cg01802073** | 14 | 54976343 | *CGRRF1* | TSS1500 | Island | 1.80 | 0.034 |
| **cg06920136** | 14 | 21945858 | *RAB2B;TOX4* | TSS1500; Body | S_Shore | 1.54 | 0.026 |
| **cg00229754** | 15 | 25018302 | | | Island | -0.55 | 0.149 |
| **cg02203881** | 15 | 42386909 | *PLA2G4D* | TSS200 | | 0.73 | 0.047 |
| **cg04980849** | 15 | 45671347 | *LOC145663;GATM* | Body; TSS1500 | S_Shore | -1.53 | 0.017 |
| **cg00024494** | 16 | 73096842 | | | N_Shore | -0.50 | 0.041 |
| **cg00073380** | 16 | 5122085 | *ALG1* | Body | Island | -0.78 | 0.038 |
| **cg00161345** | 16 | 66866226 | *NAE1* | TSS1500 | S_Shore | 2.19 | 0.014 |
| **cg00797821** | 16 | 49888944 | | | Island | -1.20 | 0.038 |
| **cg00818680** | 16 | 68881254 | *TMCO7* | Body | S_Shelf | 0.65 | 0.026 |
| **cg00933210** | 16 | 3233117 | | | Island | -2.13 | 0.045 |
| **cg01452911** | 16 | 86512095 | | | | 1.72 | 0.016 |
| **cg03910196** | 16 | 2256384 | *MLST8* | Body | S_Shore | 1.55 | 0.031 |
| **cg05839818** | 16 | 85634031 | | | | 2.56 | 0.017 |
| **cg00354884** | 17 | 1057541 | *ABR* | Body; 5'UTR | | 1.75 | 0.016 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **cg00927777** | 17 | 1960199 | *HIC1* | Body | Island | -0.92 | 0.026 |
| **cg00320243** | 19 | 18314691 | *RAB3A* | 5'UTR | Island | -1.91 | 0.038 |
| **cg00337020** | 19 | 2241089 | *SF3A2* | 5'UTR | S_Shelf | 1.04 | 0.028 |
| **cg00201670** | 21 | 46221900 | *UBE2G2* | TSS200 | Island | -1.95 | 0.043 |
| **cg00015530** | 8 | 145052950 | *PARP10* | Body | S_Shore | 0.32 | 0.244 |
| **cg00019877** | 9 | 115773738 | *LOC169834* | 5'UTR | N_Shore | 1.82 | 0.024 |
| **cg00546248** | 1 | 2412743 | *PLCH2* | Body | S_Shore | -0.45 | 0.186 |
| **cg00778190** | 1 | 95068029 | | | | 0.54 | 0.101 |
| **cg01059116** | 1 | 54163070 | *GLIS1* | 5'UTR | | 0.31 | 0.048 |
| **cg00164941** | 2 | 73523056 | | | S_Shelf | -0.59 | 0.11 |
| **cg00729049** | 2 | 171679402 | *GAD1* | Body | Island | -1.10 | 0.031 |
| **cg00405769** | 4 | 187539852 | *FAT1* | Body | | -0.57 | 0.075 |
| **cg00866215** | 4 | 54956702 | | | N_Shore | 0.64 | 0.123 |
| **cg05684300** | 4 | 102267366 | *PPP3CA* | Body | N_Shore | 0.94 | 0.061 |
| **cg00099094** | 5 | 2334514 | | | | -0.23 | 0.189 |
| **cg00106739** | 5 | 177657311 | *AGXT2L2* | Body | N_Shore | 0.97 | 0.068 |
| **cg00580497** | 5 | 1153746 | | | N_Shore | 0.81 | 0.122 |
| **cg00043444** | 6 | 134384706 | | | | -0.23 | 0.633 |
| **cg00155619** | 6 | 33173278 | *HSD17B8* | Body | S_Shore | 0.97 | 0.08 |
| **cg00491418** | 6 | 112575482 | *LAMA4* | 5'UTR | Island | -1.72 | 0.066 |
| **cg01427769** | 6 | 31830600 | *NEU1* | 5'UTR; 1stExon | Island | -2.0 | 0.016 |
| **cg12427444** | 6 | 31923064 | *RDBP* | Body | N_Shelf | -0.49 | 0.497 |
| **cg00036614** | 7 | 47093842 | | | S_Shore | 0.33 | 0.378 |
| **cg00064840** | 7 | 4784485 | *FOXK1* | Body | N_Shore | 1.42 | 0.068 |
| **cg00106345** | 7 | 27138396 | | | S_Shore | -0.39 | 0.213 |
| **cg00129269** | 7 | 127228222 | *ARF5* | TSS1500 | Island | -0.92 | 0.068 |
| **cg00301256** | 7 | 82653693 | *PCLO;PCLO* | Body | | -0.57 | 0.264 |
| **cg02474076** | 7 | 127223255 | *GCC1* | Body | N_Shore | 1.47 | 0.126 |
| **cg00161683** | 8 | 95563352 | *KIAA1429* | Body | N_Shelf | 1.89 | 0.027 |
| **cg00166750** | 8 | 102138594 | | | Island | -1.21 | 0.06 |
| **cg00328935** | 8 | 142517474 | *FLJ43860* | TSS200 | | -0.34 | 0.359 |
| **cg00405554** | 8 | 65490897 | *LOC401463* | TSS1500 | N_Shore | 1.34 | 0.084 |
| **cg00449899** | 8 | 53626436 | *RB1CC1* | 5'UTR | Island | -1.34 | 0.039 |
| **cg00592695** | 8 | 134545685 | *ST3GAL1* | 5'UTR | | -0.92 | 0.094 |
| **cg00598335** | 8 | 11627530 | *NEIL2* | 5'UTR; 1stExon | Island | 0.43 | 0.489 |
| **cg00661018** | 8 | 145769086 | *KIAA1688* | Body | N_Shore | 1.90 | 0.053 |
| **cg00556719** | 9 | 108457484 | *TMEM38B* | Body | Island | 1.98 | 0.075 |
| **cg00152946** | 10 | 130338781 | | | Island | -1.53 | 0.068 |
| **cg00338996** | 11 | 45828080 | *SLC35C1* | Body | S_Shore | 0.75 | 0.06 |
| **cg00478064** | 11 | 118271813 | *ATP5L* | TSS1500 | N_Shore | 1.61 | 0.09 |
| **cg12079699** | 11 | 131564481 | *NTM* | Body | S_Shelf | -0.77 | 0.101 |
| **cg00104348** | 12 | 10772424 | *STYK1* | 3'UTR | | -0.70 | 0.149 |
| **cg00216961** | 12 | 6054731 | *ANO2* | Body | | -0.91 | 0.189 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **cg00240719** | 12 | 132169130 | | | N_Shore | -0.71 | 0.121 |
| **cg00216180** | 13 | 114778713 | *RASA3* | Body | Island | 0.88 | 0.213 |
| **cg00247571** | 13 | 50707065 | | | N_Shore | -1.15 | 0.027 |
| **cg01729020** | 14 | 55035889 | *SAMD4A* | Body | S_Shore | 0.90 | 0.25 |
| **cg00376910** | 15 | 89765061 | *RLBP1* | TSS200 | | -1.12 | 0.054 |
| **cg00153543** | 16 | 84259830 | *KCNG4* | Body | S_Shelf | 0.54 | 0.219 |
| **cg00261912** | 16 | 89777763 | *C16orf7* | Body | Island | 0.47 | 0.054 |
| **cg00489954** | 16 | 28855387 | *TUFM* | Body | N_Shore | 0.001 | 0.999 |
| **cg00701514** | 16 | 1183462 | | | S_Shelf | -0.84 | 0.075 |
| **cg00911376** | 16 | 84274417 | *KCNG4* | TSS1500 | S_Shelf | 0.67 | 0.162 |
| **cg02455706** | 16 | 84918851 | *CRISPLD2* | Body | | -0.78 | 0.032 |
| **cg00416882** | 18 | 44495671 | *PIAS2* | Body | N_Shore | 0.88 | 0.075 |
| **cg00417323** | 18 | 29307959 | | | S_Shelf | 0.76 | 0.093 |
| **cg00486358** | 18 | 54309442 | | | S_Shelf | 0.54 | 0.178 |
| **cg01080902** | 19 | 44039870 | *ZNF575* | 3'UTR | S_Shore | 1.19 | 0.06 |
| **cg00397859** | 20 | 13768987 | *C20orf7* | Body | S_Shelf | 0.67 | 0.065 |
| **cg00399938** | 22 | 20118646 | *ZDHHC8* | TSS1500 | Island | 0.44 | 0.25 |
| **cg01133262** | 22 | 38863905 | *KDELR3* | TSS200 | Island | 0.87 | 0.226 |

CHR$^{\$}$= chromosome number
UCSC*= University of South Carolina