Re-annotating the EPICv2 manifest with genes, intragenic features, and regulatory elements

Bethan Mallabar-Rimmer[1], Philippa Wells[1], Alice Franklin[1], Jonathan Mill[1], Amy P Webster[1]

Affiliations: [1]University of Exeter Medical School, University of Exeter, UK

Abstract

The Illumina Infinium MethylationEPIC v2.0 BeadChip (EPICv2 array) is a microarray for assessment of the human epigenome. Sites on the EPICv2 array are annotated with an open-source file provided by Illumina, the EPICv2 manifest. Of the 923,452 unique genomic sites targeted by the EPICv2 array, the Illumina manifest identifies just 214,808 as mapping to a gene, excluding many sites located within a gene body.

Based on the genomic coordinates of probes, we have mapped each site assayed on the Illumina EPICv2 array using publicly available data, comprehensively annotating affiliated genes and regulatory elements. We have found that a total of 700,392 EPICv2 array sites are located within a gene body (exon, intron, or UTR) according to the GENCODE Human release 47 (GENCODEv47) database. 509,940 of these sites were not annotated as being within a gene in the Illumina EPICv2 manifest, primarily because the Illumina manifest does not annotate introns – 498,407 of the excluded sites, or 97.74%, are located within the intron of at least one transcript.

The Illumina EPICv2 manifest annotates 358,539 sites as being within 1500bp of a transcription start site (TSS). Using a distance-based approach, we have labelled 267,183 sites as being within promoter distance of a gene (<1500bp upstream or <500bp downstream of the TSS), and 140,123 sites as being within enhancer distance (1501-5000bp upstream of the TSS, excluding sites located within a gene body).

We re-annotated the EPICv2 manifest using GENCODEv47 data to label intragenic features, and a distance-based approach to label the regulatory genome. We also include a column indicating whether a site is located in any promoter or enhancer, according to the GeneHancer database. The re-annotated manifest additionally labels which sites are required for the Horvath DNA Methylation Age Calculator and MethylDetectR epigenetic clocks, to facilitate data preparation for these tools.

In conclusion, we have re-annotated the EPICv2 manifest, allowing more complete assessment of EPICv2 sites associated with gene bodies and regulatory regions during the interpretation of epigenetic studies. The re-annotated manifest is publicly available – see the Data Availability section of this article.

Introduction

The Illumina Infinium MethylationEPIC v2.0 BeadChip (EPICv2 array), released in 2023, is a BeadChip microarray developed for genome-wide assessment of DNA methylation. (1) Previous versions of the array (HumanMethylation450 and EPIC v1.0) rank among the most widely-used technologies for studies of DNA methylation in humans. (2) The EPICv2 array includes 937,690 probes, profiling DNA methylation at 923,452 unique genomic sites (as some probes are controls that do not target a genomic site and some sites are targeted by multiple probes), whereas the previous EPICv1 array included 866,554 probes. (3)

The EPICv2 manifest is an open-source file provided by Illumina, which annotates each probe on the EPICv2 array with technical details about the probe (such as its sequence and Infinium chemistry type), as well as information about the genomic features overlapping the probe's target site (such as whether the site is located within a gene, promoter, and/or transcription start site, TSS). (4,5) This genomic feature annotation is crucial to the interpretation of all results from studies conducted using EPICv2. However, comparison of the EPICv2 manifest to publicly-available genome annotation data accessed from the GENCODE Human release 47 (GENCODEv47) database (6), reveals that over half of sites profiled by the EPICv2 array are present in a gene (i.e. within an exon, intron, or UTR) according to data derived from GENCODEv47 but were not labelled as being located in a gene in the manifest. The majority of these sites are located in introns.

We used GENCODEv47 data to annotate which sites in the manifest are located within genes, transcripts, and intragenic features (including introns, exons, coding sequences or CDSs, and untranslated regions or UTRs). We used a distance-based approach to annotate which sites fall within the regions upstream of genes where regulatory activity is most likely to occur (between 500bp downstream of the TSS and 1500bp upstream of the TSS for the promoter-associated region, and between 1501bp to 5000bp upstream of the TSS for the enhancer-associated region). We annotated which sites are located within enhancer or promoter elements according to the GeneHancer database, although due to restrictions on GeneHancer data distribution, specific IDs of promoters/enhancers and the genes regulated by these elements have not been provided. (7,8) Instead, our annotation uses a binary true/false measure to indicate whether a site is present in a regulatory element according to GeneHancer, and users can find information about the specific regulatory element and associated genes via the UCSC Genome Browser (9) and GeneCards website. (10,11) Finally, we included columns in the manifest annotating which EPICv2 sites are required for upload to the Horvath (12,13) and MethylDetectR (14,15) epigenetic clocks, to facilitate data preparation when using these tools.

Results

Comparison of Illumina's original EPICv2 manifest and the re-annotated manifest

The Illumina EPICv2 manifest annotates 542,165 sites (or 58.71% of all sites targeted by the EPICv2 array) as being located in any gene or regulatory feature – this includes sites mapped to exons, UTRs, the region <1500bp upstream of the TSS, promoters, and enhancers. Notably, the Illumina EPICv2 manifest does not label any sites located within introns. In our version of the manifest, we annotate the aforementioned features as well as

introns. This difference in included features, and the fact that the databases used for annotation differ between our version of the manifest and Illumina's, mean that 866,859 sites (93.87% of all sites on the EPICv2 array) are annotated to a gene or regulatory feature in our version of the manifest.

1. Comparison of gene, transcript, and intragenic feature annotation

The Illumina EPICv2 manifest uses data derived from both the UCSC RefGene and GENCODEv41 databases to annotate sites located within a gene body or <1500bp upstream of a TSS. (4) Counting only those sites which Illumina maps to a gene body (not <1500bp upstream of a TSS) – i.e. those labelled as being in either an exon or UTR in the "UCSC_RefGene_Group" and/or "GencodeV41_Group" columns of the manifest – 216,671 EPICv2 probes (profiling DNA methylation at 214,808 unique genomic sites) were mapped to a gene by Illumina.

We used GENCODEv47 data to label 705,969 EPICv2 probes (targeting 700,392 sites) as mapping to within a gene body (defined in this analysis as the location between transcription start and end sites). Our annotation encompasses exons, introns, and UTRs; in genes, RNAs, and pseudogenes. Annotation of intraexonic features (UTRs and CDSs) is only included for sites located in the exon of a protein-coding gene, as GENCODEv47 does not provide this annotation for other gene types. If an EPICv2 site is located within multiple overlapping transcripts, our annotation labels all transcripts. See Table 1 for a detailed description of features included in the re-annotated manifest, and Figure 1 for annotation of an example site in the manifest.

In total, 513,735 probes (targeting 509,940 sites) were not labelled as being within a gene body by Illumina in the original EPICv2 manifest, but are labelled as being in a gene body in our annotation based on GENCODEv47 data. This is 55.22% of all 923,452 genomic sites assayed by the array. 97.74% of the sites newly labelled as being within a gene in our version of the manifest, or 498,407 sites, are located in the intron of at least one transcript.

Additionally, 24,437 probes (profiling 24,356 genomic sites) were annotated as being within a gene body by Illumina, but are not in a gene body according to our annotation using GENCODEv47 data. 23,372 of these sites, or 95.96%, were labelled as being in a gene body according to the "GencodeV41_Group" column, whereas 16,136 of the sites were in a gene body according to the "UCSC_RefGene_Group" column. Therefore, the sites which are no longer mapped to a gene in our annotation may reflect differences between v41 and v47 of the GENCODE database.

2. Comparison of regulatory element annotation

Illumina annotate 360,385 probes (targeting 358,539 sites) as located within 1500bp upstream of a TSS – this includes all sites labelled with the features "TSS200" or "TSS1500" in the "GencodeV41_Group" and/or "UCSC_RefGene_Group" columns. Additionally, 113,864 probes (targeting 113,342 sites) are labelled as promoter-associated in the "Regulatory_Feature_Group" column. This includes both cell type specific and general promoters. (4) Illumina also annotate 244,757 probes (targeting 242,240 sites) as within an

enhancer, either according to the FANTOM5 database, or based on GENCODE-derived enhancer annotation for a previous version of the Infinium methylation array, the HumanMethylation450 BeadChip. (4)

A direct comparison between the original Illumina annotation and our annotation is not possible, as we have looked 500bp downstream of the TSS as well as 1500bp upstream to identify sites within promoter distance of a gene. We used a window of 1501-5000bp upstream of the TSS to identify sites within enhancer distance, excluding any sites located within a gene body, as the majority of non-tissue-specific enhancers are intergenic. (16) We also used the GeneHancer database (which integrates multiple other databases and sources of evidence, including FANTOM5, to predict the location of regulatory elements) as an alternative method of annotating sites within a promoter or enhancer. (7)

In summary, in the re-annotated manifest, 268,417 probes (targeting 267,183 sites) are labelled as being within promoter distance of a gene, i.e. located <1500bp upstream or <500bp downstream of a TSS. There were 111,093 sites annotated as being <1500bp upstream of the TSS in Illumina's annotation but not in ours. On the other hand, 19,737 sites are labelled as being <1500bp upstream or <500bp downstream of a TSS in our annotation, but were not associated with a TSS in Illumina's annotation.

We annotated 141,418 probes (targeting 140,123 sites) as within enhancer distance of a gene (1501-5000bp upstream of the TSS and not in the body of any gene). We also annotated 302,730 probes (targeting 300,341 sites) as being in either a promoter or enhancer according to the GeneHancer database.

3. Annotation of epigenetic clock sites

The Horvath DNA Methylation Age Calculator uses DNA methylation-based biomarkers of aging to predict the epigenetic age of samples, and is applicable to multiple cell/tissue types. (12) The original Horvath epigenetic clock used 353 loci to predict epigenetic age (12), but the more recent online uploader, which outputs the predictions of multiple epigenetic clocks, uses 94,394 loci. (17) The Horvath uploader technically allows for the upload of EPICv2 data – however, only 82,656 of the 94,394 sites are included on the EPICv2 array. We have added a column to the novel manifest – "Horvath_Clock_Site" – labelling these sites with "TRUE" and all other EPICv2 sites as "FALSE", allowing methylation data to be more easily filtered for upload to the Horvath clock. The 82,656 sites are read by 83,747 probes – i.e. there are 892 replicate probes, which the user will need to account for in their data prior to upload. Example methods for dealing with replicate probes are described in Table 1.

The MethylDetectR epigenetic clock outputs predictions of epigenetic age based on blood tissue, as well as predictions of multiple metabolic measures including BMI, alcohol consumption, smoking behaviour, etc. (14) Similar to the Horvath uploader, input to the MethylDetectR uploader is DNA methylation data (e.g. in the form of Beta values) at 15,190 sites (18), of which 13,780 are included in the EPICv2 array. These have been labelled with "TRUE" in the "MethylDetectR_Clock_Site" column. 210 sites are read by replicate probes which will need to be de-duplicated – see Table 1.

Format of the re-annotated manifest

The re-annotated manifest includes 22 new columns, described in Table 1. To avoid including redundant information, we have removed columns from the manifest which Illumina used to annotate sites located in genes or within 1500bp upstream of the TSS. These columns include: “UCSC_RefGene_Group”, “UCSC_RefGene_Name”, “UCSC_RefGene_Accession”, “GencodeV41_Group”, “GencodeV41_Name”, and “GencodeV41_Accession”. However, a version of the manifest including both our annotations and Illumina’s annotations is available for comparison – see Data Availability. Figure 1 shows example re-annotated sites, with details on how to interpret the annotation.

Table 1

Column Name	Description
GENCODEv47_Gene_Name	Name(s) of any gene(s) which the site targeted by the probe is located in, according to the GENCODEv47 database.
GENCODEv47_Gene_ID	Same as previous, using ENSEMBL gene IDs (beginning ENSG…) rather than gene names.
GENCODEv47_Gene_Type	Gene biotypes, including protein coding, long non-coding RNAs, pseudogenes, etc. – see (19) for a list of biotypes and definitions.
GENCODEv47_Gene_Strand	Whether the gene is transcribed in the + or -direction.
GENCODEv47_Transcript_Name	Name(s) of any transcripts(s) which the site is located in, according GENCODEv47.
GENCODEv47_Transcript_ID	IDs of transcripts (beginning ENST…)
GENCODEv47_Feature_Type	Classifies which intragenic feature(s) the site is located in. For sites located in protein-coding genes, features labelled in this column include: intron, exon CDS, or exon UTR. For sites in all other types of genes, features labelled include: intron or exon. This is because the more granular level of detail about features within exons (CDSs and UTRs) is only provided by GENCODEv47 for protein-coding genes.
GENCODEv47_Feature_Exon_Number	Which exact exon(s) the site falls in, e.g. exon1, exon3, exon22, etc.
GENCODEv47_Feature_Type_Specific	A column which combines the above two columns of GENCODEv47_Feature_Type and GENCODEv47_Feature_Exon_Number, to give users a quick overview of which exact feature(s) a site is located in. E.g. “exon18_CDS” means a site is located in the coding region of exon 18 of a protein-coding transcript. Whereas “exon2_in_RNA” means the site is located in exon 2 of an RNA transcript (the GENCODEv47_Gene_Type column included in the re-annotated manifest would provide more detail on whether the RNA is a microRNA, long non-coding RNA, etc.)
GENCODEv47_Gene_Start GENCODEv47_Gene_End GENCODEv47_Transcript_Start GENCODEv47_Transcript_End GENCODEv47_Feature_Start GENCODEv47_Feature_End	The start and end coordinates of any gene(s), transcript(s) and feature(s) such as introns or exons which the site is located in, according to GENCODEv47.
DB_Element_Type	DB is an abbreviation of distance-based. Sites located within 1500bp upstream or within 500bp downstream of any gene’s TSS are within promoter distance of a gene and labelled as “Promoter_2000bp” in this column (as we looked within a 2000bp window surrounding the TSS). Sites located between 1501bp and 5000bp upstream of the

	TSS (excluding sites located within any gene body) are within proximal enhancer distance, and are labelled as “Enhancer_5000bp”.
DB_Element_Gene_Name	If a site is classified as promoter-associated or enhancer-associated due to its proximity to a TSS, described above, this column indicates the name of the gene which the TSS belongs to. Only the nearest gene to the site is used.
DB_Element_Gene_ID	As above, but the gene is labelled with ENSEMBL IDs (beginning ENSG…)
DB_Element_Gene_Type	Describes the gene biotype, as in (19)
In_GeneHancer	TRUE if a site is located within any predicted double-elite promoter or enhancer, according to the GeneHancer database, FALSE otherwise. Double-elite refers to regulatory elements where both the existence of the element and its ability to regulate associated genes are supported by two or more sources of evidence. (7)
Horvath_Clock_Site MethylDetectR_Clock_Site	TRUE if a site is required for upload to the Horvath DNA Methylation Age Calculator or MethylDetectR platforms respectively, FALSE otherwise. Note that neither epigenetic clock was originally built using EPICv2 data (12,14), and although there is evidence that previous Infinium methylation arrays have included replicate probes where multiple probes target the same genomic site, EPICv2 is the first Infinium methylation array to indicate which probes are replicates through its probe naming system. (2,20) Therefore, although our annotation can be used to filter down EPICv2 data to only the sites required by these epigenetic clocks, the user will then need to remove any replicate probes from the data. Existing methods to do this include the epicv2clean function in the wateRmelon R package (21), mLiftOver in the sesame package (22), and our epicv2refine function (which removes replicate probes identified as low quality by Peters et al. (2) then averages any remaining, and is best suited to cancer data (23)).

Table 1: *Columns added to the re-annotated EPICv2 manifest, and description of the information provided by each column. See Methods for further detail on the source of these annotations.*

Figure 1

A

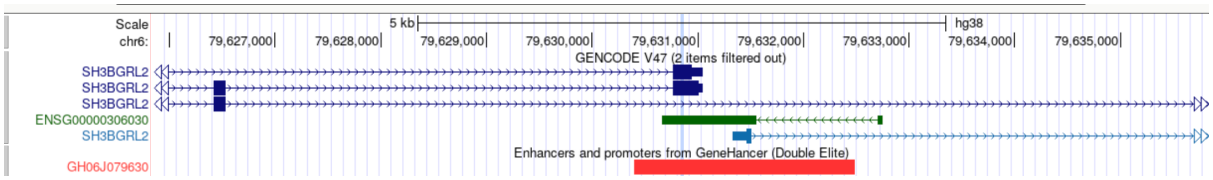
IlmnID	GENCODEv47_Gene_Name	GENCODEv47_Gene_Type
cg16433357_TC21	ENSG00000306030;SH3BGRL2;SH3BGRL2;SH3BGRL2	lncRNA;protein_coding;protein_coding;protein_coding

IlmnID	GENCODEv47_Transcript_Name	GENCODEv47_Feature_Type_Specific
cg16433357_TC21	ENST00000814882;SH3BGRL2-209;SH3BGRL2-212;S…	exon2_in_RNA;exon3_CDS;exon4_CDS;intron

IlmnID	DB_Element_Type	DB_Element_Gene_Name	In_GeneHancer
cg16433357_TC21	Promoter_2000bp	SH3BGRL2	TRUE

IlmnID	Horvath_Clock_Site	MethylDetectR_Clock_Site
cg16433357_TC21	FALSE	FALSE

B



C

IlmnID	UCSC_RefGene_Group	UCSC_RefGene_Name	GencodeV41_Group	GencodeV41_Name
cg16433357_TC21	TSS1500	SH3BGRL2	TSS1500	ENSG00000198478.8

IlmnID	Regulatory_Feature_Group	450k_Enhancer	Phantom5_Enhancers
cg16433357_TC21		FALSE	

Figure 1.

A: Screenshot of the re-annotated manifest, showing annotation of site cg16433357_TC21. Only 9 of 22 new columns added to the manifest are shown for easier interpretation. As highlighted in the red boxes, this site is located in exon 2 of one transcript of long non-coding RNA, *ENSG00000306030.* It is also located in 3 different overlapping transcripts of the protein-coding gene *SH3BGRL2,* as shown by the yellow, green and blue boxes. For two of these transcripts (highlighted in yellow and green), the site is located in the coding region of exons 3 and 4 respectively, whereas the site is in an intron of the third transcript (highlighted in blue).

The "DB_Element_Type" and "DB_Element_Gene_Name" columns indicate that the site is within promoter distance of the *SH3BGRL2* gene – i.e. it is either <500bp downstream of its TSS or <1500bp upstream. While this initially appears to contradict the site being labelled as within exons 3 or 4 of *SH3BGRL2*, viewing the site with UCSC genome browser (Figure 1B) confirms this annotation is correct; the site is located within 3 overlapping transcripts of this gene, and <1500bp upstream of a 4th transcript of the same gene.

The "In_GeneHancer" column shows this site is within at least one promoter or enhancer in the GeneHancer database. The last two columns indicate that data from this site is not required as input to either the Horvath or MethylDetectR epigenetic clocks.

B: The region 5000bp either side of site cg16433357_TC21 viewed with the UCSC Genome Browser. This confirms that the site overlaps 3 transcripts of gene *SH3BGRL2* – being in located in an exon of two of the transcripts and the intron of another. The site is located <1500bp upstream of the TSS of a 4th transcript of *SH3BGRL2*, hence why it was labelled as being within promoter distance of this gene. It is also located within exon 2 of RNA *ENSG00000306030*. Our annotation labelled this site as being within a GeneHancer regulatory element. Viewing the region with UCSC Genome Browser clarifies that this is element GH06J079630, a promoter that regulates genes including *ELOVL4*, *LCA5*, and *SH3BGRL2*. (24)

C: Original annotation of site cg16433357_TC21 in the Illumina EPICv2 manifest. Illumina's annotation correctly indicates that this site is within 1500bp upstream of a transcript of *SH3BGRL2* (*ENSG00000198478* is the ENSEMBL ID of this gene (25)), but omits firstly that the site is promoter-associated, and secondly that the site is located within the gene body of 3 additional transcripts of *SH3BGRL2* and 1 transcript of *ENSG00000306030.*

Discussion

The re-annotated manifest significantly increases the percentage of sites on the EPICv2 manifest mapped to a gene or regulatory feature, offering numerous advantages to researchers.

The original Illumina EPICv2 manifest annotated 58.71% of sites profiled by the EPICv2 array as being located in a gene or regulatory feature, whereas in our version of the manifest 93.87% of sites are annotated to a gene or regulatory feature. The original manifest lacked annotations for numerous array sites, primarily those located in introns, but also some sites in exons and the region immediately upstream of the TSS. This meant that sites of interest – for example, differentially methylated positions (DMPs) where methylation levels vary between case and control groups in comparative epigenetic studies such as epigenome-wide association studies (EWAS) – would frequently be inaccurately labelled as intergenic, or not in any regulatory feature. Our re-annotation amends this, allowing more complete assessment of the biological impact of differential methylation at sites of interest measured by the EPICv2 array. Our re-annotation particularly improves annotation of the non-coding genome, vital because methylation at promoters and enhancers regulates transcription, and methylation at the exon-intron boundary is thought to affect alternative splicing. (26)

More EPICv2 sites are annotated as within a gene or regulatory feature in our annotation, meaning researchers can more easily attain a list of genes associated with sites of interest resulting from epigenetic studies. This will improve the accuracy of subsequent pathway analysis, which relies on a list of genes as input.

The improved coverage and accuracy of EPICv2 annotation in our version of the manifest will also facilitate integrative multi-omic studies. We have labelled which EPICv2 sites are located within promoter/enhancer distance of the TSS, or in a regulatory region according to the GeneHancer database. Differential methylation of sites in regulatory elements may indicate varying levels of transcription and perhaps translation of associated genes, which could be confirmed with multi-omic studies integrating transcriptomic or proteomic data. Furthermore, accurate annotation of which genomic features EPICv2 sites are located in will improve integration of EPICv2 data with genomic data.

In the re-annotated manifest, 75.84% of sites are mapped to within a gene body (i.e. between transcript start and end site, encompassing UTRs, exons and introns). This is a surprisingly high figure considering Illumina describes the EPICv2 array as extensively profiling CpG islands and regulatory elements, as well as genes. (1) Comparatively, Illumina labels 39.70% of array sites (366,627 sites) as being within a CpG island. The high proportion of sites in genes relative to CpG islands may be because we include all transcripts of genes, not just the primary/main transcript, which will increase the number of sites mapped to a gene. It is important to note that there is significant overlap between CpG islands and gene bodies, and a site may be in both. In Illumina's annotation, 34.53% of sites in a CpG island are also labelled as being within a gene body – in our annotation, this figure rises to 73.09%. Therefore, annotating an increased number of sites as being within a gene does not compromise the EPICv2 array's coverage of CpG islands and the regulatory genome.

Limitations

The re-annotated manifest is based primarily on data from the latest version of GENCODE, which is subject to change with future updates. It is therefore likely that the manifest will undergo changes in future, both in terms of updating the annotation based on database updates, and adding additional features.

Data distribution restrictions meant that we could not provide detail on GeneHancer regulatory element names, or associated genes, as part of the manifest. Instead we have indicated whether or not a site is present in a regulatory element, indicating to users whether it would be useful to seek further information via UCSC Genome Browser or the GeneCards website. In future, we aim to increase the level of detail regarding the annotation of promoters and enhancers, which may include integrating other data sources.

Conclusion

We have re-annotated the EPICv2 manifest, providing an up-to-date and more comprehensive framework for interpreting DNA methylation data. Integrating GENCODEv47 data reveals that a significant number of intragenic EPICv2 array sites were previously unannotated as being in a gene in the original EPICv2 manifest, particularly within introns. The re-annotation also expands identification of promoter and enhancer regions, using a distance-based approach to identify sites within regulatory distance of the TSS, and labelling which sites are present in a promoter or enhancer using the GeneHancer database. Furthermore, the inclusion of epigenetic clock-specific sites allows for streamlined data preparation for the Horvath DNA Methylation Age Calculator and MethylDetectR clocks. The re-annotated manifest offers a valuable resource for researchers, facilitating more accurate interpretation of the results of studies conducted using the EPICv2 array, and providing a foundation for future updates and improvements in epigenetic research.

Method

The original EPICv2 manifest is available as part of the Infinium MethylationEPIC v2.0 Product Files from the Illumina website. (5) All analysis and re-annotation was conducted in the R programming language, version 4.3.2. (27)

Intragenic/Gene Body Annotation

The GENCODE Human release 47 basic gene annotation, for genome version GRCh38.p14, was downloaded from the GENCODE website. (26, 6) This annotation includes names, IDs, and coordinates of genes, transcripts, and exons for the whole human genome. When re-annotating the manifest, all EPICv2 sites located within a gene body were labelled with the IDs/names of any genes and transcripts they were located in, as well as whether the site was in an intron or exon, and the exon number if applicable. The GENCODEv47 basic gene annotation includes coordinates of features within exons (CDSs and UTRs) for protein coding genes only. Therefore, any EPICv2 sites within the exon of a protein coding gene were also annotated with this more granular level of detail, as being in a CDS or UTR.

The GENCODEv47 basic gene annotation does not include information about the location of introns. Therefore, we calculated intron coordinates for all genes as starting 1bp after the end coordinate of the preceding exon and 1bp before the start coordinate of the subsequent exon.

Sites in the EPICv2 manifest were annotated as being within a feature (such as a gene, transcript, or intragenic feature such as intron/exon/CDS/UTR) if 1. the genomic location of the site (found in the MAPINFO column of the manifest) fell between the start and end coordinates of the feature, and 2. the site and feature were located on the same chromosome. Figure 1 shows an example annotation for a probe, the target site of which is located in multiple overlapping transcripts. Table 1 describes annotation in more detail in the new version of the manifest.

Distance-based Regulatory Feature Annotation

DNA methylation proximal to the TSS of a gene is more likely to affect expression levels of that gene – although this effect is primarily noted for the core promoter region (±50 base pair (bp) around the TSS) (29) regulatory elements can span thousands of bp (30). For this reason, as well as annotating which EPICv2 array sites are located within a gene body, we also annotated which sites are located in the region surrounding or upstream of a TSS. Using gene start coordinates accessed from GENCODEv47, EPICv2 sites located either 1. within 1500bp upstream of a gene's TSS, or 2. up to 500bp downstream of the TSS, were labelled as being within promoter distance of that gene. Sites located between 1501-5000bp upstream of a gene's TSS, excluding sites located within any gene body, were annotated as being within enhancer distance of the gene.

GeneHancer Regulatory Feature Annotation

GeneHancer is a database of regulatory features, part of the GeneCards database, which integrates multiple other databases to identify promoters/enhancers and predict their target genes. (7) GeneHancer double-elite annotation data was accessed via UCSC Table Browser. (9) The double-elite annotation includes only the regulatory elements where both the existence of the regulatory element, and its association with any linked genes, are supported by at least two sources of evidence. (7) Because it is not currently possible to download the entire database at once, data was downloaded one chromosome at a time from UCSC Table Browser, and only the section of the chromosome spanning sites read by the EPICv2 array was downloaded.

Although the GeneHancer database includes the genomic coordinates and unique IDs of enhancers and promoters, as well as the genes which these elements regulate, data from GeneHancer cannot be redistributed in full or part without the permission of LifeMap Sciences. Therefore, when re-annotating the manifest, we did not include annotation of which specific promoters/enhancers EPICv2 sites are located in, or linked genes. Instead we have included a binary column called “In_GeneHancer”, indicating whether or not a site is present within any promoter or enhancer according to the GeneHancer database. If a site of interest (for example, a site predicted to be significant following differential methylation analysis) is labelled “TRUE” in this column, this indicates that the site is located within a double-elite regulatory element. More information about the element and genes it regulates can be found online, e.g. by inputting the site’s name or genomic coordinates into UCSC Genome Browser and viewing the GeneHancer track.

Annotation of sites used in the Horvath DNA Methylation Age Calculator and MethylDetectR epigenetic clocks

EPICv2 sites required for upload to the Horvath DNA Methylation Age Calculator were identified by matching the “EPICv1_Loci” column in the original EPICv2 manifest, to the names of CpG

sites in the file “datMiniAnnotation4_fixed.csv”, available to download from the Clock Foundation website. (17) Sites required for upload to MethylDetectR were identified by matching the “EPICv1_Loci” column to the CpGs listed in the “Truncate_to_these_CpGs.csv” file available from (18) .

Comparison of the original Illumina EPICv2 manifest and the re-annotated EPICv2 manifest

The EPICv2 array reads DNA methylation level at 923,452 genomic sites, using 937,690 probes. The EPICv2 manifest is a CSV file with one row per probe. The number of unique sites targeted by the array was calculated by joining the “MAPINFO” column of the manifest (which gives the genomic location of the site being read by each probe) to the “CHR” column (the chromosome of the site), and counting the number of unique values, excluding any probes labelled as being on ‘chromosome 0’ or with NA values in these columns.

The original EPICv2 manifest uses UCSC RefSeq and GENCODEv41 data to annotate which sites are located in genes, transcripts, exons, UTRs or upstream of a TSS. (4) Any site labelled as being in a UTR or exon in either the “UCSC_RefGene_Group” or “GencodeV41_Group” columns of the original manifest was counted as being annotated to a gene body by Illumina. Note that the original Illumina EPICv2 manifest does not include annotation of any introns. In the re-annotated manifest, we have removed the UCSC_RefGene and GencodeV41 annotations, replacing these with our annotation based on the latest GENCODEv47 release. Any site with a non-empty value in the new “GENCODEv47_Feature_Type” column was counted as being annotated to a gene body in the novel manifest.

The number of sites labelled as being upstream of a TSS by Illumina was counted by looking for the phrase “TSS” in either the UCSC_RefGene_Group” or “GencodeV41_Group” columns. This includes sites labelled as being in the “TSS200” or “TSS1500” (i.e., within 200bp or 1500bp upstream of the TSS respectively). We compared this to the number of sites annotated with the term “Promoter_2000bp” in the “DB_Element_Type” column of the re-annotated manifest. Sites annotated as being in an enhancer by Illumina were identified by looking for a value of “TRUE” in the “450k_Enhancer” column or any non-empty value in the “Phantom5_Enhancers” column. This was compared to the number of sites annotated with “Enhancer_5000bp” in the “DB_Element_Type” column.

Data Availability

The re-annotated manifest is available as a .rds file for import into R, and a zipped .csv file, from Zenodo: https://zenodo.org/records/14933469

The version we recommend for most purposes is “EPICv2_reannotated_manifest_v1.0” which includes 22 new columns annotating genes and regulatory features, and excludes redundant gene/feature annotation columns from the original Illumina EPICv2 manifest.

Another version of the manifest, “EPICv2_reannotated_manifest_forcomparison_v1.0” is available from the above link, which includes all old and new columns. This allows for comparison of Illumina’s annotation with our re-annotation, and reproducibility of the findings reported in this article.

Contributions

BMR wrote the manuscript, re-annotated the EPICv2 manifest with genes/transcripts/intragenic features using GENCODEv47 data, added the GeneHancer annotation, added the annotation of which sites are required for upload to epigenetic clocks, formatted the manifest for readability and uploaded it publicly.

PW, BMR and AW conceptualised and planned the project of re-annotating the manifest.

PW re-annotated the EPICv2 manifest with the distance-based promoter and enhancer annotation, and will add further annotations related to regulatory elements in future.

AW, PW, AF, and JM proofread the manuscript. AW supervised the project.

References

1. Illumina. Infinium MethylationEPIC v2.0 Kit [Internet]. 2025 [cited 2025 Feb 24]. Available from: https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html

2. Peters TJ, Meyer B, Ryan L, Achinger-Kawecka J, Song J, Campbell EM, et al. Characterisation and reproducibility of the HumanMethylationEPIC v2.0 BeadChip for DNA methylation profiling. BMC Genomics. 2024 Mar 6;25(1):251.

3. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016 Oct 7;17(1):208.

4. Illumina. Infinium MethylationEPIC v2.0 Manifest Column Headings | Illumina Knowledge [Internet]. 2024 [cited 2025 Feb 24]. Available from: https://knowledge.illumina.com/microarray/general/microarray-general-reference_material-list/000009183

5. Illumina. Infinium MethylationEPIC v2.0 Product Files [Internet]. 2024 [cited 2025 Feb 24]. Available from: https://emea.support.illumina.com/downloads/infinium-methylationepic-v2-0-product-files.html

6. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE 2025: reference gene annotation for human and mouse. Nucleic Acids Res. 2025 Jan 6;53(D1):D966–75.

7. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017 Jan 1;2017:bax028.

8. LifeMap Sciences. GeneCards. [cited 2025 Feb 24]. Terms of use | GeneHancer Data Request. Available from: https://www.genecards.org/Guide/DatasetRequest

9. Perez G, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2025 update. Nucleic Acids Res. 2025 Jan 6;53(D1):D1243–9.

10. GeneCards - Human Genes | Gene Database | Gene Search [Internet]. [cited 2025 Feb 24]. Available from: https://www.genecards.org/

11. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Curr Protoc Bioinforma. 2016 Jun 20;54:1.30.1-1.30.33.

12. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013 Dec 10;14(10):3156.

13. Horvath S, The Clock Foundation. Home | DNA Methylation Age Calculator [Internet]. 2023 [cited 2025 Feb 24]. Available from: https://dnamage.genetics.ucla.edu/home

14. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. Wellcome Open Res. 2020;5:283.

15. Hillary RF, Marioni RE. MethylDetectR [Internet]. [cited 2025 Feb 24]. Available from: https://shiny.igc.ed.ac.uk/MethylDetectR/

16. Borsari B, Villegas-Mirón P, Pérez-Lluch S, Turpin I, Laayouni H, Segarra-Casas A, et al. Enhancers with tissue-specific activity are enriched in intronic regions. Genome Res. 2021 Aug;31(8):1325–36.

17. Brooke R. datMiniAnnotation4_fixed.csv [Internet]. 2023 [cited 2024 Feb 24]. Available from: https://dnamage.clockfoundation.org/importanthints-page

18. Hillary RF, Marioni RE. MethylDetectR - A Translational Tool for Methylation-Based Health Profiling [Internet]. Zenodo; 2022 [cited 2025 Feb 24]. Available from: https://zenodo.org/records/7154750

19. The GENCODE group. GENCODE. 2024 [cited 2025 Feb 24]. Gene/Transcript Biotypes in GENCODE & Ensembl. Available from: https://www.gencodegenes.org/pages/biotypes.html

20. Kaur D, Lee SM, Goldberg D, Spix NJ, Hinoue T, Li HT, et al. Comprehensive Evaluation of The Infinium Human MethylationEPIC v2 BeadChip. Epigenetics Commun. 2023;3(1):6.

21. Pidsley R, Wong CC, Volta M, Lunnon K, Gorrie-Stone TJ, Touleimat N, et al. wateRmelon: Illumina DNA methylation array normalization and metrics [Internet]. 2025. Available from: https://bioconductor.org/packages/release/bioc/html/wateRmelon.html

22. Chen BH, Zhou W. mLiftOver: harmonizing data across Infinium DNA methylation platforms. Bioinformatics. 2024 Jul 1;40(7):btae423.

23. Mallabar-Rimmer B. epicv2refine [Internet]. 2024 [cited 2025 Feb 24]. Available from: https://github.com/bethan-mallabar-rimmer/epicv2refine

24. LifeMap Sciences. GeneCards. [cited 2025 Mar 4]. GH06J079630 related genes - GeneCards Search Results. Available from: https://www.genecards.org/Search/Keyword?queryString=GH06J079630

25. Gene: SH3BGRL2 (ENSG00000198478) - Summary - Homo_sapiens - Ensembl genome browser 113 [Internet]. [cited 2025 Mar 4]. Available from: https://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000198478;r=6:79537185-79703655

26. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012 Jul;13(7):484–92.

27. R Core Team. R: A Language and Environment for Statistical Computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/

28. Mudge JM, Carbonell-Sala S, Diekhans M, Martinez JG, Hunt T, Jungreis I, et al. GENCODE. 2024 [cited 2025 Feb 24]. GENCODE - Human Release 47. Available from: https://www.gencodegenes.org/human/

29. Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. Nat Rev Genet. 2020 Feb;21(2):71–87.

30. Hong J, Rhee JK. Genomic Effect of DNA Methylation on Gene Expression in Colorectal Cancer. Biology. 2022 Sep 23;11(10):1388.