Supplemental Material

METHODS

Study population

For our epigenome-wide screen, we obtained blood from 106 Caucasian children with cow's milk allergy (CMA) and 77 non-allergic Caucasian controls as defined below from the Chicago Food Allergy Study (the discovery sample), which has been previously described in detail.[1, 2] Briefly, eligible families had at least one biological child with FA and were willing to participate in the study. For each enrolled participant, the following procedures were completed: 1) questionnaire interview by trained research staff to obtain information on home environment, diet, lifestyle, history of food allergy (FA) and other allergic diseases; 2) allergy skin prick testing (SPT) to 9 food- and 6 aero- allergens; and 3) collection of venous blood samples. For each child, we also collected a detailed history of their clinical allergic reaction associated with ingestion of specific foods. The Institutional Review Board (IRB) of Ann & Robert H. Lurie Children's Hospital of Chicago and Johns Hopkins Bloomberg School of Public Health (JHBSPH) approved the study protocol.

For replication purposes, we obtained two independent samples. The first sample consisted of 5 Caucasian cases with allergy to both cow's milk and peanut and 20 positive controls with allergy to peanut but not to cow's milk (see detailed description below for phenotypic classification) enrolled from the same Chicago Food Allergy Study (the Chicago replication sample). These subjects were independent of the discovery sample. The second sample consisted of 140 African-American children from the prospective Boston Birth Cohort (BBC)[3] including 8 cases with CMA and 132 normal controls (the Boston replication sample). Unlike the samples obtained from the Chicago Food Allergy Study, the Boston sample had DNA methylation (DNAm) measured in cord blood (prior to clinical expression of disease). The IRBs at the Boston Medical Center and JHBSPH approved the BBC study protocol.

Phenotype definition

As reported previously,[2, 4] we defined CMA cases in this study using the following stringent criteria: 1) a convincing history of symptoms indicative of an allergic reaction within 2 hours of ingestion of cow's milk; and 2) clear evidence of sensitization defined as having a specific IgE (sIgE) ≥ 0.35 kU/L to cow's milk and/or a positive SPT to cow's milk with mean wheal diameter (MWD) ≥ 3 mm greater than the saline control. Normal controls were defined as a child who had neither a clinical allergic reaction nor evidence of sensitization to any of the common foods (peanut, egg white, cow's milk, soy, wheat, walnut, fish, shellfish, and sesame seed). In the Chicago replication sample, positive controls were defined as a child who had neither a clinical allergic reaction nor evidence of sensitization to cow's milk, but was allergic to peanut.

Skin prick test (SPT) measurement

In the Chicago Food Allergy Study, SPT to 9 food allergens (cow's milk, egg white, soybean, wheat, peanut, English walnut, sesame seed, fish mix [cod, flounder, halibut, mackerel, tuna], and shellfish mix [clam, crab, oyster, scallops, shrimp]) and six aeroallergens (two dust mites [Dermatohagoides teronyssinus, Dermatophagoides farina], cat hair, dog epithelia, cockroach mix, and Alternaria tenius), plus negative (50% glycerinated saline) and positive (histamine, 1.0 mg/mL) controls (Greer, Lenoir, NC, USA) was performed using a Multi-Test II device (Lincoln Diagnostics) as previously described in detail.[2, 4] Data were excluded if the MWD for the negative control was ≥3 mm, or if MWD for the positive control was <3 mm, or if the difference of the positive minus the negative control was <3 mm. A positive SPT was defined as a MWD ≥3 mm for a specific allergen. No SPT measurements were performed in the BBC.

Specific IgE (sIgE) measurement

In the Chicago Food Allergy Study, sIgE for 9 food allergens (egg white, sesame, peanut, soy, milk, shellfish, walnut, cod fish and wheat) and 6 aeroallergens (Dermatophagoides pteronyssinus and

Dermatophagoides farinae, cat dander, dog dander, German cockroach and Alternaria alternata) were measured using the Phadia ImmunoCAP system (Phadia US Inc., Portage, MI, USA), by the Clinical Immunology Laboratory at Lurie Children's, a CLIA-certified laboratory for the ImmunoCAP assay. The detection limit for each sIgE was <0.1 kU_A/L and the reporting range was 0.1 - 100 kU_A/L. In the BBC, sIgE for 8 food allergens (egg white, peanut, soy, milk, shrimp, walnut, cod fish and wheat) was measured at Quest Diagnostics using the Phadia ImmunoCAP system. The detection limit for food sIgE was <0.35 kU_A/L.

DNAm measurement and quality control steps

In the discovery stage, case and control DNA samples (50ng/uL) were evenly distributed across each plate to minimize batch effects, which were then shipped to the Northwestern University Genomics Core for DNAm profiling. Briefly, 0.5 μg of genomic DNA was bisulfite-converted using the EZ-96 DNA Methylation™ Gold Kit, and DNAm levels at 485,512 loci were measured using the Infinium HumanMethylation450 BeadChips (450K), according to the manufacturer's instructions.

A raw intensity file (.idat) for each sample was processed and several quality control steps were performed using the 'minfi'[5] Bioconductor package, as we reported previously.[2] First, we examined the 450K control probes to assess bisulfite conversion, extension, hybridization, staining, specificity and others. Second, we computed the median for both Meth and Unmeth signals for each array and displayed them in a scatter plot. This approach clearly identified one outlier sample with a median $\log_2$ intensity value <10.5, which was removed from further analyses. Third, we removed 748 loci that had a detection p-value (a measure of probe performance) > 0.01 in 10% or more of the samples. Fourth, we removed 21,524 CpG sites that had an annotated single nucleotide polymorphism (SNP, minor allele frequency >0.01) at the measured and neighboring locus (+/- 1bp depending on probe strand orientation), and removed 27,513 CpG sites previously reported to be cross-reactive.[6] After these quality control steps, DNAm of 435,642 sites (including 11,222 X-chromosome sites) from 182 samples were available for subsequent data analyses.

Using the ‘minfi’ package[5], a stratified quantile normalization procedure was applied to the raw data and normalized beta values (β), ranging from 0-1 for 0% to 100% methylated, were obtained. M values (logit transformed β) were also computed. Both the β and M values were ComBat-transformed[7] using the ‘sva’ package[8] with the array number as the surrogate for the batches. ComBat-transformed M values were utilized for downstream statistical analyses since they represent a more normal distribution than β values. ComBat-transformed β values were used for plotting purposes since they are more intuitive. The same protocols for DNAm measurement and data cleaning were applied to the two replication samples.

Empirical estimation of blood cell composition

To account for potential differences in the proportions of cells that comprised the blood,[9-11] we empirically estimated the proportion of CD8 positive T cells, CD4 positive T cells, natural killer cells, B cells, monocytes and granulocytes as previously described by Houseman et al.[10] for all of the discovery and replication samples, using the ‘minfi’ package[5].

Statistical analyses in the discovery stage

To identify differentially methylated positions (DMPs) associated with CMA in the discovery sample, we used the ‘limma’ package[12] to fit a linear regression model for ComBat-transformed M value at each CpG site as a function of CMA status (1=case, 0=control), adjusting for potential confounders including gender, age group (coded as 0=<2 years, 1=2-6 years, 2=6-10 years, 3=≥10 years), breastfeeding history (yes/no), parental history of FA (0=none, 1=one parent, 2=both parents with FA), cell type proportions for granulocytes, monocytes, B cells, NK cells, CD4+ T cells, and CD8+ T cells, as well as genetic ancestry (represented by the first two principal components based on genome-wide SNP data, as we reported previously[2]). Bonferroni correction was applied to account for multiple testing (p<1.15E-07). For CpG sites on the X-chromosome, gender-specific analyses were also performed.

As a complementary approach to single-site DMP analyses, we searched for differentially methylated regions (DMRs) associated with CMA using the bumphunting() function[13] in the 'minfi' package. Briefly, each locus was a priori assigned to a region such that any two neighboring loci separated by 300bp or less were assigned to the same region. Next, for each CpG site, we estimated the difference in average combat-transformed M value between CMA cases and controls, adjusting for the same covariates as described above. Significance was assigned with the application of a bootstrapping method (n=1000 times). CMA candidate regions with a family-wise error (FWER) <0.05 were considered significant. All statistical analyses were performed using R-3.1.1 and Bioconductor 2.12.

We then performed a KEGG pathway analysis to identify biological pathways with enrichment of genes that exhibited differential DNAm between 106 CMA cases and 76 controls, using WebGestalt[14] and with the ~20,000 annotated genes of the studied 435,642 CpG sites as the background. We used a Bonferroni adjusted p-value <0.05 as our significance threshold.

Statistical analyses in the replication samples

We decided to validate a subset of identified CMA-associated DMPs in our replication analyses due to the small number of samples. These subsets were selected based on one or more of the following criteria: 1) significant DMPs (p <1.15E-07) with an absolute DNAm difference >5%; and/or 2) significant DMPs (p <1.15E-07) that annotated to genes relevant to the TH_1-TH_2 pathway, a critical pathway in allergy development. In both replication samples, a linear regression model, similar to the one applied in the discovery stage, was fitted to explore the association of each DMP with CMA. Due to a limited sample size, the potential confounders that we considered in the Chicago sample included child's age and cell type proportions. In the Boston sample, the confounders that we adjusted included cell type proportions, parental history of FA, and gestational age group (due to a significant association between gestational age and cord blood DNAm level[15, 16]). We used an FDR corrected p-value threshold (FDR <0.05) to account for multiple testing in each replication.

For the CMA-associated DMPs that were validated in at least one replication sample, we then explored their biological relevance ,using the web tool, EpiExplorer[17].

Reference:

1. Tsai HJ, Kumar R, Pongracic J, Liu X, Story R, Yu Y, et al. Familial aggregation of food allergy and sensitization to food allergens: a family-based study. Clin Exp Allergy 2009; 39:101-9.
2. Hong X, Hao K, Ladd-Acosta C, Hansen KD, Tsai HJ, Liu X, et al. Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. Nat Commun 2015; 6:6304.
3. Hong X, Wang G, Liu X, Kumar R, Tsai HJ, Arguelles L, et al. Gene polymorphisms, breast-feeding, and development of food sensitization in early childhood. J Allergy Clin Immunol 2011; 128:374-81 e2.
4. Hong X, Caruso D, Kumar R, Liu R, Liu X, Wang G, et al. IgE, but not IgG4, antibodies to Ara h 2 distinguish peanut allergy from asymptomatic peanut sensitization. Allergy 2012; 67:1538-46.
5. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014; 30:1363-9.
6. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 2013; 8:203-9.
7. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007; 8:118-27.
8. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012; 28:882-3.
9. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 2014; 15:R31.
10. 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; 13:86.
11. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One 2012; 7:e41361.
12. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3:Article3.
13. Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int J Epidemiol 2012; 41:200-9.
14. Wang J, Duncan D, Shi Z, Zhang B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Res 2013; 41:W77-83.
15. Cruickshank MN, Oshlack A, Theda C, Davis PG, Martino D, Sheehan P, et al. Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. Genome Med 2013; 5:96.
16. Lee H, Jaffe AE, Feinberg JI, Tryggvadottir R, Brown S, Montano C, et al. DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. Int J Epidemiol 2012; 41:188-99.

17. Halachev K, Bast H, Albrecht F, Lengauer T, Bock C. EpiExplorer: live exploration and global analysis of large epigenomic datasets. Genome Biol 2012; 13:R96.

Table E1. Population characteristics of the 182 Caucasian children included in the discovery stage

Variable	Normal controls N=76	CMA cases N=106	p-value[a]
Child's age (years), mean±SD	5.5±4.0	4.2±2.7	0.010
Boy, n (%)	31 (40.8)	67(63.2)	0.003
Parent history of food allergy, n (%)			0.023
Paternal only	11 (14.5)	19 (17.9)	
Maternal only	7 (9.2)	21 (19.8)	
Both	0	5 (4.7)	
Neither	58 (76.3)	61 (57.6)	
Current maternal smoking, n (%)	20 (26.3)	18 (17.0)	0.127
Current paternal smoking, n (%)	14 (18.4)	19 (17.9)	0.932
Preterm birth, n (%)	8 (10.5)	9 (8.5)	0.632
C-section, n (%)	22 (28.9)	35 (33.0)	0.559
Pets in the home, n (%)	43 (56.6)	49 (46.2)	0.168
Any breastfeeding, n (%)	67 (88.2)	101 (95.3)	0.075
Maternal consumption of cow's milk product during breastfeeding, n (%) [b]			0.585
≤ 2 days	2 (4.4)	5 (5.6)	
3-5 days	9 (19.6)	20 (22.2)	
≥ 6 days	35 (76.1)	62 (68.9)	
Stop consumption during the first 3 months	0	3 (3.3)	
Antibiotic use during the 1st year of life			0.293
Yes	36 (47.4)	38 (35.9)	
No	38 (50.0)	65 (61.3)	
Unsure	2 (2.6)	3 (2.8)	
sIgE to cow's milk (kU/L), median (25th -75th)[c]	0	23.1 (6.2-59.6)	<0.001
SPT to cow's milk (mm), median (25th -75th)[d]	0	10.5 (8.5-14.0)	<0.001
Other types of food allergy	0	49 (46.3)	<0.001
Egg allergy	0	36 (34.0)	<0.001
Peanut allergy	0	7 (6.6)	<0.001

CMA: cow's milk allergy. DNAm: DNA methylation. SD: standard deviation. sIgE: specific IgE. SPT: skin prick test.

[a] Comparison between CMA cases and controls was performed via chi-square test and t-test, respectively, for categorical and continuous variables.

[b] This analysis was limited to 146 children (46 normal controls and 90 CMA cases) with any breastfeeding and aged < 7 years.

[c] Four children (1 control and 3 cases) had missing data on milk-specific IgE.

[d] 34 children (6 controls and 28 cases) had missing data on milk SPT.

Table E2. Population characteristics of the Caucasian replication sample from the Chicago Food Allergy Study and of the African-American replication sample from the prospective BBC birth cohort

	Caucasian replication sample			African-American replication sample		
	Controls	Cases	p	Controls	Cases	p
n	20	5		132	8	
Child's age (years), mean±SD [a]	9.1±2.5	7.2±4.1	0.331	3.0±3.3	3.8±2.1	0.312
Boys, n (%)	13 (65.0)	4 (80.0)	0.520	81(61.4)	5 (62.5)	0.949
Preterm birth, n (%)	5 (25.0)	0	0.211	57 (43.2)	2 (25.0)	0.312
Current maternal smoking, n (%)	5 (25.0)	1 (20.0)	0.815	24(18.2)	2 (25.0)	0.630
Current paternal smoking, n (%) [b]	5 (25.0)	1 (20.0)	0.461	41(31.0)[b]	3 (37.5)	0.703
Parental history of food allergy, n (%) [c]	7 (35.0)	4 (80.0)	0.070	23 (18.9)[c]	4 (57.1)[c]	0.023
Any breastfeeding, n (%)	19 (95.0)	5 (100.0)	0.610	100(75.8)	6 (75.0)	0.961
Eczema, n (%)	13 (65.0)	4 (80.0)	0.520	0	4 (50.0)	<0.001
Hay fever, n (%)	7 (35.0)	5 (100.0)	0.009	24 (18.2)	6 (75.0)	<0.001
Asthma, n (%)	8 (40.0)	5 (100.0)	0.016	30 (22.7)	5 (62.5)	<0.002
Other types of food allergy, n (%)	20(100.0)[d]	5 (100.0)	1.00	0	3 (37.5)	<0.001

DNAm: DNA methylation. SD: standard deviation.

[a] Child's age when cow's milk allergy was defined.

[b] 15 controls had missing data on current paternal smoking.

[c] 10 controls and 1 case had missing data on parental history of FA.

[d] All the non-CMA controls were allergic to peanut.

Table E3. KEGG pathways with enrichment of genes that exhibit differential DNA methylation in children with CMA

Pathway	N of genes			Ratio[d]	Raw p-value[e]	Adjusted p-value[f]	Observed genes
	Total[a]	O[b]	E[c]				
Butirosin and neomycin biosynthesis	5	3	0.10	31.32	6.8E-05	0.004	*HK1, HK2, HK3*
Starch and sucrose metabolism	43	6	0.79	7.62	0.0001	0.003	*HK1, HK2, HK3, GPI, AGL, GUSB*
Fructose and mannose metabolism	35	5	0.67	7.80	0.0004	0.011	*HK1,HK2, HK3, PFKFB3, FB1*

CMA: cow's milk allergy; *HK1*: *hexokinase 1; HK2: hexokinase 2*; *HK3*: *hexokinase 3*; *GPI: glucose-6-phosphate isomerase*; *AGL: amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase*; *GUSB: glucuronidase, beta*; *PFKFB3:* 6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3; *FB1: TCF3 (E2A) fusion partner.*

[a]The total number of the reference genes in each pathway; [b]the total number and [c]the expected number of genes in each pathway that exhibited differential DNA methylation in children with CMA.

[d]Ratio = the number of observed genes (O) / the number of expected genes (E).

[e]Raw p-value from hypergeometric test.

[f]Bonferroni correction was applied to adjust for multiple testing.

Table E4. Biological relevance[a] of the eight CMA-associated DMPs that were validated in at least one replication sample.

DMP	CHR	Gene	Location	CPG island	DHS	Repeat element	LAD[d]	TFBS[e]	Polycomb repressed gene	Enhancer
cg16386158	2	*IL1RL1*	TSS1500			SINE		CTCF		√
cg13316148	2	*STAT4*	Body				√	CTCF	√	√
cg08404225	3	*IL5RA*	5'UTR				√			√
cg26787239	5	*IL4*	TSS1500				√	CTCF		√
cg09377531	8	*TRAPPC9*	Body			LINE				
cg11770323[b]	13	*NDFIP2*	Body		√ [c]			CTCF		√
cg18550847[b]	14	*EVL*	3'UTR	√						√
cg06040872	17	*CCL18*	Body						√	

CHR: chromosome; DHS: DNaseI hypersensitive sites; DMP: differentially methylated position; LAD: lamina associated domains; LINE: Long interspersed nuclear element; SINE: short interspersed nuclear element; TFBS: Transcription factor binding sites. TSS: transcription start sit; UTR: untranslated region.

[a] Biological relevance was analyzed using the web tool, EpiExplorer.

[b] Overlapped with a conserved region

[c] In mammary epithelial cells.

[d] LAD: lamina associated domains. An event of hypomethylation of these regions may cause or reflect a fault of the mechanisms that are central to the development and conservation of normal states of differentiation and tissue-specific patterns of gene expression.

[e] CTCF: CCCTC binding factor, a transcriptional regulator which may protect downstream DNA from upstream methylation activities

Fig Legend

Fig E1. A genomic cluster in the *EVL* gene that hypomethylated in CMA children compared to normal controls. 1a). The upper panel shows DNA methylation levels on the y-axis and genomic location on the x-axis. Each point represents the methylation level at a specific CpG site for each CMA case (in blue) and control (in black). The lower panel displays the methylation difference between cases and controls for the six involved CpG sites (cg26415437, cg16409452, cg14084609, cg06756385, cg18550847 and cg01000631);* p < 1.15E-07 for the methylation difference between cases and controls**.** 1b). DNA methylation correlation for the genomic cluster shown in fig E1a. The measure of correlation coefficient (r) among each pair of CpG is shown graphically, with blue representing high positive correlation.

Fig E2. Scatter plots of DNA methylation levels at 8 CMA-associated DMPs that were validated in at least one replication sample. (A) Plot for each DMP in the discovery sample (n=182); (B) Plot for each DMP in the Chicago replication sample; (C) Plot for each DMP in the Boston replication sample. The red segment denotes the median DNA methylation levels within each subset of subjects.

Figure E1a.

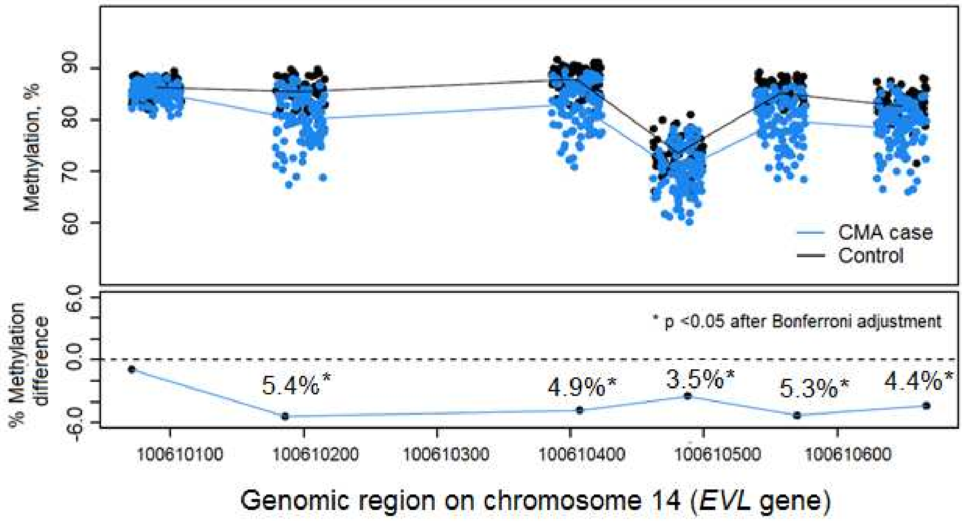


Figure E1b.

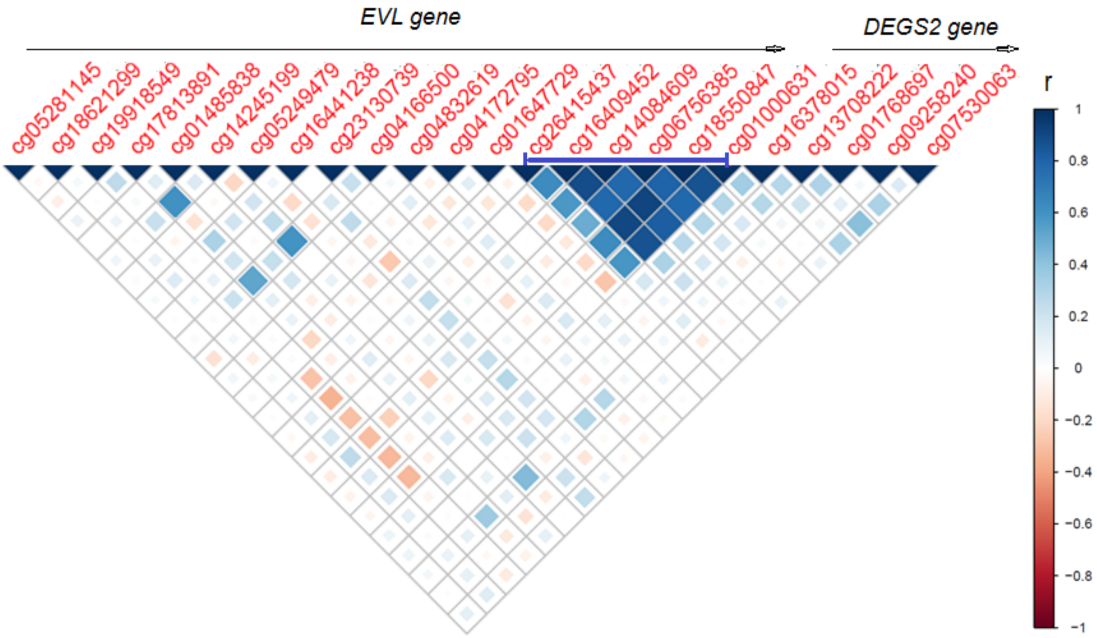
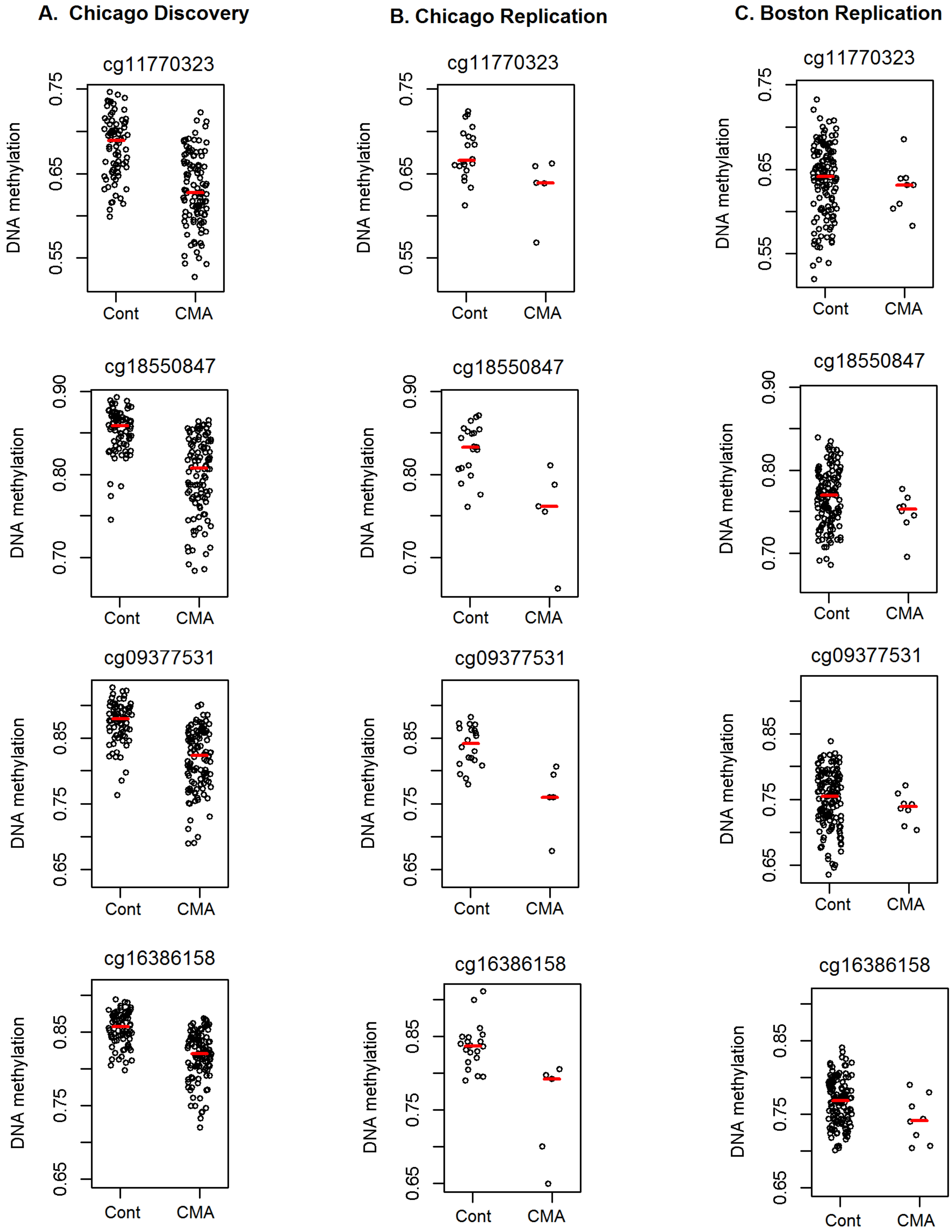


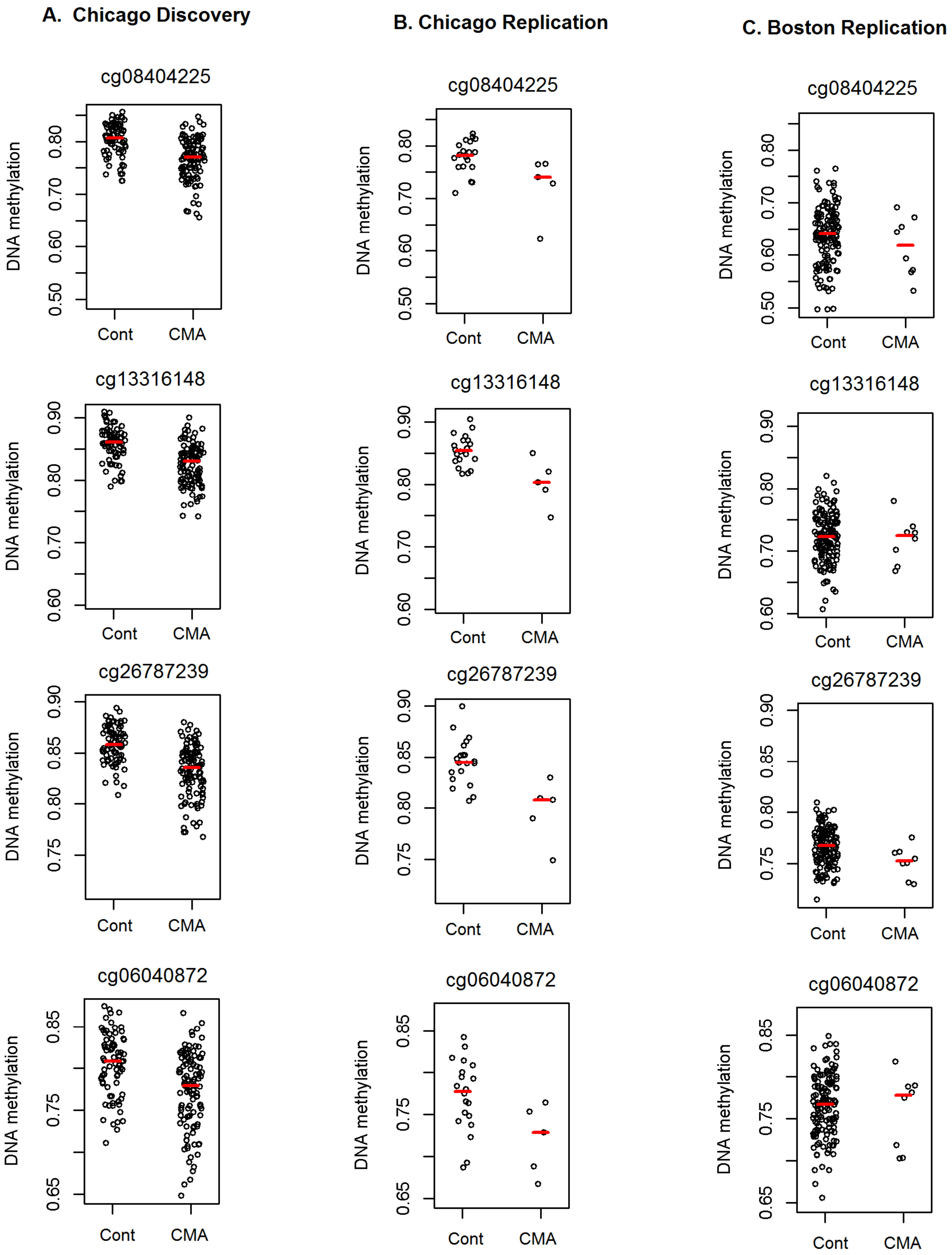
Figure E2



A. Chicago Discovery

B. Chicago Replication

C. Boston Replication


A. Chicago Discovery
B. Chicago Replication
C. Boston Replication
cg08404225
cg13316148
cg26787239
cg06040872
DNA methylation
0.50
0.60
0.70
0.80
0.90
0.65
0.75
0.85
Cont
CMA