Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue

Jianxin Shi, Crystal N. Marconett, Jubao Duan, Paula L. Hyland, Peng Li, Zhaoming Wang, William Wheele, Beiyun Zhou, Mihaela Campan, Diane S. Lee, Jing Huang, Weiyin Zhou, Tim Triche, Laufey Amundadottir, Andrew Warner, Amy Hutchinson, Po-Han Chen, Brian S.I. Chung, Angela C. Pesatori, Dario Consonni, Pier Alberto Bertazzi, Andrew W. Bergen, Mathew Freedman, Kimberly D. Siegmund, Benjamin P. Berman, Zee Borok, Nilanjan Chatterjee, Margaret A. Tucker, Neil E. Caporaso, Stephen J. Chanock, Ite A. Laird-Offringa & Maria Teresa Landi

The genetic regulation of the human epigenome is not fully appreciated. Here we describe the effects of genetic variants on the DNA methylome in human lung based on methylation-quantitative trait loci (meQTL) analyses. We report 34,304 cis- and 585 trans-meQTLs, a genetic–epigenetic interaction of surprising magnitude, including a regulatory hotspot. These findings are replicated in both breast and kidney tissues and show distinct patterns: cis-meQTLs mostly localize to CpG sites outside of genes, promoters and CpG islands (CGIs), while trans-meQTLs are over-represented in promoter CGIs. meQTL SNPs are enriched in CTCF-binding sites, DNaseI hypersensitivity regions and histone marks. Importantly, four of the five established lung cancer risk loci in European ancestry are cis-meQTLs and, in aggregate, cis-meQTLs are enriched for lung cancer risk in a genome-wide analysis of 11,587 subjects. Thus, inherited genetic variation may affect lung carcinogenesis by regulating the human methylome.
DNA methylation plays a central role in epigenetic regulation. Twin studies have suggested that DNA methylation at specific CpG sites can be heritable; however, the genetic effects on DNA methylation have been investigated only in brain tissues, adipose tissues and lymphoblastoid cell lines. Most studies were based on the Illumina HumanMethylation27 array, which has a low density and mainly focuses on CpG sites mapping to gene promoter regions. While the functional role of DNA methylation in non-promoter or non-CpG Island (CGI) regions remains largely unknown, evidence shows roles in regulating gene splicing and alternative promoters, silencing of intragenic repetitive DNA sequences, and predisposing to germline and somatic mutations that could contribute to cancer development. Notably, a recent study suggests that most DNA methylation alterations in colon cancer occur outside of promoters or CGIs, in so-called CpG island shores and shelves, and the Cancer Genome Project has reported high mutation rates in CpG regions outside CGI in multiple cancers. Although expression QTLs (eQTLs) have been extensively studied in different cell lines and tissues, the minimal overlap observed between cis-acting meQTLs and eQTLs (≈5–10%) emphasizes the necessity of mapping meQTLs that may function independently of nearby gene expression. This might reveal novel mechanisms for genetic effects on cancer risk, particularly since many of the established cancer susceptibility SNPs map to non-genic regions.

Lung diseases constitute a significant public health burden. About 10 million Americans had chronic obstructive pulmonary disease in 2012 (ref. 16) and lung cancer continues to be the leading cancer-related cause of mortality worldwide. To provide functional annotation of SNPs, particularly those relevant to lung diseases and traits, we systematically mapped meQTLs in 210 histologically normal human lung tissues using Illumina Infinium HumanMethylation450 BeadChip arrays, which provide a comprehensive platform to interrogate the DNA methylation status of 485,512 cytosine targets with excellent coverage in both promoter and non-promoter regions (Fig. 1a), CGI and non-CGI regions (Fig. 1b) and gene and non-gene regions. Thus, our study enables the characterization of genetic effects across the methyome in unprecedented detail. Moreover, since DNA methylation exhibits tissue-specific features, we investigated whether similar meQTLs could be identified in other tissues.

**Results**

**Identification of cis-acting meQTLs.** We profiled DNA methylation for 244 fresh-frozen histologically normal lung samples from non-small cell lung cancer (NSCLC) patients from the Environment and Genetics in Lung cancer Etiology (EAGLE) study. A subset of 210 tissue samples that passed quality control and had germline genotype data from blood samples was used for meQTL analysis. The analysis was restricted to 338,456 autosomal CpG probes after excluding those annotated in repetitive genomic regions or that harboured genetic variants. The distribution of methylation levels differed strongly across distinct types of genomic regions (Supplementary Fig. 1a,b). Consistent with previous studies, CpG sites in promoter or CGI regions were largely unmethylated while those in other regions were largely methylated (Supplementary Fig. 1a,b).

**Figure 1 | cis-meQTL structural characteristics.** (a) Distribution of CpG probes and corresponding cis-meQTL numbers and proportions in gene and non-gene regions. meQTLs were detected based on EAGLE lung normal tissue samples (n = 210). (b) Distribution of CpG probes and corresponding cis-meQTL numbers and proportions in CpG islands (CGIs), shores (<2 kb from the boundary of CGI), shelves (2–4 kb from the boundary of CGI) and the remaining region or ‘open sea’. (c) The strongest cis-association is between SNP rs10090179 and CpG probe cg19504605, P = 1.5 × 10−73, t-test. The SNP explains 79.8% of the phenotypic variance. The box plots show the distribution of the methylation levels in each genotype category with error bars representing the 25 and 75% quantiles. (d,e) The x coordinate is the average s.d. of methylation levels for CpG probes in each category. The y coordinate is the proportion of CpG probes detected with cis-meQTLs. The proportion of methylation probes detected with cis-meQTLs varied across categories, ranging from 4.0% for CGIs in 1st exons to 15.7% for south shores in non-gene regions.
We performed cis-meQTL analysis for each methylation trait by searching for SNPs within 500 kb of the target CpG-site in each direction (1 Mb overall). The genetic association was tested under an additive model between each SNP and each normalized methylation probe, adjusting for sex, age, plate, population stratification and methylation-based principal component analysis (PCA) scores. Controlling FDR at 5% (\(P = 4 \times 10^{-5}\)), we detected cis-meQTLs for 34,304 (10.1% of 338,456) CpG probes (Supplementary Table 1), mapping to 9,330 genes. A more stringent threshold (\(P = 6 \times 10^{-6}\) at FDR = 1%) detected cis-meQTLs for 27,043 CpG probes, mapping to 8,479 genes. Moreover, with a 200-kb window (100 kb from both sides) instead of a 1-Mb window we detected 40,650 cis-meQTLs (\(P = 2 \times 10^{-4}\)), controlling for FDR = 5%. The methylation distribution in CpG sites detected with meQTLs differed substantially from those without meQTLs (Supplementary Fig. 1a,b). The peak SNPs were equally distributed on either side of the target CpG sites with a median distance (\(\Delta\)) of 11.8 kb. The proportion of explained phenotypic variance (\(h^2\)) ranged from 7.7 to 79.8% (Supplementary Fig. 1c) and inversely to the variability of the cis-residues, and for regulation of tissue-specific patterns of methylated cytosine dinucleotides in 3'-UTR regions, where microRNA target sites are typically located, showed an opposite trend in both cis- and trans-meQTLs (Fig. 2b).

In 62.8% of the trans associations, the SNPs involved were also detected to have cis-acting effects. We investigated whether trans associations were mediated by these cis-regulated proximal CpG sites (Fig. 2c,d). We found that 30 and 166 trans associations had full and partial mediation, respectively, while 389 had no significant mediation. The trans associations involving SNPs in gene desert regions are less likely to be mediated by proximal CpG probes (15.7 versus 34.3%; \(P = 0.0067\), Fisher’s exact test). To obtain mechanistic insight into the trans associations showing mediation effects (\(n = 196\)), we used the DAVID tool\(^{23}\) to characterize the function of genes harbouring the mediating cis-CpG probes. The analysis was performed for 115 genes after excluding the major histocompatibility complex (MHC) region because of long-range complex LD patterns. The GO analysis revealed three top gene categories with nominal significance involved in DNA methylation regulation, including GTPase-activity related genes (\(P = 0.004\), Fisher’s exact test), genes regulating transcription (\(P = 0.02\)) and genetic imprinting (\(P = 0.04\), Fisher’s exact test, Supplementary Table 2).

Notably, 106 trans SNPs with \(P < 2.5 \times 10^{-10}\) were associated with multiple distal CpG probes, suggesting that they are multi-CpG regulators. In particular, we detected one master regulatory SNP, rs12933229 located at 16p11.2, in a very large intron of the NPIPL1 gene, which was associated with the methylation of CpG sites annotated to five genes on different chromosomes (Fig. 2a, Supplementary Fig. 2 and Supplementary Table 3). These associations were partially mediated by a proximal CpG probe cg06871736. All five trans associations were replicated in TCGA. The trans associations show a consistent direction, with the ‘C’ allele associated with higher methylation levels. All five regulated target sites are in CGIs, and three are in gene promoter regions. We evaluated the association with gene expression for these three CpG probes, using 28 TCGA histologically normal lung tissue samples with RNA sequencing data. Based on this limited sample size, two of the target genes, PABPC4 and STARD3, showed decreased expression with increased methylation (FDR = 10%).

**Table 1 | Replication of EAGLE lung meQTLs in TCGA histologically normal tissue samples.**

| Tissue | N | All cis associations in EAGLE (34,304 associations, \(P < 4.0 \times 10^{-5}\)) | Strong cis associations in EAGLE (12,083 associations, \(P < 1.0 \times 10^{-10}\)) | All trans associations in EAGLE (585 associations, \(P < 2.5 \times 10^{-10}\)) |
|--------|---|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
|        |   | Consistent direction FDR < 0.05 | Consistent direction FDR < 0.05 | Consistent direction FDR < 0.05 |
| Lung   | 65 | 32,128 (93.7%) | 22,441 (65.4%) | 11,250 (99.3%) | 11,229 (92.9%) | 556 (95.2%) |
| Breast | 87 | 30,391 (88.6%) | 18,762 (54.7%) | 11,640 (96.3%) | 9,987 (82.7%) | 561 (96.1%) |
| Kidney | 142| 30,975 (90.3%) | 23,984 (70.0%) | 11,634 (96.3%) | 10,783 (89.2%) | 558 (95.5%) |

\(N\) is the sample size in replication studies. FDR was calculated based on single-sided \(P\)-values.
Enrichment of meQTLs in DNA regulatory regions. SNPs associated with complex diseases in genome-wide association studies (GWAS) or with eQTLs have been reported to be enriched in ENCODE-annotated regulatory regions. These include DNaseI hypersensitivity sites, CTCF-binding factor (CTCF) binding sites and regions enriched in active and repressive histone modification marks. The large number of meQTLs detected in our study, both cis and trans, enabled us to systematically investigate their enrichment in regulatory regions. We performed enrichment analysis using Chip-Seq data in small airway epithelial cells (SAEC) from the ENCODE project for histone marks, CTCF occupancy and DNaseI hypersensitivity sites, and histone marks in primary human alveolar epithelial cells (hAEC) from our own laboratory. Compared with the ‘control’ SNP set not associated with the methylation of CpG sites (with minor allele frequency (MAF) and CpG probe density matched with meQTL SNPs), the meQTL SNPs were strongly enriched for sites of CTCF, DNaseI hypersensitivity and histone marks (H3K4me3, H3K9-14Ac and H3K36me3) associated with active promoters, enhancers and active transcription, and to a lesser extent for the repressive mark H3K27me3 (Table 2). Enrichment of all regulatory regions became stronger with increasing significance of association, with the exception of the H3K27me3 repressive mark (Fig. 3). Using SAEC CTCF ChIP data, we found that meQTL SNPs or associated SNPs in high LD located within CTCF consensus sequences can affect allele-specific binding of CTCF (see two examples in Supplementary Figs 3 and 4).

Lung cancer risk SNPs affect methylation in human lung tissue. To determine whether the identified meQTLs might provide functional annotation to the established genetic associations with lung cancer risks, we examined SNPs in five genomic regions reported to be associated with lung cancer risk in GWAS of populations of European ancestry: 15q25.1 (refs 30–32) (CHRNA5–CHRNA3–CHRNB4), 5p15.33 (refs 20,33,34), 6p21.33 (ref. 33) (BAT3, most strongly associated with squamous cell

Figure 2 | trans-meQTL structural characteristics. (a) Circos plot for trans-meQTLs. The outer rim shows the log10 P-values Manhattan plots of trans-meQTL associations. The innermost network depicts spokes between all trans-meQTL SNPs and their target CpG sites. The red spokes show a master regulatory SNP rs12933229 located at 16p11.2 associated with methylation of CpG sites located in CGIs annotated to five genes. (b) Proportion of CpG probes detected with cis-meQTLs and trans-meQTLs across gene regions. The asterisks ‘*, **, ***’ indicate t-test P<0.05, 0.01 and 0.0001, respectively, for the comparison between CGI and non-CGI regions. CGI regions are strongly enriched with trans-meQTLs, while non-CGI regions are enriched with cis-meQTLs. CpG sites in 3' UTR regions show an opposite trend. (c) The association between a SNP denoted as G and a distal CpG-site B may be mediated through a proximal CpG-site A. (d) For each trans-association (G, B) pair, the dots show their marginal versus partial correlation coefficients upon conditioning on the proximal A CpG probes. Analysis was based on 210 samples. Reduction of correlation coefficients by conditioning on A suggests the magnitude of the mediation effect.
Table 2 | Chromatin marks are enriched on meQTL SNPs.

| Cell line | Mark | Control proportion (%) | cis only Proportion (%) | cis only Fold change | trans only Proportion (%) | trans only Fold change | cis + trans Proportion (%) | cis + trans Fold change |
|-----------|------|------------------------|------------------------|---------------------|--------------------------|------------------------|--------------------------|-------------------------|
| SAEC      | CTCF | 11.8                   | 35.3                   | 3.0                 | 29.6                     | 2.5                    | 45.4                     | 3.8                     |
|           | DNasel | 25.4                   | 54.0                   | 2.1                 | 45.8                     | 1.8                    | 59.6                     | 2.3                     |
|           | H3K27me3 | 20.4                   | 34.1                   | 1.7                 | 25.4                     | 1.2                    | 42.9                     | 2.1                     |
|           | H3K4me3 | 4.8                    | 29.7                   | 6.2                 | 18.0                     | 3.8                    | 39.9                     | 8.3                     |
|           | H3K36m3 | 13.4                   | 36.8                   | 2.7                 | 22.8                     | 1.7                    | 45.4                     | 3.4                     |
| hAEC      | H3K27me3 | 17.5                   | 25.3                   | 1.4                 | 15.6                     | 0.9                    | 33.2                     | 1.9                     |
|           | H3K4me3 | 7.6                    | 37.0                   | 4.9                 | 25.0                     | 3.3                    | 54.9                     | 7.2                     |
|           | H3K9-14Ac | 17.3                  | 47.6                   | 2.8                 | 32.3                     | 1.9                    | 65.3                     | 3.8                     |

meQTL SNPs were enriched in chromatin marks, including CTCF binding sites, DNasel hypersensitive sites and histone marks from SAEC from ENCODE and hAEC from our laboratory. A SNP is determined to be related with a regulatory region if the SNP or any LD-related SNP ($r^2 \geq 0.8$) resides in the ChIP-Seq peaks of the regulatory regions. Enrichment for cis-meQTL SNPs without trans effects (‘cis only’), trans-meQTL SNPs without cis effects (‘trans only’) and SNPs with both trans and cis effects (‘cis + trans’). The baseline proportion (control set) was calculated based on SNPs not associated with meQTLs and with minor allele frequencies and local CpG probe-density matching to the meQTL SNPs. The fold changes were calculated using the control set as baseline.

Figure 3 | Chromatin marks are increasingly enriched on meQTL SNPs with larger effect sizes. (a) We split cis-meQTL SNPs into five categories according to the meQTL association strength ($P > 10^{-7}$, $10^{-7} > P > 10^{-10}$, $10^{-10} > P > 10^{-15}$, $10^{-15} > P > 10^{-20}$, and $P < 10^{-20}$). A SNP is determined to be related with a regulatory region if the SNP or any LD-related SNP ($r^2 \geq 0.8$) resides in the ChIP-Seq peaks of the regulatory regions. Regulatory elements include CTCF-binding sites, DNasel hypersensitive sites and histone marks from SAEC from ENCODE and hAEC from our laboratory. For each P-value category, we calculated the proportions of cis-meQTL SNPs related with regulatory regions. The figures show that the proportions of cis-meQTL SNPs related with regulatory regions increase with the significance of meQTL associations except for the repressive mark H3K27me3.

carcinoma or SQ, 12p13.3 (ref. 35) (RAD52 for SQ) and 9p21.3 (ref. 36) (CDKN2A/CDKN2B, particularly for SQ). The GWAS SNPs at 15q25.1 were reported to be associated with total expression levels and multiple isoforms of CHRNA5 in normal lung tissue samples37,38. The GWAS SNPs at the other four loci have not been reported to be associated with the total expression of nearby genes. Consistently, we did not observe an association in RNA-seq data from TCGA lung normal tissue samples ($n = 59$), although a detailed investigation of alternative promoters, splice sites and allele-specific gene expression in larger studies is warranted. Here we investigated whether these SNPs contributed to lung cancer risk with epigenetic regulation by examining their associations with DNA methylation levels.

The top GWAS SNPs located at 15q25.1, 5p15.33, 6p21.33 and 12p13.3 were all strongly associated with the methylation of the nearby CpG probes and the associations were replicated in TCGA lung data (Fig. 4). Importantly, five of the six GWAS SNPs at these loci, excluding the RAD52 locus, were also the SNPs with the strongest association with the corresponding CpG probes. For the cg22937753 probe located in the RAD52 locus, another SNP, rs724709, with weak correlation with the GWAS SNP ($r^2 = 0.1$) had the strongest association with meQTL. All involved CpG sites are located within gene bodies (which may affect gene splicing39) or the 3'UTR regions. No meQTL was detected for 9p21.3 (Supplementary Fig. 5), possibly because of fewer CpG dinucleotide probes available in this gene region on the Illumina platform. The location of these lung cancer GWAS-associated CpG sites might identify which genes within the relevant regions are more likely associated with the risk SNPs, something that is particularly important for regions with complex LD structure, as the MHC region on 6p21. In MHC, two GWAS SNPs in complete LD ($r^2 = 1$), rs3117582 (BAT3) and rs3131379 (MSH5), were most strongly associated with the methylation of CpG sites located nearby of MSH5 (involved in DNA mismatch repair and meiotic recombination process), suggesting that MSH5 ($P = 5.4 \times 10^{-13}$, t-test) is more likely to be involved in lung carcinogenesis than BAT3 ($P = 8.8 \times 10^{-5}$, t-test) or that the SNP closer to MSH5 (rs3131379) is more likely to be the SNP most responsible of the GWAS association with lung cancer risk (Fig. 4b). Our meQTL data also show that rs3131379 trans-regulated the methylation level of CpG probe cg12093005, located in the body of FBRS11 at 12q24 ($P_{EAGLE} = 4.0 \times 10^{-9}$, $P_{TCGA} = 7.2 \times 10^{-4}$ and $P_{combined} = 5.4 \times 10^{-11}$, t-test). Thus, this known GWAS locus might affect lung cancer risk through a gene located on a different chromosome.

Of note, on the 15q25.1 locus, two independent lung cancer risk SNPs, rs2036534 and rs1051730, were associated with CpG probes not linked with CHRNA5 expression. In Supplementary Fig. 6, we show that the two SNPs jointly regulated another methylation probe cg22563815 within the CHRNA5 promoter, which is associated with CHRNA5 expression. This extends and further confirms the complex regulatory pattern with multiple SNPs previously observed for this locus45.

Most subjects in the analyses were smokers ($n = 206$). Adjustment for smoking status (former and current) or intensity (pack/years) did not change the results.

cis-meQTLs are enriched in lung SQ risk. We investigated whether the identified cis-meQTL SNPs were enriched in the National Cancer Institute (NCI) lung cancer GWAS including 5,399 cases and 5,848 controls of European ancestry19. To focus
on potentially new genetic risk associations, we excluded the top lung cancer GWAS SNPs mentioned above and their surrounding regions. We tested the enrichment by examining whether the GWAS $P$-values for the LD-pruned cis-meQTL SNPs deviated from the uniform distribution, that is, no enrichment. When all cis-meQTL SNPs were analysed together, we detected a strong enrichment for overall lung cancer risk ($P<10^{-4}$, based on 10,000 permutations), which was primarily driven by the enrichment in SQ ($P<10^{-4}$, based on 10,000 permutations) (Fig. 5a). The genomic control $\lambda$ values based on genome-wide SNPs showed that the type-I error rates of our enrichment test were not inflated ($\lambda=1.01$ and 1.00, for overall lung cancer and SQ, respectively). Stratified analyses further enriched the definition of cis-meQTL SNPs regulating CpG sites mapping to north shore (Fig. 5b) and gene body (Fig. 5c) regions (see Supplementary Fig. 7 for the quantile–quantile plot). These gene bodies and north shores were enriched for genes involved in cancer pathways ($P=2.5 \times 10^{-4}$, Fisher’s exact test), and particularly those in NSCLC pathway (for example, AKT1, MAPK1, RASSF5, and so on, Supplementary Table 4). In contrast, cis-meQTLs related with CGI regions or promoters were not enriched with the risk of overall lung cancer or any lung cancer subtype, further emphasizing the need to comprehensively study the methylene to identify functional mechanisms for GWAS findings and identify new genetic loci.

As the meQTL SNPs affecting CpG sites in gene body/non-CGI regions were mostly enriched for SQ risk (Fig. 5d), we performed further analysis in this category by integrating the ENCODE SAEC data. We chose SAEC data because this cell type may be involved in SQ development. We restricted enrichment analysis to the ‘regulatory’ meQTL SNPs, which localized in the CTCF binding regions, DNaseI hypersensitive sites or histone marks (H3K27me3, H3K4me3 and H3K36me3) or had at least one LD SNP ($r^2 \geq 0.95$) residing in these regions. The strong enrichment in SQ was driven by SNPs overlapping with CTCF binding sites ($P<10^{-4}$, based on 10,000 permutations) or the repressive mark H3K27me3 ($P<10^{-4}$, based on 10,000 permutations) (Fig. 5e). The enrichment test was not significant after excluding the SNPs overlapping with these regulatory regions ($P=0.14$, based on 10,000 permutations).

Replication of meQTLs in TCGA breast and kidney tissues. To explore the tissue specificity of the genetic effects on DNA methylation, we examined whether the meQTLs detected in EAGLE lung tissue data could be replicated in TCGA breast ($n=87$) or kidney ($n=142$) histologically normal tissue samples, the only two organs to date with data available for a large number of normal tissues of European ancestry. Results are provided in Table 1 and Supplementary Fig. 8. For both cis- and trans-meQTLs, a large proportion of associations had the same direction of EAGLE meQTLs in both breast and kidney samples. For cis associations, 54.7 and 70.0% were replicated with FDR $=5\%$ based on single-sided $P$-values in two data sets, respectively. For the strong cis associations with $P<10^{-5}$ in EAGLE, the replication rates increased to 82.7 and 89.2% in the two data sets. For trans associations, 83.4 and 86.4% were replicated in breast and kidney samples, respectively. The detected master regulator (Fig. 2a) was strongly replicated in both data sets (Supplementary Table 3). Interestingly, some cis-meQTLs, but not trans-meQTLs, had an opposite but very strong association ($P<10^{-5}$) in breast ($n=7$) or kidney ($n=58$) compared with the EAGLE lung data, a phenomenon previously reported in a cell type-specific eQTL study46.

Discussion

We found that inherited genetic variation profoundly and extensively impacts DNA methylation in target organs. Based on high-density methylation arrays in a large sample size, we identified 34,304 cis-meQTLs and 585 trans-meQTLs, one to two
orders of magnitude larger compared with previous studies.\textsuperscript{3-5,7} meQTLs involved nearly half of the autosomal genes, of which 9,330 in cis and 373 in trans, with 9,525 unique genes in total. We show that \(~10\%\) of the cis-meQTLs were affected by at least two SNPs independently. Moreover, we detected a master regulator SNP associated with the methylation levels of five CpG probes on different chromosomes, demonstrating the existence of regulatory hotspots for DNA methylation, as previously shown for eQTL.\textsuperscript{41,42} Most meQTLs were replicated in independent histologically normal lung tissue samples from TCGA. We also showed a high similarity of genetic control on DNA methylation across different tissues. Our findings show that genetic effects on DNA methylation are extensive in scale and complex in structure across the whole genome and suggest a series of important biological implications.

First, our results show that the genomic architecture surrounding cis- and trans-meQTLs is distinct. cis-meQTLs are very large in number, impact predominantly the CpG sites mapping to non-gene regions, and when they occur in genes, are mostly in non-promoter and non-CpG island regions. In contrast, trans-meQTLs are rarer, mainly affect promoter CGI regions, and may be associated with distal CpG sites through the mediation effect of proximal CpG sites.

We found preliminary evidence that the cis-CpG sites mediating the trans-meQTL associations were enriched for genes involved in methylation regulation, such as genes encoding for GTPase or proteins involved in genetic imprinting. GTPase-related gene pathways appear to modulate expression of DNA methyltransferases.\textsuperscript{43} Methylation-induced expression changes of these genes may result in further methylation changes of other genes (that is, in trans). Moreover, a noncoding RNA within the intron of KCNQ1, a key gene regulating genetic imprinting, can influence chromatin three-dimensional (3D) structure via a protein complex including DNA methyltransferase proteins.\textsuperscript{44,45} These findings suggest intricate mechanisms for trans-regulating effects through proximal methylation.

cis-meQTLs may affect cancer risk. To understand the functional consequences of GWAS loci is challenging and multiple principles for post-GWAS functional characterization of genetic loci have been proposed, including the exploration of epigenetic mechanisms.\textsuperscript{46} In our study, the top GWAS lung cancer loci were strongly associated with methylation levels of CpG sites in nearby gene bodies through cis-regulation, and adjusting for smoking status or intensity did not change the results. Furthermore, SNPs affecting the DNA methylation of gene bodies (which are typically methylated) were also collectively associated with the risk for SQ after excluding the established GWAS loci, and were enriched for genes in cancer pathways. In contrast, no enrichment was observed for SNPs affecting the methylation of gene promoters or CGI regions, which are typically not methylated in normal tissues. This suggests a potential novel mechanism for genetic effects on cancer risk.

Figure 5 | Enrichment of cis-meQTL SNPs for lung cancer risk. Analysis based on NCI lung cancer GWAS data (5,739 cases and 5,848 controls). P-values were produced based on 10,000 permutations. AD, SQ and SC represent adenocarcinoma, squamous cell carcinoma and small cell carcinoma. (a) Enrichment was tested using all cis-meQTL SNPs after LD pruning. (b,c) Strong enrichments were observed for cis-meQTL SNP associated with CpG probes annotated to north shores and gene body regions for SQ. (d) The enrichment in (c) was driven by the cis-meQTLs SNPs impacting CpG probes in non-CpG islands. (e) The enrichment in (d) is driven by the SNPs (or their LD SNPs with \(r^2 > 0.95\)) overlapping with CTCF binding sites or H3K27me3 mark regions.
In fact, gene body-enriched cis-meQTLs outside CGI regions may increase the risk for germline and somatic mutations due to their increased propensity to become mutated\textsuperscript{41,42}. Upon spontaneous hydrolytic deamination, methylated cytosine residues turn into thymine, which are less likely to be efficiently repaired than the uracils that result from deamination of unmethylated cytosine residues. For example, \~25\% of mutations in TP53 in cancers are thought to be due to epigenetic effects\textsuperscript{47}. Indeed, analyses of adipose meQTLs in whole-blood samples\textsuperscript{6} might be explained by cancer-associated CpG island promoter hypermethylation variation may have a profound impact on carcinogenesis by frequent G→T mutations enriched for CpG dinucleotides outside CGI regions, suggesting a role for methylated cytosine since CGI, as we confirmed, are usually unmethylated\textsuperscript{48}. A similar signature was recently observed in other tumours\textsuperscript{14}. Thus, inherited genetic variation may have a profound impact on carcinogenesis by regulating the human methylome.

We observed a high similarity of genetic control on DNA methylation across tissues. Since tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns\textsuperscript{49}, a natural question is whether the genetic regulation of methylation is tissue-specific. While the tissue specificity of eQTLs has been investigated for a few tissues\textsuperscript{50} for cis-meQTL only a recent investigation was conducted\textsuperscript{8}, showing that 35.7\% of 88,751 cis-meQTLs detected in 662 adipose samples were replicated in \~200 whole-blood samples. We found that a large proportion of meQTLs in EAGLE lung samples, particularly those with large effect sizes, were robustly replicated in breast and kidney tissue samples from TCGA, suggesting a high similarity of genetic regulation of methylation across these tissues and related impact on somatic mutation rates\textsuperscript{14,48}. The lower replication rate of adipose meQTLs in whole-blood samples\textsuperscript{6} might be explained by the heterogeneity of different cell types in whole blood and by their more liberal \(P\)-value threshold \((8.6 \times 10^{-4})\), which led to the identification of a large number of weak cis-meQTLs.

Compared with cis-regulation, trans-eQTL regulation is typically considered to be more complex, has smaller effect sizes and is more difficult to be replicated even in the same tissue. However, in our study the lung trans-meQTLs are highly reproducible in TCGA lung, breast and kidney tissues. Notably, this similarity allows mapping meQTLs with substantially improved power by borrowing strength across tissues\textsuperscript{51}.

MeQTL SNPs are strongly associated with multiple epigenetic marks. Chromatin regulators play a role in maintaining genomic integrity and organization\textsuperscript{52}. We found that meQTL SNPs were strongly enriched for DNase hypersensitive sites, and sequences bound by CTCF or modified histones. SNPs could affect these epigenetic marks by several mechanisms, such as by affecting the core recognition sequences (exemplified for rs2816057 on chromosome 1 for CTCF), causing loss or gain of a CpG within a binding region, which, when methylated, could affect binding\textsuperscript{53}, or altering the binding sequence for interacting factors\textsuperscript{53}.

CTCF could cause changes in epigenetic marks through its multiple key roles, including genome organization through mediating intra- and inter-chromosomal contacts\textsuperscript{54,55}, the regulation of transcription by binding between enhancers and promoters\textsuperscript{54,56}, and the regulation of splicing, which may impact tissue specificity during tissue development\textsuperscript{57}. These changes can impact regulation of distant genes, and not the genes proximal to the SNPs that would be typically investigated in eQTL studies. This may be one reason for the previously observed lack of correlation between eQTLs and meQTLs\textsuperscript{5,47}. Future large studies integrating SNP profiles, the DNA methylome and transcriptome data through tissue developmental stages will hopefully shed light on this possibility.

There may be a myriad of other DNA-binding factors whose binding is directly or indirectly affected by SNPs. For example, among the histone marks, the strongest enrichment of meQTLs in our study was for H3K4me3 in both SAEC and hAEC cell types. As H3K4me3 is the chromatin mark primarily associated with regulatory elements at promoters, this suggests a strong influence of meQTLs on regulating gene activity. Unfortunately, transcription factor binding data in SAEC or hAEC are not available, so we could not test whether SNPs in their core sequence could affect the deposition of epigenetic marks, for example, by recruiting DNA methyltransferases\textsuperscript{57}. It will be important to obtain ChIP data from relevant primary cells for numerous DNA-binding regulatory factors to further elucidate the mechanisms whereby meQTLs and other SNP-affected epigenetic marks arise.

In conclusion, we show here that genetic variation has a profound impact on the DNA methylome with implications for cancer risk, tissue specificity, and chromatin structure and organization. The meQTL data (Supplementary Data 1 and 2) attached to this manuscript provide an important resource for studying genetic–DNA methylation interactions in lung tissue.

\textbf{Methods}  
\textbf{Sample collection.} We assayed 244 fresh-frozen paired tumour and non-involved lung tissue samples from Stage 1 to IIIA non-small cell lung cancer (NSCLC) cases in EAGLE study\textsuperscript{18}. EAGLE includes 2,100 incident lung cancer cases and 2,120 population controls enrolled in 2002–2005 within 216 municipalities of the Lombardy region of Italy. Cases were newly diagnosed primary cancers of lung, trachea and bronchus, verified by tissue pathology (67.0\%), cytology (28.0\%) or review of clinical records (5.0\%). They were 35–79 years of age at diagnosis and were recruited from 13 hospitals that cover \~80\% of the lung cancer cases from the study area. The study was approved by local and NICI Institutional Review Boards, and all participants signed an informed consent form. Lung tissue samples were snap-frozen in liquid nitrogen within 20 min of surgical resection. Surgeons and pathologists were together in the surgery room at the time of resection and sample collection to ensure correct sampling of tissue from the tumour, the area adjacent to the tumour and an additional area distant from the tumour (1–5 cm).

The precise site of tissue sampling was indicated on a lung drawing and the pathologists classified the samples as tumour, adjacent lung tissue and distant non-involved lung tissue. For the current study, we used lung tissue sampled from an area distant from the tumour to reduce the potential effects of field cancerization.

\textbf{DNA methylation profiling and data quality control.} Fresh-frozen lung tissue samples remained frozen while \~30 mg was subsampled for DNA extraction into a lysis buffer included 2.0 ml microcentrifuge tubes. Lysates for DNA extraction were generated by incubating 30 mg of tissue in 1 ml of 0.2 mg ml\textsuperscript{-1} Proteinase K (Ambion) in DNA Lysis Buffer (10 mM Tris–Cl (pH 8.0), 0.1 M EDTA (pH 8.0) and 0.5% (w/v) SDS) for 24 h at 56 °C with shaking at 850 rpm in Thermomixer R (Eppendorf). DNA was extracted from the generated lysate using the QIAamp DNA Blood Maxi Kit (Qiagen) according to the manufacturer’s protocol. Bisulfite treatment and Illumina Infinium HumanMethylation450 BeadChip assays were performed by the Southern California Genotyping Consortium at the University of California Los Angeles (UCLA) following Illumina’s protocols.

This assay generates DNA methylation data for 485,512 cytosine targets (482,421 CpG and 3091 CpH) and 65 SNP probes for the purpose of data quality control. Raw methylated and unmethylated intensities were background-corrected, and dye-bias-equalized, to correct for technical variation in signal between arrays. For background correction, we applied a normal-exponential convolution, using the intensity of the Infinum \(I\) probes in the channel opposite their design to model the noise-specific signal\textsuperscript{58}. Normalization was then equalized by subtracting the background from the ratio of the average red and green fluorescing normalization control probes. Both methods were conducted using the methylumi package in Bioconductor version 2.11.

For each probe, DNA methylation level is summarized as a \(B\) value, estimated as the fraction of signal intensity obtained from the methylated beads over the total signal intensity. Probes with detection \(P\)-values of \>0.05 were considered not significantly different from background noise and were labelled as missing. Methylation probes were excluded from meQTL analysis if any of the following criteria were met: on X/Y chromosome, annotated in repetitive genomic regions, annotated to harbour SNPs and missing rate \>5%. As the \(B\) values for the 65 SNP probes are expected to be similar in matched pair of normal and tumour tissues, we performed PCA using these 65 SNP probes to confirm the labelled pairs. We then performed PCA using the 5000 most variable methylation probes with \(B\) value \>0.02 and found that the normal tissues were clustered together and well separated from the tumour tissues. We further excluded five normal tissues that were relatively close to the tumour cluster. From the remaining 239 normal tissue samples, we analysed 210 with genotype data from a previous GWAS of lung cancer\textsuperscript{20}.
Genotype data and genetic association analysis. The blood samples were genotyped using the Illumina HumanHap550K SNP arrays in EAGLE GWAS30. The SNPs with MAF > 0.05% were removed from further analysis. The genotype association analysis was then performed using PLINK and R, adjusted for the top three PCA scores based on GWAS SNPs to control for potential population stratification.

Identification of cis-meQTLs. For each CpG methylation probe, the cis region was defined as being <500 kb upstream and downstream from the target CpG-site (1 Mb total). A methylation trait was detected to harbour a cis-meQTL if any SNP in the cis region had a SNP-CpG nominal association P-value less than P0, where P0 was set to control FDR at 5% by permutations. Here we determined the permutation procedure to choose P0 to control FDR at 5%. For a given P0, let N(P0) be the total number of CpG probes with detected cis-meQTLs and N0(P0) the expected number of CpG probes falsely determined to have cis-meQTLs. FDR is defined as N(P0)/N0(P0). The key is to estimate N0(P0) under the global null hypothesis that no cis-meQTLs exist. We randomly permuted the genomewide across subjects for 100 times, keeping the correlation structure of the 338,456 methylation traits in each permutation. Then, N0(P0) was estimated as the average number of methylation traits that were detected to harbour cis-meQTL SNPs with nominal P < P0. Control FDR at 5% requires P0 = 4.0 × 10−5. The same procedure was applied to detect trans-meQTLs.
meQTL SNPs affect CTCF binding. We found that meQTL SNPs are strongly enriched in CTCF consensus sequences. We used SAEC data from ENCODE to test whether meQTL heterozygous SNPs directly affect CTCF binding by disrupting the CTCF recognition sites. P-values were calculated based on a binomial distribution (Binom(N, 0.5)). Here N is the total number reads covering the SNPs. Raw sequencing data (fastq format) from SAEC cells were generated at the University of Washington as part of the ENCODE project and downloaded from the UCSC genome browser. Raw data was aligned to the hg19 genome using CLC genomics workbench (v 5.5.1), parsing out data with <80% contiguous alignment to the genome and duplicate reads in excess of 10 copies. We used the CTCFBSDB 2.0 programme to predict whether the meQTL SNPs or their LD SNPs (r^2 ≥ 0.8) were within CTCF peaks and then examined in SAEC whether CTCF exhibited allele-specific binding. As common SNPs are more likely to be heterozygous, we only looked for SNPs with MAF ≥ 0.4. Here we present two such examples. Systematic investigation of all meQTL SNPs that are heterozygous in SAEC is warranted once more samples with genotypic data are available.

References
1. Kaminsky, Z. A. et al. DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. 41, 240–245 (2009).
2. Heijmans, B. T., Kremer, D., Tobi, E. W., Boomsma, D. I. & Slagboom, P. E. Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. Hum. Mol. Genet. 20, 457–467 (2007).
3. Gibbs, J. R. et al. Abundant quantitative trait loci exist for DNA methylation variation in regulatory DNA. Nature 466, 253–257 (2010).
4. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the biology of epigenetic parasites. Trends Genet. 13, 335–340 (1997).
5. Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an endogenous mutagen in the human Ldr receptor and P53 Genes. Science 249, 1288–1290 (1990).
6. Shen, H. & Laird, P. W. Interplay between the cancer genome and epigenome. Cell 153, 38–46 (2013).
7. Laurent, L. et al. Dynamic changes in the human methylation during differentiation. Genome Res. 20, 320–331 (2010).
8. Maunakea, A. K. et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466, 253–257 (2010).
9. Yoder, J. A., Walsh, C. P. & Bestor, T. H. The role of DNA methylation in transcriptional silencing. EMBO J. 15, 600–607 (1996).
10. Alexandrov, L. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
11. Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics. Nat. Rev. Genet. 12, 277–282 (2011).
12. Schiller, J. S., Lucas, J. W., Ward, B. W. & Perego, J. A. Summary health statistics for U.S. adults: National Health Interview Survey, 2010. Vital. Health. Stat. 10, 1–207 (2012).
13. American Cancer Society. Cancer Facts & Figures 2013 (American Cancer Society, 2013).
14. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13, 484–492 (2012).
15. Landi, M. T. et al. Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. BMC Public Health 8, 203 (2008).
16. Sabin, L. R., Delas, M. J. & Hannon, G. J. The many influences of acetylcholine receptor subunit genes on 15q25. Nat. Genet. 44, 638–642 (2012).
17. Rus et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res. 22, 1680–1688 (2012).
18. Marconet, R. E. et al. The accessible chromatin landscape of the human genome. Nature 489, 75–82 (2012).
19. Amos, C. I. et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat. Genet. 40, 616–622 (2008).
20. Landi, M. T. & Hainaut, P. TP53 mutations in lung cancer and peripheral arterial disease. Nature 452, 638–642 (2008).
21. Wang, Y. et al. Common SNP 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat. Genet. 40, 1407–1409 (2008).
22. McKay, I. D. et al. Lung cancer susceptibility locus at 5p15.33. Nat. Genet. 40, 1404–1406 (2008).
23. Shi, J. et al. Inherited variation at chromosome 12p13.33, including RAD52, influences the risk of squamous cell lung carcinoma. Cancer Discov. 2, 131–139 (2012).
24. Timofeeva, M. N. et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum. Mol. Genet. 21, 4493–4500 (2012).
25. Tekpli, X. et al. Functional effect of polymorphisms in 15q25 locus on CHRNA5 RNA, bulk DNA adducts and TP53 mutations. Int. J. Cancer 132, 1811–1820 (2013).
26. Selove, F. S. et al. Multiple isoforms and differential allelic expression of CHRNA5 in lung tissue and lung adenocarcinoma. Carcinogenesis 34, 1281–1285 (2013).
27. Slukka, S. et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479, 74–79 (2011).
28. Fairfax, B. P. et al. Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat. Genet. 44, 502–510 (2012).
29. Small, K. S. et al. Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat. Genet. 43, 561–564 (2011).
30. Heinig, M. et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature 467, 460–464 (2010).
31. Wu, Y. et al. Interplay between menin and K-Ras in regulating lung adenocarcinoma. J. Biol. Chem. 287, 40003–40011 (2012).
32. Korostowski, L., Sedlak, N. & Engel, N. The Kcnq1ot1 long non-coding RNA affects chromatin conformation and expression of Kcnq1, but does not regulate its imprinting in the developing heart. PLoS Genet. 8, e1002956 (2012).
33. Sabín, L. R., Delas, M. J. & Hannon, G. J. Dogma derailed: the many influences of acetylcholine receptor subunit genes on 15q25. Nat. Genet. 44, 638–642 (2012).
34. Freedman, M. L. et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat. Genet. 43, 513–518 (2011).
35. Olivier, M., Hölstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harb. Perspect. Biol. 2, a001008 (2010).
36. Pleasance, E. D. et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463, 184–190 (2010).
37. Sabin, L. R., Delas, M. J. & Hannon, G. J. Doma detailing the many influences of RNA on the genome and disease. Genome Res. 13, 561–564 (2012).
38. Flute, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. PLoS Genet. 9, e1003486 (2013).
39. Papamichos-Chronakis, M. & Peterson, C. L. Chromatin and the genome integrity network. Nat. Rev. Genet. 14, 62–75 (2013).
40. Chernukhin, I. V. et al. Physical and functional interaction between two pluripotent proteins, the Y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF. J. Biol. Chem. 275, 29915–29921 (2000).
41. Handoko, L. et al. CTCF-mediated functional chromatin interaction in pluripotent cells. Nat. Genet. 43, 630–638 (2011).
42. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. Cell 137, 1179–1211 (2009).
43. Bell, A. C., West, A. G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98, 387–396 (1999).
44. Brenner, C. et al. Myc represses transcription through recruitment of DNA methyltransferase complex, MBD3. EMBO J. 24, 336–346 (2005).
45. Triche, Jr T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 41, e90 (2013).
Acknowledgements
This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, MD, USA (http://biowulf.nih.gov). We are grateful to the EAGLE participants and the large number of EAGLE collaborators (listed in http://dceg.cancer.gov/eagle), The Cancer Genome Atlas project for the genotype and methylation data and the ENCODE project for the regulatory region data. This work was supported by the Intramural Research Program of NIH, NCI, Division of Cancer Epidemiology and Genetics and, in part, by the Norris Comprehensive Cancer Center core grant (P30CA014089) from NCI, the Transdisciplinary Research in Cancer of the Lung (TRICL) and the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network (U19CA148127). A.W., Z.W., W.Z. and A.H. were also funded by the NCI, NIH (HSN26120080001E). I.A.L.-O. and Z.B. were also funded by NIH grants (1 R01 HL114094, 1 P30 H101258, and R37HL062569-13), Whittier Foundation and Hastings Foundation. Z.B. was also funded by the Ralph Edgington Chair in Medicine. C.N.M. was funded by ACS/Canary postdoctoral fellowship (PFTED-10-207-01-SIED).

Author contributions
M.T.L. conceived the study. I.A.L.-O. supervised DNA methylene analysis. J.S. performed EAGLE, TCGA and ENCODE genetic analyses. C.N.M. performed allele-specific binding analyses. J.D. contributed to genetic analyses and performed GO analyses. J.S., P.L., I.A.L.-O. and M.T.L. performed quality control analyses. A.C.P., D.C., P.A.B., A.W.B., N.E.C. and M.T.L. conducted the EAGLE study and provided tissue samples. A.W. and A.H. prepared the tissue samples for the analyses. B.Z. and Z.B. isolated and cultured alveolar epithelial cells. T.T. and K.D.S. performed methylation normalization. Z.W. and W.W. performed LD analyses. J.S., C.N.M., J.D., P.L.H., M.C., D.S.L, J.H., P.-H.C, B.S.I.C., W.Z., I.A., M.F., B.P.B., N.C., M.A.T., S.J.C., I.A.L.-O. and M.T.L. contributed to the data interpretation. J.S. and M.T.L. wrote the manuscript. All authors participated in the discussion and reviewed the manuscript.

Additional information
Accession codes: Genotype data have been deposited in dbGAP under accession code phs000093.v2.p2. Methylation data have been deposited in dbGAP under accession code GSE52401.

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

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

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

How to cite this article: Shi, J. et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat. Commun. 5:3365 doi: 10.1038/ncomms4365 (2014).