nature COMMUNICATIONS

ARTICLE

Received 2 Dec 2013 | Accepted 8 Sep 2014 | Published 20 Oct 2014

DOI: 10.1038/ncomms6195

The meta-epigenomic structure of purified human stem cell populations is defined at *cis*-regulatory sequences

N. Ari Wijetunga[1,*], Fabien Delahaye[2,*], Yong M. Zhao[2], Aaron Golden[1], Jessica C. Mar[3], Francine H. Einstein[2] & John M. Greally[1]

The mechanism and significance of epigenetic variability in the same cell type between healthy individuals are not clear. Here we purify human CD34+ haematopoietic stem and progenitor cells (HSPCs) from different individuals and find that there is increased variability of DNA methylation at loci with properties of promoters and enhancers. The variability is especially enriched at candidate enhancers near genes transitioning between silent and expressed states, and encoding proteins with leukocyte differentiation properties. Our findings of increased variability at loci with intermediate DNA methylation values, at candidate 'poised' enhancers and at genes involved in HSPC lineage commitment suggest that CD34+ cell subtype heterogeneity between individuals is a major mechanism for the variability observed. Epigenomic studies performed on cell populations, even when purified, are testing collections of epigenomes, or meta-epigenomes. Our findings show that meta-epigenomic approaches to data analysis can provide insights into cell subpopulation structure.

[1] Center for Epigenomics and Department of Genetics (Division of Computational Genetics), Albert Einstein College of Medicine, 1301 Morris Park Avenue, Price 322, Bronx, New York 10461, USA. [2] Department of Obstetrics and Gynecology and Women's Health, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Price 322, Bronx, New York 10461, USA. [3] Department of Computational and Systems Biology, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Price 322, Bronx, New York 10461, USA. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to F.H.E. (email: francine.einstein@einstein.yu.edu) or to J.M.G. (email: john.greally@einstein.yu.edu).

Variation in epigenetic marks defines specific cell types in an organism[1,2]. Epigenome-wide association studies examine epigenetic variability within the same cell type or tissue in different individuals, to assess the role of the epigenome in those individuals with a specific disease or other phenotype[3–6]. In addition to epigenomic variability studied among different cell types in an individual or that in the same cell type among phenotypically different individuals, epigenomic variability occurring in the same cell type among healthy individuals is also now being studied[7–13]. The mechanism and functional consequences of this type of epigenetic variability remain unclear. Such variability has been found in plants[14,15] and has been described as 'interindividual' differential methylation[12] occurring at 'epipolymorphic' loci that characteristically have intermediate DNA methylation levels[16]. The potential for stochasticity to drive at least part of this epipolymorphism of DNA methylation has been proposed[16], and finds support from studies of allelic exclusion in the central nervous system of mouse[17], monoallelic expression in neural stem cells[18] and studies of heritability of DNA methylation in cloned ovarian carcinoma cells[19]. However, the proportions of genes at which these stochastic events are implicated is low (1–2%)[17,18], indicating that other processes are likely to be involved.

Underlying genetic polymorphism has been demonstrated to be a contributor to DNA methylation variability[10,12,13,20]. Such genetic effects are unlikely to be the only influence, as monozygotic human twins[12,20] and inbred mice[8,21] also manifest epigenetic variability that cannot be attributed to DNA sequence differences. Some studies have linked DNA methylation variation with transcriptional consequences at nearby genes[8,10,12,13]. Some of the variability observed in a study of peripheral blood leukocytes has been explained in terms of cell subtype effects[9], although that study's quantification of neutrophil, lymphocyte and monocyte percentages lacked the finer resolution cell subtype discrimination demonstrated in a later study to have effects on DNA methylation[22]. It has been shown that clinically normal cervical epithelial samples from women who proceed to develop cervical neoplasia within 3 years have increased variability of DNA methylation[11]. Although this specific example reflects an underlying pathological process, epigenetic variability has also been proposed to be stochastic in origin and to influence normal phenotypic variability[8]. Supplementation of methyl donors in the diet of isogenic mice has been observed to increase the variability of DNA methylation in liver samples, suggesting to the authors a mechanism for disease or evolutionary selection[21]. The epigenetic variability observed in human CD14+ monocytes has been found to remain over the course of years, despite the short lifespan of these cells, indicating that the variability is encoded in leukocyte stem or progenitor cells[12]. Here we focus on using DNA methylation assays to define the loci with epigenetic variability in CD34+haematopoietic stem and progenitor cells (HSPCs) purified from neonatal cord blood. We used the results of chromatin immunoprecipitation sequencing (ChIP-seq) studies of the same cell type by the Roadmap Epigenomics programme to annotate the CD34+ HSPC genome empirically, so that we could define where epigenetic variability occurs in these cells, gaining insights into why the variability is occurring in seemingly identical cell types from different healthy individuals.

Results

Identifying variably DNA methylated loci in CD34+ HSPCs. We used two sources of DNA methylation data, one from the Roadmap Epigenomics programme, publicly available reduced representation bisulphite sequencing (RRBS)[23] data on mobilized CD34+ HSPCs from 7 adults, and the second generated by our group, using CD34+ HSPCs isolated from cord blood from 29 phenotypically normal neonates assayed using the HELP-tagging assay[24]. Despite the differences in how each of these assays measures DNA methylation, both showed increased variability at loci with intermediate methylation values (Fig. 1), consistent with previous observations[16].

We continued our analyses based on the HELP-tagging data, which are derived from a greater number of samples and from neonates, who have less potential for manifesting age-associated variability than adults[25]. As HELP-tagging is based on the use of methylation-sensitive restriction enzymes[24], we were able to use the results from the methylation-insensitive MspI control enzyme to estimate the degree of technical variability, and a permutation analysis of the HpaII-derived data also showed enrichment of the observed variability over expected background levels (Supplementary Fig. 2a). A number of loci with differing degrees of variability were chosen for bisulphite PCR, using seven of the samples that had been tested using HELP-tagging as well as eight independent samples. These amplicons were combined for each

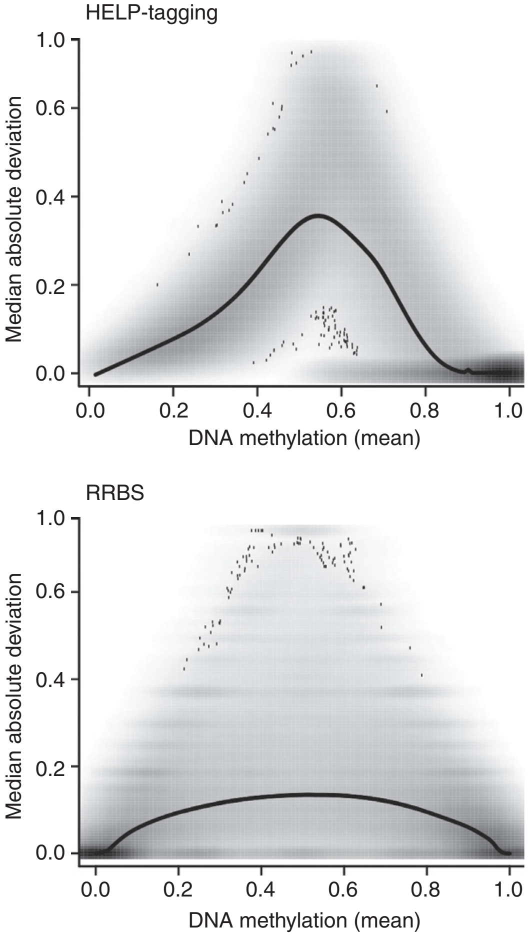
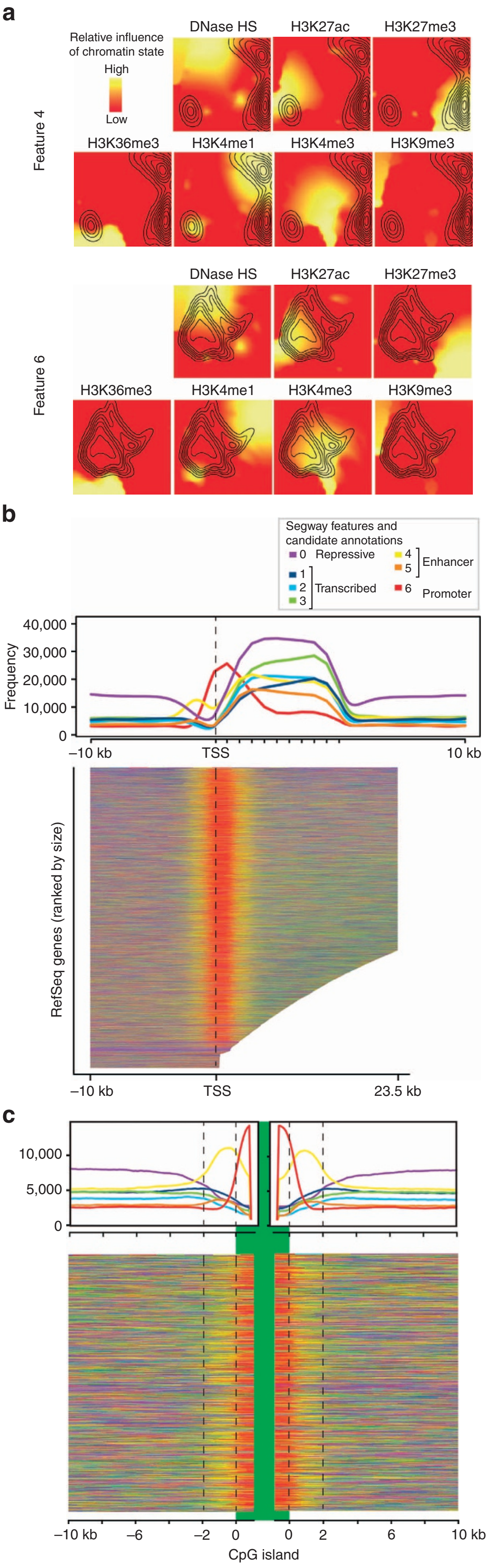


Figure 1 | DNA methylation variability is increased at loci of intermediate methylation. The MAD for DNA methylation values in CD34+ HSPCs measured by HELP-tagging (top, 29 individuals) or RRBS (bottom, 7 individuals) are shown as a function of mean DNA methylation across all of the samples tested. Although HELP-tagging usually plots DNA methylation with a zero value to indicate complete methylation, we inverted the scale on this occasion to make the two plots comparable. The number of loci is reflected by the grey shading. The line shown indicates the mean MAD value and reveals for both data sets increased variability of DNA methylation at loci with intermediate values.



individual and used to generate Illumina libraries, allowing targeted massively parallel sequencing of the bisulphite-converted DNA. The results confirm that DNA methylation variability is enriched at loci with variability measures above the threshold attributable to technical variability or chance (Fig. 2b).

Inferring the effects of DNA sequence variability. To test whether the variability we observed could be accounted for by genomic sequence polymorphism, we segregated the variability at loci overlapping common single-nucleotide polymorphisms (SNPs, minor allele frequencies $\geq 1\%$, 7.6% of sites tested) from the remaining majority of the genome. A Kolmogorov–Smirnov test (K–S test) showed significantly increased levels of variability of DNA methylation at these polymorphic loci ($P < 2x10^{-16}$), indicating that genetic influences are contributing to the variability observed (Supplementary Fig. 3). There are two ways that local sequence variability can influence DNA methylation variability. The site being tested in the DNA methylation assay can itself be a sequence variant, as cytosine to thymine transitions at CG dinucleotides represent a frequent source of SNPs due to the increased mutability of methylcytosine[26], leading to the failure of methylation-sensitive restriction enzymes to cut or the misleading appearance of bisulphite-mediated conversion at these sites. The second mechanism is for SNPs *in cis* to the tested site influencing DNA methylation, as has previously been shown[27]. We find that the K–S test is no longer significant ($P = 0.1563$) at sites tested even within the immediate flanking 10 bp of the common SNPs (Supplementary Fig. 3). We infer that although genetic variability is influential at the tested loci themselves, there exists a substantial amount of epigenetic variability in the remaining majority of loci in the genome, and that local genetic polymorphism is not likely to be the sole cause of the epigenetic variability observed, consistent with the conclusions of prior studies[8,12,20,21].

Mapping functional elements in CD34+ HSPCs. To determine whether epigenetic variability was occurring at regulatory sites with possible functional consequences, we took advantage of public chromatin mapping data for CD34+ HSPCs generated by the Roadmap Epigenomics programme (Supplementary Table 5). The DNase hypersensitivity and ChIP-seq data create combinatorial patterns that have previously been exploited to define functional elements in the genome[28]. We processed the Roadmap data using an adaptation of an imaging signal processing algorithm[29], to define the locations of chromatin constituents with minimal data transformation (Supplementary Fig. 4). These chromatin constituent locations were then used to generate a self-organizing map (SOM)[30], and to map candidate regulatory elements using the Segway algorithm[28] (Supplementary Fig. 5). The individual Segway features were then overlaid as contour

Figure 2 | Empirical annotation of the CD34+ HSPC genome based on chromatin features reveals candidate *cis*-regulatory element locations. (**a**) A contour plot of the regions within the SOM where Segway features 4 (above) and 6 (below) enrich, showing feature 4 to be composed of loci where H3K4me1 and H3K27me3 occur, while the loci composing feature 6 contain the H3K4me3 and H3K27ac modifications. Consistent with these findings, **b** shows feature 6 (red) to be enriched at the TSS for a metaplot (top) and a heat map (below) of all RefSeq genes, indicating promoter characteristics, while feature 4 (yellow) flanks this region and is consistent with enhancers in a poised state. In **c**, similar metaplot (top) and heat map (below) representations of the 2-kb flanking CpG islands demonstrate strong enrichment in feature 4, indicating that these 'CpG island shores' in fact represent candidate enhancers in this cell type.

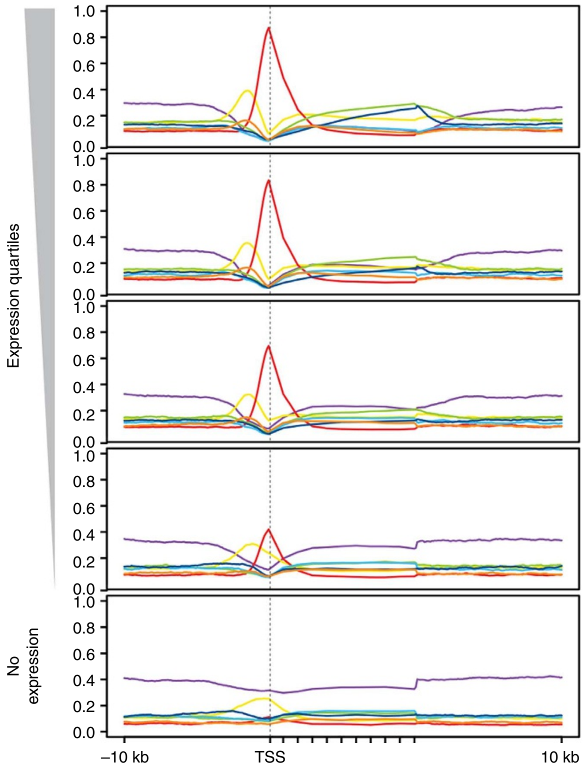


Figure 3 | Transcriptional relationships of Segway features. A RefSeq metaplot for the Segway features divided by expression quartile shows that features 1–3 enrich in the bodies of genes as transcription increases, at the expense of feature 0, which appears to represent repressed chromatin. Feature 6 is strongly enriched at canonical TSSs, flanked by an enrichment of feature 4 and, to a lesser extent, feature 5, which have chromatin signatures indicative of enhancer function.

plots onto the SOM, which clusters in two-dimensional space loci with similar genomic characteristics, allowing intuitive visualization of the major contributors to each feature (Fig. 2a and Supplementary Fig. 6). Of the multiple chromatin states for which each feature is enriched, feature 6 has the H3K4me3 enrichment, indicating promoter function, features 4 and 5 both have marks indicative of enhancer function (H3K4me1 and H3K27ac, respectively), features 1–3 have the H3K36me3 enrichment typical of transcribed sequences, while feature 0 in enriched for heterochromatic marks (H3K9me9 and H3K27me3). We also created a metaplot of these new annotations relative to all RefSeq genes in the genome (Supplementary Fig. 7), showing that Segway feature 6 is strikingly enriched at transcription start sites (TSSs), flanked by enrichment for feature 4 and, to a lesser degree, feature 5 (Fig. 2b). Features 1–3 are enriched in gene bodies and feature 0 at intergenic sequences. Statistical testing of the enrichment of features 4 and 6 in their windows of peak frequencies compared with their distributions over all RefSeq genes and flanking regions showed significance ($P<0.001$ for each). CpG islands and their immediate flanking sequences have previously been related to 'stochastic' DNA methylation variability[8] and gene expression regulation[31]. The Segway annotations demonstrate that although the bodies of CpG islands are enriched for the candidate promoter (feature 6) sequences, the ±2 kb flanking region, generally described as its 'shore', is strikingly enriched for feature 4 (Fig. 2c). Both achieve statistical significance ($P<0.001$) when compared with their distributions over all CpG islands (feature 6) or flanking regions (feature 4). Finally, stratifying the RefSeq genes by expression quartile in CD34+ HSPCs reveals the transcriptional dependencies of the Segway annotations (Fig. 3). We conclude that the Segway annotations define candidate promoters (feature 6), enhancers (features 4 and 5), transcribed regions (features 1–3) and repressed chromatin (feature 0) for CD34+ HSPCs.

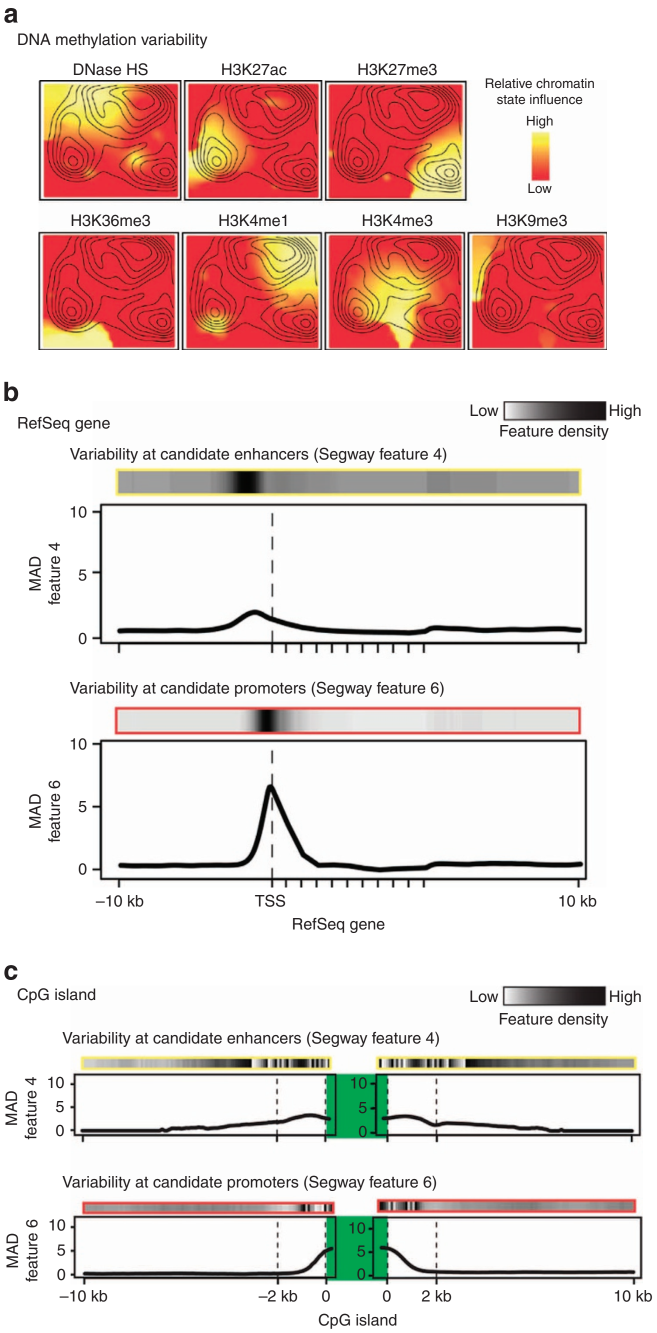


Figure 4 | DNA methylation variability is enriched at candidate enhancers and promoters at TSSs of RefSeq genes and at CpG islands and shores. (**a**) Enrichment of variability of DNA methylation is marked at loci with H3K4me1, H3K27ac and H3K27me3 marks in particular. (**b**) A RefSeq metaplot with feature density indicated by the grey shading above and within the graph, and mean variability for features 4 (yellow, top) and 6 (red, bottom) depicted, with increased variability distributing where each mark is maximally located. The significance of the enrichment is shown at the depicted peak *P*-value location. Analysis of CpG islands (**c**) shows variability in flanking regions (shores) associated with the presence of feature 4.

Variable DNA methylation is enriched at functional elements. With the genome annotated for functional elements in a cell type-specific manner, we then tested the associations between genomic annotations and the loci with increased variability in DNA methylation. In Fig. 4a, we show the strongest associations for highly variable loci within clustered SOM space to be with H3K27ac, H3K27me3 and H3K4me1. Figure 4b also shows enrichment of variability at the TSS of RefSeq genes for feature 6 and immediately upstream at feature 4, both significant at $P<0.001$. Figure 4c shows enrichment in variability at the proximal part of CpG island shores for feature 6 and more extensively into the CpG island shore for feature 4, both also significant at $P<0.001$. A complementary SOM analysis using the published ChromHMM annotations of the human genome[32] reveals consistent results (Supplementary Fig. 8). DNA methylation variability is therefore enhanced at candidate *cis*-regulatory sequences (promoters and enhancers) and the epigenetic variability previously observed for CpG island shores[8,31] is reflective of this general characteristic of enhancers. Common SNPs are not enriched in density in any of the features (Supplementary Fig. 9) and therefore are unlikely to be the major reason for selective enrichment of epigenetic variability in these specific genomic contexts. If the variability of DNA methylation occurs at loci with potential transcriptional regulatory properties, it raises the question of whether variability occurs selectively near genes with specific transcriptional activities. We find that all levels of expression have comparable levels of epigenetic variability at promoter sequences. Genes expressed at the lowest levels in the genome are those with selective enrichment of epigenetic variability at nearby candidate enhancers (Fig. 5), a significant inverse relationship ($P=10^{-8}$) using the Jonckheere trend test.

Variable DNA methylation is enriched at haematopoietic genes. The finding of increased variability at loci with intermediate values of DNA methylation (Fig. 1) cannot be explained by individual cells having intermediate methylation values, as DNA methylation is either present or absent on an allele; thus, a cell can only have 0%, 50% or 100% methylation at a locus with two parental alleles in a diploid cell. For one individual to have an intermediate value such as 30% and another 60% methylation at the same locus, there has to be, within the pool of cells tested, differing subpopulations of methylated alleles present. The parsimonious explanation for such allelic subpopulation differences is that different proportions of cell subtypes are present in the individuals tested. The CD34+ HSPC is a well-studied cell type, recognized to have ∼15 cellular subtypes bearing the surface marker in different lineages and stages of early haematopoietic differentiation[33]. If these subtypes differ in proportion between the 29 subjects tested, the loci where DNA methylation patterns are associated with specific CD34+ cell subtypes would be expected to be the most variable, whereas genes expressed in all cell types (housekeeping genes) should be invariant in terms of DNA methylation. We show in Supplementary Table 6 and Supplementary Fig. 10 that DNA methylation variability at candidate enhancers is enriched for leukocyte-specific networks but not housekeeping genes, consistent with a CD34+ cell subtype model. Candidate promoters, which have comparable levels of DNA methylation variability at all expression levels (Fig. 5), have, as expected, equivalent levels of variability for housekeeping and haematopoietic genes.

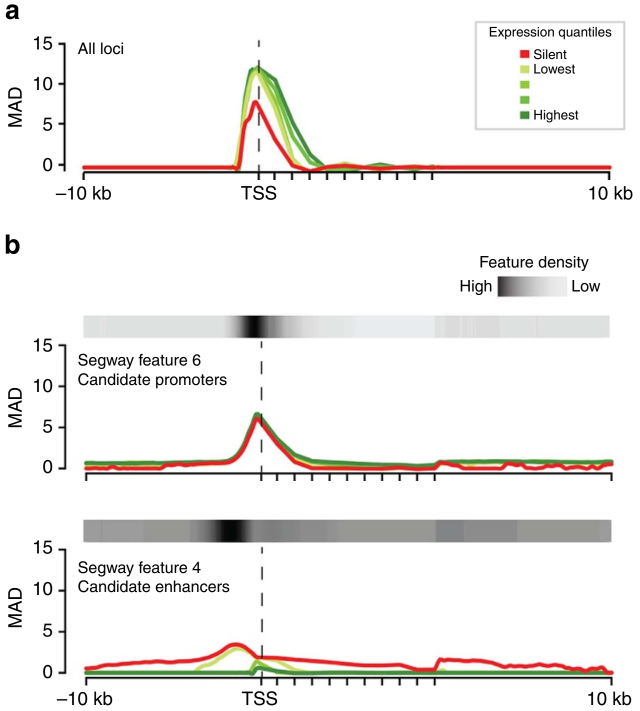


Figure 5 | Variability of DNA methylation at candidate enhancer sequences discriminates genes expressed at lower levels. (**a**) The overall pattern of DNA methylation variability at RefSeq genes broken down by expression quantile, showing differences at silent compared with expressed genes at TSS. (**b**) No such differences are measurable when testing candidate promoters (Segway feature 6, top), whereas candidate enhancers (Segway feature 4, bottom) show increased variability for DNA methylation for genes that are either silent or expressed at the lowest quartile.

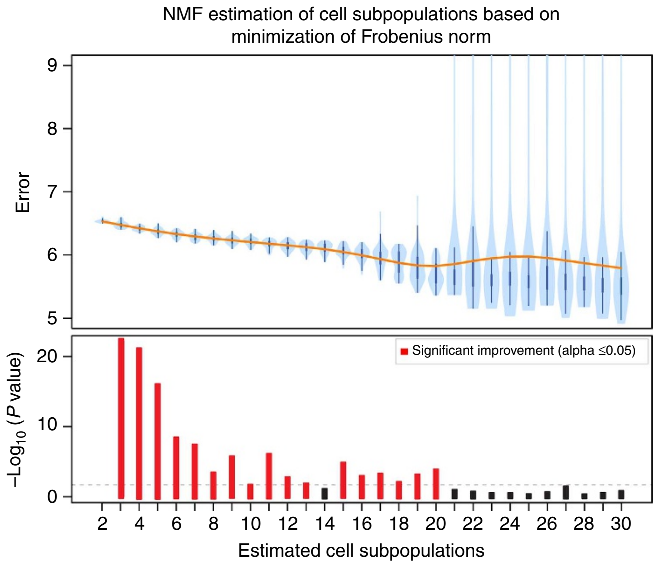


Figure 6 | NMF of DNA methylation profiles shows evidence for 13–20 subpopulations within the CD34+ HSPC population. The upper plot shows a smooth spline (orange) and value distribution (blue) for the Frobenius norm as a function of increasing cell subpopulation number, with the lower plot representing the *P*-value testing (two sample *t*-test) whether the difference between the successive simulations is significant. We observe two points at which the subsequent change is insignificant, at values 13 and 20, suggesting that the number of subpopulations differing in DNA methylation profiles within the CD34+ HSPC population is within this range.

Using variability information to quantify cell subtypes. Variability of gene transcription levels in cell samples from multiple individuals has allowed patterns to be identified that predict the numbers of cell subtypes present. We adapted one of the approaches used for these transcriptional variability studies, non-negative matrix factorization (NMF)[34,35], to our DNA methylation variability data to estimate the number of cell subtypes in our purified CD34+ HSPCs. In Fig. 6, we show the NMF output to predict ~13–20 cell subtypes, consistent with the ~15 distinct types of cells that have previously been described to express the CD34 cell surface marker[33].

Discussion

This study explored the epigenomic variability between normal healthy individuals occurring in a purified cell type, homogenous for the same cell surface marker. In common with prior studies, we found DNA methylation to vary between individuals, especially for those loci with intermediate DNA methylation values[16], and increased variability at regulatory regions near the genes expressed at the lowest levels in the genome[12]. Our empirically based annotation of regulatory elements in the CD34+ HSPC genome allowed us to define candidate promoter and enhancer sites, revealing the latter in particular to have DNA methylation variability associated with the lowest levels of gene expression. The genes at which this enhancer-associated variability is enriched are those encoding proteins with properties associated with leukocyte function. Because of the necessity for intermediate DNA methylation values to require allelic subpopulations with distinct methylation states, we interpret the epigenomic variability to be due to DNA methylation-sensitive enhancers changing their epigenetic states as nearby genes switch their transcription on or off in different CD34+ HSPC cell subtypes. The presence of differing proportions of these cell subtypes in the different individuals studied and the consequent differences in allelic proportions of methylated DNA at these *cis*-regulatory sites together drive the variability observed.

In keeping with this model of allelic and cell-subtype heterogeneity, the chromatin signature of feature 4, that of 'poised' enhancers[36,37], is defined by the co-occurrence in the genome of activating H3K4me1 and repressive H3K27me3 marks. As cells commit to the use of the enhancer, the locus is activated and marked by H3K27ac accumulation[38] and the loss of H3K27me3 (ref. 36). Implicit in the idea of a poised regulatory element is its capability to commit to an active or inactive state by choosing one of two pre-existing states encoded in the nucleosome, as demonstrated by sequential ChIP experiments defining bivalent chromatin domains in pluripotent mouse cells[39]. Poised enhancers have not been reported to have been tested by sequential ChIP, making it possible that the activating and repressive marks are not encoded within the same nucleosome, but occur on different alleles in the cell population studied. This would be consistent with the presence of distinct subtypes of cells within the population tested by ChIP-seq when defining these poised enhancers. Our finding of variability and intermediate levels of DNA methylation at these loci with the features of poised enhancers is suggestive of epigenetic heterogeneity among the alleles in the cell population. The presence of the activating and repressive marks of these enhancers on different alleles would be consistent with a mosaic model for epigenetic regulatory marks in the population, rather than a poised state encoded by different marks within the same nucleosome.

Such a model also suggests that there is a relationship between DNA methylation and histone states at the candidate enhancers. One of our findings is that the mosaic candidate enhancers are enriched at what have been generally described as CpG island shores, the ±2 kb flanking CpG islands[31]. One such shore at the *GLT1* gene has previously been found to encode an enhancer that is, when silenced, enriched in DNA methylation and H3K27me3 but depleted in H4ac[40]. For this individual example, it appears that the DNA methylation and histone modifications are associated with enhancer function, potentially representing enhancers more generally. The *GLT1* CpG island shore demonstrates co-localization of DNA methylation and H3K27me3, an association that has been shown to occur outside but not within CpG islands[41,42], using ChIP for H3K27me3 followed by bisulphite sequencing to detect the DNA methylation state on the alleles bearing the H3K27me3 nucleosomes. The more general finding that increased DNA methylation at CpG island shores correlates with decreased gene expression[31] supports a model of these sites representing DNA methylation-sensitive enhancers where Polycomb-mediated H3K27me3 modification adds to the silencing of these regulatory elements.

We conclude that the epigenome of a cell population is, in fact, a collection of epigenomes, or a meta-epigenome, reflecting the presence of subpopulations of cell subtypes even in highly purified cell samples. This study's use of CD34+ HSPCs was fortuitous, as this cell type is extremely well studied, and already recognized to be composed of multiple cell subtypes expressing the CD34 surface marker[33]. It is possible that the multipotent nature of the CD34+ HSPC cells is associated with greater epigenetic variability than more lineage-committed cell types, as suggested previously[43]. The prediction based on the observation of epigenetic variability in monocytes that HSPCs should be epigenetically variable between individuals[12] is supported by our findings, although their findings could also be re-interpreted to be due to the long-term persistence of varying subtypes of monocytes in their studied populations. We propose that purified CD34+ HSPCs are not likely to be unique in being composed of multiple cell subtypes, and that detailed study of any presumed uniform cell population will reveal subpopulation heterogeneity. The results indicate that reference epigenomes based on the analysis of single or limited numbers of samples will contain epigenetically variable loci at which marks are unstable, and the co-localization of chromatin states at the same genomic position cannot reliably be taken to infer the presence of the epigenetic marks on the same alleles.

We show that the meta-epigenomic structure in the cell population can be exploited to estimate the number of cell subtypes present (using an approach such as NMF[34,35]) and their functional characteristics (by studying the properties of the genes located in proximity to the epigenetically variable *cis*-regulatory elements). Such meta-epigenomic analytical approaches could find an early application in cancer research. There are now consistent observations that epigenetic variability exists within cancer cells that have been extensively purified from contaminating cell types[44] and even in non-neoplastic cervical epithelial cells in women who later develop cervical neoplasms[11]. These observations have been described to involve 'stochastic variation' of epigenetic regulation[44], but the CD34+ HSPC observations add a further layer of complexity, requiring that we understand cell-subtype structure within the tested cell population before we can define any additional epigenetic variability as stochastic or disease associated. This is a far-reaching issue, because although there now exist approaches that attempt to account for cell subtype heterogeneity within mixed cell populations tested using epigenome-wide assays, implicit in those approaches is that the subtypes of cells contributing to the variability can be distinguished histologically or by the use of cell-sorting approaches[22,45–48]. What we show here is that even in histologically identical sorted cells, there exist subpopulation structures that continue to influence the results of epigenome-wide assays, and that the current approaches that rely on the use of sorted subpopulations of cells potentially do not go far enough to capture these influences.

The current study also has significant implications relevant to the interpretation of the epigenome-wide association study results. If a locus is found to change DNA methylation to the moderate extent typical of the results of non-cancer disease studies[5], a distinction needs to be made between changes at loci that are normally invariant between individuals (those at the extremes of DNA methylation) and loci normally variable between individuals (intermediate methylation levels). In the former case, epigenetic changes must be occurring in some cells within the population studied. Changes at loci that are normally epigenetically variable may, however, be due to changes in cell-subtype proportions between the individuals tested and not involve epigenetic changes occurring in any of the cells in the population tested. We increasingly recognize the influence of cell subtypes to be an issue in epigenomic studies of mixed cell types such as peripheral blood leukocytes[22] but the same concern arises even in purified cell populations, which are likely to have unrecognized cell-subtype structure. The development of meta-epigenomics as a field of study is an essential early step towards improvement of our design and interpretation of the results of epigenome-wide assays.

Methods

Sample collection. The Supplementary Data section provides greater detail about the samples and methods used in this study. The sample collection component to the study was approved by our institutional review board and was in accordance with Health Insurance Portability and Accountability Act regulations. Cord blood samples were obtained at the time of delivery of healthy, non-anomalous neonates with normal growth based on birth weight and ponderal index nomograms (Supplementary Table 1). CD34+ HSPCs were purified from the cord blood using magnetic immunosorting, with purity confirmed by flow cytometry.

Molecular assays. DNA from the purified cells was used for DNA methylation assays, the HELP-tagging assay[24] for genome-wide analysis and bisulphite PCR amplicon sequencing for verification and validation studies. HELP-tagging was performed on genomic DNA from the frozen CD34+ HSPCs, digested to completion by either HpaII or MspI, following which the digested DNA was ligated to two custom adapters containing Illumina adapter sequences, an EcoP15I recognition site and the T7 promoter sequence. Using EcoP15I, we isolated sequence tags flanking the sites digested by each enzyme, methylation-sensitive HpaII or methylation-insensitive MspI, followed by massively parallel sequencing of the resulting libraries (Illumina Technology). HpaII profiles were obtained for each sample ($n=29$), calculating methylation scores using a previously generated MspI human reference, which was also used to determine the degree of technical variability in the assay, using three replicates. For targeted bisulphite sequencing, we bisulphite-converted 200 ng of DNA using the Zymo EZ-96 Methylation-Lightning Kit. After separate PCR amplification of ten target regions (primers listed in Supplementary Table 2), we pooled the amplicons in equal ratios and generated Illumina libraries using robotic automation (Tecan). In total, 15 libraries were multiplexed on the Illumina Miseq for 250 bp paired-end sequencing. Bisulphite conversion efficiency was calculated as the per cent conversion of cytosines in a non-CG context (Supplementary Table 3). Massively parallel sequencing was performed using the Illumina HiSeq 2000 for HELP-tagging and the Illumina MiSeq was used for bisulphite PCR amplicon studies.

Measurement of DNA methylation variability. An overview of all of the data sources, processing and analysis is shown in Supplementary Fig. 1. Variability of DNA methylation was calculated using the median absolute deviation (MAD) value, previously used to define variably-methylated regions[49]. The MAD calculation is less sensitive to outlying values, giving a more robust and conservative variance estimate. MAD values were calculated from our HELP-tagging data (29 neonates) and from RRBS data generated on the same cell type by the Roadmap in Epigenomics (7 adults).

Verification and validation assays. We performed verification studies at 10 separate loci on 7 of the 29 samples tested using HELP-tagging, and for validation we added 8 new CD34+ HSPC samples from similarly healthy infants with normal growth. Using the DNA methylation value for each assayed HpaII locus, we calculated the MAD of DNA methylation for both validation and verification data sets, and observed an increase in variability when background technical variation (defined by control MspI HELP-tagging representations) was exceeded (Supplementary Fig. 2). The MAD distribution was calculated genome wide, at HpaII sites overlapping common SNPs (minor allele frequencies $\geq 1\%$) and at the 10 bp immediately flanking these common SNPs (Supplementary Fig. 3). K–S testing was performed to measure the significance of differences of distributions of MAD values.

Chromatin state data analysis. The youngest individual studied by the Roadmap in Epigenomics from whom RRBS, chromatin and transcriptional studies had been performed was chosen for further analysis. Wiggle track data representing DNase hypersensitivity and ChIP-seq for H3K4me3, H3K4me1, H3K9me3, H3K27me3, H3K27ac and H3K36me3 were collected from the Roadmap in Epigenomics web resource (http://www.roadmapepigenomics.org/data). All wiggle tracks were converted to bigwig format using the UCSC Genome Browser utility *wigToBigWig* version 4 (http://hgdownload.cse.ucsc.edu/downloads.html). Subsequently, the utility available through the UCSC Genome Browser *bigWigAverageOverBed* (http://hgdownload.cse.ucsc.edu/downloads.html) was used to calculate the sum of the ChIP-seq signals over 100 bp genomic intervals spanning the 22 autosomal and 2 sex chromosomes, a resolution which we believe is sufficient to characterize chromatin states, being smaller than an individual nucleosome, while usually including no more than 0-1 HpaII sites per window. ChIP-seq signals summed over 30,956,785 intervals were generated and formatted in bedGraph format.

The Roadmap in Epigenomics data are provided as raw signals and not as defined peaks. To avoid imposing excessive processing on these data, we used as simple and intuitive an approach as possible. The ChIP-seq bedGraph files were log-transformed to exaggerate highly positively skewed signal density, a prerequisite for the recursive kernel density learning framework for robust foreground object segmentation approach[29]. This image processing technique relies on removing background until the remaining signal is bimodal and approximately Poisson distributed. Gaussian kernel density was estimated using the *density* function in R, resulting in multiple modes of signal density that are increasingly smaller. As signal intensity was derived from a ChIP-seq read count, the measure should be an approximate Poisson distribution, and we aimed to eliminate low-signal intensity signal modes iteratively. Using the *turnpoints* function within the *pastecs* library in R, we recursively identified the modes of signal intensity and set signal thresholds based on the maximum mode, which was also generally the leftmost mode. The algorithm ran until, at most, two signal modes remained and the resulting distributions were approximately Poisson distributed. The results are shown in Supplementary Fig. 4, with the stepwise approach for the H3K4me1 signal illustrated as an example of the process.

Annotation of functional elements in CD34+ HSPCs. One hundred base pair windows were then classified as having the chromatin state or not. Windows of chromosome 1 were used to train the Segway algorithm[28], which then annotated seven features genome wide. A subset of 100 bp windows containing HpaII sites tested using HELP-tagging was chosen for SOM construction[30]. All SOM analysis was completed using the Java SOMToolbox from Institute of Software Technology and Interactive Systems at the Vienna University of Technology (http://www.ifs.tuwien.ac.at/dm/somtoolbox/). Out of 30,956,785 one hundred base pair genomic intervals, 1,520,684 intervals overlapping 1,696,696 HpaII sites were chosen to reduce the dimensionality of the data set and greatly reduce the required computation. An input data matrix was created, where rows represent the 1,520,684 vectors defined by genomic intervals and the columns represent the observations for the investigated tracks (that is, processed ChIP-seq signal). In total, two SOMs were constructed, one representing the chromatin states from the Roadmap in Epigenomics and the other representing the ChromHMM annotation[32], choosing a map size to yield 100–200 interval vectors per map unit to reduce the required computation while generating maps of sufficient resolution to aid in further analysis. For all SOMs, the standard SOM algorithm by Kohonen was employed, using the default SOMToolbox settings of learnrate = 0.7 and randomSeed = 11. Java code was implemented over ~120 h using a high performance computing cluster and 100 GB of virtual memory.

We performed an analysis to define CD34+ HSPC genomic annotations using the Segway genomic segmentation approach[28]. Using the *segway* package, annotations were generated from the seven processed chromatin state signal bedGraph files. A Segway segmentation of the genome was created by training on chromosome 1, using the results to annotate the whole genome by requesting seven labels, allowing each chromatin state signal to vary independently from the others. Furthermore, we required at least 1,000 bp segments, a 500-bp ruler and a 500-bp ruler scale, which had the effect of smoothing across the segmentation, which we found to be excessively sensitive to varied signal at the default settings. Segway was completed requesting ten simultaneous runs, over which maximum likelihood estimations regarding chromatin state were performed. Segway code was implemented using the high performance computing cluster with the training step taking ~72 h and the identify step taking ~2 h using 40 GB of virtual memory.

Mapping epigenetic variability to functional annotations. To interpret our genomic annotations (Segway features) and epigenomic variability (DNA methylation MAD values), we created contour plots showing enrichment within map units within the SOMs. The enrichment within SOMs was determined by a proportion test, specifically asking whether the observed proportion of a feature within a map unit was significantly greater than the expected proportion given the distribution of Segway features overall. A cutoff of 23.09, the 98.5th centile of overall MAD, was used to dichotomize MAD into high and low variable states (for reference, note that ln (23.09) = 3.14, the cutoff shown in Supplementary Fig. 2). From calculated proportions of highly variable 100 bp intervals within the SOM units, we performed a proportion test, specifically testing whether the observed proportion of a highly variable interval within a map unit was significantly ($\alpha = 0.05$) greater than the expected proportion, given the overall proportion of highly variable intervals. The *MASS* library in R along with the *kde2d* and *contour* functions were used to represent the density of variability-enriched map units with a contour plot over SOMs.

We analysed as metaplots RefSeq gene and CpG island annotations, and the 10-kb regions flanking these annotations, allowing us to study the relationship between MAD values and these genomic elements. The bodies of RefSeq genes and CpG islands were split into deciles to be able to compare genes of varying lengths, and the 10-kb flanking region was separated into 100 bp windows. Gene coordinates were rounded to the nearest 100 bp to ensure that sequences were not represented twice. Segway features were divided into 100 bp intervals and matched with 100 bp windows or matched with RefSeq gene deciles, allowing more than one Segway feature to match a particular window or decile. For each 100 bp window or decile, the frequency of each Segway feature was calculated and plotted.

Relating DNA methylation variability to gene expression. To test the relationship of the Segway-derived functional annotations and DNA methylation variability with transcription, we used the CD34+ HSPC RNA-seq data from the same individual from whom chromatin data were obtained (RO_01549, Supplementary Table 4, GEO accession number SRA010036). Of the 17 available RNA-seq runs, we used SRR453391, corresponding to 16,000,000 reads and 2.4 Gb of sequence. Reads were quality controlled using *FASTX-Toolkit* v0.0.13, with *fastq_quality_trimmer* trimming nucleotides with quality lower than 3 and removing sequences shorter than 17 bp. Reads were aligned to the human genome using *GSNAP* version 2012-07-20 requesting at most ten alignments for multiple aligned reads. *SAMtools* v0.1.8 was used to convert the alignment to BAM format. The *Cufflinks* v2.02 programme *Cuffdiff* v2.0.2 was used to calculate fragments per kilobase per million reads values for RefSeq genes, employing normalization by the upper quartile of the number of fragments mapping to individual loci and default weighting of multiple aligned reads based on the number of alignments. Fragments per kilobase per million reads (FPKM) values were used to separate genes into those that were not expressed (8,963 genes) and those that were expressed by quartile of expression (7,872 genes per quartile). Expression information was then linked to the annotated RefSeq gene body deciles and 10 kb flanking regions, thus allowing the stratification of Segway features overlapping 100 bp windows and gene body deciles by gene expression.

DNA methylation variability (MAD values), Segway features and gene expression levels were studied relative to RefSeq genes, dividing the bodies of the genes into deciles, to allow comparisons for genes of different sizes, and extending the analysis using the 100 bp windows to flank the gene body 10 kb upstream and downstream. A similar approach to study the margins and flanking regions of CpG islands was also performed.

DNA methylation variability tests at functional annotations. We tested whether the peak loci for enrichment of 100 bp windows for features 4 and 6 reached statistical significance. The peak window for feature 4 is at −1,500 bp upstream from RefSeq TSSs, where it comprises 31.50% of the features (compared with 15.28% genome wide), while feature 6 is at peak enrichment at −100 bp, comprising 56.60% of annotated features (compared with 9.27% genome wide). A one-way proportion test for each feature at the peak locations shows significance for enrichment for both features ($P < 0.001$). We then tested whether the variability of DNA methylation at these windows of peak feature enrichment was also significantly increased. We compared the MAD values for each feature at these peak enrichment sites with those values at the same number of windows randomly selected from either RefSeq genes, showing with a one-way Wilcoxon rank sum test that variability for DNA methylation at these loci was also significantly increased for features 4 and 6 ($P < 0.001$). Using the same analytical approach, feature 6 was found to be significantly enriched within CpG islands and feature 4 in the ±2 kb CpG island shores ($P < 0.001$). We also tested whether the observed trend of increased DNA methylation variability at feature 4 associated with decreased gene expression levels was significant using the non-parametric Jonckheere trend test. The trend was significant at $P = 10^{-8}$.

Gene properties associated with DNA methylation variability. To interpret DNA methylation variability observations, we asked the question whether the increased variability we observed at candidate promoters and enhancer sequences (Segway features 6 and 4, respectively) was occurring non-randomly at genes with known functions. For each RefSeq gene, Segway feature 6 (promoters) overlapping the TSS and Segway feature 4 (enhancers) occurring within 5 kb flanking the TSS were isolated. We calculated the median MAD over these features. The MAD over promoters and enhancers was dichotomized using the 23.09 value (98.5th percentile) allowing genes to be characterized as having high and low variability over both promoters and enhancers. The Broad Institute's Gene Set Enrichment Analysis web applet (http://www.broadinstitute.org/gsea/) performs a hypergeometric/Fisher's exact test on gene list supplied from the Molecular Signatures Database v3.1 (http://www.broadinstitute.org/gsea/msigdb/), to identify pathways differentially enriched for high- and low-variability enhancers using an false discovery rate *q*-value <0.05 (Supplementary Table 6). We further demonstrate this association using Reactome pathways isolated through the Pathway Commons (http://www.reactome.org/static_wordpress/about/ and http://www.pathwaycommons.org). Gene pathways were visualized in Cytoscape v2.8.3 with edges representing the physical interactions between nodes stored in the GeneMANIA v3.2 plugin. DNA methylation variability was high for promoters for both housekeeping (KEGG ribosome) and leukocyte-specific (leukocyte transendothelial migration) genes at candidate promoters (Segway feature 6), but substantially decreased or absent at housekeeping and not leukocyte-specific genes at candidate enhancer loci (Segway feature 4, Supplementary Fig. 11).

Non-negative matrix factorization. To infer the number of cell subtypes present in the CD34+ HSPC population, we used the NMF approach that has previously been applied to transcriptomic data[34,35]. NMF has been employed successfully in deconvolving gene expression data[35], but it has not previously been applied to DNA methylation data sets. The goal of an NMF algorithm is to deconvolve a matrix *V*, a $(n \times p)$ matrix, to find an approximation of the matrices *W* and *H* such that:

$$V \approx WH, \tag{1}$$

where W and H are $(n \times r)$ and $(r \times p)$ non-negative matrices. The rank of the matrix (r) should be >0 and at least 2, to represent 2 subpopulations. As the methylation outcome is binary and the existing NMF algorithms allow DNA methylation to vary without this constraint, the estimated matrix, W′, may not be interpretable directly, and improvement of the technique may allow the methylation pattern of the individual subpopulations to be estimated. We applied an existing NMF algorithm to understand the presence of subpopulations within our data set but did not interpret the specific values within W′, rather we focused on the difference between the actual and simulated data sets.

Using the R package *deconf*, we varied the matrix rank and estimated the matrices W′ and H′ such that:

$$V \approx W'H' \tag{2}$$

The distance between the original data set V and V′ was calculated as the Frobenius norm:

$$||V' - V||_F^2 = F \tag{3}$$

The process was repeated 100 times per value of *r*, subsetting the data to 10,000 HpaII sites to make the algorithm computationally tractable. A plot of the distribution of *F* as a function of increasing *r* (cell subpopulations) is shown (Fig. 6). In addition, the distribution of Frobenius norms for each cell subpopulation was compared with the preceding cell subpopulation with a two sample *t*-test, testing for a difference in distribution of Frobenius norms between successive simulation levels. As estimating additional subpopulations will always explain additional variability, a smooth spline was fit to the data to look for inflection points, indicating a local minimum in the Frobenius norm when related to cell subpopulations.

References

1. Shen, Y. *et al.* A map of the *cis*-regulatory sequences in the mouse genome. *Nature* **488,** 116–120 (2012).
2. Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23,** 555–567 (2013).
3. Satterlee, J. S., Schubeler, D. & Ng, H. H. Tackling the epigenome: challenges and opportunities for collaboration. *Nat. Biotechnol.* **28,** 1039–1044 (2010).
4. Bell, J. T. & Spector, T. D. A twin approach to unraveling epigenetics. *Trends Genet.* **27,** 116–125 (2011).
5. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nat. Rev. Genet.* **12,** 529–541 (2011).
6. Verma, M. Epigenome-Wide Association Studies (EWAS) in Cancer. *Curr. Genomics* **13,** 308–313 (2012).
7. Flanagan, J. M. *et al.* Intra- and interindividual epigenetic variation in human germ cells. *Am. J. Hum. Genet.* **79,** 67–84 (2006).
8. Feinberg, A. P. & Irizarry, R. A. Evolution in health and medicine Sackler colloquium: stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl Acad. Sci. USA* **107**(Suppl 1): 1757–1764 (2010).

9. Talens, R. P. *et al.* Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. *FASEB J.* **24,** 3135–3144 (2010).
10. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12,** R10 (2011).
11. Teschendorff, A. E. *et al.* Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* **4,** 24 (2012).
12. Gemma, C. *et al.* Inactive or moderately active human promoters are enriched for inter-individual epialleles. *Genome Biol.* **14,** R43 (2013).
13. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2,** e00523 (2013).
14. Vaughn, M. W. *et al.* Epigenetic natural variation in Arabidopsis thaliana. *PLoS Biol.* **5,** e174 (2007).
15. Schmitz, R. J. & Ecker, J. R. Epigenetic and epigenomic variation in Arabidopsis thaliana. *Trends Plant Sci.* **17,** 149–154 (2012).
16. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat. Genet.* **44,** 1207–1214 (2012).
17. Wang, J. *et al.* Dual DNA methylation patterns in the CNS reveal developmentally poised chromatin and monoallelic expression of critical genes. *PLoS ONE* **5,** e13843 (2010).
18. Jeffries, A. R. *et al.* Stochastic choice of allelic expression in human neural stem cells. *Stem Cells* **30,** 1938–1947 (2012).
19. Teng, M. *et al.* The influence of cis-regulatory elements on DNA methylation fidelity. *PLoS ONE* **7,** e32928 (2012).
20. Gervin, K. *et al.* Extensive variation and low heritability of DNA methylation identified in a twin study. *Genome Res.* **21,** 1813–1821 (2011).
21. Li, C. C. *et al.* A sustained dietary change increases epigenetic variation in isogenic mice. *PLoS Genet.* **7,** e1001380 (2011).
22. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).
23. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454,** 766–770 (2008).
24. Suzuki, M. *et al.* Optimized design and data analysis of tag-based cytosine methylation assays. *Genome Biol.* **11,** R36 (2010).
25. Bocker, M. T. *et al.* Genome-wide promoter DNA methylation dynamics of human hematopoietic progenitor cells during differentiation and aging. *Blood* **117,** e182–e189 (2011).
26. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287,** 560–561 (1980).
27. Gertz, J. *et al.* Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. *PLoS Genet.* **7,** e1002228 (2011).
28. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–476 (2012).
29. Zhu, Q., Song, Z., Xie, Y. & Wang, L. A novel recursive Bayesian learning-based method for the efficient and accurate segmentation of video with dynamic background. *IEEE Transact. Image Process.* **21,** 3865–3876 (2012).
30. Kohonen, T. *Self-Organizing Maps* (Springer-Verlag New York, Inc, 2001).
31. Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.* **41,** 178–186 (2009).
32. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9,** 215–216 (2012).
33. Stella, C. C. *et al.* CD34-positive cells: biology and clinical relevance. *Haematologica* **80,** 367–387 (1995).
34. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).
35. Gaujoux, R. & Seoighe, C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.* **12,** 913–921 (2012).
36. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2011).
37. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res.* **21,** 1273–1283 (2011).
38. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107,** 21931–21936 (2010).
39. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
40. Perisic, T., Holsboer, F., Rein, T. & Zschocke, J. The CpG island shore of the GLT-1 gene acts as a methylation-sensitive enhancer. *Glia* **60,** 1345–1355 (2012).
41. Brinkman, A. B. *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res.* **22,** 1128–1138 (2012).
42. Statham, A. L. *et al.* Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. *Genome Res.* **22,** 1120–1127 (2012).
43. Pujadas, E. & Feinberg, A. P. Regulated noise in the epigenetic landscape of development and disease. *Cell* **148,** 1123–1131 (2012).
44. Hansen, K. D. *et al.* Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.* **43,** 768–775 (2011).
45. Koestler, D. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types: a validation analysis. *Epigenetics* **8,** 816–826 (2013).
46. Koestler, D. C., Christensen, B. C., Marsit, C. J., Kelsey, K. T. & Houseman, E. A. Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. *Stat. Appl. Genet. Mol. Biol.* **12,** 225–240 (2013).
47. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31,** 142–147 (2013).
48. Montano, C. M. *et al.* Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol.* **14,** R94 (2013).
49. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41,** 200–209 (2012).

Acknowledgements

Support for this project was provided by Roadmap Epigenomics R01 HD063791 (Einstein/Greally). Support was also provided by Einstein's Medical Student Training Program award (to N.A.W., NIH NIGMS T32 GM007288) and Einstein's Center for Epigenomics, including the Epigenomics Shared Facility and Computational Epigenomics Group.

Author contributions

N.A.W. and F.D. performed experiments, analysed data and wrote manuscript. Y.M.Z. performed experiments. J.C.M. and A.G. provided guidance and performed analytical approaches. F.H.E. and J.M.G. designed study, analysed data, wrote manuscript.

Additional information

Accession codes: All HELP-tagging data generated have been deposited in the Gene Expression Omnibus (GEO) database under the accession code GSE49836.

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: Wijetunga, N. A. *et al.* The meta-epigenomic structure of purified human stem cell populations is defined at *cis*-regulatory sequences. *Nat. Commun.* 5:5195 doi: 10.1038/ncomms6195 (2014).