Syrian Arab Republic
Ministry of Higher Education
and Scientific Research
Syrian Virtual University

الجمهورية العربية السورية
وزارة التعليم العالي والبحث العلمي
الجامعة الافتراضية السورية

DNA Methylation Patterns of Behavioural Disorders: A Cross-trait Analysis

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science in Bioinformatics

Student | Saeed Omar Saado

Supervisor | Dr. Yanal Ahmad Alkuddsi

2025

Declaration

I hereby declare that the research presented in this dissertation, titled *(DNA Methylation Patterns of Behavioural Disorders: A Cross-trait Analysis)*, is entirely my own work and has not been submitted, in whole or in part, for any other degree or qualification at this or any other university or educational institution. All results and work presented are a product of my personal efforts, conducted under the guidance of my supervisor. Any external information or results referenced have been duly cited with appropriate acknowledgment of their original sources and authors in the text and the reference list.

إهداء

إلى والديّ العزيزين، شكرًا جزيلاً على دعمكم اللامحدود، الذي لم يقتصر على الدعم الأبوي والمادي، بل أنه تجاوز ذلك ليشمل الثقة بالنفس والتفاؤل بكل خير. لقد كنتم مصدر إلهام حقيقي لي وما زلتم تلهموني في كل فصل من حياتي. شكرًا لكونكم السند والمعلم والمربي.

إلى ابنتي ليان، أريد أن أهدي هذا العمل لكِ أيضًا لأنكِ أكبر دافع لي في الحياة. أنتِ المثال والرمز الذي يذكرني في التغلب على الصعوبات والإيمان بقدرة الله عز وجل.

شكر وتقدير

بداية أتوجه بالشكر والتقدير لمشرف المشروع، الدكتور ينال القدسي على إخلاصه وتفانيه في تقديم المساعدة والمشورة طوال فترة العمل. كما أود توجيه شكر وعرفان للأستاذة الذي كان لهم دور بارز في الإعطاء المميز والذي ساهم في تعزيز الفهم الأكاديمي والتطبيق العملي وإيصال الأفكار، وأخص بالذكر كل من الدكتور مجد الجمالي، والدكتور رؤوف حمدان والدكتور ياسر خضرا والدكتورة رنوة السيد، والدكتورة لمى يوسف، والدكتور باسم عصفور.

كما أتوجه بالشكر لبقية الأساتذة وفريق عمل الجامعة السورية الافتراضية على دعم مشروع التعليم المستمر والافتراضي الذي كان له دور كبير في دخولي برنامج التخصص وإنجاز هذا العمل.

وأخيرا أتوجه بالشكر والعرفان إلى كل من ساندني من أصدقاء وزملاء قاموا بتوفير الأسباب من اجل إتمام هذا العمل.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

1	**CpG**	Cytosine - Linking Phosphate - Guanine
2	**BS-Seq**	Bisulfite Sequencing
3	**WGBS**	Whole Genome Bisulfite Sequencing
4	**PCR**	Polymerase Chain Reaction
5	**DMRs**	Differentially Methylated Regions
6	**DMPs**	Differentially Methylated Probes
7	**SeSAME**	SEnsible Step-wise Analysis of DNA MEthylation BeadChips
8	**pOOBAH**	P-value with out-of-band (OOB) array hybridization
9	**LIMMA**	Linear Models for Microarray and Omics Data
10	**FDR**	False Discovery Rate
11	**Funnorm**	Functional Normalization
12	**SWAN**	Subset Within Array Normalization
13	**DLD**	Developmental Language Disorder
14	**Early Alz**	Early-onset Alzheimer's disease
15	**EOAD**	Early-onset Alzheimer's disease
16	**Fam Alz**	Familial Alzheimer Disease
17	**FTP GRN**	Genetic Frontotemporal Dementia (GRN Mutation)
18	**FTP MAPT**	Genetic Frontotemporal Dementia (MAPT Mutation)
19	**ELA**	Early Life Adversity
20	**SAD**	Social Anxiety Disorder
21	**SAD ELA**	Social Anxiety Disorder and Early Life Adversity
22	**TSD**	Total Sleep Deprivation
23	**QC2 or QC (part2)**	Quality Control After *SeSAME*
24	**Max Abs. Diff.**	Maximum Absolute Deviation From the Mean
25	**WM**	Weighted Mean Average
26	**IQR**	Interquartile Range
27	**STD**	Standard deviation
28	**TRPQC**	Threshold Probe Quality Control

Abstract (Arabic Version)

الملخص

إن مثيلة الحمض النووي آلية أساسية في علم الوراثة اللاجيني، وهي عملية تتأثر بالعوامل البيئية والأمراض التي تسبب تغييرات في أنماط المثيلة على سلاسل الحمض النووي. يمكن أن تؤثر هذه التغييرات على النمط الظاهري دون تغيير تسلسل النيوكليوتيدات، مما يؤدي عادة إلى كبت الجينات من خلال فرط المثيلة في مناطق المحفزات promoter regions. كما تؤدي التغييرات المستمرة في المثيلة إلى الطفرات، مما يحفز الباحثين لدراسة المؤشرات الحيوية للمثيلة المرتبطة بالأمراض، وتمثل السرطانات الأمراض الأكثر دراسة في هذا المجال. أما الاضطرابات الأخرى فتحظى باهتمام أقل، وكذلك العوامل البيئية. إن الدافع من هذه الدراسة منطلق من المبدأ القائل بأن الوقاية خير من العلاج، خاصة عندما تكون آليات بعض الاضطرابات غير واضحة، مع عدم وجود عوامل خطر محددة. تمثل الحالات النفسية واضطرابات السلوك مثالاً على هذه المشكلة، حيث يفتقر معظمها إلى علاجات حاسمة. علاوة على ذلك، فإن تنوع الأعراض وتداخلها يجعل التشخيص أكثر تعقيدًا. غالبًا ما يترافق مع الأدوية المتاحة لهذه الاضطرابات آثار جانبية، خاصة عند استخدامها لفترات طويلة. وفي ضوء هذه التحديات، استلهمنا فكرة البحث عن العوامل البيئية المحتملة لهذه الاضطرابات أو عن مؤشرات لاجينية Epigenetic مترافقة بين اضطرابات مختلفة، من أجل التعرف أكثر على العوامل المساهمة في تطورها، بهدف تحسين جودة حياة الأفراد دون الاعتماد بالضرورة على الأدوية، أو على الأقل تقليل استخدامها إلى الحد الأدنى. بالإضافة إلى ذلك، يمكن أن يساعد تحديد العوامل البيئية المحتملة أو المؤشرات الحيوية اللاجينية في زيادة الوعي حول تشخيص المرض باكرا والمساهمة في تقليل معدلات الإصابة. تم تحليل بيانات Micro Array لـ 9 أنماط ظاهرية مختلفة (إجمالي حجم العينة 125 فرد) بهدف البحث عن تباينات مهمة على مستوى مثيلة الحمض النووي، مع إثراء النتائج جينيًا قبل المقارنة عبر المجموعات. من الناحية التقنية، تم إجراء مراجعة شاملة للأدبيات حول طرق المعالجة المبدئية للبيانات. وبالتالي، تم اختيار الأدوات المستخدمة في المعالجة بناءً على أحدث التطورات في الطرائق والتوصيات. تم استخدام *SeSAMe* من R Bioconductor لأسلوبها الحديث (*pOOBAH*) في حساب قيم P values بالإضافة إلى خاصية تقنيع masking شاملة تقوم باستثناء الإشارات الغير الموثوقة بناء على عوامل متعددة، هذا وقد أعطت *SeSAME* أفضل النتائج حين مقارنتها بالأدوات الأخرى في الأدبيات البحثية. تبع ذلك جولة إضافية من خطوات ضبط جودة البيانات التي قام بها الباحث والتي هدفت إلى تقليل نسبة التعديل على البيانات إلى الحد الأدنى وضمان جودة عالية. تم الكشف عن المناطق متباينة المثيلة Differentially Methylated Regions (DMRs) باستخدام *Limma*مع تطبيق تصحيح FDR. بينما كانت التغيرات في مستويات المثيلة التي تم اكتشافها في هذه الدراسة محصورة بشكل عام على مناطق صغيرة من الحمض النووي (Single Probes)، لوحظ اشتراك هذه المناطق في المستوى الجيني. لقد أظهرت مناطق مختلفة تتبع كل من الجينات *CD81* و*MAD1L1* ترافقا في انخفاض المثيلة عند مجموعة تأخر اللغة التنموي (DLD) ومجموعة أو مجموعتين من مرضى الزهايمر. ومن المثير للاهتمام، هو ملاحظة زيادة التعبير الجيني ل *CD81* لدى مرضى الزهايمر في دراسة أخرى استخدمت عينات مأخوذة من نسيج القشرة الأمامية للدماغ، الأمر الذي يجعل الدراسة التي قمنا بها، إلى جانب دراسات أخرى، تعطي أدلة إضافية تدعم إمكانية استخدام المؤشرات الحيوية في الدم المحيطي كمؤشرات تعكس الوظائف الدماغية. من ناحية أخرى، تمت مناقشة *MAD1L1*

بشكل متكرر في الأدبيات العلمية التي تتناول مثيلة هذا الجين في أنماط نفسية وبيئية، الأمر الذي يشجعنا على استهداف هذا الجين تحديدا في الدراسات القادمة وذلك لدوره المحتمل في تفسير الأعراض العصبية والنفسية. من ناحية أخرى فإن نتائجنا تؤكد على حاجة الباحثين إلى دليل موحد في تحليل مثيلة الحمض النووي، الأمر الذي سيعزز اتساق النتائج وموثوقيتها خصوصا في الدراسات التي تتناول اضطرابات تفتقر إلى التغيرات الكبيرة في نمط المثيلة.

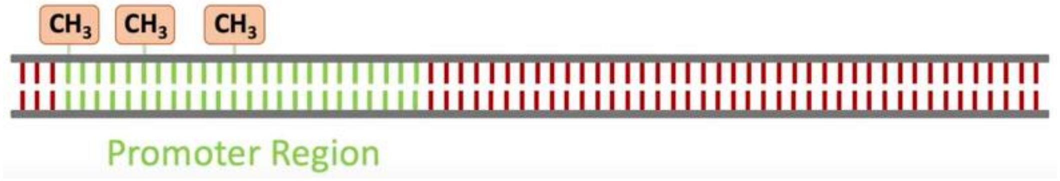
Abstract

DNA methylation is a key mechanism in epigenetics, influenced by environmental factors and disorders that cause changes in methylation patterns across DNA strands. These changes can affect phenotype without altering nucleotide sequences, often silencing genes through hypermethylation at promoter regions. Persistent methylation alterations may lead to mutations, prompting researchers to study DNA methylation biomarkers associated with diseases, with cancers being the most extensively investigated. Other disorders receive much less attention, and so do environmental factors. The motive of this study originates from the principle that prevention is better than treatment, particularly when the mechanisms of certain disorders remain unclear, with no established risk factors. Psychiatric conditions and behavioural disorders exemplify this issue, as most lack definitive treatments. Moreover, the diversity and overlap of symptoms complicate diagnosis further. Available medications for these conditions often come with side effects, especially when used long-term. This has inspired us to explore potential environmental contributions to such disorders, with the goal of improving individuals' quality of life without necessarily relying on medication, or at least minimizing its use. In addition, identifying potential environmental factors or epigenetic biomarkers can raise awareness about disease prognosis and ultimately help reduce incidence rates. Micro array data for 9 different phenotypes (Total 145 Subjects) were analysed for significant DMRs, with the results being gene-enriched prior to cross-comparison. From a technical standpoint, a comprehensive literature review on data preprocessing methods was conducted. Consequently, the tools used for analysis were selected based on recent advancements and literature recommendations. *SeSAMe* from R Bioconductor was employed for its modern p-value calculation method and comprehensive QC masking. This was followed by additional customized QC steps to minimize imputation of masked values and ensure high data quality. DMRs were detected using the *Limma* package, with FDR correction applied. While the alteration in methylation levels detected in this study was generally limited to single-probe differentiation, several DMPs were shared at the gene level. Both *CD81* and *MAD1L1* exhibit hypomethylation associated with DLD and one or two Alzheimer's groups. Interestingly, *CD81* has been reported as upregulated in Alzheimer's candidates in a study using prefrontal cortex tissue samples. Therefore, our research, along with other studies, provides further evidence supporting the potential of peripheral blood biomarkers in reflecting neurological symptomatology. *MAD1L1*, on the other hand, has been frequently discussed in existing literature regarding the methylation of the same gene in psychiatric and environmental contexts. These findings should encourage further investigation of *MAD1L1* to explore its potential role in neuropsychiatric symptoms. The study also emphasizes the need for standardized methods tailored to specific cell types or phenotypes. Such standardization would improve result consistency and enhance the reliability of DNA methylation analysis, particularly for diseases that lacks global methylation changes.

Introduction

1. Importance of DNA Methylation

DNA methylation plays a pivotal role in epigenetics. Environmental factors and disorders can contribute to alterations in methylation levels over the DNA strand, more specifically on cytosines molecules. This alteration can change phenotype without changing a single nucleotide. In general, a typical mechanism to explain the effect methylation in functional biology is the down regulation (silencing) of genes as a result of covalently bonded – methyl groups with cytosines (Figure 1), especially when binding occurs at the promoter site.



DNA Methylation ➔ Inactivation of Genes

Figure 1. An example of gene inactivation resulting from epigenetic modification.

The alteration of methyl levels over cytosines are not limited to hypermethylation but also representing in hypomethylation. For example, hypomethylation of tandem repeats contribute to carcinogenesis and chromosomal rearrangements (Choi et al., 2009). While these alterations are tissue-specific, scientists often seek associations between different tissues. This is crucial for leveraging feasible tissues, such as blood or buccal cells, to identify reliable markers (e.g., differential genes or regions) associated with disorders affecting less accessible tissues, such as those in neuro disorders. However, the scope of DNA methylation is not limited to disorders, but also extends to environmental factors and overall quality of life. Studies have investigated sleep, stress, diet, exercise and other factors to check if a factor can contribute or prevent certain disorder. Therefore, most of the studies follows case-control study design.

2. DMRs

When studying a case versus control group in terms of methylation levels, the researcher aims to find differential methylated regions (DMRs) or differentially methylated probes (DMPs) in case group that expresses either hypermethylation or hypomethylation compared to control group. If a differentiation is detected, these regions undergo gene enrichment analysis and other mapping steps in order to interpret the results and extract meaningful findings that can relate to the differences between phenotypes (case versus control). DMRs can be expressed as global changes in methylation levels as it is the case for cancers, or as a localized alteration within specific part of DNA such as it is the case with behavioural disorders.

3. Sequencing Technique (BS-Seq)

A frequent occurrence of Cytosine and Guanine over a part of DNA is called CpG site (p stands for phosphate which represents the phosphodiester between C and G), and clusters of these CpG sites are called CpG islands (Takeshima & Ushijima, 2018). These islands are often located near the promoter regions in 40% - 50% of human genes and therefore plays an important regulatory role (Juo et al., 2014, Elango & Yi, 2011).

To detect DMRs, a differentiation between methylated cytosine and unmethylated cytosine needs to take place. Therefore, a technique known as bisulfite sequencing (BS-Seq) is used to add sodium bisulfite to DNA sample to convert unmethylated cytosines into Uracil. Methylated cytosines on the other hand remains unconverted. This conversion allows for detecting methylation patterns on single base pair resolution (Bibb et al., 2017). BS-Seq technique is considered gold standard in DNA methylation studies. After conversion of unmethylated cytosines into Uracil, PCR is taking place and the Uracil eventually converted to Thymine. It is important to understand that the original Thymine (T) can be distinguished from the Cytosine-converted Thymine through comparing the untreated DNA with the treated DNA.

4. Platforms and Application

After bisulfite conversion, scanners with fluorescence technology are used to detect the methylated and unmethylated cytosines. In general, methylation micro arrays have gained popularity due to their cost and time efficiency compared to Whole-Genome Bisulfite Sequencing (WGBS). Majority of experiments are performed using Illumina Infinium platforms (Table 1). Examples of other less common platforms are Agilent arrays, ex: Agilent-023795 Human DNA Methylation Microarray 244k (platfrom id: GPL10878). There are also custom platforms which are built to target specific regions based on the study interest, ex: UHN Microarray Centre Human 8.1K CpG island microarray (Platform id: GPL10342). Among the previous platforms, Infinium by Illumina, specifically 450K and EPIC (v1.0) are the most common. In this study, the literature focuses on human methylation using Infinium by Illumina (specifically 450K and EPIC v1.0) unless stated otherwise. The numbers (27K, 244K, 450K, etc..) represents the number of probes used in each platform. And of course, higher number of probes corresponds to wider coverage.

Platform	Marker Count	Release Year	Reference
Infinium Human Methylation 27K BeadChip	~ 27K markers	2008	www.illumina.com, HumanMethylation27 product support files
Infinium Human Methylation 450K BeadChip	~ 450K markers	2011	www.illumina.com, Infinium HumanMethylation450K V1.2 Product Files, n.d.
Infinium MethylationEPIC v1.0	~ 850K markers	2016	www.illumina.com, Infinium MethylationEPIC v1.0 product Files, n.d.
Infinium MethylationEPIC v2.0	~ 930K markers	2024	www.illumina.com, Infinium MethylationEPIC v2.0 Product Files

Table 1. Infinium platforms

The widest coverage among current platforms is provided by Illumina Infinium Methylation EPIC v2.0 which is recently released (in 2024) and therefore it is still not quite available compared to EPIC v1.0. Therefore, in this study, all EPIC platforms refer to EPIC v1.0 version (850K) unless stated otherwise.

5. Signal Reads (IDAT files)

The output of scanners represents in methylation signal (M) and unmethylated signal for each probe. In Illumina platforms, these signals are directly stored in IDAT format. The output of methylation array experiment is two IDAT files per sample (2 IDATs for each individual), one for the green channel that measures the methylated signal (*sampleID_Grn.idat*), and the other for red channel which measures the unmethylated signal (*sampleID_Red.idat*) (Introduction to DNA Methylation Analysis — methylprep 1.6.5 documentation, n.d). Therefore, if an experiment includes samples from 8 individuals, the output of the experiment would be 16 IDAT files. Fig. 2 shows how methylation signals look like after processing pair of IDAT files for one sample. As shown in Fig. 2, each probe has 4 columns that represents methylation signals:

1- MG ➔ Methylated signal from the green channel (retrieved from _Grn.idat)
2- UG ➔ Unmethylated signal from the green channel (retrieved from _Grn.idat)
3- MR ➔ Methylated signal from the red cannel (retrieved from _Red.idat)
4- UR ➔ Unmethylated signal from the red channel (retrieved from _Red.idat)

6. Beta Values

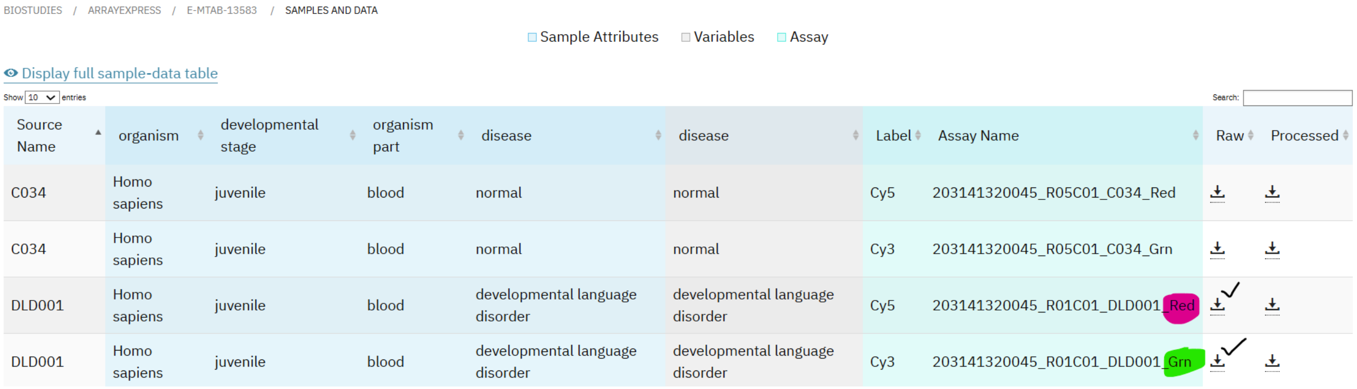
For easier statistical analysis, methylation intensities are converted into either:

- <u>M values</u> ➔ ranges between -1 and +1 where 0 means the probe is 50% methylated.

Or:

- <u>Beta Values</u> ➔ ranges between 0 – 1 where 0 means fully hypomethylated, and 1 means fully hypermethylated.

1- Example of downloading random pair of IDATs from a repository:



BIOSTUDIES / ARRAYEXPRESS / E-MTAB-13583 / SAMPLES AND DATA

□ Sample Attributes □ Variables □ Assay

Display full sample-data table

Show 10 entries Search:

Source Name	organism	developmental stage	organism part	disease	disease	Label	Assay Name	Raw	Processed
C034	Homo sapiens	juvenile	blood	normal	normal	Cy5	203141320045_R05C01_C034_Red	⤓	⤓
C034	Homo sapiens	juvenile	blood	normal	normal	Cy3	203141320045_R05C01_C034_Grn	⤓	⤓
DLD001	Homo sapiens	juvenile	blood	developmental language disorder	developmental language disorder	Cy5	203141320045_R01C01_DLD001_Red	⤓ ✓	⤓
DLD001	Homo sapiens	juvenile	blood	developmental language disorder	developmental language disorder	Cy3	203141320045_R01C01_DLD001_Grn	⤓ ✓	⤓

2- Locate the path of downloaded IDATs and read it (R Studio):

```
library(sesame)
Folder <- "~/Target/Sesame/Sample"
idat_files <- searchIDATprefixes(Folder)
idat = readIDATpair(idat_files)
head(idat)
```

3- Output:

```
    Probe_ID MG MR    UG   UR col  mask
1 cg00000029 NA NA  2890 1685   2 FALSE
2 cg00000103 NA NA  5573  618   2 FALSE
3 cg00000109 NA NA  5327  664   2 FALSE
4 cg00000155 NA NA  8023  837   2 FALSE
5 cg00000158 NA NA 10422 1165   2 FALSE
6 cg00000165 NA NA  1155 8043   2 FALSE
>
```

Figure 2. A simple example to download and read IDATs

In general, beta values are more commonly used due to their ease of interpretation and intuitive biological meaning. However, M values offer greater statistical validity (Du et al., 2010). Du et al. (2010) offers a complete guide to compare between both methods. Fig. 3 demonstrate the calculation process for each method, where M represents the maximum methylation signal, and U represents the maximum unmethylated signal detected.

$\beta = \frac{M}{(M + U + \alpha)}$	$Mval = \log_2(\frac{(M + \alpha)}{(U + \alpha)})$

Figure 3. Calculation of Beta, & M values. Typically, α is a constant which is set to 100 for β and 1 for Mval.

The terminology (Beta) is derived from the distribution curve which is similar to beta distribution (Du et al., 2010). This is because by nature, majority of CpG sites are either hyper methylated (betas are close to 1) or hypo methylated (betas close to 0) (Fig. 4).

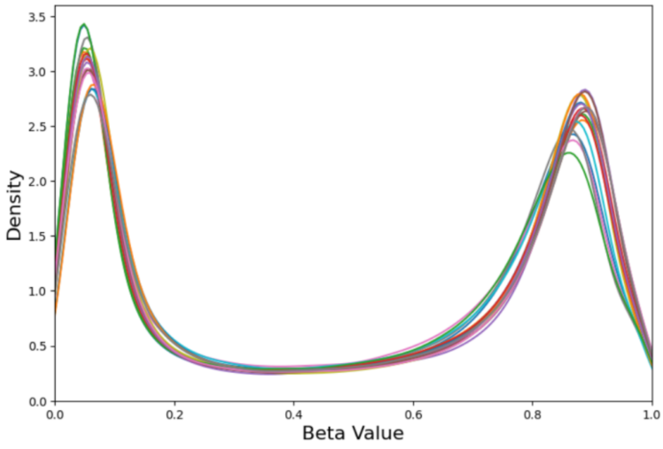


Figure 4. Multiple curves correspond to multiple samples. Platform used: Epic Array

7. Motive of the study

- Literature abundance

While DNA methylation research is abundant in cancer studies, psychiatric and behavioural disorders have received comparatively less attention. However, the increasing observations of shared epigenetic markers among psychiatric disorders has sparked interest in recent studies.

- Ambiguity of risk factors

The mechanism for many psychiatric disorders is still unknown. As a result, no specific biomarkers are available to monitor the risk of developing the disease, which makes complete prevention is unfeasible. On the other hand, investigating the role of environmental factors in developing or even reversing such conditions is important, as this approach can reduce reliance on medical intervention and ultimately avoid the side effects often associated with prolonged use of medication.

a. Increasing rate of developmental/ behavioural disorders among children

This research is further motivated by the rising rates of behavioural disorders like ADHD, and Autism among children, especially in the recent years. The developmental challenges in children are not limited to the existence of well-defined disorder, but also extends to general developmental delays that often overlap with each other’s or with other psychiatric conditions (ex: language and learning delays, social anxiety, depression, attention deficit, sleep deprivation etc..). While some conditions may improve as children grow older, their impact on schooling and social life can persist into adulthood, potentially lowering overall quality of life.

d. Complexity of behavioural disorders

The overlapping symptoms among behavioural disorders, especially those that occur during developmental ages in children, pose challenges in diagnostic accuracy. As a result, one disorder can be confused with another, particularly when symptoms are unclear or do not appear persistently.

8. Study Objective

To investigate possible epigenetic markers that may play a role in developing of certain behavioural disorders.

Chapter 1 | Literature Review

The literature review for this study can be divided into 4 sections:

1- Behavioural Disorders:
A review of behavioural disorders in terms environmental risk factors and overlapping symptoms.

2- Analysis Workflow:
Exploring the latest recommendations in terms of preprocessing methods and differential analysis.

3- Comparison of Pipelines
As an exploratory approach, comparing the resulted beta distribution curve among different R Bioconductor packages to confirm that different methods have profound effects on the results.

4- Comparison of Detection p value Calculation Methods
Four different methods are tested to check how many values are masked in each method.

The study intends to compare DMRs among multiple experiments and explore possible shared DMRs among certain phenotypes. Therefore, the majority part of the study is technical and involves statistical applications. There are numerous methods and tools to choose when performing micro array data analysis. Therefore, it was crucial to select methods that are up to date. Another challenge was the un abundance of one specific workflow of which analyse and compare multiple experiments in the domain of DNA methylation. As a result, a comprehensive review of the latest guides and protocols is carried to select the most appropriate tools that best serves our study design. Confounding factors like sex, age, race, lab conditions, can all contribute unreliability of the results. Therefore, in order to make sure that the retrieved results are related to biological differences rather than confounding factors, a conservative approach was chosen in every step of the analysis.

Several studies have concluded that using different preprocessing methods can result in significant effects on downstream analysis (Marabita et al., 2013). As a result, part of our literature review was dedicated to review the documentation of common Bioconductor libraries, apply the recommended pipeline by each library, and finally compare the beta value distribution in each one as an exploratory procedure to observe the differences on overall beta distribution. On the other hand, to evaluate the potential benefits and validity of comparing epigenetic markers across psychiatric conditions, it was also necessary to review the existing literature on these conditions. Furthermore, it was also necessary to review the relationship between environmental factors and psychiatric conditions.

1- Behavioural disorders

The interest in connecting behavioural disorders with environmental factors is not novel. For example, a review by Cassoff, J., et al. (2012) highlighted several studies suggesting potential associations between ADHD and sleep deprivation or general sleep disturbances (Cassoff, J., et al. 2012). Studies also demonstrates that effects of having inadequate sleep in childhood are not limited only to be associated solely with ADHD, but more importantly with general conditions that represents in overall cognitive function and academic performance (O'Callaghan et al., 2010). These symptoms are often observed in other disorders like ASD. For example, children with ASD often struggles in focussing on things they don't like, and also expresses impaired reasoning ability in problem solving. Other findings were presented in a study done by Van Der Heijden, K. B. et al. (2005), which highlights the effects of maladjusted circadian rhythms on children that somehow mimics ADHD symptoms such as, late nighttime, daytime fatigue, and consequently sleep disturbances (Van Der Heijden et al., 2005). On the other hand, according to National Institute of General Medical Sciences (NIGMS), circadian genes itself can be triggered by food intake, stress, and social environment (National Institute of General Medical Sciences [NIGMS], n.d.). A recent study by Han, Y. et al. (2024) has pointed that low protein diet altered peripheral clock regulation. Compared to typical developing children, children with autism (ASD) on the other hand has shown higher frequency of association with other psychiatric comorbidities like mood disorder, anxiety, depression, and even ADHD (Gurney et al., 2006, Magnuson & Constantino, 2011). Unlike ASD and ADHD disorders, anxiety and depressive symptomology was easier to correlate with environmental factors. For example, the increasing rates of depression and anxiety among US population was obvious in the period of COVID pandemic Fig. 5 (Vahratian et al., 2021). The increased prevalence of anxiety disorder during COVID pandemic was further validated in the meta-analysis conducted by Delpino, F. M. et al. (2022).

On the other hand, the overlap of symptoms among various psychiatric disorders is quite common (Bourque et al., 2024). Alomari. N. A., et al. (2022) presented several psychiatric conditions that overlap with social anxiety disorder, which often poses challenges in diagnosis. For example, the differential diagnosis of PTSD (post traumatic disorder) is very difficult as its symptoms overlap with other anxiety and mood disorders (Alomari et al., 2022). One of the recent systematic reviews has investigated the genetic and phenotypic similarities among the major psychiatric disorders (Schizophrenia, Bipolar Disorder, Major Depressive Disorder, Autism Spectrum Disorder, and Attention Deficit Hyperactivity Disorder) (Bourque et al., 2024). The review has included significant findings related to the heritability of the previously mentioned disorders, but more importantly that nearly 75% of-the significant genetic loci where shared by at least two disorders (Bourque et al., 2024, Polderman et al., 2015, Anttila et al., 2018).

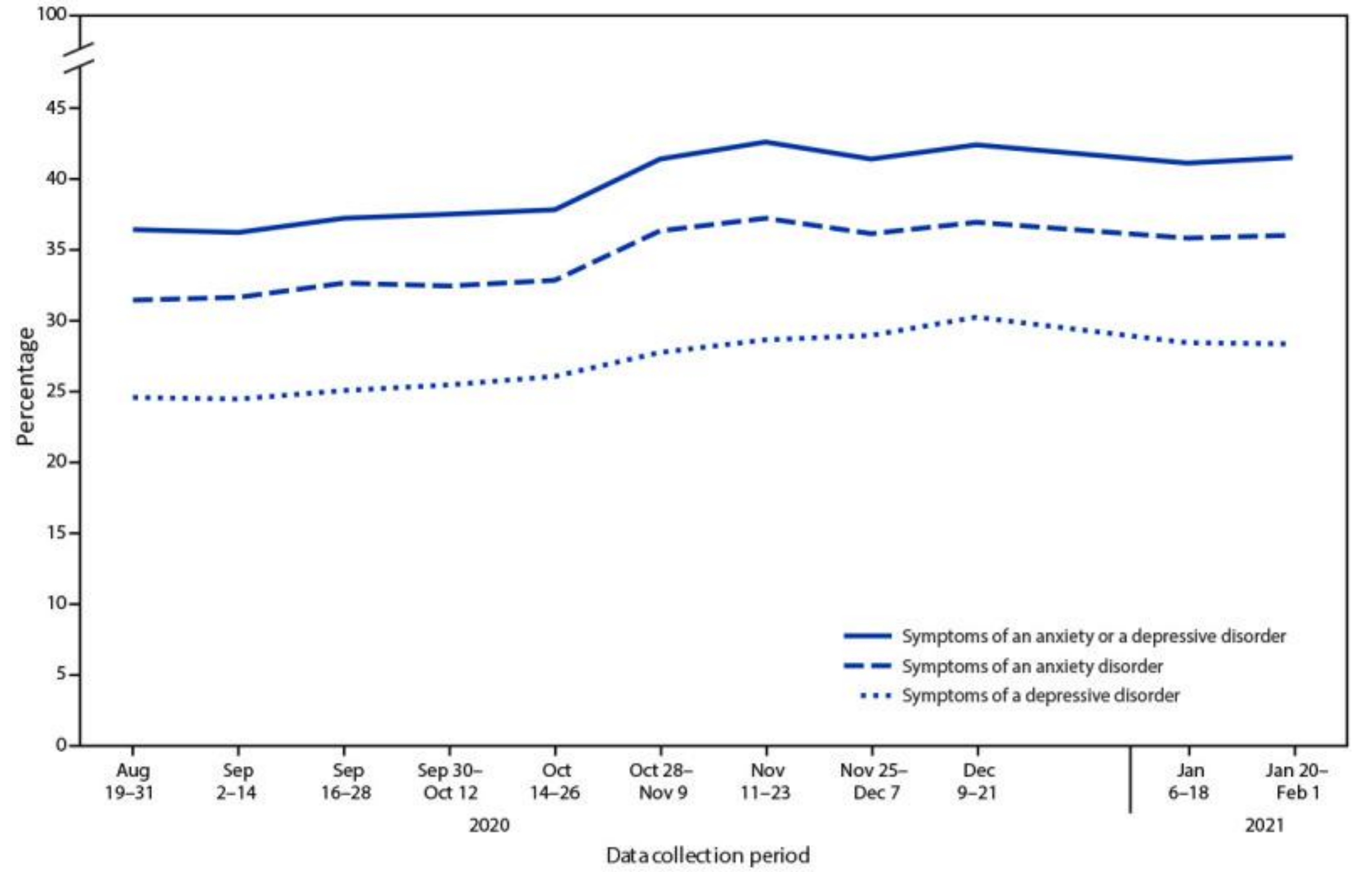


Figure 5. Increasing rate of depressive and anxiety disorders during COVID pandemic during 2020. Image from Morbidity and Mortality Weekly Report (MMWR; Vahratian et al., 2021).

2- Analysis Workflow

Sahoo, K., and Sundararajan, V. (2024) conducted the most recent comprehensive review on DNA methylation analysis methods. the review not only outlines the steps in a general DNA methylation analysis workflow but also evaluates commonly used methods at each stage. Notably, it provides a comparison for different libraries and tools used to detect DMRs. Following data collection, preprocessing raw IDAT files is identified as the initial step in the analysis workflow. Sahoo and Sundararajan (2024) highlights some common quality control procedures:

a. Filtering probes.

Ex: probes with P val > 0.05, probes with many low-quality samples, SNPs, probe with cross hybridization potential.

b. Quality control that includes background subtraction and filtering outliers.

c. Batch correction and FDR correction.

The other fundamental part of the workflow is sample normalization. According to the review, here are some common preprocessing algorithms used for sample normalization, along with their corresponding Bioconductor libraries (Table 2):

#	Preprocessing Algorithm	Bioconductor Library	Software
1	Beta mixture quantile normalization (BMIQ)	wateRmelon	R programming
2	Quantile normalization	Limma	R programming
3	Noob (Normal-exponential convolution using out-of-band probe)	Minfi	R programming
4	SQN: Subset-quantile normalization	ENmix	R programming
5	Illumina (genome studio)	NA	Illumina (genome studio)
6	Functional normalization (funNorm)	Minfi	R programming
7	SWAN: Subset-quantile normalization	Minfi	R programming

Table 2. Commonly used algorithms in normalization process

A comparison table is available in supplementary material of Sahoo and Sundararajan (2024) review that provides general information for different algorithms. There are many recommendations about which normalization method to choose (Sahoo & Sundararajan, 2024, Wang et al., 2018). For example, *FunNorm*() is often recommended in case of global methylation changes (Cancer versus Normal) (Fortin et al., 2014, K. D. Hansen & Fortin, Minfi User Guide). *PreprocessQuantile()* is the opposite where global changes are not expected (K. D. Hansen & Fortin, Minfi User Guide).

The available literature also highlights the advantages and disadvantages for different methods. For example, Illumina did not recommend quantile and loess normalization methods as it can remove biological signal (www.illumina.com ,A Patient-Centric Methylation Pipeline). Notably, Quantile-based methods are reported to be the worst in Welsh et al. (2023) study, which performs a systematic evaluation of normalization methods specifically on EPIC arrays. Welsh et al. (2023) stated that the *SeSAME* pipeline was the best among the investigated methods. Figure 6 from the same study clearly shows the variance between replicates for each method. According to Figure 6, *NOOB*, *NOOB+BMIQ*, and the *SeSAME* pipeline had the best results.

Interestingly, the standard pipeline in *SeSAME* uses normalization exponential (Norm-Exp) deconvolution parametrized by out-of-band probes. In simple terms, this method is similar to *NOOB* normalization.

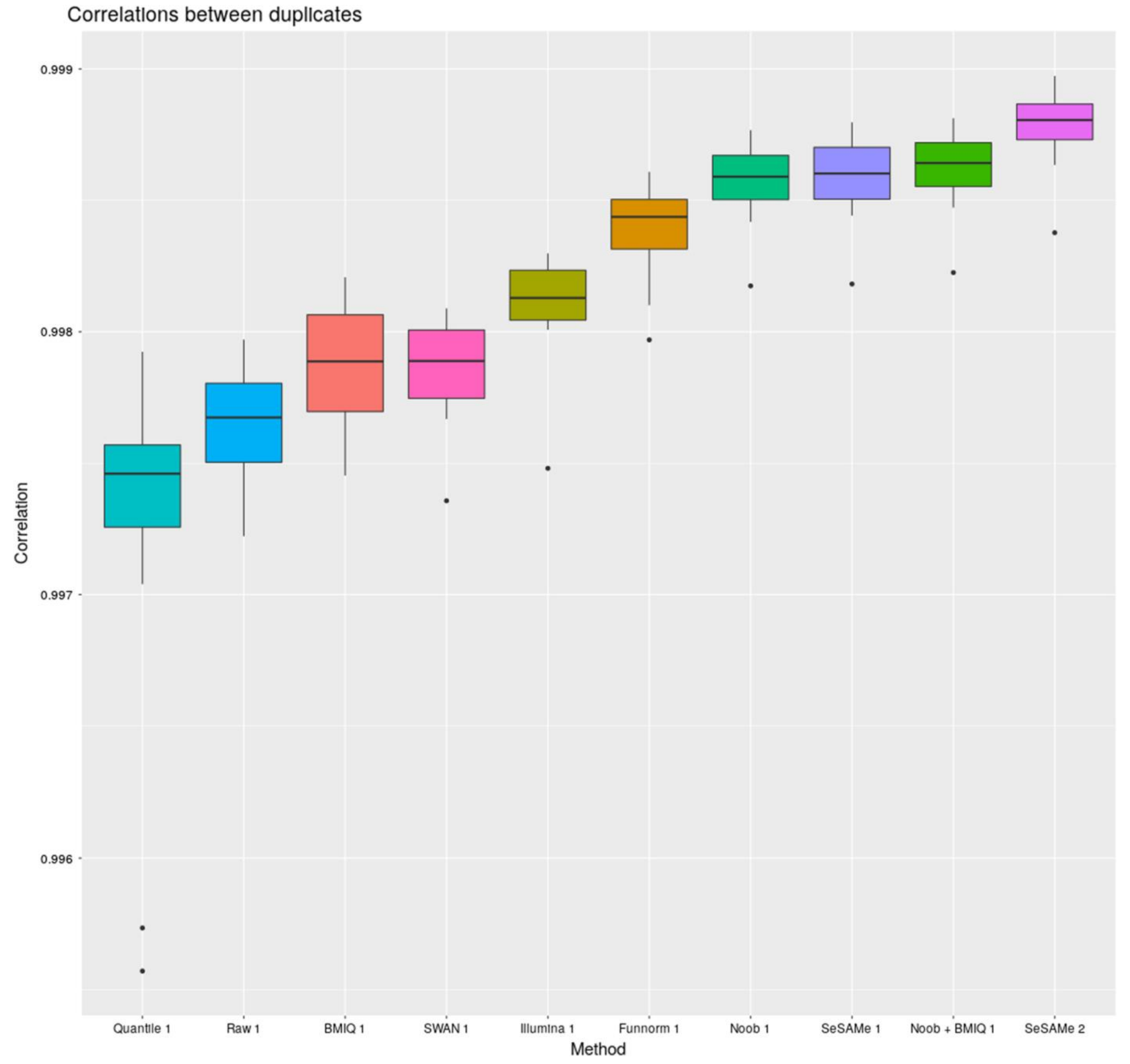


Figure 6. A comparison among different preprocessing methods. Image from (Welsh et al., 2023).

For the calculation of detection p-values, it is interesting to note that the *pOOBAH* method, originally provided by *SeSAME*, is the most up-to-date method. This method uses out-of-band probes to substantially remove technical variation while preserving biological variation (Zhou et al., 2018). Interestingly, *pOOBAH* is now also available within the *Rnbeads* package (RNBeads Reference Manual, 2024).

Furthermore, using *pOOBAH* twice in Welsh et al. (2023) comparison achieved the highest correlation among replicates compared to other methods. Based on the available information, it appears that several independent studies agree on the superiority of the *pOOBAH* method and the *SeSAME* pipeline in general. Another parameter that differs among different Bioconductor packages is the method used for dye bias correction. For example, similar to SeSAME's novel method (*pOOBAH*), the *Enmix* standard pipeline uses a novel dye bias correction method called *RELIC*, compared to the traditional methods used in other libraries (e.g., *Minfi* and *SeSAME*, which use non-linear dye bias correction). Xu et al. (2017) demonstrated the advantage of using *RELIC* compared to other methods. However, unlike *SeSAME*, we could not find additional papers that further support this novel method, and it seems that it is still not quite common among researchers.

When using the standard pipeline provided by certain packages, there are several common parameters considered to improve data quality. Table 3 lists the most important parameters that researchers need to know how to use. Table 4 lists two examples of pipelines from different packages.

Parameter	**Description**
Samples threshold	Removes samples with low-quality probes count greater than the chosen threshold.
Probe threshold	Removes probes with low-quality methylation values count (samples) greater than the chosen threshold.
Imputation of missing/unreliable values	An optional argument usually set as True or False. If true, the function will replace masked (or missing) values with various imputation methods (mean average, k nearest neighbour, using machine learning, etc.).
Outliers' detection	Outliers are detected and may or may not be replaced by other values.
P value threshold	The researcher has the option to set it to 0.05, 0.01, etc.

Table 3. Common parameters to be set by the researcher. (K. D. Hansen & Fortin, Minfi User Guide, SeSAME User Guide, 2024, Enmix User Guide, 2024, Sahoo & Sundararajan, 2024).

Package	Example of Pipeline
Minfi reference manual	*preprocessQuantile*(data, fixOutliers = **TRUE**, removeBadSamples = **TRUE**, badSampleCutoff = **0.5**, quantileNormalize = **TRUE**, stratified = **TRUE**)
SeSAME reference manual	*openSesame*(x, prep = "**QCDPB**", func = **getBetas**, min_beads = **1**) *

Table 4. Examples of Standard pipelines with parameters distinguished in bold. (K. D. Hansen & Fortin, [Minfi User Guide], (SeSAME User Guide, 2024)

* Choosing "QCDPB" parameter in SeSAME makes the function works as a wrapper for NOOB normalization + nonlinear dye bias correction + pOOBAH masking.

Typically, the output of preprocessing and QC steps is a beta value (or M value) matrix, where the header contains the probe IDs in the first column, followed by the sample IDs (Table 5). The rest of the matrix contains the corresponding beta values for each sample. These values are often calculated within the preprocessing step using the methylated (M) and unmethylated (U) signal intensities.

ProbeID	FTP_MAPT_203282450164_R07C01_Betas	FTP_MAPT_203282450165_R06C01_Betas	FTP_MAPT_203282450206_R06C01_Betas
cg00000321	0.93185736	0.452552278	0.170955216
cg00000363	0.956787254	0.245961765	0.195561541
cg00000540	0.595062545	0.814777269	0.957632846
cg00000596	0.035264643	0.407019545	0.49373275
cg00000776	0.093778138	0.236755558	0.430315058
cg00001099	0.737510528	0.530688509	0.851287559

Table 5. Example of beta value file (first 6 probes)

The matrix may or may not contain missing values, depending on the pipeline and parameters chosen. For p-values, packages like *SeSAME*, provide the ability to extract them if the researcher chooses to, which is not the case for other packages like *Minfi*, where p-values are calculated without the ability to convert them into a data frame.

As a result, p-values cannot be obtained as a standalone dataset after preprocessing. Once preprocessing and QC is done, beta values can be analysed for DMRs. For DMRs detection, *Limma* package is considered among the packages that prove its effectiveness (Sahoo & Sundararajan, 2024). The algorithm uses empirical Bayes approach.

3- Comparison of pipelines

To confirm the differences of different preprocessing methods, several attempts were made on sample ID 203141320045_R04C01_DLD004 from E-MTAB-13583 Experiment (BioStudies, n.d.). E-MTAB-13583 raw and processed methylation data is publicly available on ArrayExpress. As per the meta data provided with the experiment, the processed data is obtained after several QC steps. This was followed with normalization using *SWAN* method. In addition to raw and SWAN processed datasets provided by the experiment, we have selected the following methods to process the same file and compare accordingly (Table 6):

Package	Pipeline used	Algorithm / parameters remarks
Minfi	preprocessQuantile(data, fixOutliers = TRUE, removeBadSamples = TRUE, badSampleCutoff = 0.5, quantileNormalize = TRUE, stratified = TRUE)	Stratified quantile normalization for an Illumina methylation array.
	preprocessSWAN()	Subset-quantile Within Array Normalisation (standard pipeline from Minfi)
Enmix	mpreprocess(data, nCores=2, bgParaEst="oob", dyeCorr="RELIC", qc=TRUE, qnorm=TRUE, qmethod="quantile1", fqcfilter=FALSE, rmcr=FALSE, impute=TRUE)	RELIC is used for dye correction. Background correction + Quantile normalization method

Table 6. Pipelines used in comparison against the processed and raw data published with E-MTAB-13583 experiment.

Attempting to mimic the distribution curve in the pre-processed data provided with the experiment, we used the same algorithm (*preprocessSWAN*). Interestingly, the comparison results in different beta distributions, most likely due to different QC parameters (Figure 7). All comparisons are available in the Appendices Chapter (*Section 1 , Pipelines Comparison*).

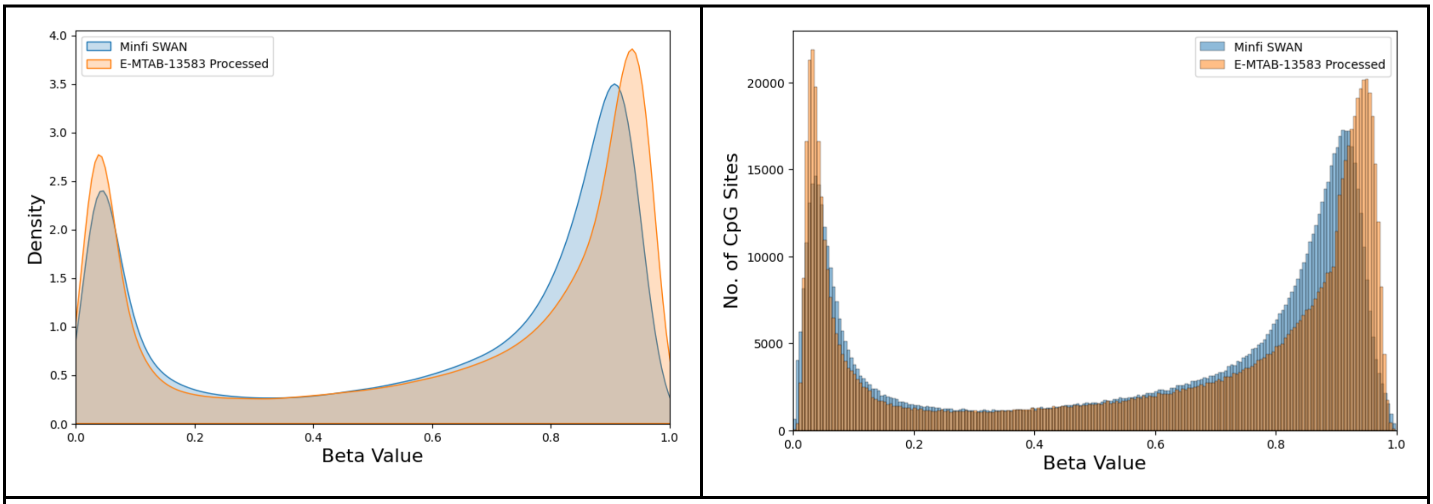


Fig 7. A comparison between preprocessSWAN() standard approach as per Minfi reference manual versus the preprocessSWAN() pipeline used in E-MTAB-13583 experiment.

4- Comparison of detection p value calculation methods

This section is dedicated to exploring the differences associated with using pOOBAH method which is originally provided by *SeSAME* package. Table 7 represent a comparison in terms of the number of masked probes based on p value 0.05 using 4 different methods. The same sample has different of number of failed probes in each method, which highlights the profound effects that can results from different methods. The source code used to output the Table 7 is available in the Appendices Chapter (*Section 2, P-value Methods Comparison*).

Method	R01C01_DLD001	R02C01_DLD002	R03C01_DLD003	#	#	R06C01_V183	R07C01_V187	R08C01_V188	Average
'pOOBAH' Method by SeSAME	19008	9975	9716	//	//	10958	8560	20587	11666.21
'detectionP (M+U)' Method by Minfi	617	345	282	//	//	508	261	530	465.9167
'oob' Method by ENmix	31538	12168	12251	//	//	11563	7366	13482	12256.17
'negative' Method by ENmix	514	263	214	//	//	428	216	424	383.1667

Table 7. A comparison of calculating dectection p value methods. P value threshold set to 0.05 in all methods across samples from the experiment E-MTAB-13583.

Chapter 2 | Methodology

1. Data Collection:

Data is collected from GEO and ArrayExpress. Experiments have been selected using the following keywords: [brain, behavior, behaviour, child, psychiat, adhd, asd, attention, autism, impulsive, sleep, stress, adversity, developmental, language, abuse]. To limit the confounding variables especially in terms of tissues, and make the included experiments compatible for cross comparison, the search was restricted to the following parameters:

- Tissue: blood, or peripheral blood
- Species: Homo Sapiens
- Study type: Methylation Profiling by Array.

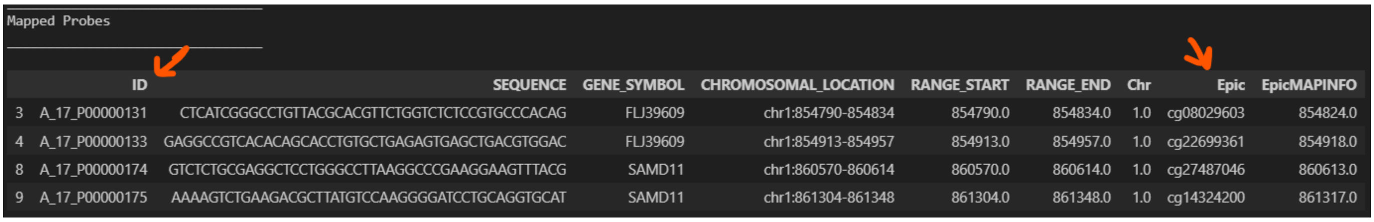
2. Platform Compatibility

Since this study attempts to compare methylation levels across different experiments, it was important to evaluate the compatibility of platforms that are provided by different manufacturers before including the corresponding experiments. Upon checking the manifests for different platforms (ex: Illumina, Agilent, etc..), it was observed that important information must be considered when comparing different platforms.

a. All Illumina Infinium platforms (27K, 450K, EPIC) uses the same length of probes (50 bp per probe).

b. A considerable number of probes are shared among different Infinium platforms. Furthermore, these probes correspond to the same genomic locations.

c. Probes in platforms manufactured by other providers (Ex: Agilent), has different IDs. Therefore, an attempt was done to map these ids to its genomic locations using the manifest provided by the manufacturer. These locations are then compared to Illumina probes in order to find matched probes which can be extracted and included in the analysis (Figure 8).



Mapped Probes

	ID	SEQUENCE	GENE_SYMBOL	CHROMOSOMAL_LOCATION	RANGE_START	RANGE_END	Chr	Epic	EpicMAPINFO
3	A_17_P00000131	CTCATCGGGCCTGTTACGCACGTTCTGGTCTCTCCGTGCCCACAG	FLJ39609	chr1:854790-854834	854790.0	854834.0	1.0	cg08029603	854824.0
4	A_17_P00000133	GAGGCCGTCACACAGCACCTGTGCTGAGAGTGAGCTGACGTGGAC	FLJ39609	chr1:854913-854957	854913.0	854957.0	1.0	cg22699361	854918.0
8	A_17_P00000174	GTCTCTGCGAGGCTCCTGGGCCTTAAGGCCCGAAGGAAGTTTACG	SAMD11	chr1:860570-860614	860570.0	860614.0	1.0	cg27487046	860613.0
9	A_17_P00000175	AAAAGTCTGAAGACGCTTATGTCCAAGGGGATCCTGCAGGTGCAT	SAMD11	chr1:861304-861348	861304.0	861348.0	1.0	cg14324200	861317.0

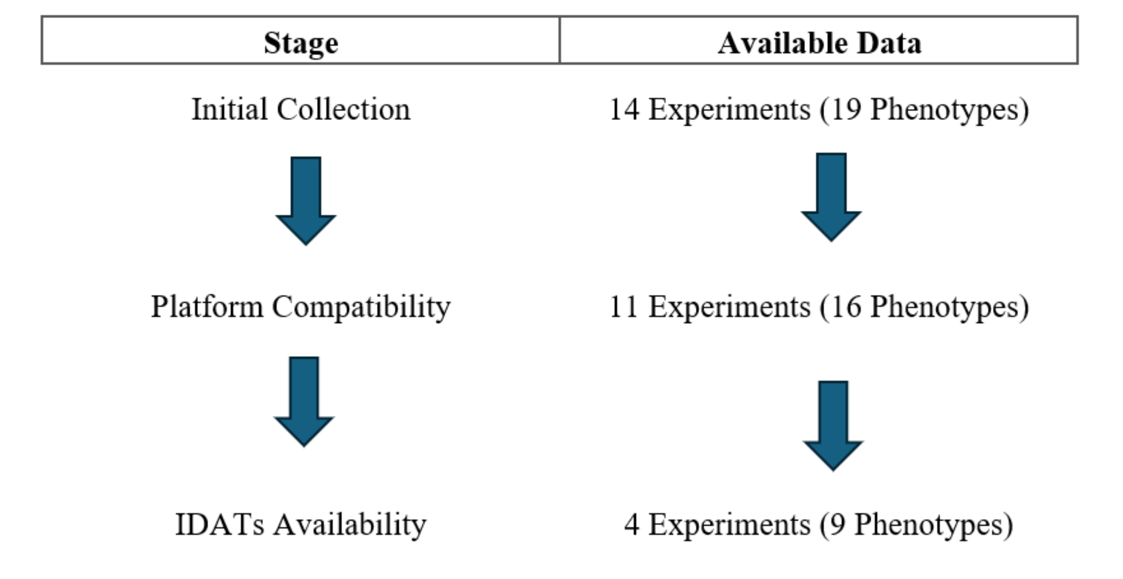
Figure 8. Attempting to map Agilent 244K probes to Illumina EPIC probes

The mapping procedure leveraged the information provided by the *MAPINFO* column from the EPIC manifest (www.illumina.com), which was consistently found to be within the range of the start and end coordinates of the corresponding probe. This information was then compared with the start and end coordinates of Agilent probes (*GEO Accession Viewer*, n.d. Platform ID GPL10878). Despite approximately 50K out of the 244K probes on the Agilent platform having 'somewhat' similar genomic regions to EPIC probes, it is important to note that the length of the probes differs (*SEQUENCE* column), with EPIC probes being 50 bp in length, while Agilent probes can range up to 200 bp.

3. Exclusion criteria

Initially, 14 studies were selected, containing methylation arrays for 19 different phenotypes. The metadata for all experiments were reviewed to verify the platform specifications and determine compatibility for cross-comparison. The majority of the experiments utilized Illumina Infinium bead chips (27K, 450K, EPIC v1.0). Therefore, it is preferable to restrict the platform of choice to Infinium platforms only. The rationale for excluding experiments from different platforms was discussed in the previous section. In brief, probes differ in length (50 bp in Illumina versus 40–200 bp in Agilent), and the genomic loci of similar probes do not always match (differences in start and end coordinates).

Another exclusion criterion was the availability of raw IDAT files, as some experiments only provide processed data, which does not align with our approach. Our method relies on the availability of original raw IDAT files to ensure consistent preprocessing using our chosen pipeline. As a result, 4 experiments remained, encompassing a total of 9 phenotypes (i.e., 9 datasets) (Figure 9). Table 8 summarizes the experiment details, and the available phenotypes considered for this study.



Stage	Available Data
Initial Collection	14 Experiments (19 Phenotypes)
Platform Compatibility	11 Experiments (16 Phenotypes)
IDATs Availability	4 Experiments (9 Phenotypes)

Figure 9. Exclusion process.

#	Array / Study	ID and Link	Data Source	Platform	Sample Size	Phenotype / Environmental Factor
1	Comparison of the methylation profiles of children with developmental language disorder and healthy control subjects	E-MTAB-13583	ArrayExpress	Illumina EPIC v1.0	12	**Developmental Language Disorder**
2	Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-wide DNA Methylation Profiles and RNA Expression	E-MTAB-4664	ArrayExpress	Illumina 450K	18	**Acute Sleep Deprivation**
3	DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity	GSE164056	GEO	Illumina EPIC v1.0	35	**Social Anxiety**
4	DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity	GSE164056	GEO	Illumina EPIC v1.0	30	**Ealry Life Adversity**
5	DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity	GSE164056	GEO	Illumina EPIC v1.0	31	**Social Anxiety & Ealry Life Adversity**

6	Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines	E-MTAB-11975	ArrayExpress	Illumina EPIC v1.0	5	**sporadic early-onset Alzheimer's disease**
7	Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines	E-MTAB-11975	ArrayExpress	Illumina EPIC v1.0	6	**Familial Alzheimer Disease**
8	Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines	E-MTAB-11975	ArrayExpress	Illumina EPIC v1.0	5	**Genetic Frontotemporal Dementia (GRN Mutation)**
9	Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal	E-MTAB-11975	ArrayExpress	Illumina EPIC v1.0	5	**Genetic Frontotemporal Dementia (MAPT Mutation)**

	dementia in brain tissue and lymphoblastoid cell lines					
10	Comparison of the methylation profiles of children with developmental language disorder and healthy control subjects	E-MTAB-13583	ArrayExpress	Illumina EPIC v1.0	12	**Healthy Controls**
11	Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-wide DNA Methylation Profiles and RNA Expression	E-MTAB-4664	ArrayExpress	Illumina 450K	18	**Healthy Controls**
12	DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity	GSE164056	GEO	Illumina EPIC v1.0	47	**Healthy Controls**
13	Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer's disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines	E-MTAB-11975	ArrayExpress	Illumina EPIC v1.0	5	**Healthy Controls**

Table 8. List of experiments (4) and available phenotypes (9 cases + 4 controls = Total of 13 datasets) that are included in the workflow.

4. Workflow Steps and Approach

Based on the literature reviewed (Chapter 1), the data preprocessing protocol significantly influences the results, particularly when the investigated phenotypes do not exhibit global changes in DNA methylation. Therefore, it was crucial to select a comprehensive and consistent protocol that could be applied across all included experiments.

4.1. Preprocessing (QC part 1)

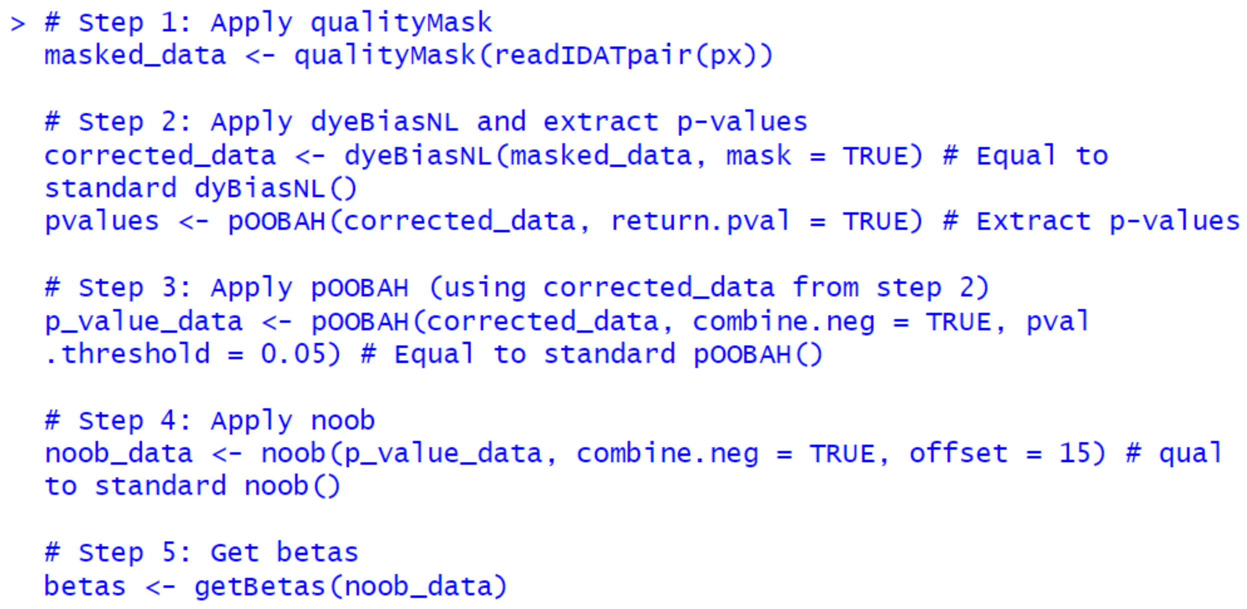
Purpose:	1-	Eliminate noise from artifact effects
	2-	Mask weak and unreliable signals
	3-	Extract masking summary (metric-wise)
	4-	Convert IDATs into betas matrix

Our package of choice was *SeSAME* for the following reasons:

- None of the included datasets has used *SeSAME*.
- Relatively new (released in 2018) and uses up to date methods (*pOOBAH* for detection p-value).
- Researchers have more control over sample/ probe exclusions.
- Conservative and comprehensive quality mask is provided. The masking procedure targets low quality signals (beta values) without removing the probes which give the choice for the researcher to check if certain samples exhibit extra number of low-quality probes.
- The quality mask provided by *SeSAME* target probes with suboptimal hybridization, multimapping, and other features like insignificant p value based on a threshold decided by the researcher, and probes with low bead count.
- P values can be extracted as a standalone dataset and used as a guide for removing bad performing samples during the workflow.
- *SeSAME* is recommended by Illumina (www.illumina.com, Infinium™ Methylation Screening Array).
- Availability of workflows and published papers that are used as a reference.(Zhou et al., 2018, Zhou et al., 2022, Welsh et al., 2023).

The main reason why we preferred to have p values after preprocessing is that we wanted to delay any correction (imputation of missing/ low quality values) after exploring the original dataset and hence we will have the confidence to impute missing/ low quality values based on clear insights.

Furthermore, p values are a used with other metrics to evaluate the overall quality of samples. This is beneficial in the workflow when dataset has lot of outliers, and the researcher wanted to check if certain sample is having distinguished number of outliers and low-quality values. The first step was to process raw IDATs using the recommended approach provided in *SeSAME* reference manual (Figure. 10), in addition to Zhou et al., (2022) study. The complete code used for the 13 datasets (9 cases and 4 controls) is available in the Appendices Chapter (*Section 3, preprocessing IDATs*).



```
> # Step 1: Apply qualityMask
  masked_data <- qualityMask(readIDATpair(px))

  # Step 2: Apply dyeBiasNL and extract p-values
  corrected_data <- dyeBiasNL(masked_data, mask = TRUE) # Equal to
  standard dyBiasNL()
  pvalues <- pOOBAH(corrected_data, return.pval = TRUE) # Extract p-values

  # Step 3: Apply pOOBAH (using corrected_data from step 2)
  p_value_data <- pOOBAH(corrected_data, combine.neg = TRUE, pval
  .threshold = 0.05) # Equal to standard pOOBAH()

  # Step 4: Apply noob
  noob_data <- noob(p_value_data, combine.neg = TRUE, offset = 15) # qual
  to standard noob()

  # Step 5: Get betas
  betas <- getBetas(noob_data)
```

Figure 10. Partial overview of the main steps involved in preprocessing IDATs with Noob and Dye Bias Correction within the SeSAMe pipeline.

After preprocessing, a masking summary is generated to identify the sample with the highest percentage of masked probes. This procedure utilizes built-in functions from the *SeSAME* package. However, we have combined all key parameters into a single function that outputs comprehensive metrics for all samples in one CSV file. This process was applied to all datasets. The complete code is provided in the Appendices Chapter (*Section 4, Masking Summary*). An example of the output is shown in Table 9. A masking summary represents a table that enables us to view some important QC statistics like percentage of masked probes using *pOOBAH* with threshold of pval = 0.05 and pval = 0.01 at the same time. The table also provides an easy way to know the percentage of probes that are masked for reasons other than p value (Table 9).

Procedure	R01C01_ DLD001	R02C01_ DLD002	#	#	R02C01_ DLD011	R03C01_ DLD012
Platform Recognized	EPIC	EPIC	//	//	EPIC	EPIC
No. of Masked Probes in The Raw Sample	0	0	//	//	0	0
Perc. Of Missing Betas in The Raw Sample	0.021128	0.010941	//	//	0.011797	0.010209
No. of Masked Probes After qualityMask()	105454	105454	//	//	105454	105454
Perc. Of Missing Betas After qualityMask()	0.139165	0.130171	//	//	0.130845	0.12959
No. of Masked Probes with dyeBiasNL()	0	0	//	//	0	0
No. of Masked Probes with pOOBAH()	19008	9975	//	//	10758	9375
Perc. Of Missing Betas As a Result of pOOBAH	0.021128	0.010941	//	//	0.011797	0.010209
No. of Masked Probes with noob()	0	0	//	//	0	0
Perc. Of Missing Betas As a result of noob()	0.021128	0.010941	//	//	0.011797	0.010209
Perc. Of Missing Betas After (qualityMask() + dyeBiasNL() + pOOBAH() + noob())	0.139165	0.130171	//	//	0.130845	0.12959
Total Masked probes	124462	115429	//	//	116212	114829

Table 9. Number/ Percentage of masked betas as a result of performing *SeSAME* standard QC mask

4.2. QC (part 2)

Note: SNP probes, control probes, ch probes, and probes on sex chromosomes are removed prior to QC (part 2). The exclusion of these probes is a standard procedure in methylation studies (e.g: Wiegand et al., 2021, Hop et al., 2020, Ramos-Campoy et al., 2024, Illumina "Infinium controls training guide", www.illumina.com) unless the researcher chose not to base on the purpose of the study.

Complete source code for QC (part 2) is available as supplementary material (*QC2*).

Importance of QC (part2):

1- Rank subjects based on outliers and quality metrics
2- Identify best and worst performing subjects
3- Exclude low-quality probes (and low-performing subjects if needed)
4- Isolate probes with extra variability
5- Detect outliers using IQR in isolated probes.
6- Mask the outliers
7- Impute the masked values using WM
8- Visualize the results and variability improvement

Purpose: Reduce artifact effects while preserving biological variability.

To maintain a conservative approach, we decided to perform another round of QC right after *SeSAME* preprocessing. This was important since none of the probes or samples were removed during preprocessing. This idea was inspired by the study 2023, welch et al which performs 2 rounds of QC with *SeSAME*, leveraging *pOOBAH* for improving the reliability if methylation values. The workflow of QC (part 2) is summarized in Table 10 which lists all the used functions step wise from 1-17. Source code is available as supplementary material (*QC2*).

This part of the workflow begins with evaluating the potential removal of low-performing samples. To achieve this, a scoring matrix is created, incorporating several quality metrics. Based on these metrics, subjects are ranked from highest to lowest quality in terms of methylation signal performance (Table 11 is the output of Step 1). Functions (2–7) check for subjects with an unusually high number of values showing the greatest absolute deviation from the mean (probe-wise). This procedure helps determine whether a specific subject contributes significantly to the majority of outliers. Combined with the scoring matrix (Table 11), these metrics facilitate the decision-making process regarding the removal or retention of certain samples (Table 12).

Step	Self-Built Functions (Python)	Process
1	pval_df = subject_score(df)	Creating a scoring matrix (Table 11)
2	abs_dif = df_abs(df1)	Calculation of absolute difference from the mean average
3	plotAbsDifference(abs_dif)	Plot the subjects to check for distinguished number of probes for certain subject
4	abs_rank = abs_dif_rank(abs_dif)	Rank subjects based on absolute deviation
5	subjectsPerformance(abs_rank, pval_df)	Returns a table that shows Max Abs. Diff. Count for each subject compared with its rank in scoring matrix.

6	Top_Scorer = '203259750077_R04C01_DLD013_Betas'
	Bad_Samples = ['203141320045_R01C01_DLD001_Betas'] *# leave it [] in case no bad subjects is determined*
	""" Decide probe_QC threshold Based on Subject Counts """
	TRPQC = (len(df1.columns)-1) * (2/3)
		Initially, set a target subject to have extra weight for imputation with mean average.
		The bas samples (if any) will be stored in list to be removed.
		A threshold to be decided (mostly 2/3 of sample size).

7	df1_removed = df1.drop(columns=[col for col in df1.columns if col in Bad_Samples])	Dropping bad sample(s) if any.
8	df2 = probe_QC(df1_removed, threshold = TRPQC, remove = True)	Removing bad probes (probes that has > TRPQC NaN values
9	qc_table_2 = mask_summary(df2)	Check masking summary after removal of bad probes
10	count_probes_with_range(df2)	Check how many probes have more than 0.3 range

11	`df3, df_remain = extract_probes_with_range(df2, threshold=0.3)`	Extract probes with range > 0.3 to a separate dataframe.

12	`plotAbsDifference(abs_dif_1)`
	
	`abs_rank_1 = abs_dif_rank(abs_dif_1)`
	
	`# Compare Against Pval Scores`
	`subjectsPerformance(abs_rank_1, pval_df)`
		Plot the subjects to check which one has extra number of outliers.
		
		Rank the subjects based on the results.
		
		Check the overall performance for each and decide if certain subjects needs to be removed.

13	`df4 = replace_outliers_withNaN(df3)`	Replace outliers with NaN. (Masking outliers as missing values)

14	`qc_table_3 = mask_summary(df3_updated)`
	`qc_table_4 = mask_summary(df4)`
	
	`outlier_inspection_table = compare2qc_tables(qc_table_3, qc_table_4)`
		Check how much beta values are masked after outlier detection

15	`df5 = append_masked_to_original(df4, df_remain_updated)`	Rejoin the isolated probes to the remaining datframe.

16	`df6 = probe_QC(df5, threshold=TRPQC, remove= True)`
	`  # Set remove = True to Remove Bad Probes`
		Remove any probes that exceeds the threshold TRPQC.

17	`df7 = impute_WM(df6, target_col=Top_Scorer, target_weight=Target_Weight, default_weight=1)`	Impute the remaining missing (masked) beta values

Table 10. All functions used in QC (part 2). The code for each function is available as Supplementary material (*DNA_Meth_Module.ipynb*)

Function in step (8) removes any probe that has masked values more than 2/3 (default threshold "TRPQC") of the sample size. This is followed by step (9) exploring the masking percentage. Functions (10-11) checks for the number of probes that has range of values (max – min) greater than 0.3 and isolate these probes to handle the outliers separately from the other probes that has range < 0.3. The advantage of this method is that any future imputation for outliers will take place only over the isolated probes rather than the entire dataset, and hence the overall adjustments are minimal. The range 0.3 is decided based on available literature which states that DMRs are considered significant when the cases are at least 0.2 greater or less than controls (Cabezón et al., 2021, Jiang et al., 2015, Van Doorn et al., 2016). Therefore, to maintain biological variability among samples, 0.3 is considered conservative. In other words, we are considering probes with a range of 0.3 or less as biologically variable.

#	ID_Betas	p-value Mean	p-val > 0.05	p-val > 0.01	Perc. of Masked Betas Resulted from SeSAME	QC Score	QC Score (Perc.)
1	R04C01_DLD013_Betas	0.00305	6844	38223	12.76%	48	100.00%
2	R03C01_DLD012_Betas	0.00342	8234	40331	12.90%	42	87.50%
3	R03C01_DLD007_Betas	0.00353	8234	43021	12.89%	38	79.17%
4	R02C01_DLD006_Betas	0.00345	8266	41007	12.91%	36	75.00%
5	R04C01_DLD008_Betas	0.0036	8236	44287	12.89%	31	64.58%
6	R03C01_DLD003_Betas	0.00357	8508	42202	12.93%	30	62.50%
7	R04C01_DLD004_Betas	0.00357	8703	41409	12.95%	28	58.33%
8	R02C01_DLD002_Betas	0.00373	8787	43791	12.95%	22	45.83%
9	R01C01_DLD010_Betas	0.00376	9562	47265	13.05%	15	31.25%
10	R02C01_DLD011_Betas	0.00381	9669	44786	13.05%	14	29.17%
11	R01C01_DLD005_Betas	0.00402	9886	49916	13.07%	8	16.67%
12	R01C01_DLD001_Betas	0.00576	17737	67254	13.90%	4	8.33%

Table 11. Scoring matrix to check samples performance.

Functions in step (12) are similar to (2-7) except that it calculates the absolute difference for the isolated probes only (probes with range > 0.3). Similarly, the subjects are ranked to check which one accounts for the most outliers (the output is similar to Table 12). Subjects are kept/ removed accordingly. It is important to consider the sample size as a small sample size will limit the ability to exclude low performing subjects.

Functions (13–16) replace (mask) all detected outliers using the IQR method with NaN, before appending the isolated probes back into the original dataset. This is followed by another round of removal of low-quality probes. Function (17) imputes the remaining masked beta values using the weighted mean method, where extra weight is given to the sample that achieve top scores throughout the entire workflow.

#	Column Name	Max Abs. Diff. Count	QC Score
1	203259750076_R02C01_DLD006_Betas	31890	75.00%
2	203259750077_R03C01_DLD012_Betas	36592	87.50%
3	203259750077_R04C01_DLD013_Betas	37745	100.00%
4	203259750076_R01C01_DLD005_Betas	47025	16.67%
5	203259750077_R01C01_DLD010_Betas	54050	31.25%
6	203141320045_R03C01_DLD003_Betas	54704	62.50%
7	203259750077_R02C01_DLD011_Betas	57670	29.17%
8	203259750076_R04C01_DLD008_Betas	67241	64.58%
9	203141320045_R04C01_DLD004_Betas	72142	58.33%
10	203141320045_R02C01_DLD002_Betas	75714	45.83%
11	203259750076_R03C01_DLD007_Betas	78858	79.17%
12	203141320045_R01C01_DLD001_Betas	232594	8.33%

Table 12. Subject performance ranking table. Lower-quality subjects are often associated with a higher count of beta values showing the greatest absolute deviation from the mean.

4.3. DMRs Detection

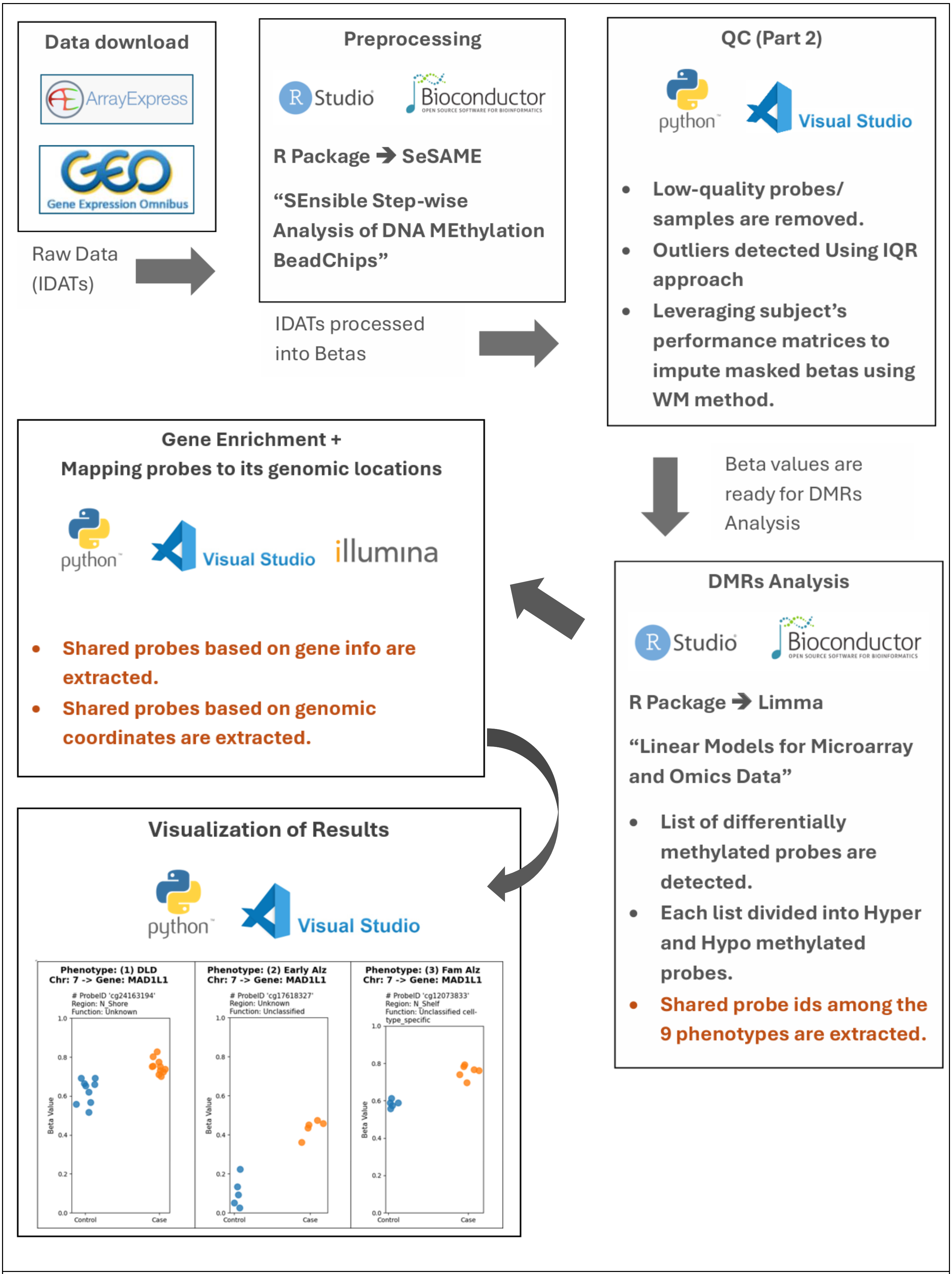
The *Limma* package was used to perform the differential analysis. *Limma* is a popular choice among researchers in DNA methylation analysis because it offers several advantages, such as using the Benjamini-Hochberg method for FDR correction. Furthermore, *Limma* has been extensively tested and used over a long period of time. This step is applied over the 9 datasets separately. The analysis of DMRs is carried out for each phenotype against its corresponding control dataset provided in the study. The code script used to run *Limma* analysis along with FDR correction is included in the Appendices Chapter (*Section 5, Limma Analysis*). The tables containing a list of differentially methylated probes for each dataset are available in supplementary material (*Limma DMPs*).

4.4. Cross Comparison

After extracting DMRs for each case dataset (a total of 9 results corresponding to 9 phenotypes), it was observed that one of the cases (TSD) resulted in zero DMRs after FDR correction. For the exploratory approach intended in this study, we decided to use the DMRs file without FDR correction for this specific dataset only (E-MTAB-4664, Total Acute Sleep Deprivation). However, this will be considered one of the limitations of the study, and any shared DMRs/ DMPs with this specific dataset will be highlighted as weak results. The cross-comparison among the 9 phenotypes was conducted based on 3 criteria (Table 13).

Criteria	Procedure
Shared Probes	This part will look for shared probe ids that are differentially methylated among the 9 datasets.
Shared Genes	Using Illumina manifests, the differentially methylated probes for each phenotype are gene enriched, and then a cross comparison among the 9 phenotypes is carried out to check for probes that are mapped to the same gene. A priority is given to probes that are mapped to promoter regions
Shared Regions	For each phenotype, using Illumina manifests, the differentially methylated probes are mapped to their genomic locations. Probes with shared regions across the 9 results are determined based on a threshold of base pair distance. Since this approach is exploratory, an initial threshold of 10,000 bp was applied, but no significant probes were identified. Therefore, the threshold was increased to 100,000 bp. Although a range of 100,000 bp is commonly used in other studies (Bondhus et al., 2022) when identifying DMRs, the results should be interpreted with caution.

Table 13. Approach to find shared methylation patterns across multiple phenotypes.



Data download

ArrayExpress

GEO Gene Expression Omnibus

Raw Data (IDATs)

Preprocessing

R Studio Bioconductor

R Package → SeSAME

"SEnsible Step-wise Analysis of DNA MEthylation BeadChips"

IDATs processed into Betas

QC (Part 2)

python Visual Studio

- Low-quality probes/ samples are removed.
- Outliers detected Using IQR approach
- Leveraging subject's performance matrices to impute masked betas using WM method.

Beta values are ready for DMRs Analysis

DMRs Analysis

R Studio Bioconductor

R Package → Limma

"Linear Models for Microarray and Omics Data"

- List of differentially methylated probes are detected.
- Each list divided into Hyper and Hypo methylated probes.
- Shared probe ids among the 9 phenotypes are extracted.

Gene Enrichment + Mapping probes to its genomic locations

python Visual Studio illumina

- Shared probes based on gene info are extracted.
- Shared probes based on genomic coordinates are extracted.

Visualization of Results

python Visual Studio

Figure 11. Workflow diagram

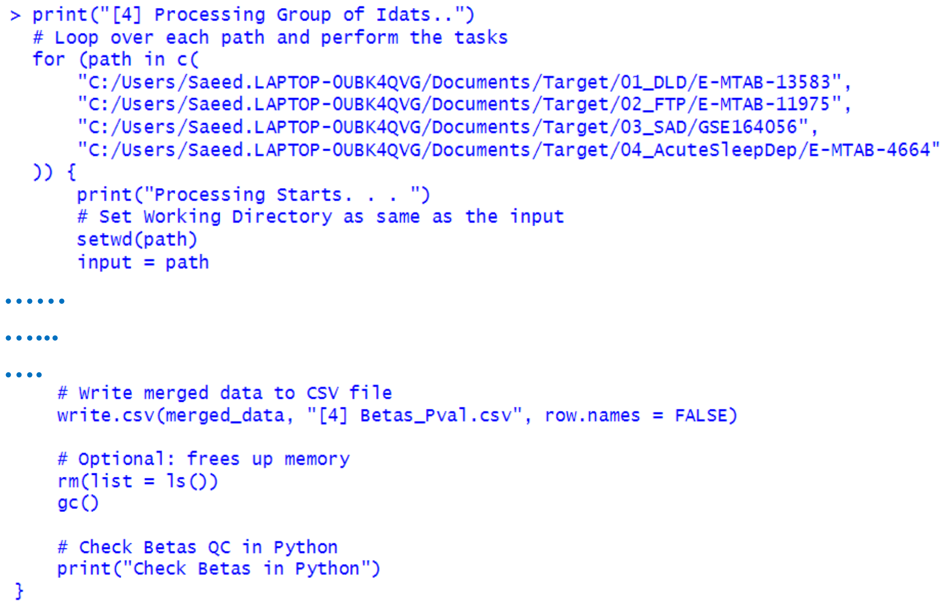
A general workflow plan is demonstrated in Fig. 11, along with the tools and sources used. This chapter is divided into 4 sections:

1. Preprocessing
2. Quality Control (QC part 2)
3. Differential Analysis
4. Cross Comparison

1. Preprocessing

All datasets (cases and controls) were processed using the *SeSAME* pipeline, as detailed in the methodology (Chapter 2). The output of this procedure comprises 13 datasets in CSV format (Figure 12).

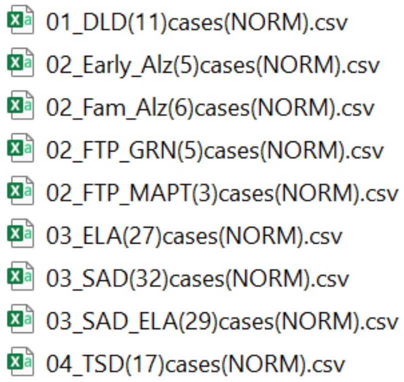
Abstract of the code (R programming) used for 4 datasets (controls)



```
> print("[4] Processing Group of Idats..")
  # Loop over each path and perform the tasks
  for (path in c(
      "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/01_DLD/E-MTAB-13583",
      "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/02_FTP/E-MTAB-11975",
      "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/03_SAD/GSE164056",
      "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/04_AcuteSleepDep/E-MTAB-4664"
  )) {
      print("Processing Starts. . . ")
      # Set Working Directory as same as the input
      setwd(path)
      input = path
......
......
....
     # Write merged data to CSV file
     write.csv(merged_data, "[4] Betas_Pval.csv", row.names = FALSE)

     # Optional: frees up memory
     rm(list = ls())
     gc()

     # Check Betas QC in Python
     print("Check Betas in Python")
 }
```

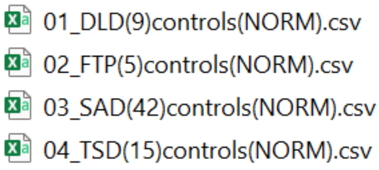
Output (Total 13 datasets)

9 cases:



01_DLD(11)cases(NORM).csv
02_Early_Alz(5)cases(NORM).csv
02_Fam_Alz(6)cases(NORM).csv
02_FTP_GRN(5)cases(NORM).csv
02_FTP_MAPT(3)cases(NORM).csv
03_ELA(27)cases(NORM).csv
03_SAD(32)cases(NORM).csv
03_SAD_ELA(29)cases(NORM).csv
04_TSD(17)cases(NORM).csv

4 controls:



01_DLD(9)controls(NORM).csv
02_FTP(5)controls(NORM).csv
03_SAD(42)controls(NORM).csv
04_TSD(15)controls(NORM).csv

Figure 12. preprocessing samples and convert raw IDATs to Betas (.csv format)

2. Quality Control (QC part 2)

Unlike preprocessing, which occurs iteratively, the second round of quality control (QC) is conducted separately for each dataset to explore and manage the removal of probes and subjects. The same steps are repeated across the 13 datasets. Table 14 summarizes the output of this section. Since all datasets undergo the same process, one complete workflow is demonstrated in detail for the 01_DLD dataset (page: 39), while the others are summarized with experiment details and normalization curves (9 cases followed by 4 controls).

#	Phenotype	Dataset Abbreviation	Probes		Samples		Percentage of Imputation	Std	
			Before	After	Before	After		Before	After
1	Developmental Language Disorder	01_DLD (cases)	866553	737275	12	11	1.67%	0.019	0.0185
2	Sporadic Early-Onset Alzheimer's Disease	02_Early_Alz (cases)	866553	712035	5	5	9.97%	0.0797	0.0665
3	Familial Alzheimer Disease	02_Fam_Alz (cases)	866553	730137	6	6	7.90%	0.0562	0.049
4	Genetic Frontotemporal Dementia (GRN Mutation)	02_FTP_GRN (cases)	866553	708698	5	5	10.30%	0.0794	0.0662
5	Genetic Frontotemporal Dementia (MAPT Mutation)	02_FTP_MAPT (cases)	866553	734277	3	3	0.97%	0.0772	0.0769
6	Early Life Adversity	03_ELA (cases)	866553	738553	30	27	2.28%	0.0185	0.0177
7	Social Anxiety Disorder	03_SAD (cases)	866553	738751	35	33	2.48%	0.0193	0.0182
8	Social Anxiety Disorder and Early Life Adversity	03_SAD_ELA (cases)	866553	738975	31	29	2.29%	0.0198	0.019
9	Total Acute Sleep Deprivation	04_TSD (cases)	486427	406311	18	17	3.08%	0.0191	0.0184
10	Healthy Controls	01_DLD (controls)	866553	736430	12	9	1.86%	0.0208	0.0187
11	Healthy Controls	02_FTP (controls)	866553	707987	5	5	1.94%	0.0703	0.0577
12	Healthy Controls	03_SAD (controls)	866553	738597	47	42	2.83%	0.0191	0.0181
13	Healthy Controls	04_TSD (controls)	486427	406446	18	17	2.24%	0.0197	0.0179

Table 14. A summary of QC (part 2) effect on number of probes, samples, and overall deviation probe wise.

Table 15 includes information about technical aspects and output data availability.

Data Availability		Notes
Detailed Workflow for 01_DLD dataset.	Pages: 37 - 44	Demonstration of detailed output
Summary and Output of remaining datasets.	Pages: 45 - 56	Beta distribution curves before and after QC (part 2)
Source Code used	Supplementary Material (QC2)	(.ipynb) files for 13 datasets (Python)*
Complete output **	Supplementary Material (*Quality Score*)	Tables for 13 datasets
	Supplementary Material (*QC2/ Subject Performance*)	Tables for 13 datasets

Table 15. Output data availability and technical information.

* Using the VS Code editor on a 16GB RAM Intel Core i5 PC, processing a single (.ipynb) file takes approximately 4 minutes for smaller sample sizes (e.g., 5 subjects in the 02_Early Alz dataset) and up to ~16 minutes for larger sample sizes (e.g., 43 subjects in the 03_SAD control dataset). This time includes generating beta value distribution plots.

** The large number of tables made it impractical to include this information in the thesis text or appendices, particularly for datasets with a high number of samples.

Dataset [1]

Comparison of the methylation profiles of children with developmental language disorder and healthy control subjects.

Released: 2024 | Link: E-MTAB-13583 < ArrayExpress < BioStudies < EMBL-EBI

➢ Experiment Details:

ID	E-MTAB-13583	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	24 Subjects (3-7 yr)
Published Article	Hypomethylation of Wnt Signaling Regulator Genes in Developmental Language Disorder, (2024). Link: https://doi.org/10.2217/epi-2023-0345	**Phenotype Sample Size**	12 Subjects (3-7 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Developmental Language Disorder**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes (.csv format)		

➢ Exclusion of SNP, Control, and Sex Chromosomes Probes:

Initial Array Size	Probes to be Excluded	Output (Before Exclusion)	Output (After Exclusion)	Updated Array Size
866553 Probes	➢ *19640 (chrX, chrY) Probes* ➢ *53 'rs' Probes* ➢ *635 'ctl' Probes*	*cg 862927* *ch 2932* *ct 635* *rs 59*	*cg 843386* *ch 2839*	*846225 Probes*

➢ Summary of Workflow and Steps:

	Raw Signals	**QC (1)**	**QC (2)**				
Order	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Steps	**Raw Beta Values**	**SeSAME Output**	**Low Performing Subjects**	**Low Performing Probes (1) ***	**Isolating & Handling Probes with range > 0.3**	**Masking Outliers in the isolated probes**	**Low Performing Probes (2) ***
No. of Probes	846225	846225	846225	**737288**	737288	737288	**737275**
No. of Subjects	12	12	**11**	11	11	11	11
Total Masked	0	1322182 (13.02%)	1204573 (12.94%)	**14982 (0.18%)**	14982 (0.18%)	**22922 (0.28%)**	**22841 (0.28%)**

Betas **							
Details	-	13.02%of Betas are masked based on SeSAME quality metrics	One subject is excluded: '203141320045_R01C01_DLD001*	Bad probes (108937) are removed.	(10432) probes of which range > 0.3 are isolated for handling outliers separately from other probes with range < 0.3	7940 outlier values are masked	Bad probes (13) are removed.

Table 16. The effect of QC (part 2) on DLD dataset. Step 7 is followed by imputation of masked betas.

* The removal of low performing probes using probeQC() function is carried twice; the first one is after removing bad samples, and the second one is after masking outlier values in step (6).

** Both number and percentage represent the total number/ percentage out of the entire array respectively.

➢ Details and Discussion of Steps (2-7)

Steps (2-4): Thresholds and QC Criteria (Determining High vs Low Performing Samples)

After *SeSAME* processing, beta values along with its corresponding p-values for all subjects were grouped together in one dataset and QC score is calculate for each subject following our methodology. Table 17 Shows the ranking for each subject where highest and lowest performing samples highlighted in green and red respectively.

Rank	ID_Betas	p-value Mean	p-val > 0.05	p-val > 0.01	Perc. of Masked Betas Resulted from SeSAME	QC Score	QC Score (Perc.)
1	203259750077_R04C01_DLD013_Betas	0.00305	6844	38223	12.76%	48	100.00%
2	203259750077_R03C01_DLD012_Betas	0.00342	8234	40331	12.90%	42	87.50%
3	203259750076_R03C01_DLD007_Betas	0.00353	8234	43021	12.89%	38	79.17%
4	203259750076_R02C01_DLD006_Betas	0.00345	8266	41007	12.91%	36	75.00%
5	203259750076_R04C01_DLD008_Betas	0.00360	8236	44287	12.89%	31	64.58%
6	203141320045_R03C01_DLD003_Betas	0.00357	8508	42202	12.93%	30	62.50%
7	203141320045_R04C01_DLD004_Betas	0.00357	8703	41409	12.95%	28	58.33%

8	203141320045_R02C01_DLD002_Betas	0.00373	8787	43791	12.95%	22	45.83%
9	203259750077_R01C01_DLD010_Betas	0.00376	9562	47265	13.05%	15	31.25%
10	203259750077_R02C01_DLD011_Betas	0.00381	9669	44786	13.05%	14	29.17%
11	203259750076_R01C01_DLD005_Betas	0.00402	9886	49916	13.07%	8	16.67%
12	203141320045_R01C01_DLD001_Betas	0.00576	17737	67254	13.90%	4	8.33%

Table 17. Subjects ranking (Quality score)

This was followed by a bar plot to visualize any subject(s) with excessive outliers (Fig. 13).

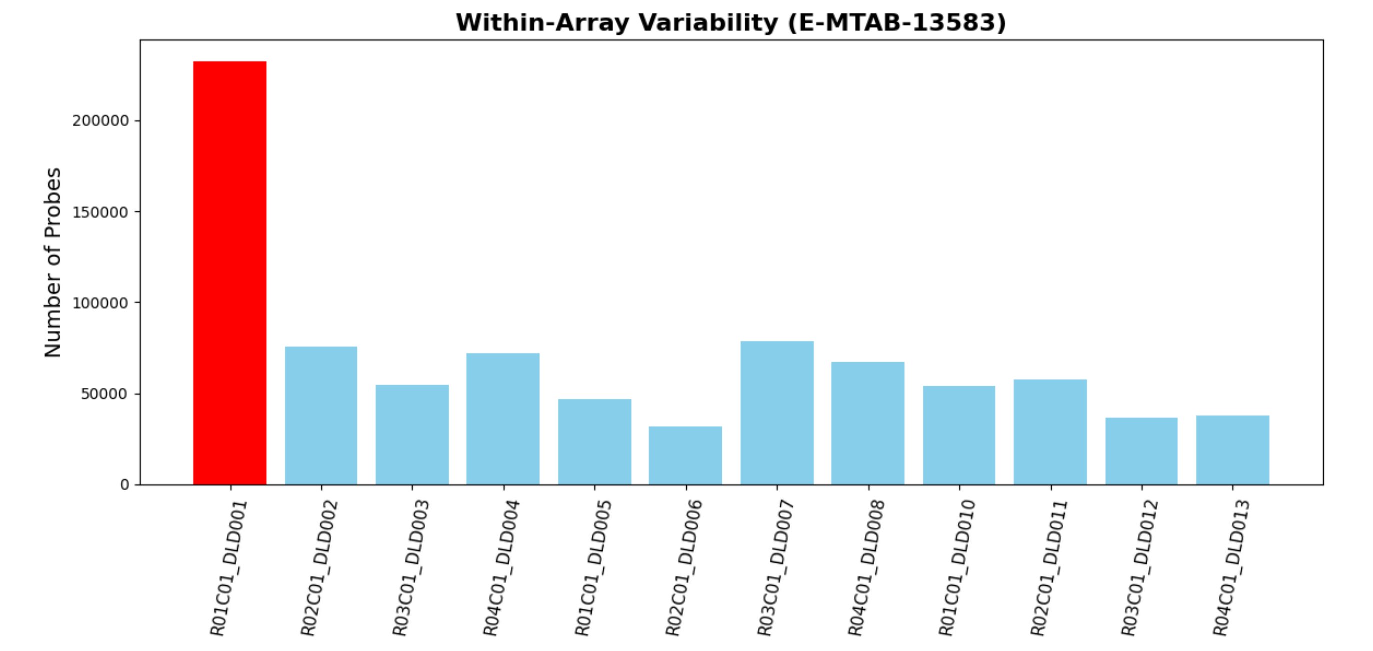


Figure 13. A bar plot showing the number of probes each subject achieves the highest absolute deviation from the mean.

Sample id (203141320045_R01C01_DLD001) achieves the highest absolute difference from the mean average in more than 200,000 probes. Together with being the lowest score in QC score table, we decided to remove the sample from downstream analysis. To validate our decision, a comparison table is used to number of outlier values observed in one sample against its QC score (Table 18). As mentioned in summary table, the removal of low performing probes take place in step (4) to ensure that no probes have exceeds the threshold of masked betas. The threshold (TRPQC) for E-MTAB-13583 is decided to be 6 (12 * 0.5 = 6).

#	ID_Betas	Max Abs. Diff. Count	QC Score	#	ID_Betas	Max Abs. Diff. Count	QC Score
1	203259750076_R02C01_DLD006	31890	75.00%	7	203259750077_R02C01_DLD011	57670	29.17%
2	203259750077_R03C01_DLD012	36592	87.50%	8	203259750076_R04C01_DLD008	67241	64.58%
3	203259750077_R04C01_DLD013	37745	100.00%	9	203141320045_R04C01_DLD004	72142	58.33%
4	203259750076_R01C01_DLD005	47025	16.67%	10	203141320045_R02C01_DLD002	75714	45.83%
5	203259750077_R01C01_DLD010	54050	31.25%	11	203259750076_R03C01_DLD007	78858	79.17%
6	203141320045_R03C01_DLD003	54704	62.50%	12	203141320045_R01C01_DLD001	232594	8.33%

Table 18. A Comparison of QC score versus maximum absolute deviation.

Steps (5-7) Outlier Detection Based on range and IQR approach:

Following our methodology, first we isolate the probes that has a range (max min) beta values > 0.3 using the following functions:

`Function:`
`count_probes_with_range(df, thresholds=[0.2, 0.3, 0.4, 0.5])`
	`Function:`
	`extract_probes_with_range(df, threshold=0.3)`

> Output: *Probes with range > 0.2: 38540 Probes (5.23%)* *Probes with range > 0.3: 10432 Probes (1.41%)* *Probes with range > 0.4: 3026 Probes (0.41%)* *Probes with range > 0.5: 833 Probes (0.11%)*	*> Output:* *Probes Above Threshold Are Successfully Isolated* *No. of Probes Isolated: 10432* *No. of Probes Not Affected: 726856*

The 10432 probes are then inspected separately to check the subject's performance for this set of probes. The bar plot shows 1 sample to have distinguished count of probes with highest absolute difference. However, the count (2236 probes) is not enough to exclude this sample when considering its QC score, the size of array and the performance of other samples (Figure 14), (Table 19).

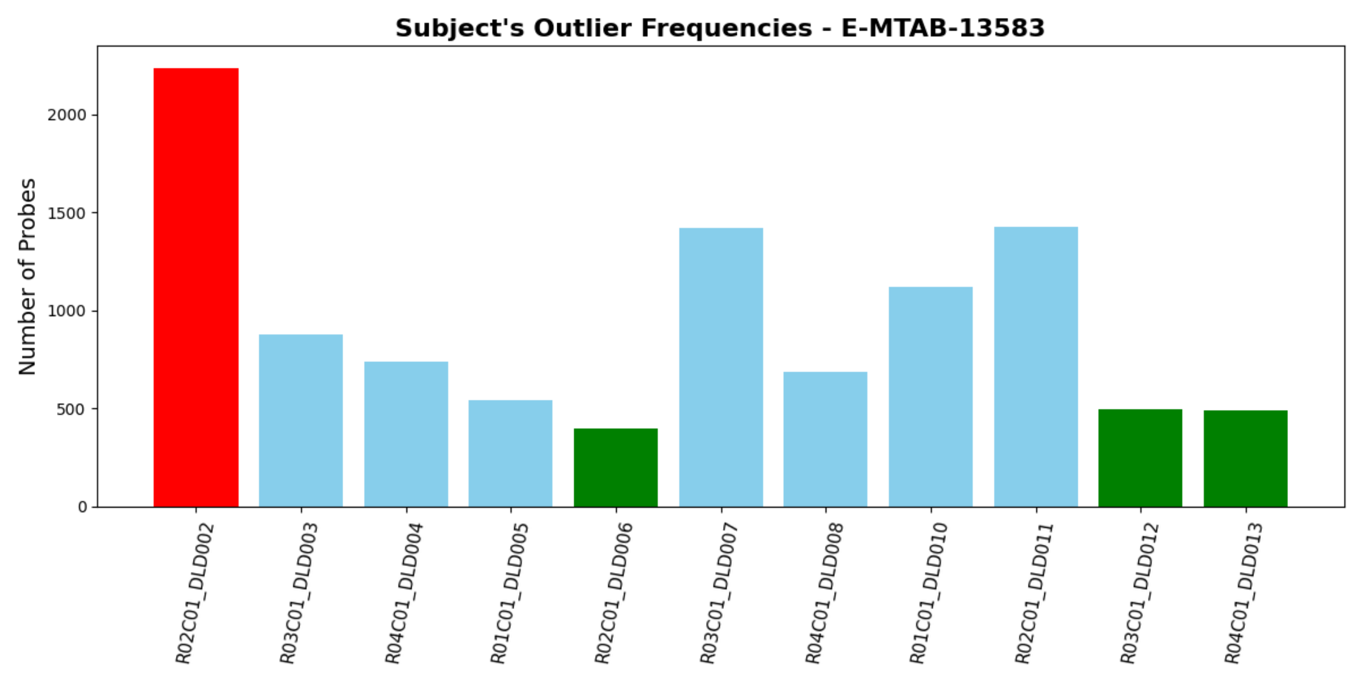


Figure 14. A bar plot comparing the outlier count for each subject based on the isolated probes using the IQR method.

On the other hand, sample id (203259750077_R04C01_DLD013) highlighted in green shows the least amount of Max Abs. Diff. in addition to being the top performing sample in terms of original QC score (Table 19). This makes it qualified to be chosen as the target sample when imputing missing values using weighted average in the final procedure of QC (2).

#	ID_Betas	Max Abs. Diff. Count	QC Score	#	ID_Betas	Max Abs. Diff. Count	QC Score
1	203259750076_R02C01_DLD006	398	75.00%	7	203141320045_R03C01_DLD003	879	62.50%
2	**203259750077_R04C01_DLD013**	**490**	**100.00%**	8	203259750077_R01C01_DLD010	1121	31.25%
3	203259750077_R03C01_DLD012	494	87.50%	9	203259750076_R03C01_DLD007	1419	79.17%
4	203259750076_R01C01_DLD005	544	16.67%	10	203259750077_R02C01_DLD011	1428	29.17%
5	203259750076_R04C01_DLD008	684	64.58%	11	**203141320045_R02C01_DLD002**	**2236**	**45.83%**
6	203141320045_R04C01_DLD004	739	58.33%				

Table 19. Highest and lowest performing subjects are highlighted in green and blue respectively.

The next step is to replace outlier values in the 10432 Probes based on IQR approach followed by re-joining them again to the existing array using below functions:

Function: `replace_outliers_withNaN(df)`	Function: `append_masked_to_original(df_masked, df_remaining)`	Validation of dimensions
> Output: *7940 Outliers have been replaced with NaN using IQR approach.*	*> Output:* *Isolated Probes Are Re-joined Back to the Remaining Array* *Check Dimensions:* *737288 Probes, 11 Samples*	*> Output: Checking Dimensions of df5.. 737288 Probes, 11 Subjects >>> Dimensions Confirmed*

After masking the 7940 outliers, the total missing betas increased to 22922 values in the whole array. Therefore, another exclusion of low performing probes took place to exclude probes that has more than 6 missing betas. This has resulted in exclusion of only 13 probes. Below is a comparison of number of probes with certain ranges before and after masking the outliers:

Before Masking Outliers	**After Masking Outliers**
Function: `count_probes_with_range(df, thresholds=[0.2, 0.3, 0.4, 0.5])`	Function: `count_probes_with_range(df, thresholds=[0.2, 0.3, 0.4, 0.5])`
> Output: *Probes with range > 0.2: 27719 Probes (6.82%)* *Probes with range > 0.3: 8481 Probes (2.09%)* *Probes with range > 0.4: 2830 Probes (0.70%)* *Probes with range > 0.5: 858 Probes (0.21%)*	*> Output:* *Probes with range > 0.2: 24604 Probes (6.06%)* *Probes with range > 0.3: 3294 Probes (0.81%)* *Probes with range > 0.4: 1052 Probes (0.26%)* *Probes with range > 0.5: 428 Probes (0.11%)*

After Step (7), the final procedure will be to impute any remaining masked betas.

Imputation of Masked Betas:

Following our methodology which uses weighted mean for imputation, we have chosen sample id (9297962042_R04C01) highlighted in green to be the target sample. A weight of 3 is decided since the sample achieves the top score in QC score and the least number of probes for Max. Abs. Diff. column. Total of 22841values were imputed with WM using *impute_WM()* function:

Parameters:	Function:

`Top_Scorer:`
`203259750077_R04C01_DLD013_Betas`
`Target_Weight: 3`
	`impute_WM(df6, target_col=Top_Scorer,`
	`target_weight=Target_Weight,`
	`default_weight=1)`

Application:

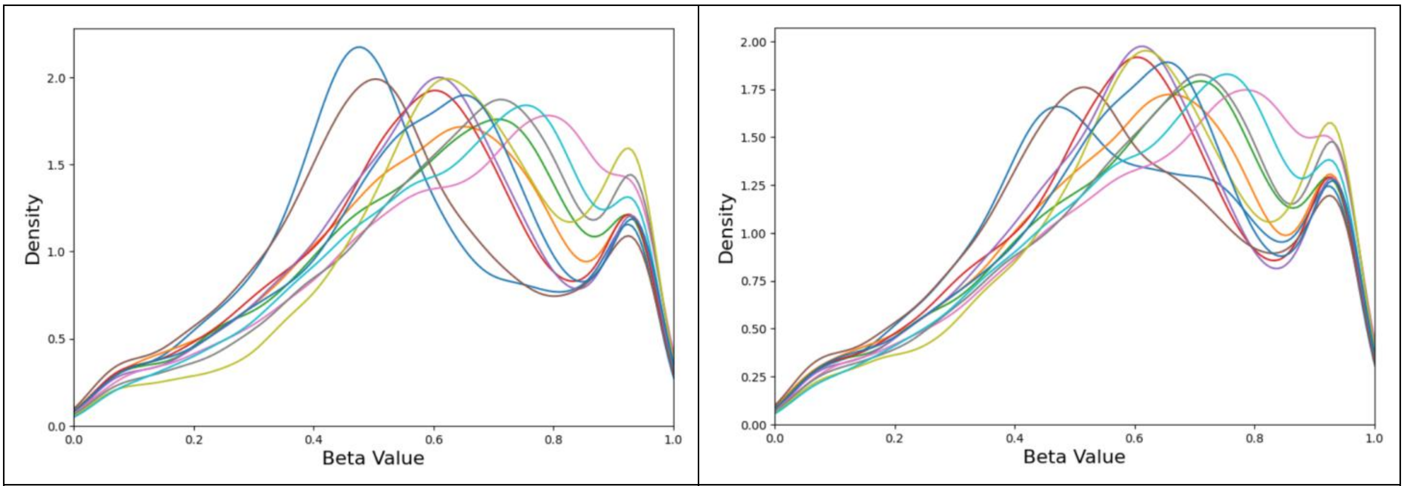
Imputation of Masked Betas	Validation of the Result Probe Wise (Before and After Imputation)	Overall effect of QC (2) Procedure on the Array:

`Function:`
`impute_WM(df6,`
`target_col=Top_Scorer,`
`target_weight=Target_Weight,`
`default_weight=1)`
	`Function:`
	`probe_QC(df, threshold=1,`
	`remove=False)`
		Subtraction of affected probes from the whole array

> Imputation: *Total Missing Betas Replaced with WM are* 22841, *(0.24% Out of Total Betas)*	*> Output:(Before Imputation) Number of probes with >= 1 masked betas: 12331 (1.67%)* --------------------------- *> Output:(After Imputation) Number of probes with >= 1 masked betas: 0 (0.00%)*	*> Output:* *No. of Probes Not Modified by Imputation: 724944 (98.33%)* *Final Dimensions 737275 Probes, 11 Subjects*

Visualization

A comparison (Visualization and quantification of the changes happened to raw data) is done through beta distribution curves. Figure 15 shows the difference in normalization that took place for the probes (10432) that were > 0.3 in range:



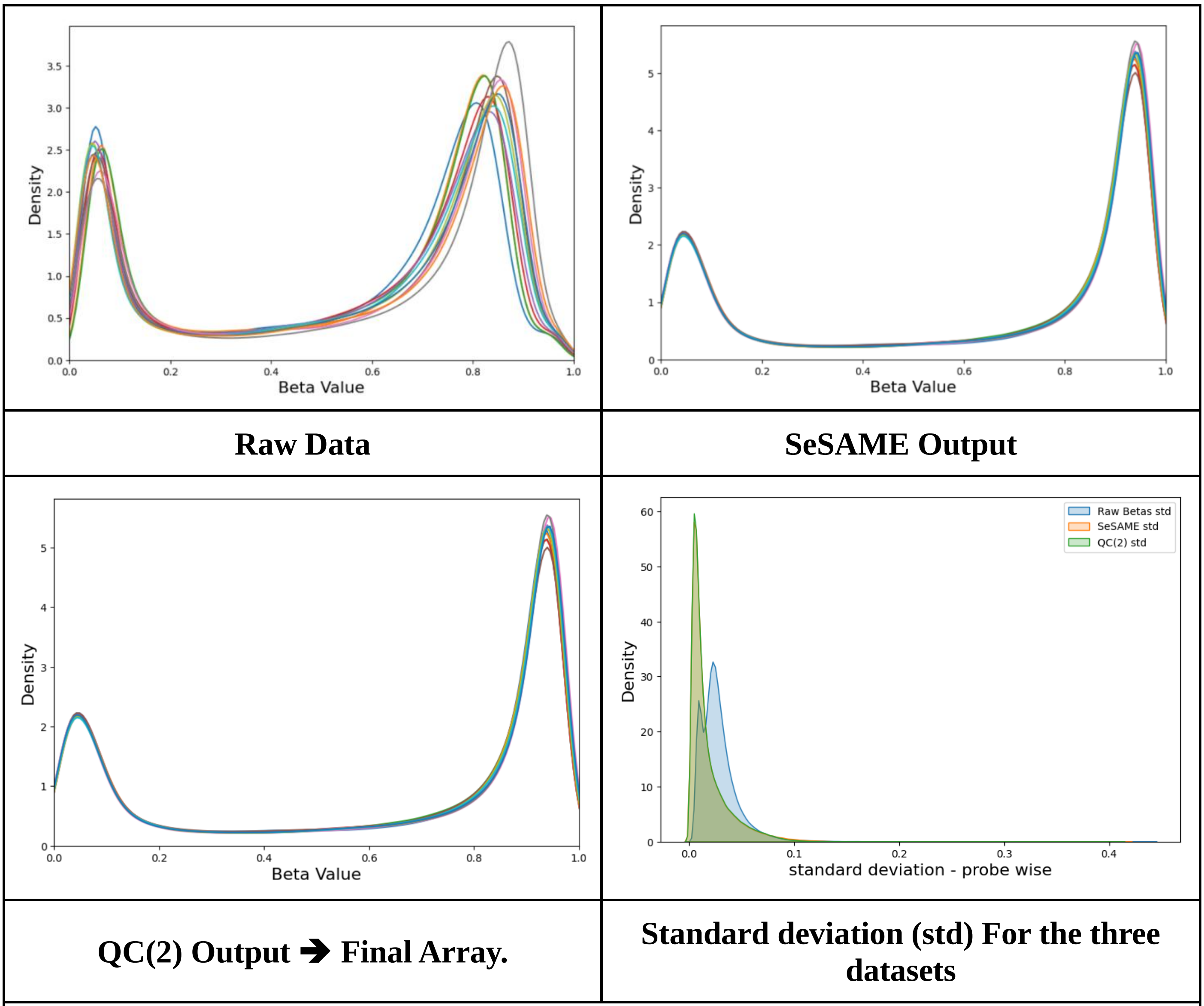
Before Imputation | **After Imputation**

Figure 15. Beta value distribution for the isolated probes before and after imputation.

Another visualization (Figure 16.) using beta distribution curves is carried for the entire array during the three stages of data:

1- Raw Data (The output from getBetas() without preprocessing)
2- Data after SeSAME.
3- Data after QC(2) (The final output after imputation).

Graphs ➔



Raw Data

SeSAME Output

QC(2) Output ➔ Final Array.

Standard deviation (std) For the three datasets

Figure 16. Beta value distribution curve for E-MTAB-13583 (01_DLD cases)

Quantification:

To quantify the difference in normalization over the three stages, we have measured the mean average of standard deviation among the entire probes in the array for each stage:

Raw Data	**After SeSAME**	**After QC(2)**

`Function:`
`raw_betas['std'] = raw_betas.filter(like='_Betas').std(axis=1)`
 	
`avg_std_raw_betas = raw_betas['std'].mean()`
	`Function:`
	`df2['std'] = df2.filter(like='_Betas').std(axis=1)`
	
	`avg_std_df2 = df2['std'].mean()`
		`Function:`
		`dfFinal['std'] = dfFinal.filter(like='_Betas').std(axis=1)`
		
		`avg_std_dfFinal = dfFinal['std'].mean()`

Output: *Average std in raw data:* ***0.0293***	*Output:* *Average std after SeSAME:* ***0.0190***	*Output:* *Average std after QC(2):* ***0.0185***

The results match the distribution curves we see in Figures (15, 16), since the amount of decrease from **0.0293** in raw data to **0.0190** in data after *SeSAME* is more noticeable than the decrease from **0.0190** in *SeSAME* to **0.0185** in data after QC(2). However, the slight improvement in variability reduction observed in QC(2) compared to the *SeSAME* output is satisfactory, as the procedure aims to minimally adjust weak and unreliable values without introducing extreme modifications, thereby preserving biological variability.

Remaining datasets (12) are summarized with experimental details and beta value distribution curves (pages: 48 - 60). The adjustments that took place in QC(2) for each dataset is previously mentioned in Table 14.

Dataset [2-1] Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines.

Released: 2024 | Link: E-MTAB-11975 < ArrayExpress < BioStudies < EMBL-EBI

- **Experiment Details:**

ID	E-MTAB-11975	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	64 Subjects (31-92 yr), (prefrontal cortex tissue is excluded ➔ 24 Subjects (40-76 yr)
Published Article	Genome-Wide DNAMethylation in Early-Onset-Dementia Patients Brain Tissue and Lymphoblastoid Cell Lines, (2024). Link: https://doi.org/10.3390/ijms25105445	**Phenotype Sample Size**	**5 Subjects** (52-63 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Sporadic Early-Onset Alzheimer's Disease**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	No		

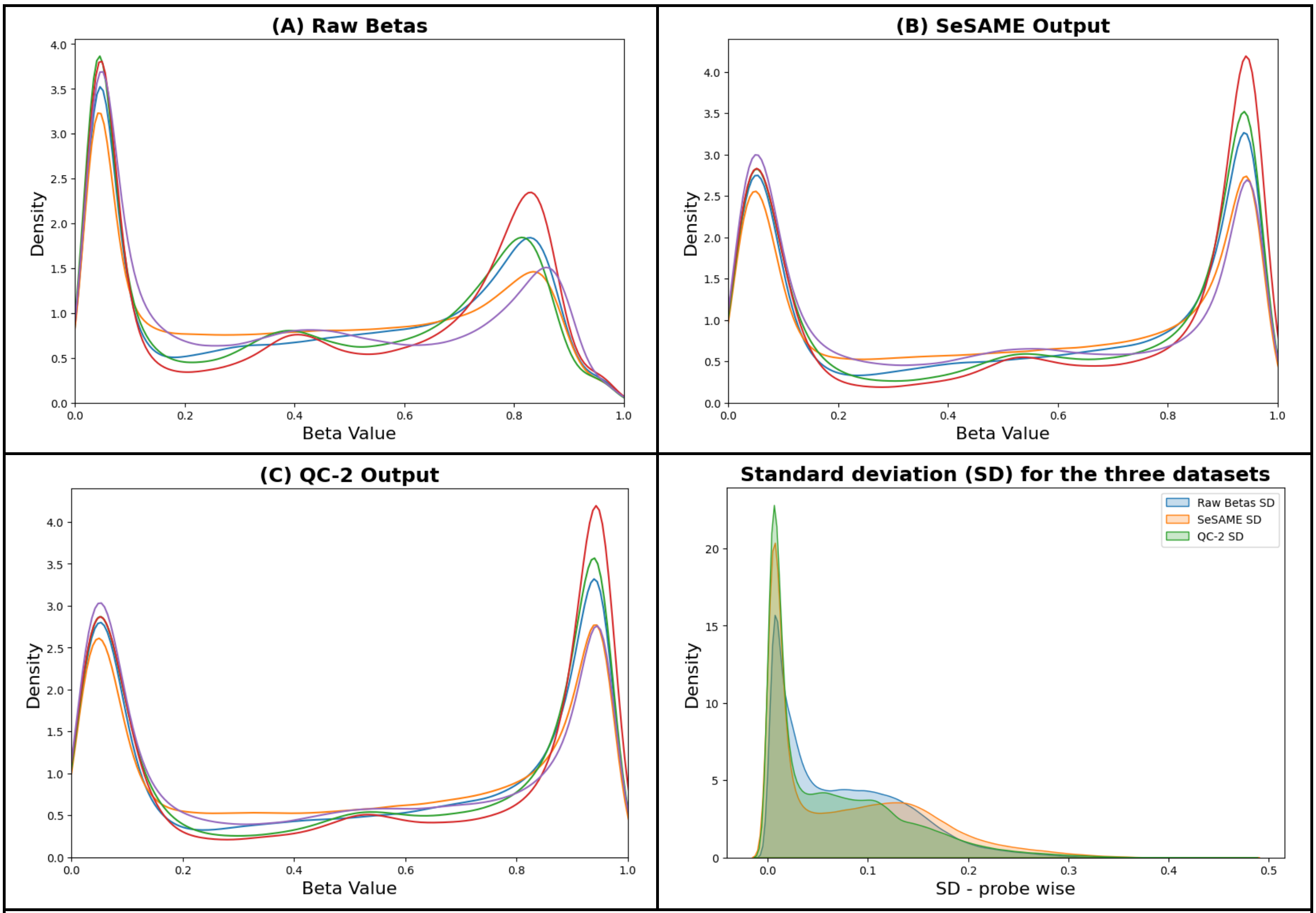


Figure 17. Beta value distribution curve curves for E-MTAB-11975, dataset 02_Early_Alz (cases).

Dataset [2-2]

Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines.

Released: 2024 | Link: E-MTAB-11975 < ArrayExpress < BioStudies < EMBL-EBI

- **Experiment Details:**

ID	E-MTAB-11975	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	64 Subjects (31-92 yr), (prefrontal cortex tissue is excluded ➔ **24 Subjects (40-76 yr)**
Published Article	Genome-Wide DNAMethylation in Early-Onset-Dementia Patients Brain Tissue and Lymphoblastoid Cell Lines, (2024). Link: https://doi.org/10.3390/ijms25105445	**Phenotype Sample Size**	6 Subjects (42-59 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Familial Alzheimer Disease**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	No		

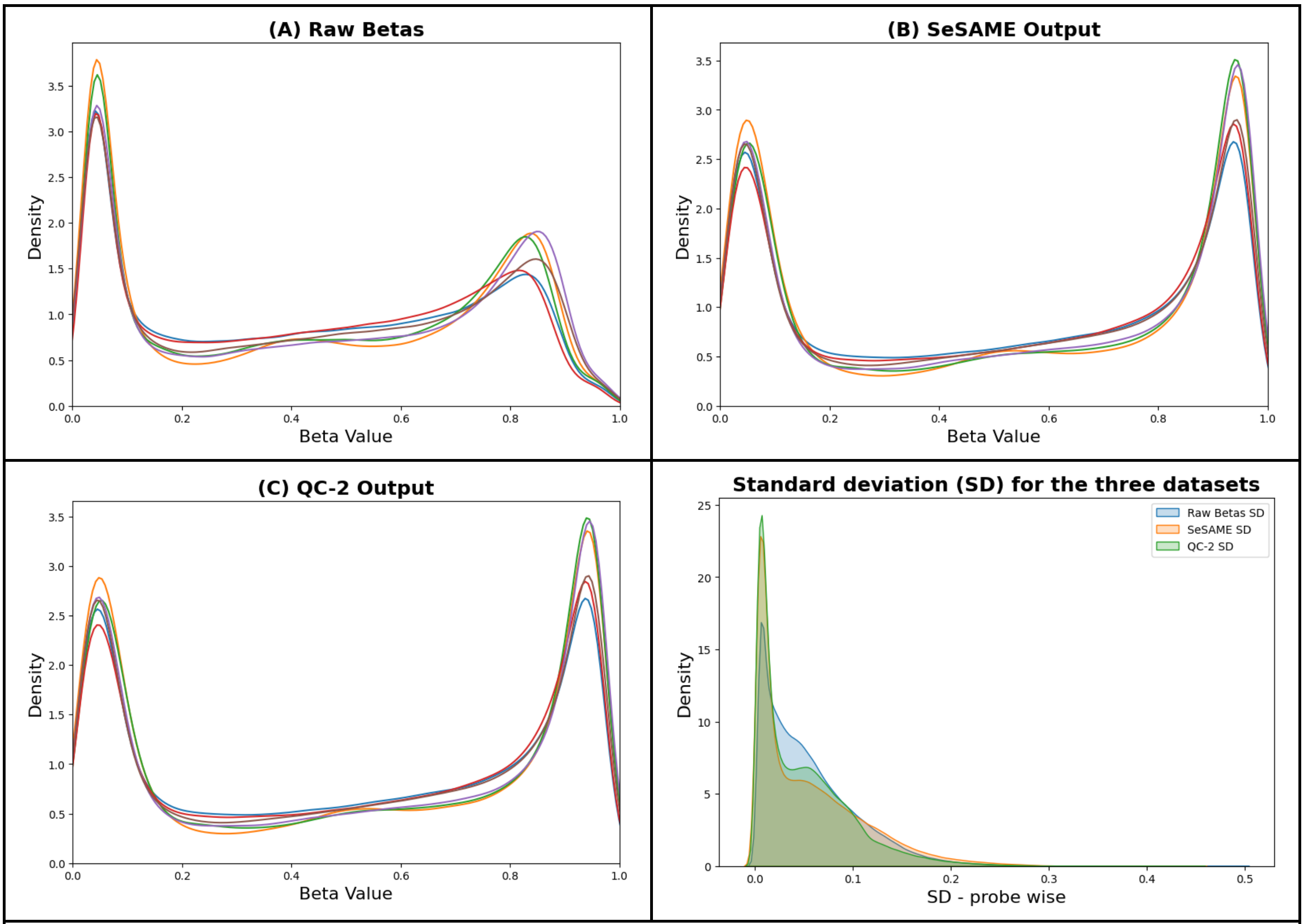


Figure 18. Beta distribution curves for E-MTAB-11975, dataset 02_Fam_Alz (cases).

Dataset [2-3]

Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines.

Released: 2024 | Link: E-MTAB-11975 < ArrayExpress < BioStudies < EMBL-EBI

➢ Experiment Details:

ID	E-MTAB-11975	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	64 Subjects (31-92 yr), (prefrontal cortex tissue is excluded ➔ 24 Subjects (40-76 yr)
Published Article	Genome-Wide DNAMethylation in Early-Onset-Dementia Patients Brain Tissue and Lymphoblastoid Cell Lines, (2024). Link: https://doi.org/10.3390/ijms25105445	**Phenotype Sample Size**	**5 Subjects** (54-63 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Genetic Frontotemporal Dementia (GRN Mutation)**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	No		

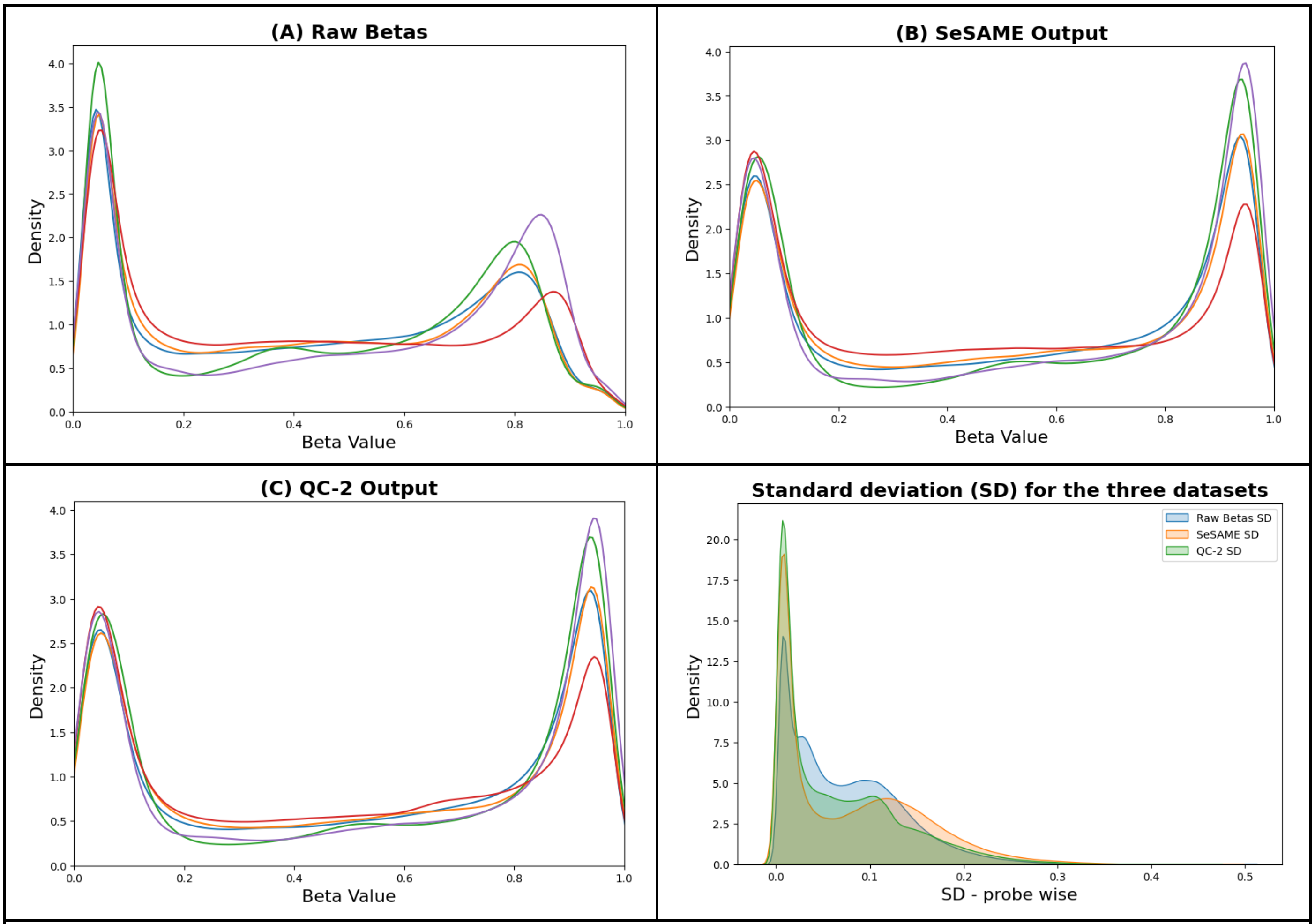


Figure 19. Beta value distribution curves for E-MTAB-11975, dataset 02_FTP_GRN (cases)

Dataset [2-4]

Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines.

Released: 2024 | Link: E-MTAB-11975 < ArrayExpress < BioStudies < EMBL-EBI

- **Experiment Details:**

ID	E-MTAB-11975	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	64 Subjects (31-92 yr), (prefrontal cortex tissue is excluded ➔ 24 Subjects (40-76 yr)
Published Article	Genome-Wide DNAMethylation in Early-Onset-Dementia Patients Brain Tissue and Lymphoblastoid Cell Lines, (2024). Link: https://doi.org/10.3390/ijms25105445	**Phenotype Sample Size**	**5 Subjects** (54-63 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Genetic Frontotemporal Dementia (MAPT Mutation)**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	No		

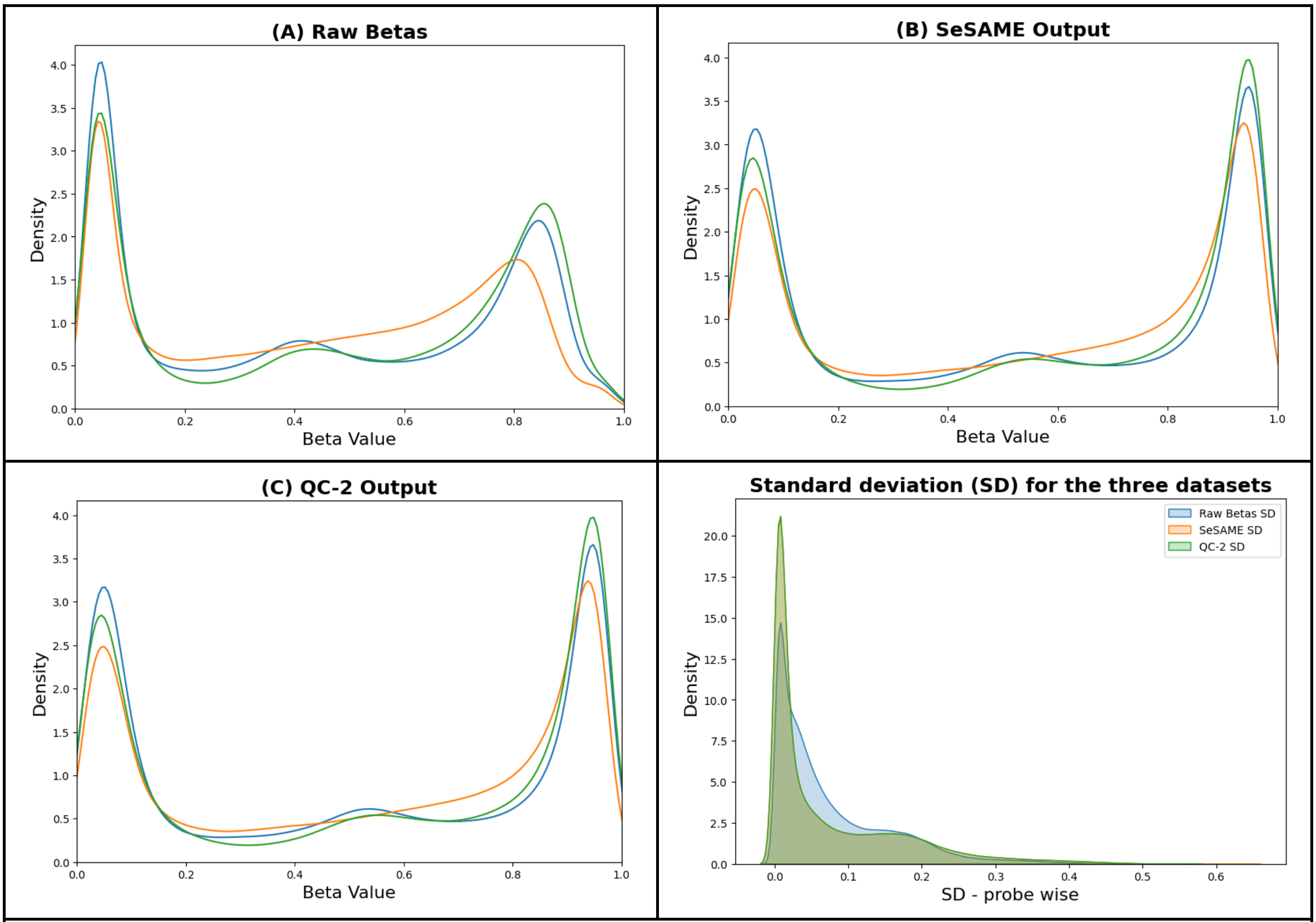


Figure 20. Beta value distribution curves for E-MTAB-11975, dataset 02_FTP_MAPT (cases)

Dataset [3-1]

DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity.

Released: 2021 | Link: [GSE164056 < Accession Display < GEO < NCBI](GSE164056)

- **Experiment Details:**

ID	GSE164056	**No. of Probes (CpG Sites)**	866553 Probes
Source	GEO Accession Viewer (National Center for Biotechnology Information NCBI)	**Total Participants in Experiment**	143 Subjects (19-50 yr),
Published Article	DNA methylation differences associated with social anxiety disorder and early life adversity, (2021). Link: https://doi.org/10.1038/s41398-021-01225-w	**Phenotype Sample Size**	**30 Subjects** (19-50 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Early Life Adveristy**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes		

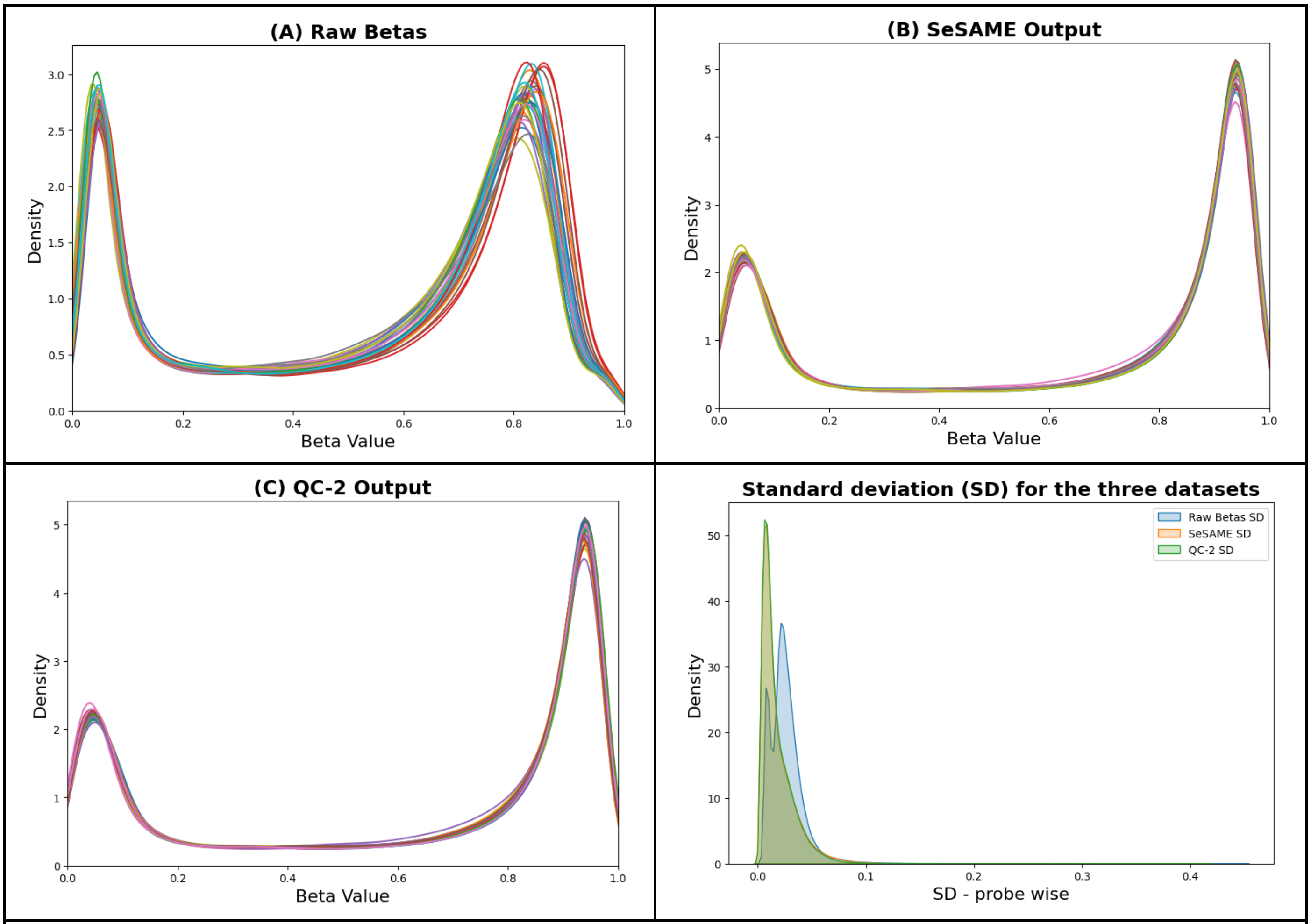


Figure 21. Beta value distribution curves for GSE164056, dataset 03_ELA (cases).

Dataset [3-2]

DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity.

Released: 2021 | Link: [GSE164056 < Accession Display < GEO < NCBI](GSE164056)

➢ Experiment Details:

ID	GSE164056	**No. of Probes (CpG Sites)**	866553 Probes
Source	GEO Accession Viewer (National Center for Biotechnology Information NCBI)	**Total Participants in Experiment**	143 Subjects (19-50 yr),
Published Article	DNA methylation differences associated with social anxiety disorder and early life adversity, (2021). Link: https://doi.org/10.1038/s41398-021-01225-w	**Phenotype Sample Size**	**35 Subjects** (19-37 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Social Anxiety Disorder**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes		

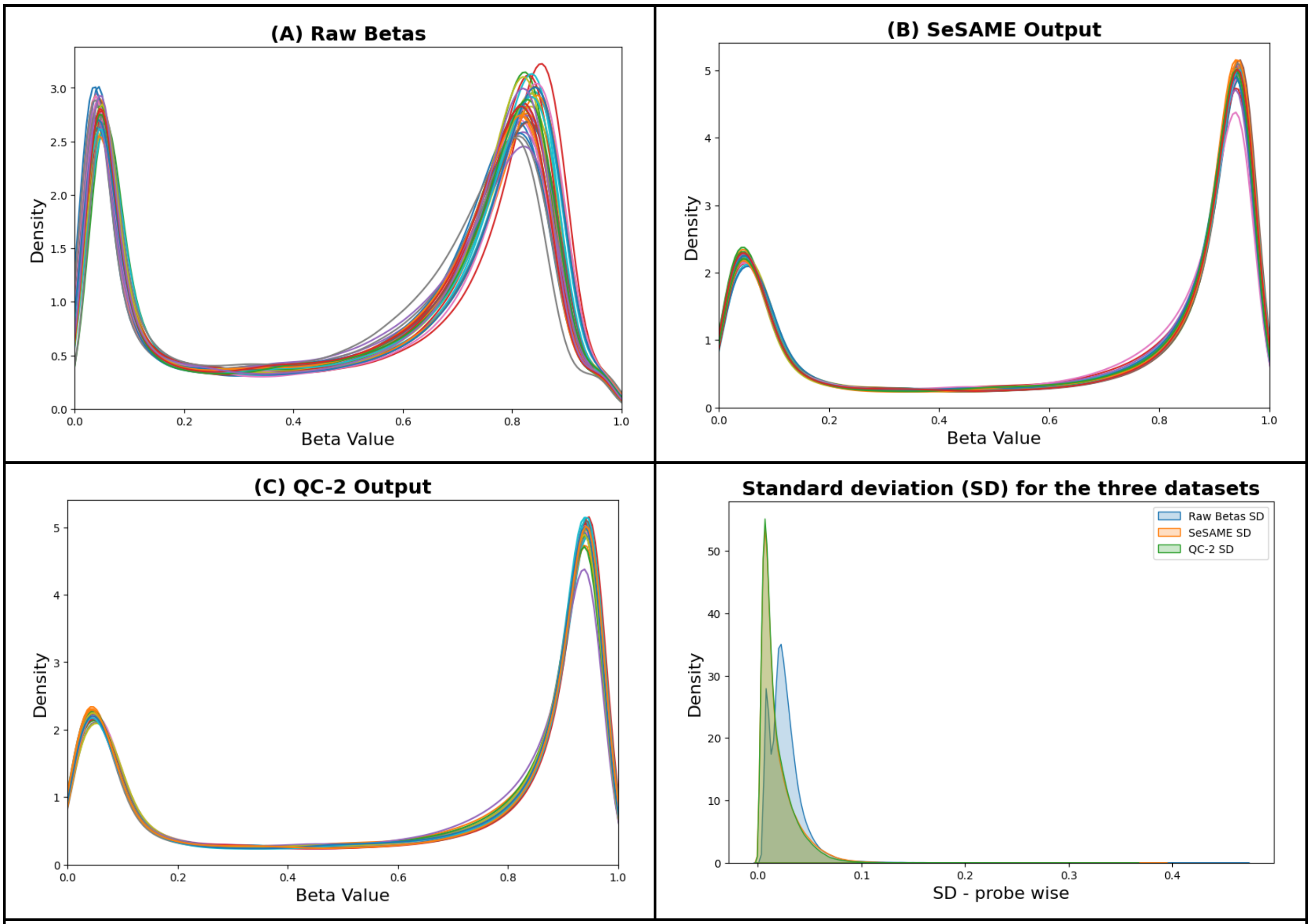


Figure 22. Beta value distribution curves for GSE164056, dataset 03_SAD (cases).

Dataset [3-3]

DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity.

Released: 2021 | Link: GSE164056 < Accession Display < GEO < NCBI

➢ **Experiment Details:**

ID	GSE164056	**No. of Probes (CpG Sites)**	866553 Probes
Source	GEO Accession Viewer (National Center for Biotechnology Information NCBI)	**Total Participants in Experiment**	143 Subjects (19-50 yr),
Published Article	DNA methylation differences associated with social anxiety disorder and early life adversity, (2021). Link: https://doi.org/10.1038/s41398-021-01225-w	**Phenotype Sample Size**	**31 Subjects** (19-45 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Social Anxiety Disorder & Early Life Adversity**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes		

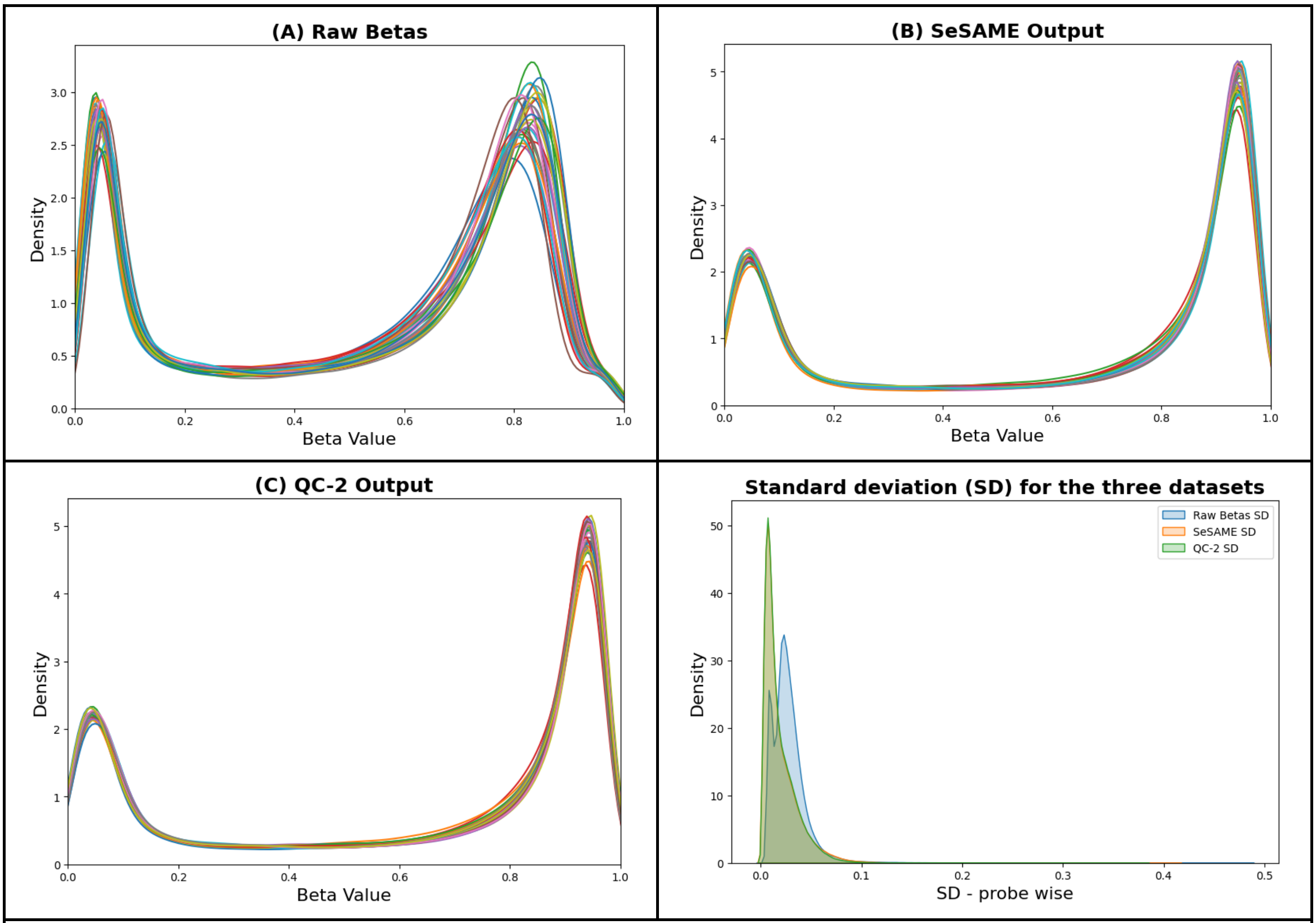


Fig. 23. Beta value distribution curves for GSE164056, dataset 03_SAD_ELA (cases).

Dataset [4]

Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-wide DNA Methylation Profiles and RNA Expression.

Released: 2016 | Link: E-MTAB-4664 < ArrayExpress < BioStudies < EMBL-EBI

➢ Experiment Details:

ID	E-MTAB-4664	**No. of Probes (CpG Sites)**	486427 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	36 Subjects (19-31 yr)
Published Article	Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-Wide DNA Methylation Profiles and RNA Expression, (2016). Link: https://doi.org/10.1089/omi.2016.0041	**Phenotype Sample Size**	**18 Subjects** (19-31 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Total Acute Sleep Deprivation**
Platform Used	Illumina Infinium HumanMethylation450 BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)		
Processed Data Available	No	**Organism Part**	Peripheral Blood

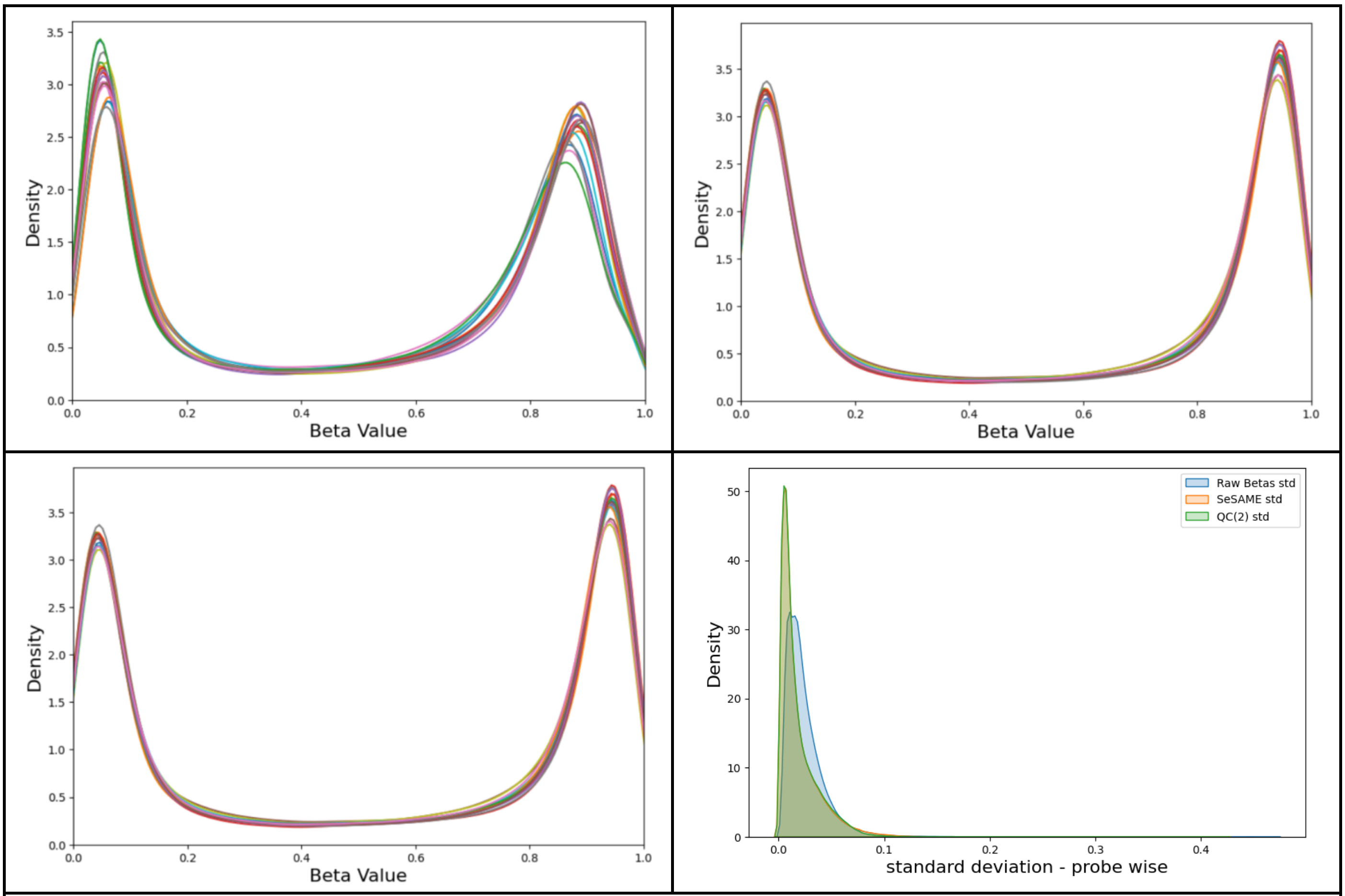


Figure 24. Normalization curves for E-MTAB-4664, dataset 04_TSD (cases)

Dataset [1] - Controls

Comparison of the methylation profiles of children with developmental language disorder and healthy control subjects.

Released: 2024 | Link: E-MTAB-13583 < ArrayExpress < BioStudies < EMBL-EBI

- **Experiment Details:**

ID	E-MTAB-13583	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	24 Subjects (3-7 yr)
Published Article	Hypomethylation of Wnt Signaling Regulator Genes in Developmental Language Disorder, (2024). Link: https://doi.org/10.2217/epi-2023-0345	**Phenotype Sample Size**	**12 Subjects** (3-7 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Healthy Controls**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes (.csv format)		

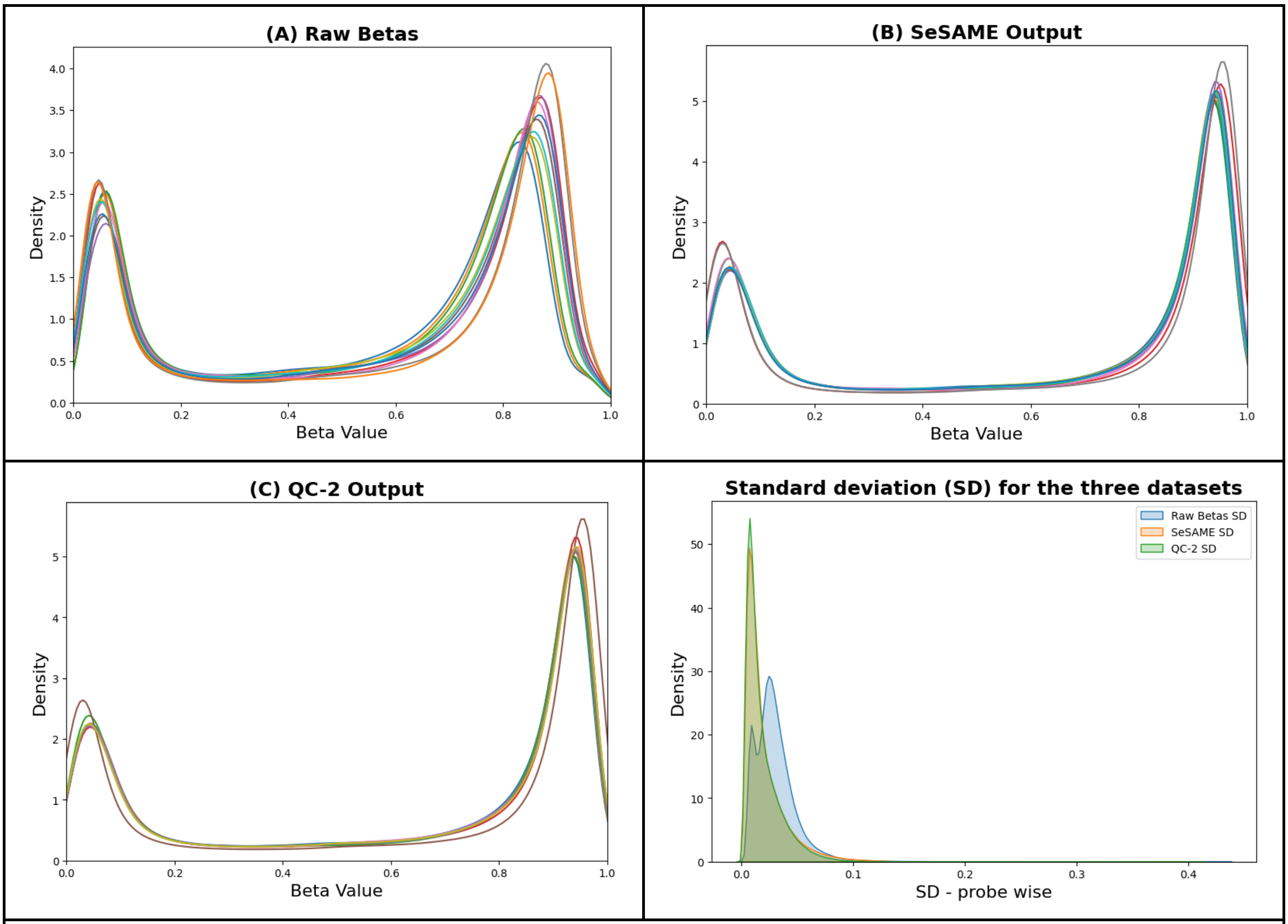


Figure 25. Beta value distribution curves for E-MTAB-13583, dataset 01_DLD (controls).

Dataset [2] - Controls

Genome-wide DNA methylation analysis identifies epigenetic differences in Alzheimer’s disease and frontotemporal dementia in brain tissue and lymphoblastoid cell lines.

Released: 2024 | Link: [E-MTAB-11975 < ArrayExpress < BioStudies < EMBL-EBI]

- **Experiment Details:**

ID	E-MTAB-11975	**No. of Probes (CpG Sites)**	866553 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	64 Subjects (31-92 yr), (prefrontal cortex tissue is excluded ➔ 24 Subjects (40-76 yr)
Published Article	Genome-Wide DNAMethylation in Early-Onset-Dementia Patients Brain Tissue and Lymphoblastoid Cell Lines, (2024). Link: https://doi.org/10.3390/ijms25105445	**Phenotype Sample Size**	**5 Subjects** (40-65 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Healthy Controls**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	No		

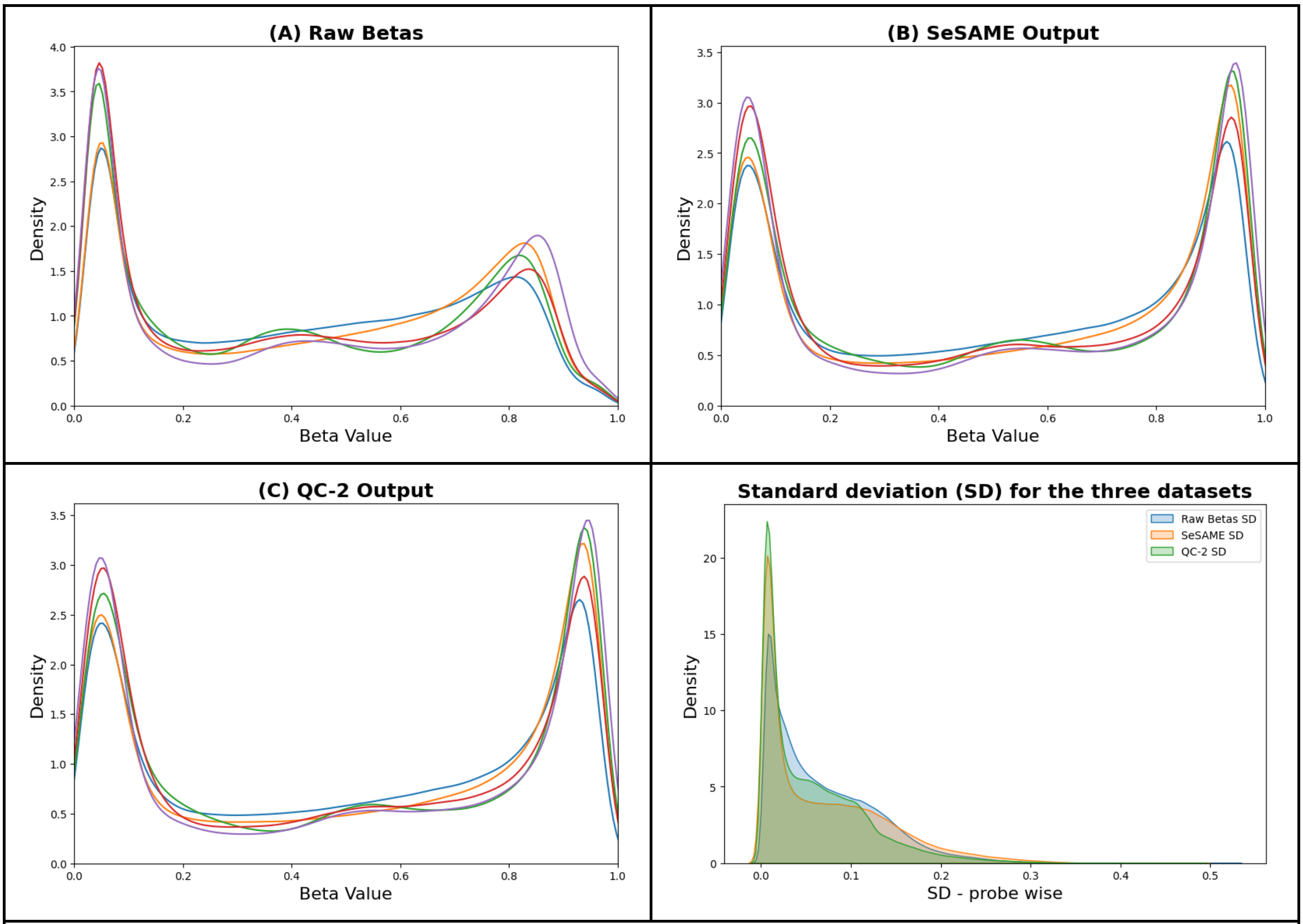


Figure 26. Beta value distribution curves for E-MTAB-11975, dataset 02_FTP (controls).

Dataset [3] - Controls

DNA Methylation Differences Associated with Social Anxiety Disorder and Early Life Adversity.

Released: 2021 | Link: GSE164056 < Accession Display < GEO < NCBI

➢ Experiment Details:

ID	GSE164056	**No. of Probes (CpG Sites)**	866553 Probes
Source	GEO Accession Viewer (National Center for Biotechnology Information NCBI)	**Total Participants in Experiment**	143 Subjects (19-50 yr)
Published Article	DNA methylation differences associated with social anxiety disorder and early life adversity, (2021). Link: https://doi.org/10.1038/s41398-021-01225-w	**Phenotype Sample Size**	**47 Subjects** (19-42 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Healthy Controls**
Platform Used	Illumina - Human Infinium Methylation EPIC BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)	**Organism Part**	Peripheral Blood
Processed Data Available	Yes		

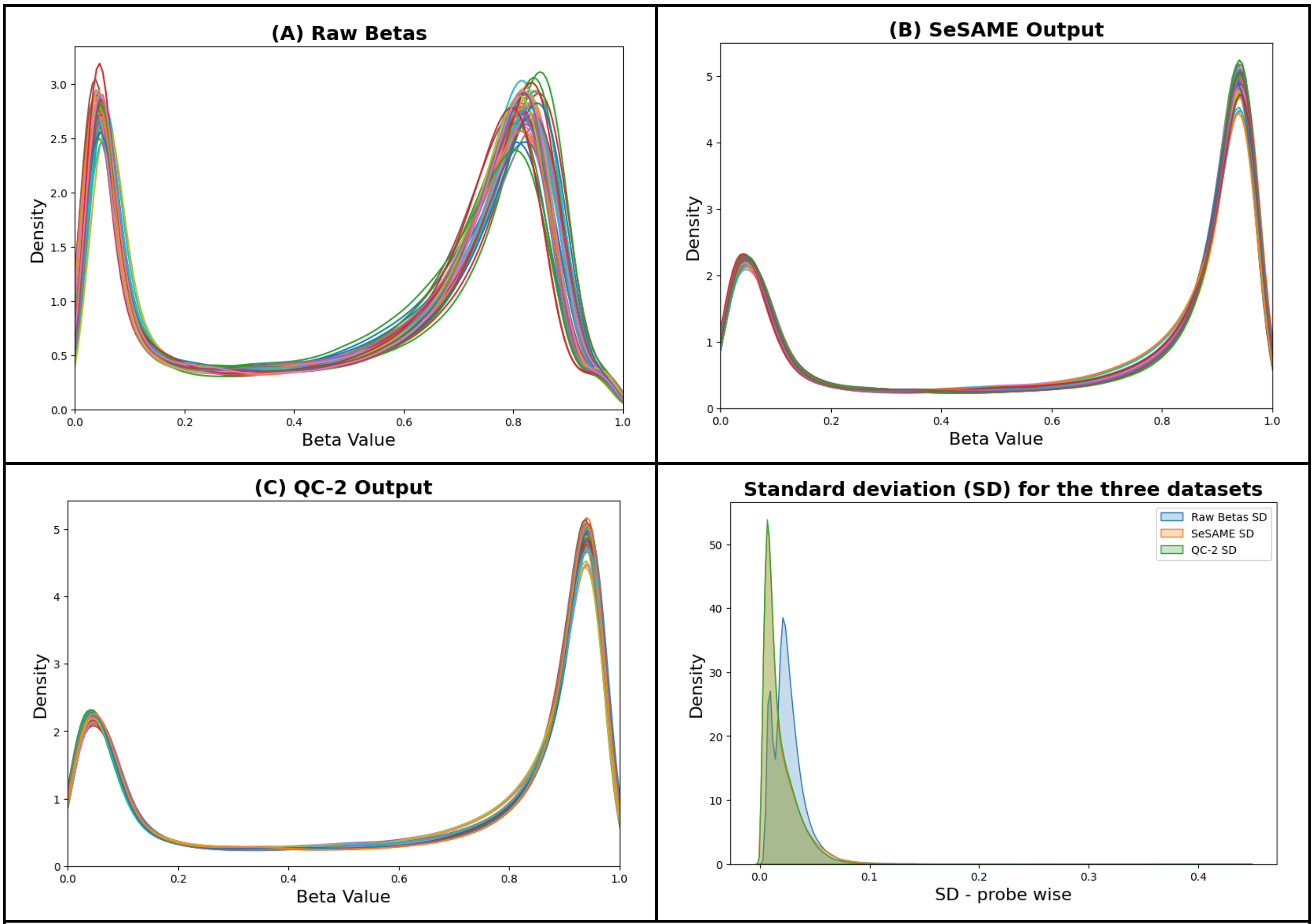


Figure 27. Beta value distribution curves for GSE164056, dataset 03_SAD (controls).

Dataset [4] - Controls

Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-wide DNA Methylation Profiles and RNA Expression.

Released: 2016 | Link: [E-MTAB-4664 < ArrayExpress < BioStudies < EMBL-EBI]()

- **Experiment Details:**

ID	E-MTAB-4664	**No. of Probes (CpG Sites)**	486427 Probes
Source	ArrayExpress (BioStudies, EMBL's European Bioinformatics Institute)	**Total Participants in Experiment**	36 Subjects (19-31 yr)
Published Article	Epigenomics of Total Acute Sleep Deprivation in Relation to Genome-Wide DNA Methylation Profiles and RNA Expression, (2016). Link: https://doi.org/10.1089/omi.2016.0041	**Phenotype Sample Size**	**18 Subjects** (19-31 yr)
Experiment Type	Methylation Profiling by Array	**Phenotype**	**Healthy Controls**
Platform Used	Illumina Infinium HumanMethylation450 BeadChip	**Species**	Homo sapiens
Raw Data Available	Yes (.idat format)		
Processed Data Available	No	**Organism Part**	Peripheral Blood

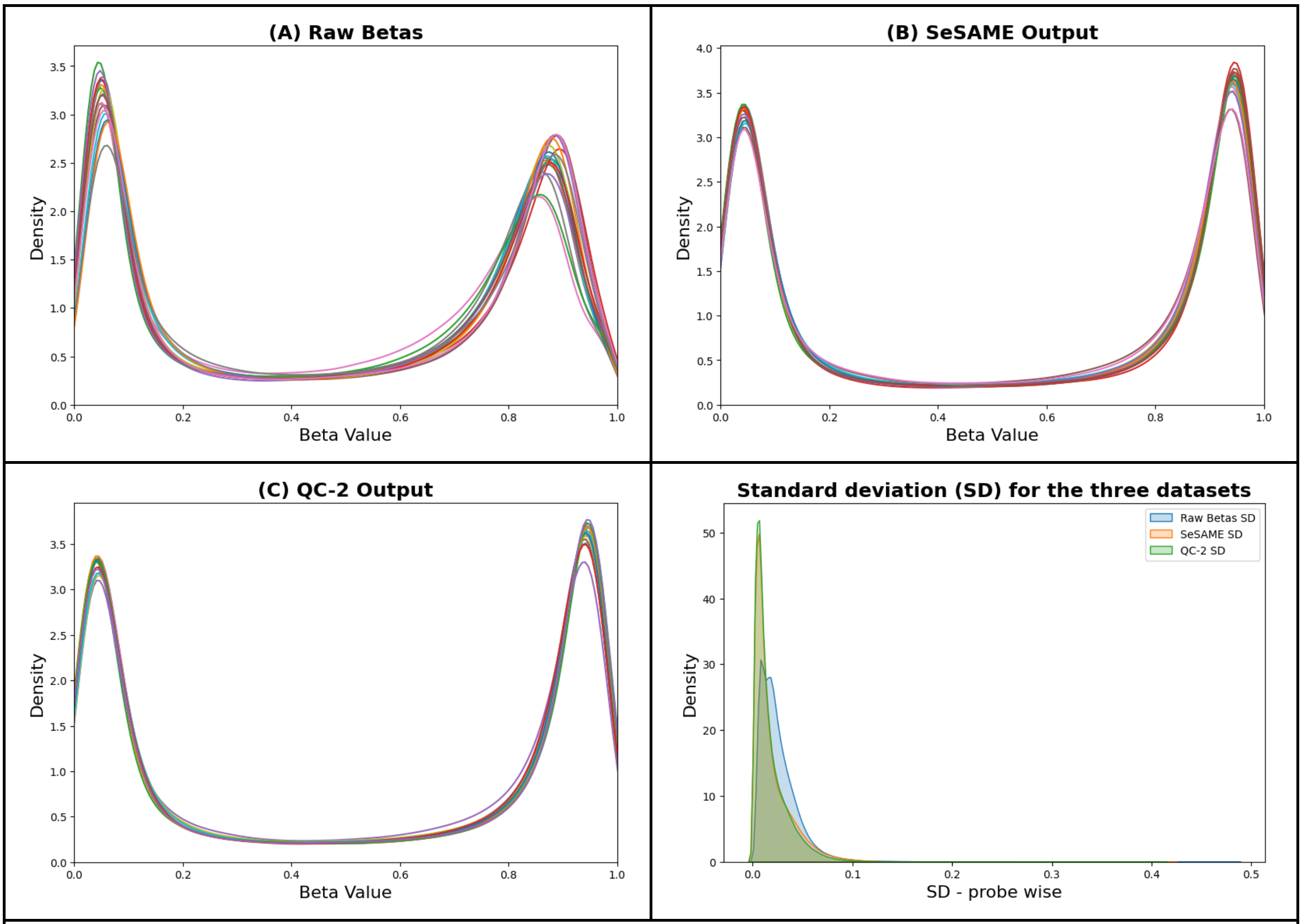


Fig. 28. Beta value distribution curves for E-MTAB-4664, dataset 04_TSD (controls).

3. Differential Analysis

Each phenotype dataset is analysed against its corresponding control dataset. A threshold of 0.05 for FDR adjusted p values is used to subset the initial amount of differentially methylated probes for each dataset. Table 18 summarizes the resulted output of this procedure.

File name	No. of probes with P val < 0.05	File name	No. of probes with adjusted P val < 0.05
01 DLD Results.csv	102588	01 DLD Significant Probes.csv	746
02 Early Alz Results.csv	32375	02 Early Alz Significant Probes.csv	489
03 Fam Alz Results.csv	38551	03 Fam Alz Significant Probes.csv	45
04 FTP GRN Results.csv	37751	04 FTP GRN Significant Probes.csv	265
05 FTP MAPT Results.csv	50554	05 FTP MAPT Significant Probes.csv	5
06 ELA Results.csv	54620	06 ELA Significant Probes.csv	42
07 SAD Results.csv	75303	07 SAD Significant Probes.csv	33
08 SAD_ELA Results.csv	42821	08 SAD_ELA Significant Probes.csv	9
09 TSD Results.csv	15484	09 TSD Significant Probes.csv	0

Table 18. The initial differentially methylated probes are filtered to include only those that passes the FDR correction of 0.05.

All DMPs extracted from the 9 datasets are available in the supplementary material (Limma DMPs). The DMPs for each phenotype are mapped to their genomic information using Illumina manifests (HG19) and are provided in a separate folder within the supplementary material (*Probe Info*).

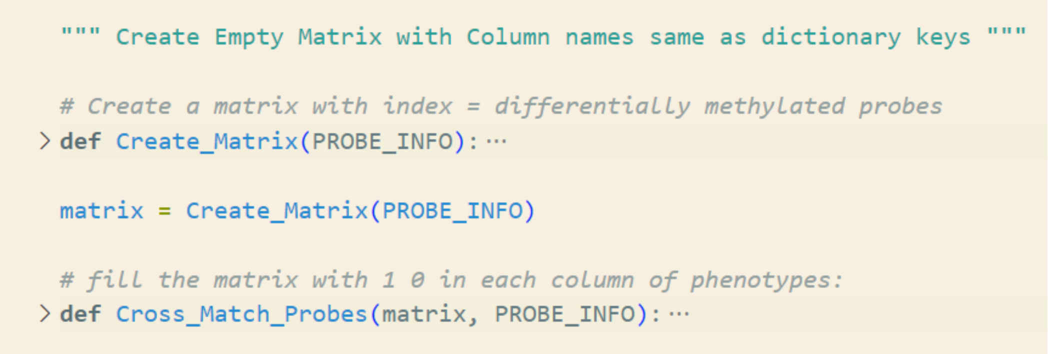
Note: As mentioned in the methodology (Chapter 2), the DMPs included from the 04_TSD dataset were not FDR-corrected, as no probes passed the FDR correction. Therefore, the initially detected differentially methylated probes were used for analysis.

4. Cross Comparison

Following the methodology in (chapter 2), the comparison is carried out over three sections:

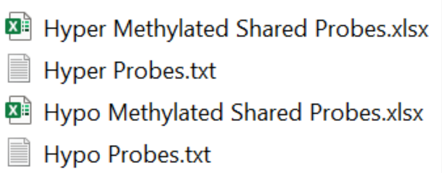
A. Shared probes

A matrix of 1 and 0 values is created to check which phenotype contains a certain probe (1 means probe is found in the list of DMPs for that phenotype). All differentially methylated probes from all phenotypes are combined into one column, while phenotypes represent the remaining columns, as demonstrated in (Tables 20, 21). This procedure is performed twice: once for hypermethylated probes (Table 20) and once for hypomethylated probes (Table 21).



```
""" Create Empty Matrix with Column names same as dictionary keys """

# Create a matrix with index = differentially methylated probes
def Create_Matrix(PROBE_INFO):···

matrix = Create_Matrix(PROBE_INFO)

# fill the matrix with 1 0 in each column of phenotypes:
def Cross_Match_Probes(matrix, PROBE_INFO):···
```

Output*:



Hyper Methylated Shared Probes.xlsx
Hyper Probes.txt
Hypo Methylated Shared Probes.xlsx
Hypo Probes.txt

*Availability: Supplementary Material (*Shared Probes*)

ProbeList	(1) DLD	(2) Early Alz	(3) Fam Alz	(4) FTP GRN	(5) FTP MAPT	(6) ELA	(7) SAD ELA	(8) SAD	(9) TSD	phenoCount
cg09945813	0	1	0	1	0	0	0	0	0	2
cg21374153	0	0	1	1	0	0	0	0	0	2
cg26187194	0	1	0	1	0	0	0	0	0	2

Table 20. Hypermethylated probes that are shared among phenotypes

ProbeList	(1) DLD	(2) Early Alz	(3) Fam Alz	(4) FTP GRN	(5) FTP MAPT	(6) ELA	(7) SAD ELA	(8) SAD	(9) TSD	phenoCount
cg15454820	0	1	1	0	1	0	0	0	0	3
cg04586579	0	1	1	1	0	0	0	0	0	3
cg25782229	0	1	0	1	0	0	0	0	0	2
cg24291747	0	1	0	1	0	0	0	0	0	2
cg06952310	0	1	0	1	0	0	0	0	0	2
cg20022454	0	0	0	0	0	1	0	1	0	2
cg08259796	0	0	0	0	0	1	1	0	0	2
cg15370054	0	1	0	1	0	0	0	0	0	2
cg20824804	0	0	1	1	0	0	0	0	0	2
cg07529625	0	1	0	1	0	0	0	0	0	2
cg16288713	0	1	1	0	0	0	0	0	0	2
cg00531088	0	0	0	0	0	1	1	0	0	2
cg07257571	0	1	0	1	0	0	0	0	0	2
cg22579590	0	1	0	1	0	0	0	0	0	2
cg11335335	0	1	1	0	0	0	0	0	0	2

Table 21. Hypomethylated probes that are shared among phenotypes

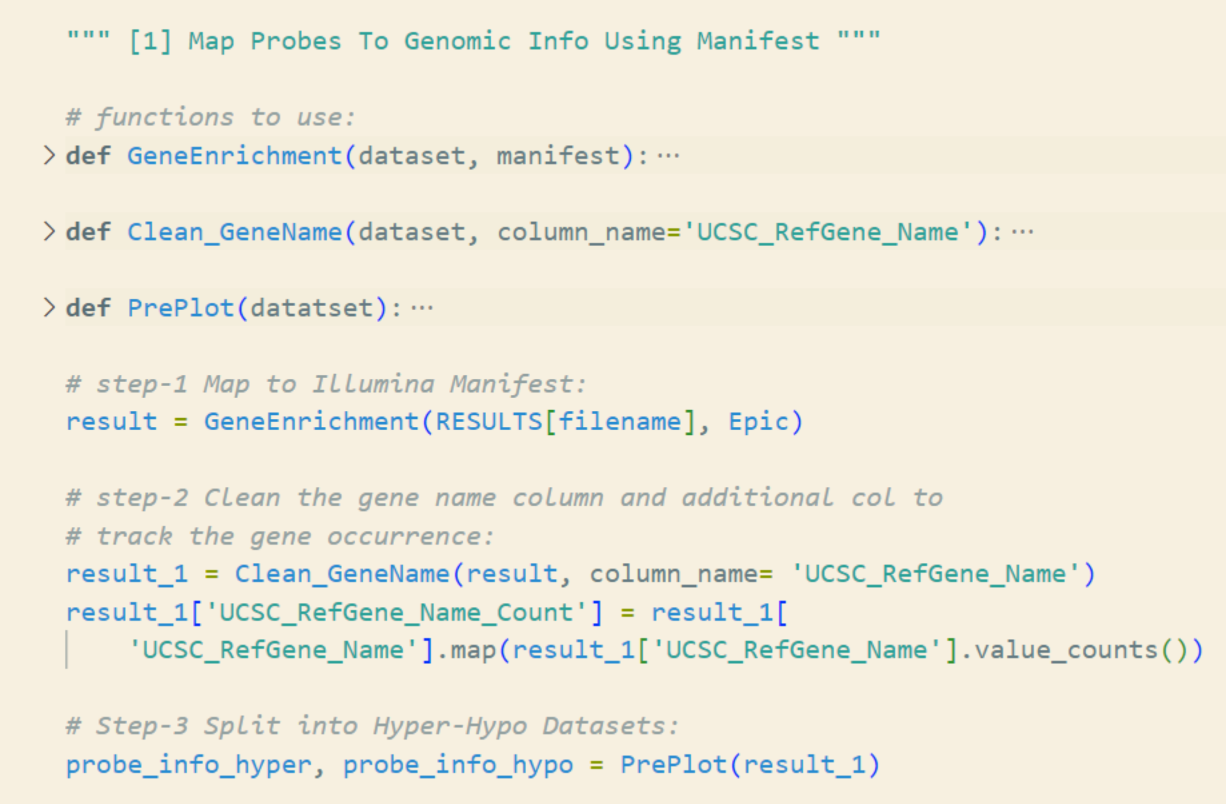
The column *phenoCount* is created to track the number of phenotypes that are associated with one probe. Since TSD dataset was excluded from FDR correction, a conservative approach to include only probes that are shared with TSD and at least 2 phenotypes. However, none of the probes achieves such criteria (Tables 20, 21).

Source code and output for the three sections (A. Shared probes, B. Shared Genes, C. shared Regions) is available in supplementary material (Folders: *Shared Probes, Shared Genes, Shared Regions*)

Next, shared genes...

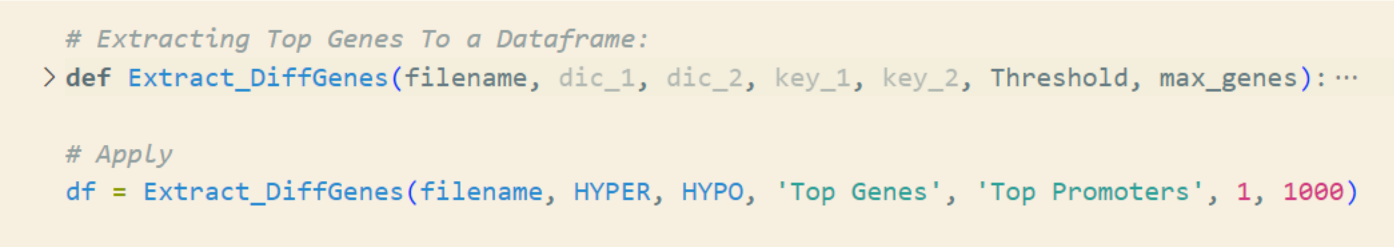
B. Shared Genes

Differentially methylated probes were mapped and sorted based on gene occurrence (i.e. genes that are mapped to multiple probes are ranked higher).



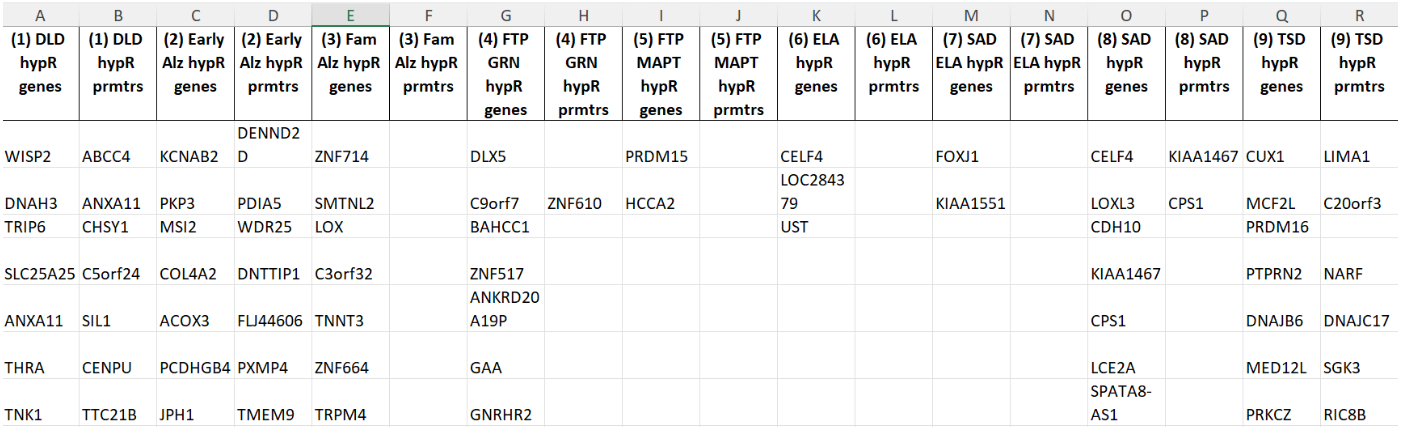
```
""" [1] Map Probes To Genomic Info Using Manifest """

# functions to use:
def GeneEnrichment(dataset, manifest):⋯

def Clean_GeneName(dataset, column_name='UCSC_RefGene_Name'):⋯

def PrePlot(datatset):⋯

# step-1 Map to Illumina Manifest:
result = GeneEnrichment(RESULTS[filename], Epic)

# step-2 Clean the gene name column and additional col to
# track the gene occurrence:
result_1 = Clean_GeneName(result, column_name= 'UCSC_RefGene_Name')
result_1['UCSC_RefGene_Name_Count'] = result_1[
    'UCSC_RefGene_Name'].map(result_1['UCSC_RefGene_Name'].value_counts())

# Step-3 Split into Hyper-Hypo Datasets:
probe_info_hyper, probe_info_hypo = PrePlot(result_1)
```

Two lists of genes are extracted for each phenotype: hypermethylated and hypomethylated genes.



```
# Extracting Top Genes To a Dataframe:
def Extract_DiffGenes(filename, dic_1, dic_2, key_1, key_2, Threshold, max_genes):⋯

# Apply
df = Extract_DiffGenes(filename, HYPER, HYPO, 'Top Genes', 'Top Promoters', 1, 1000)
```

This procedure is done for the 9 phenotypes. Fig. 29 shows an abstract of the final output. In the next step, all available genes are combined under one column in order to create a matrix that shows gene availability in each phenotype (Table 22 for hypermethylated genes combined datasets, and Table 23 for hypomethylated genes combined dataset).



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
(1) DLD hypR genes	**(1) DLD hypR prmtrs**	**(2) Early Alz hypR genes**	**(2) Early Alz hypR prmtrs**	**(3) Fam Alz hypR genes**	**(3) Fam Alz hypR prmtrs**	**(4) FTP GRN hypR genes**	**(4) FTP GRN hypR prmtrs**	**(5) FTP MAPT hypR genes**	**(5) FTP MAPT hypR prmtrs**	**(6) ELA hypR genes**	**(6) ELA hypR prmtrs**	**(7) SAD ELA hypR genes**	**(7) SAD ELA hypR prmtrs**	**(8) SAD hypR genes**	**(8) SAD hypR prmtrs**	**(9) TSD hypR genes**	**(9) TSD hypR prmtrs**
WISP2	ABCC4	KCNAB2	DENND2D	ZNF714		DLX5		PRDM15		CELF4		FOXJ1		CELF4	KIAA1467	CUX1	LIMA1
DNAH3	ANXA11	PKP3	PDIA5	SMTNL2		C9orf7	ZNF610	HCCA2		LOC284379		KIAA1551		LOXL3	CPS1	MCF2L	C20orf3
TRIP6	CHSY1	MSI2	WDR25	LOX		BAHCC1				UST				CDH10		PRDM16	
SLC25A25	C5orf24	COL4A2	DNTTIP1	C3orf32		ZNF517								KIAA1467		PTPRN2	NARF
ANXA11	SIL1	ACOX3	FLJ44606	TNNT3		ANKRD20A19P								CPS1		DNAJB6	DNAJC17
THRA	CENPU	PCDHGB4	PXMP4	ZNF664		GAA								LCE2A		MED12L	SGK3
TNK1	TTC21B	JPH1	TMEM9	TRPM4		GNRHR2								SPATA8-AS1		PRKCZ	RIC8B

Figure 29. Abstract of the output after gene enrichment for all phenotypes (hyper methylated genes/ promoters). Complete Output is available in supplementary material (*Shared Genes*)

Genes	DLD	Early Alz	Fam Alz	FTP GRN	FTP MAPT	ELA	SAD ELA	SaD	TSD	phenoCount
MAD1L1	1	1	1	0	0	0	0	0	0	3
PTPRN2	0	1	0	0	0	0	0	0	1	2
EBF3	0	1	0	0	0	0	0	0	1	2
SLC39A11	0	1	0	1	0	0	0	0	0	2
DIP2C	0	1	0	0	0	0	0	0	1	2
CAMTA1	0	0	0	1	0	0	0	0	1	2
CELF4	0	0	0	0	0	1	0	1	0	2
MED12L	1	0	0	0	0	0	0	0	1	2
IFRD1	0	1	0	0	0	0	0	0	1	2
HDAC4	0	0	0	1	0	0	0	0	1	2
DENND2D	1	1	0	0	0	0	0	0	0	2
KIAA0182	1	0	0	0	0	0	0	0	1	2
MSI2	0	1	0	0	0	0	0	0	1	2
N4BP1	1	1	0	0	0	0	0	0	0	2
GATAD2A	1	1	0	0	0	0	0	0	0	2
ASAP2	1	0	0	0	0	0	0	0	1	2
PRDM16	1	0	0	0	0	0	0	0	1	2
FGFR2	0	0	1	0	0	0	0	0	1	2
GPD2	0	1	0	0	0	0	0	0	1	2
KCNQ1	1	1	0	0	0	0	0	0	0	2
JADE1	1	1	0	0	0	0	0	0	0	2
IER3	1	0	0	0	0	0	0	0	1	2
NRP1	1	0	0	0	0	0	0	0	1	2

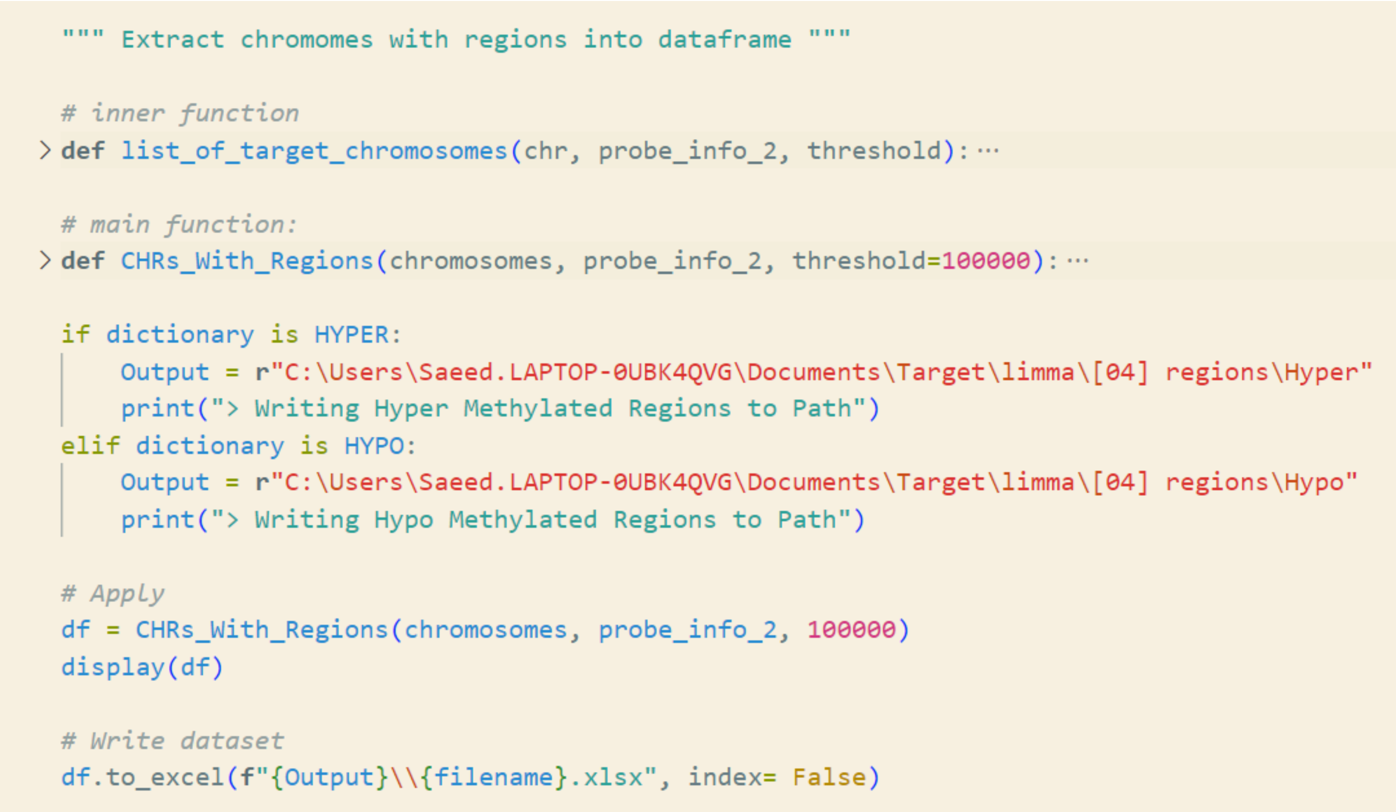
Table 22. Hypermethylated genes that are shared among phenotypes.

Genes	DLD	Early Alz	Fam Alz	FTP GRN	FTP MAPT	ELA	SAD ELA	SaD	TSD	phenoCount
CD81	1	1	0	1	0	0	0	0	0	3
SDR42E1	0	1	1	1	0	0	0	0	0	3
MAD1L1	1	1	0	0	0	0	0	0	1	3
LOC101928708	0	0	1	1	0	0	0	0	0	2
TRPS1	0	0	0	1	0	0	1	0	0	2
LPCAT1	0	1	0	0	0	0	0	0	1	2
RNF219	0	1	1	0	0	0	0	0	0	2
SLC6A16	1	1	0	0	0	0	0	0	0	2
RPS6KA2	1	0	0	0	0	0	0	0	1	2
PIP4K2A	0	1	0	1	0	0	0	0	0	2
FOXN3	0	1	0	1	0	0	0	0	0	2
MUT	0	0	0	0	0	1	1	0	0	2
CNTNAP2	1	0	1	0	0	0	0	0	0	2
ARHGAP26	1	0	0	1	0	0	0	0	0	2
HOOK2	0	1	0	1	0	0	0	0	0	2
ADARB2	1	0	0	0	0	0	0	0	1	2
WT1	0	1	0	1	0	0	0	0	0	2
SPATA4	0	1	0	1	0	0	0	0	0	2
CCDC26	0	1	0	1	0	0	0	0	0	2
NCAN	0	1	0	1	0	0	0	0	0	2
EPM2AIP1	0	0	0	0	0	1	1	0	0	2
SP100	1	0	0	1	0	0	0	0	0	2
DRD4	0	1	1	0	0	0	0	0	0	2
HIVEP3	1	1	0	0	0	0	0	0	0	2

Table 23. Hypomethylated genes/ promoters that are shared among phenotypes.

C. Shared Regions

After mapping the differentially methylated probes to their genomic coordinates, clusters of regions were detected using a maximum threshold of 100,000 bp. A data frame that contains chromosome number in the first column is created and the clusters of regions are appended as columns. An abstract of the output is available in Table 24. Each phenotype has two datasets similar to Table 24 (one for hyper- and one for hypomethylated regions). The mean average for each pair of coordinates (e.g., min_coord_1 and min_coord_2) is calculated for each dataset. Datasets from 9 phenotypes are then combined into two datasets: hypermethylated regions and hypomethylated regions (Supplementary Material: *Shared Regions*).



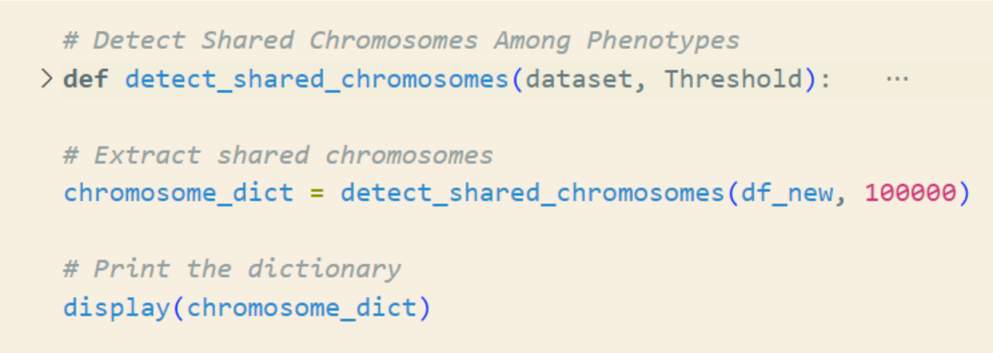
```
""" Extract chromomes with regions into dataframe """

# inner function
def list_of_target_chromosomes(chr, probe_info_2, threshold):⋯

# main function:
def CHRs_With_Regions(chromosomes, probe_info_2, threshold=100000):⋯

if dictionary is HYPER:
    Output = r"C:\Users\Saeed.LAPTOP-0UBK4QVG\Documents\Target\limma\[04] regions\Hyper"
    print("> Writing Hyper Methylated Regions to Path")
elif dictionary is HYPO:
    Output = r"C:\Users\Saeed.LAPTOP-0UBK4QVG\Documents\Target\limma\[04] regions\Hypo"
    print("> Writing Hypo Methylated Regions to Path")

# Apply
df = CHRs_With_Regions(chromosomes, probe_info_2, 100000)
display(df)

# Write dataset
df.to_excel(f"{Output}\\{filename}.xlsx", index= False)
```

chr	min_coord_1	max_coord_1	min_coord_2	max_coord_2	min_3,4, etc..	max_3,4, etc..
1	6420713	6454339	17766917	17829087	//	//
2	47077192	47077388			//	//
3	100581781	100594209	150996563	151036761	//	//
12	26348011	26349129	32638669	32654929	//	//
6	14117402	14117423	75898357	75922610	//	//
16	2058189	2058701	21161842	21162212	//	//
7	7142996	7162892	100463206	100464145	//	//
10	81946545	81965771	82247853	82265445	//	//
5	10522197	10601638			//	//
17	7283774	7283897	26683926	26699551	//	//
4	55618746	55650446			//	//
19	1068561	1074425	46932069	46946599	//	//
14	51288521	51288740	104345945	104397864	//	//
15	72409092	72409169	86296229	86296274	//	//
8	134307728	134369320			//	//
9	130860583	130866500			//	//
20	43343304	43343997			//	//
18	48404401	48404491			//	//

Table 24. Abstract of output for one dataset that shows the detected clusters in each chromosome.

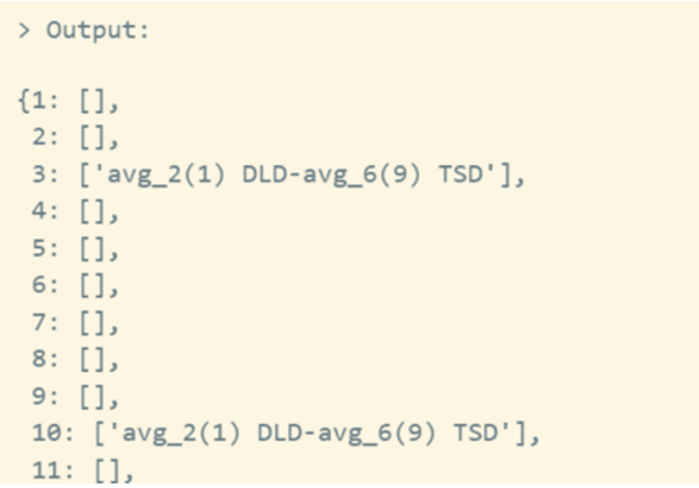
Each dataset is then inspected for regions that share the same chromosome and are located within a pre-determined threshold of 100K base pairs*. The output** of detected regions is represented in a dictionary where chromosomes are the keys and phenotype names are the values. If any pairs are detected, another function is used to get the corresponding coordinates***. An example of this procedure (from hypermethylated regions) is demonstrated below:

*Detection of phenotypes that shares close regions

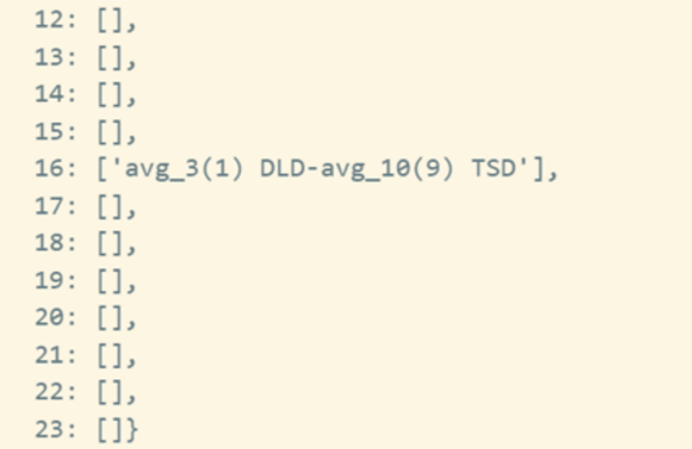


```
# Detect Shared Chromosomes Among Phenotypes
> def detect_shared_chromosomes(dataset, Threshold):   ···

# Extract shared chromosomes
chromosome_dict = detect_shared_chromosomes(df_new, 100000)

# Print the dictionary
display(chromosome_dict)
```

**Output of previous code:

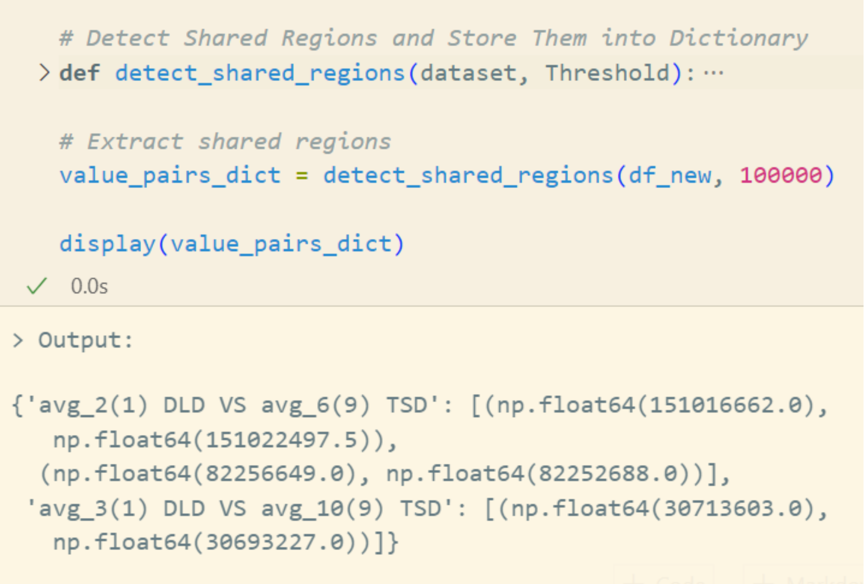


```
> Output:

{1: [],
 2: [],
 3: ['avg_2(1) DLD-avg_6(9) TSD'],
 4: [],
 5: [],
 6: [],
 7: [],
 8: [],
 9: [],
 10: ['avg_2(1) DLD-avg_6(9) TSD'],
 11: [],
```



```
 12: [],
 13: [],
 14: [],
 15: [],
 16: ['avg_3(1) DLD-avg_10(9) TSD'],
 17: [],
 18: [],
 19: [],
 20: [],
 21: [],
 22: [],
 23: []}
```

***Next, Extract the coordinates of detected phenotypes…



```
# Detect Shared Regions and Store Them into Dictionary
> def detect_shared_regions(dataset, Threshold):···

# Extract shared regions
value_pairs_dict = detect_shared_regions(df_new, 100000)

display(value_pairs_dict)
✓ 0.0s
```

```
> Output:

{'avg_2(1) DLD VS avg_6(9) TSD': [(np.float64(151016662.0),
   np.float64(151022497.5)),
  (np.float64(82256649.0), np.float64(82252688.0))],
 'avg_3(1) DLD VS avg_10(9) TSD': [(np.float64(30713603.0),
   np.float64(30693227.0))]}
```

How significant probes are selected?

From each of the previous sections (A. Shared probes, B. Shared genes, C. Shared regions), a set of significant probes/ genes is extracted based on the following criteria:

- Overall logFC value which represent the magnitude of differentiation.
- Number of probes mapping to a specific gene.
- Number of phenotypes sharing a specific probe/gene.
- Number of probes/phenotypes sharing a similar region.

The top differentiated candidates are presented in the results section, where the probes are visualized with detailed information. All initial results are provided in the supplementary material, along with the source code for reproducibility (Supplementary Material: *Limma DMPs, Probe Info*).

End of Chapter 3 | Workflow
Next: Chapter 4 | Results and Discussion

1. Results

To distinguish between hypermethylated and hypomethylated probes, figures were highlighted in green in sections 1.1, 1.2, 1.3 for hypermethylation and highlighted in yellow in sections 1.4, 1.5, and 1.6 for hypomethylation.

1.1. Hypermethylated probe ids that are shared among 2 or more phenotypes:

The maximum match detected in hyper methylated probes (Table 20) was two phenotypes. The probes that achieve this match are 'cg09945813', 'cg21374153', 'cg26187194' as shown in Figures (30, 31, 32) respectively.

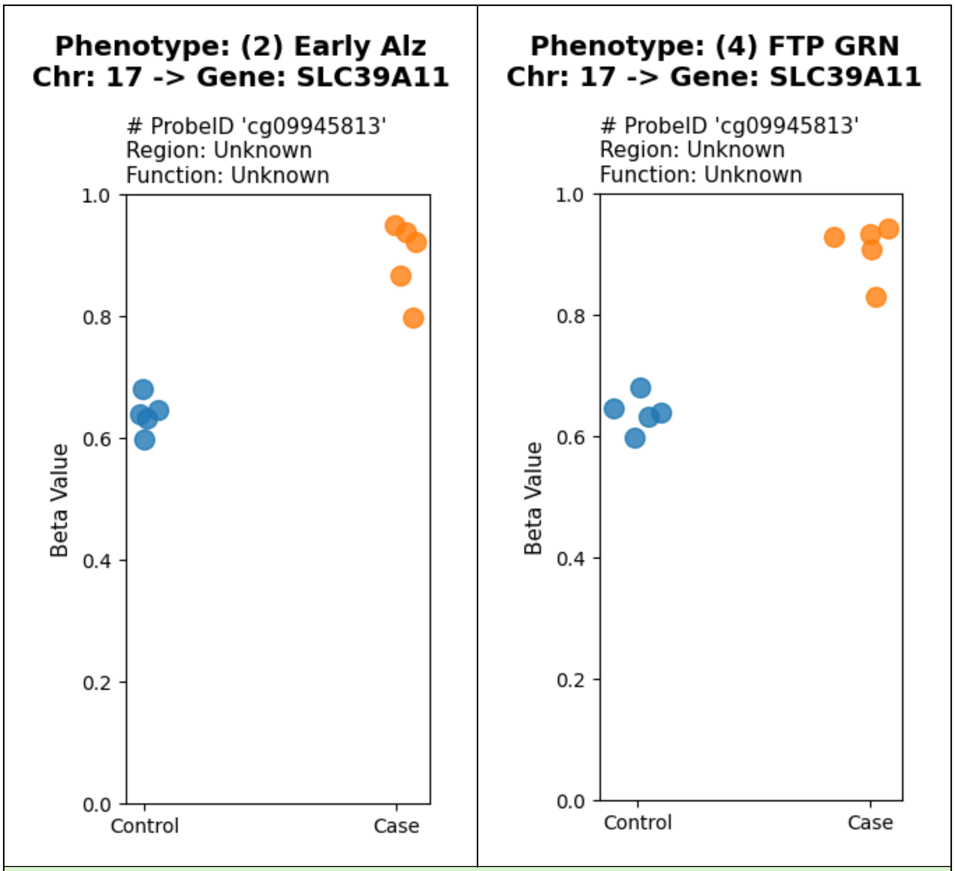


Figure 30. Probe ID 'cg09945813'

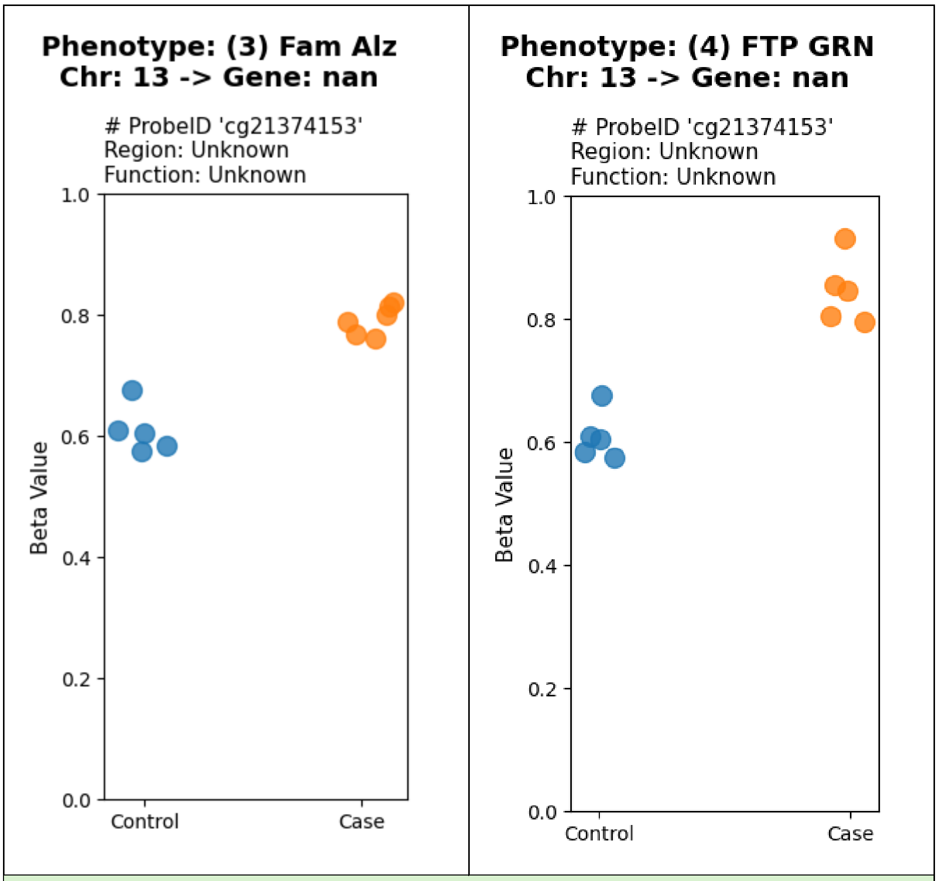


Figure 31. Probe ID 'cg21374153'

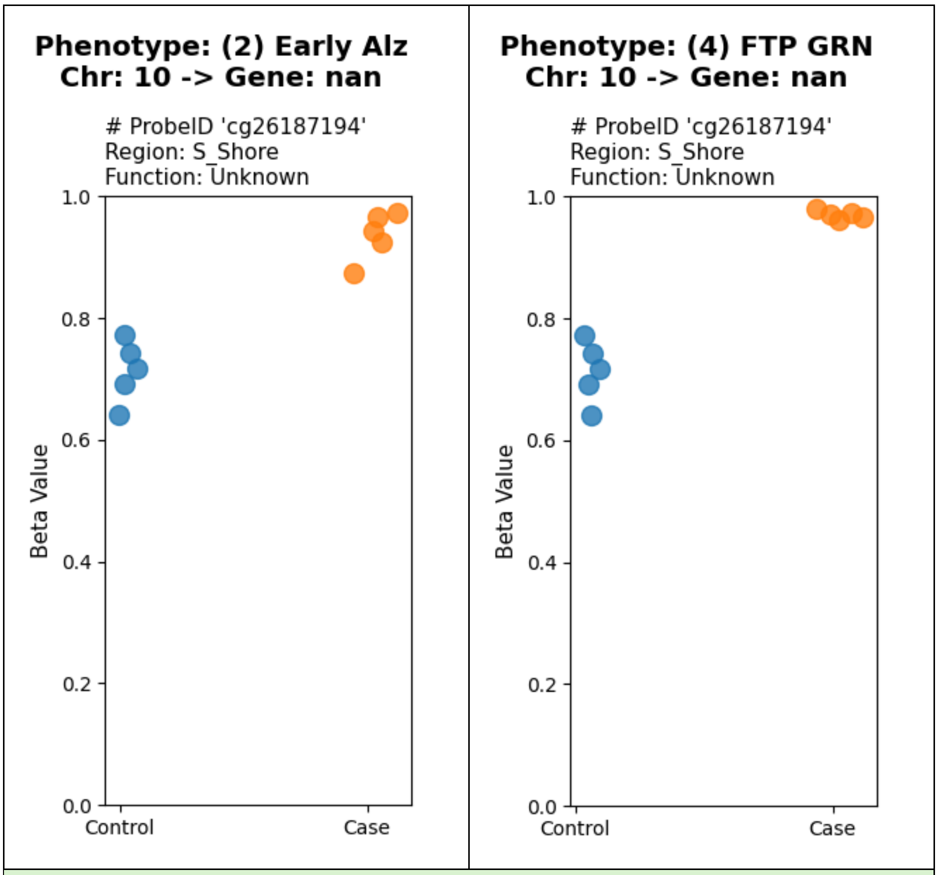


Figure 32. Probe ID 'cg26187194'

1.2. Hypermethylated probes that share the same gene among 2 or more phenotypes: Figures (33, 34, 35, 36).

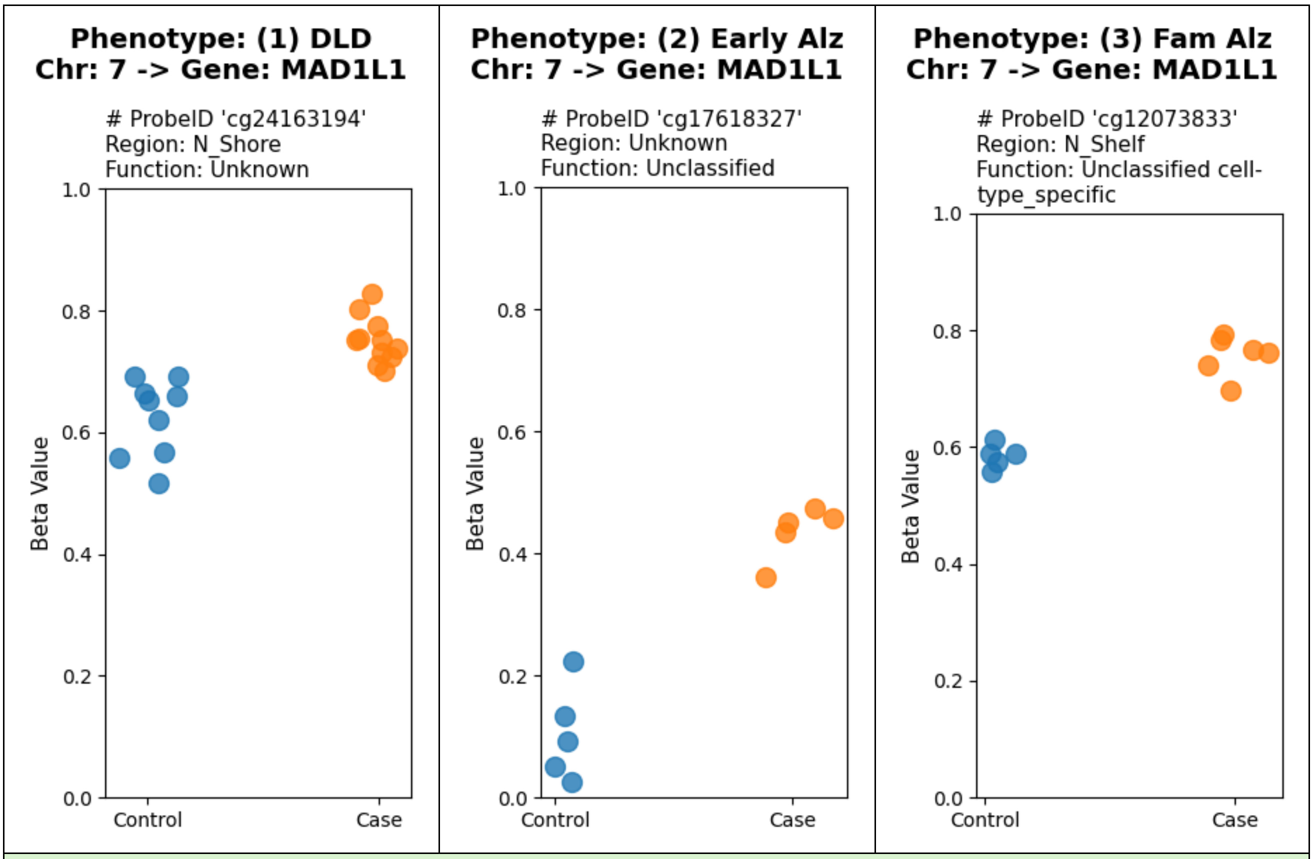


Figure 33. Gene: MAD1L1

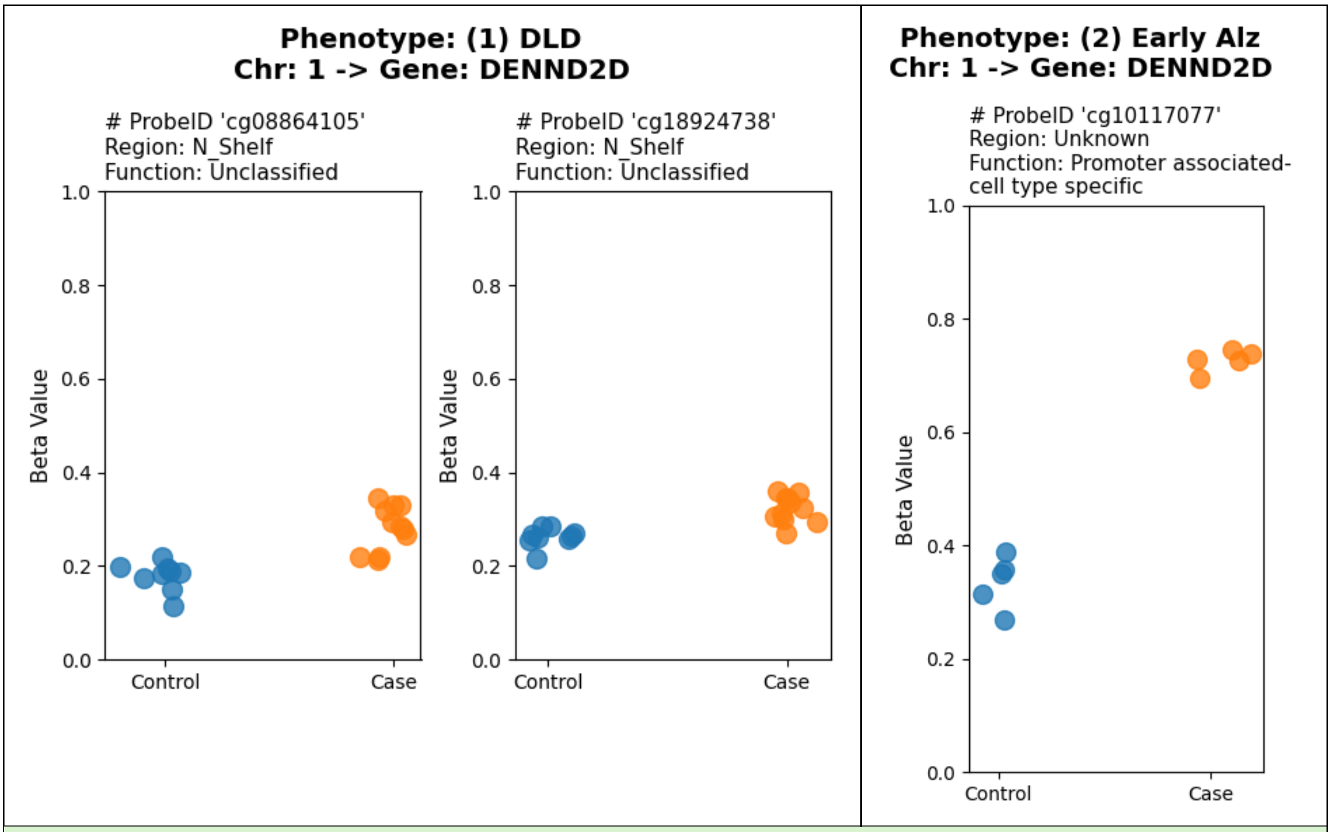


Figure 34. Gene: DENND2D

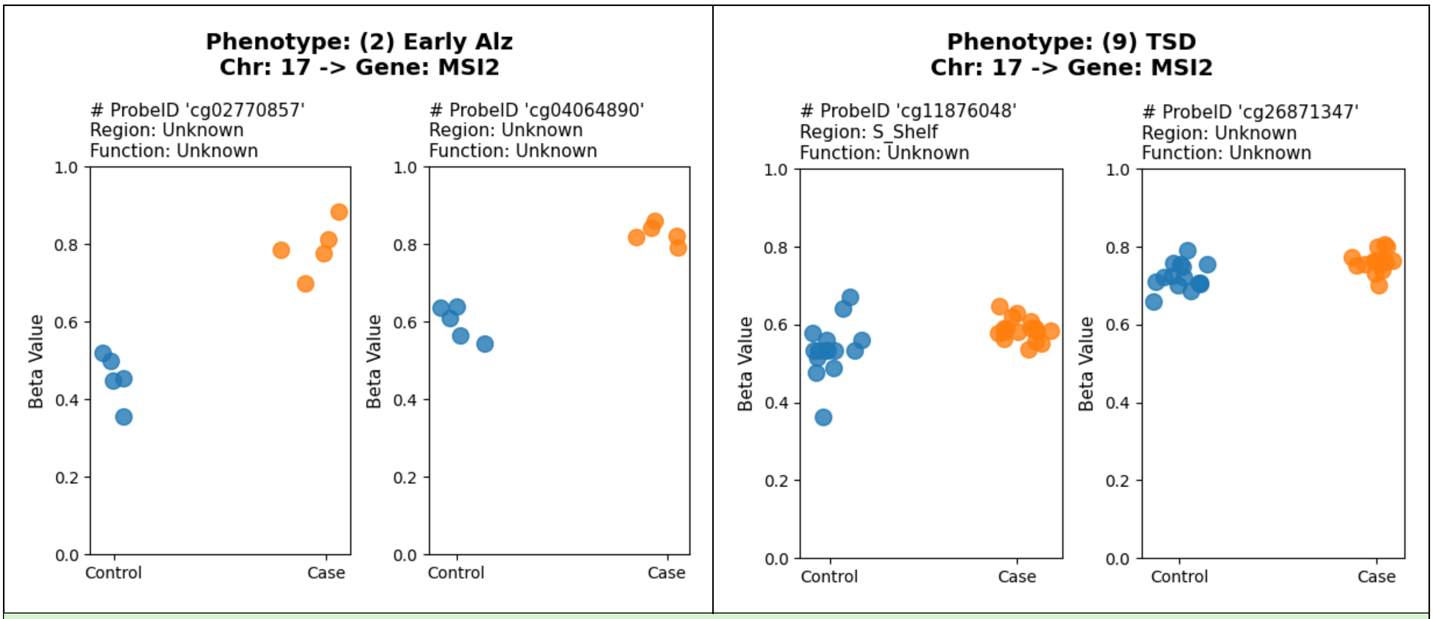


Figure 35. Gene: MSI2

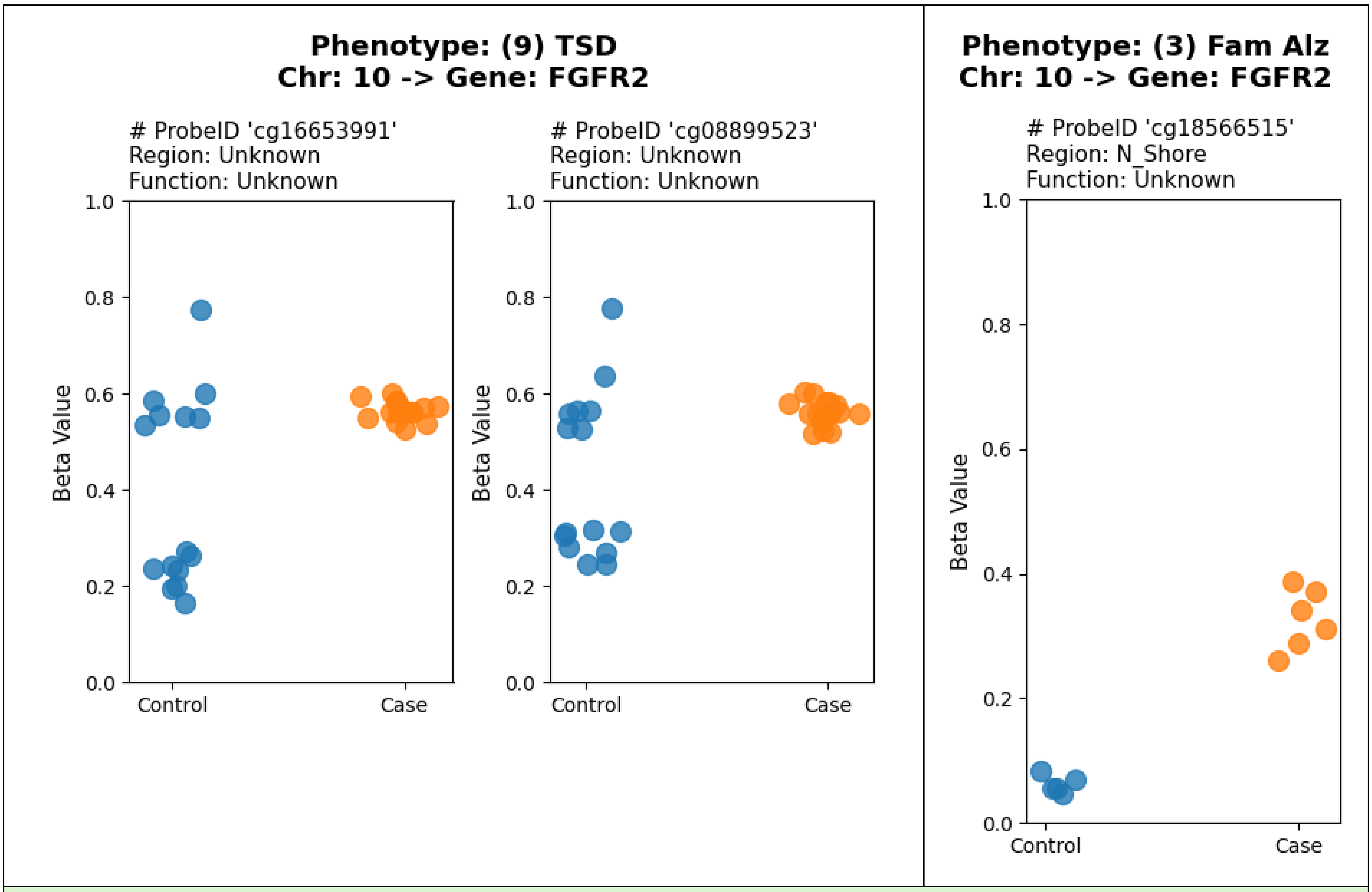


Figure 36. Gene: FGFR2

1.3. Hypermethylated probes in nearby regions (< 100K bp) shared between two or more phenotypes: (Figures 37, 38, 39)

Chromosome 3

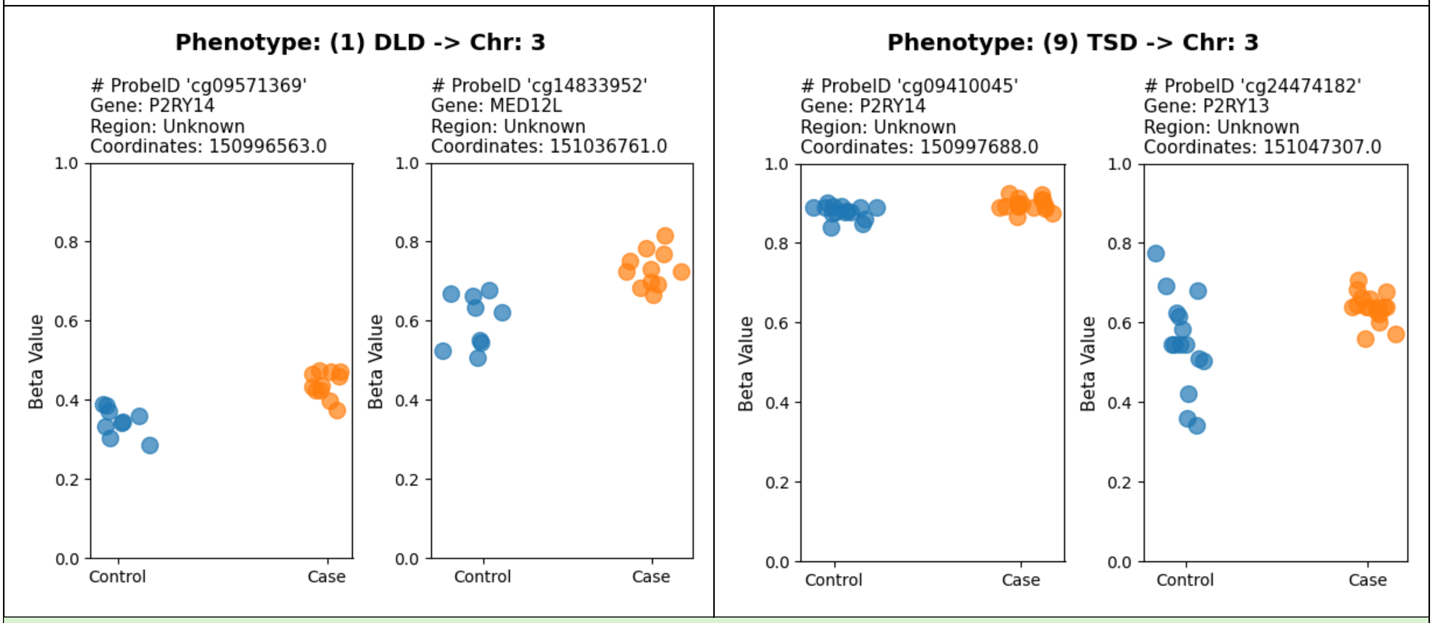


Figure 37. Shared probes within the genomic loci (chr3:151,016,761-151,077,307)

Chromosome 10

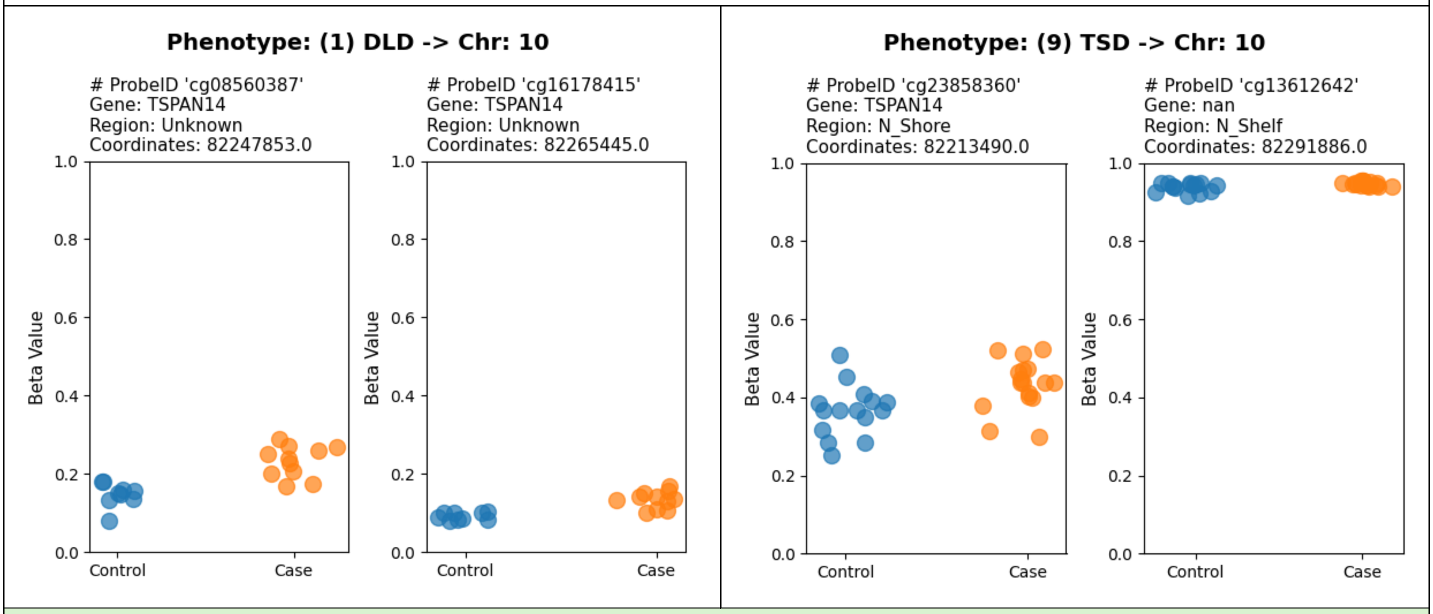


Figure 38. Shared probes within the genomic loci (chr10:822,478,53-822,918,86)

Chromosome 16

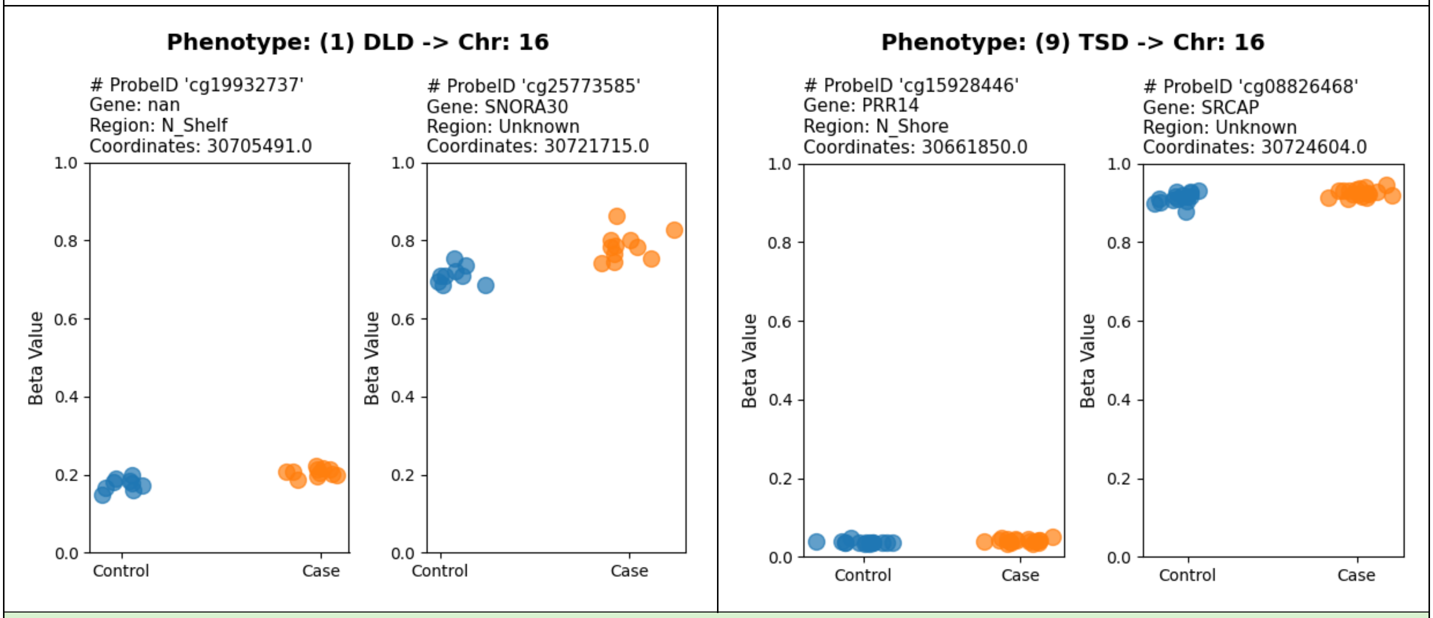


Figure 39. Shared probes within the genomic loci (chr16:307,054,91-307,246,04)

Next: Hypomethylated probes (Sections: 1.4, 1.5, 1.6)

1.4. Hypomethylated probe IDs that are shared among 2 or more phenotypes:

The maximum match detected in hypo methylated probes (Table 21) was three phenotypes. The probes that achieve this match are 'cg15454820', 'cg04586579' as shown in Figures (40, 41) respectively.

Figures (40, 41) ➔

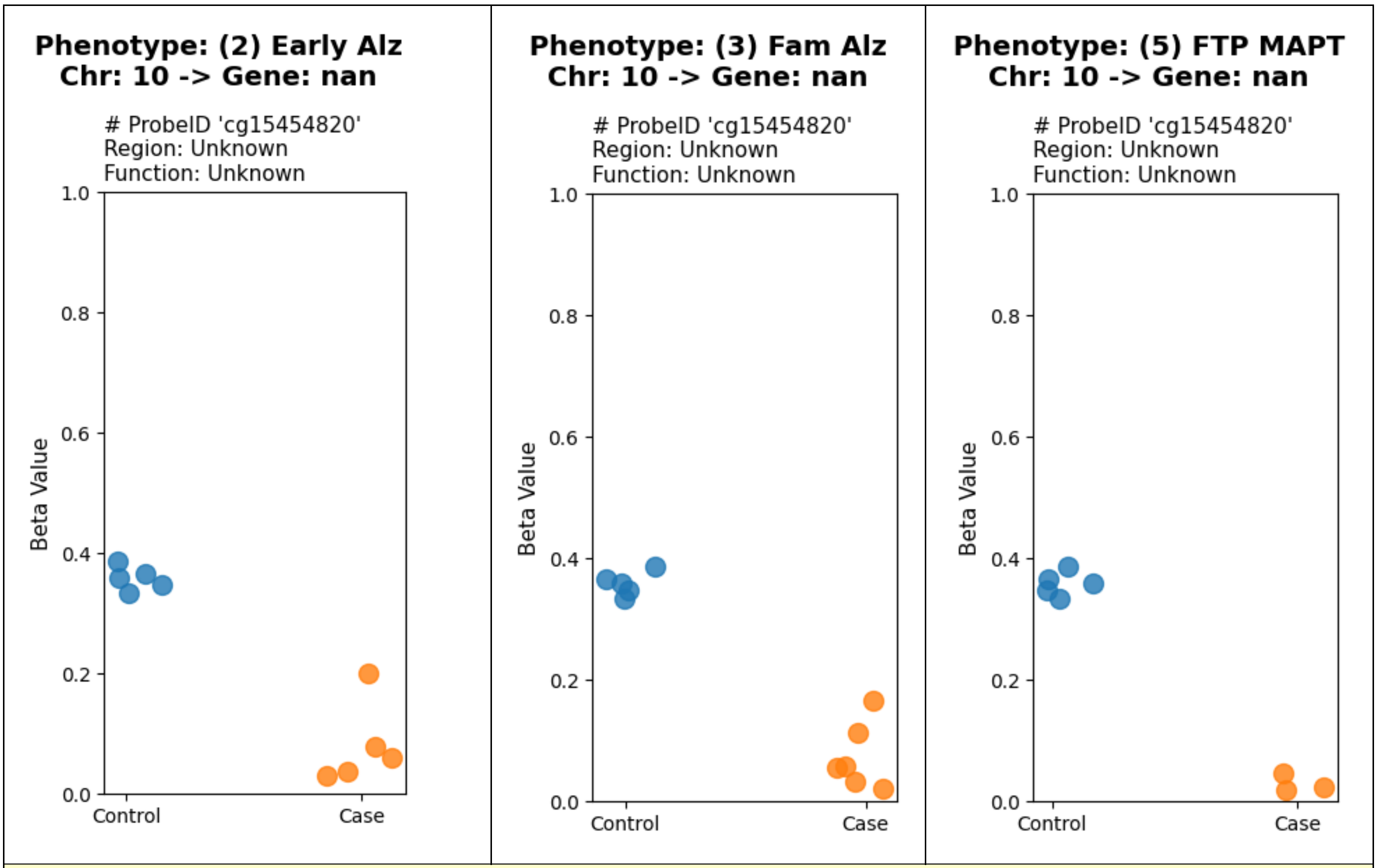


Figure 40. Probe ID 'cg15454820'

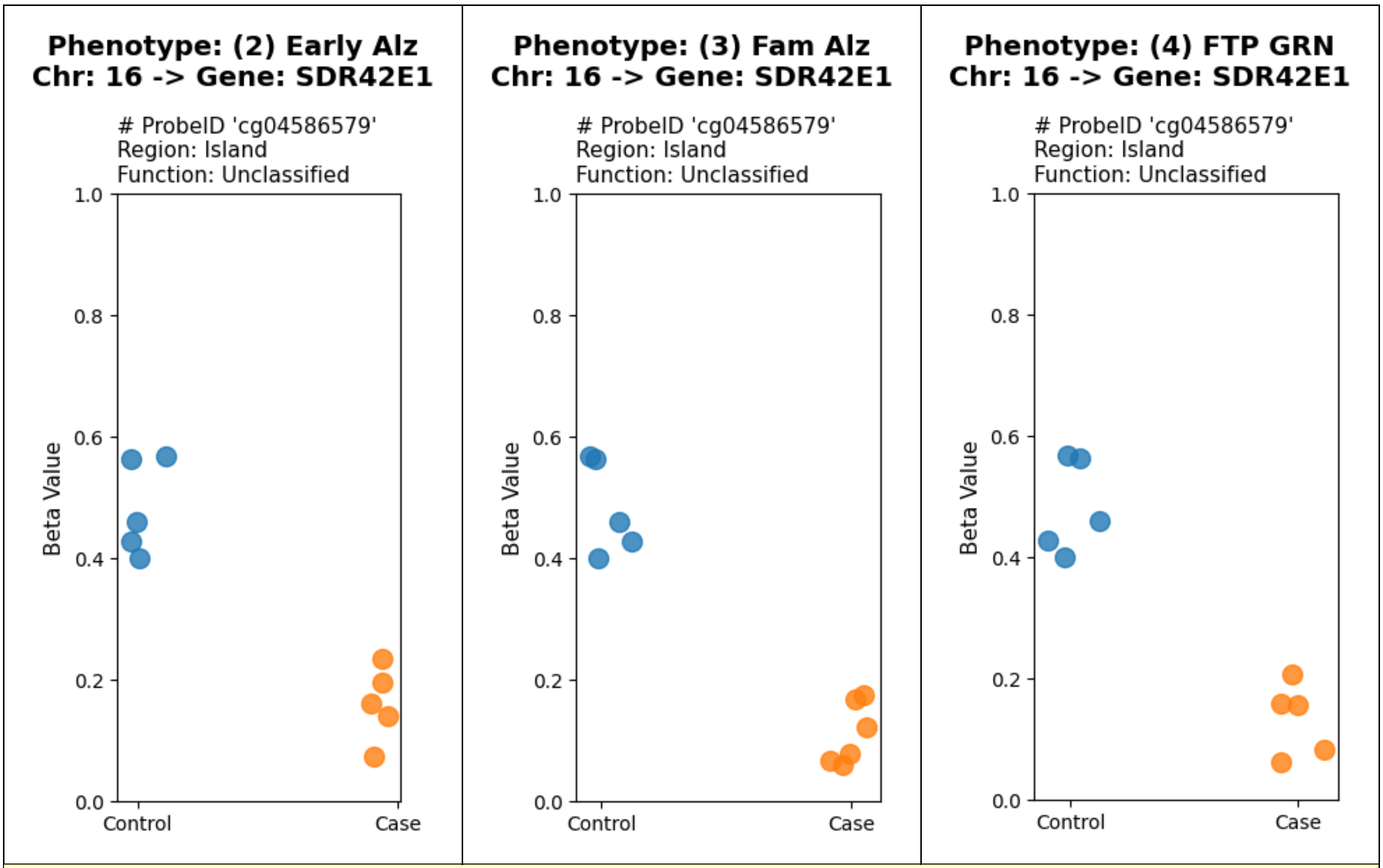


Figure 41. Probe ID 'cg04586579'

1.5. Hypomethylated probes that share the same gene among 2 or more phenotypes: Figures (42, 43, 44, 45).

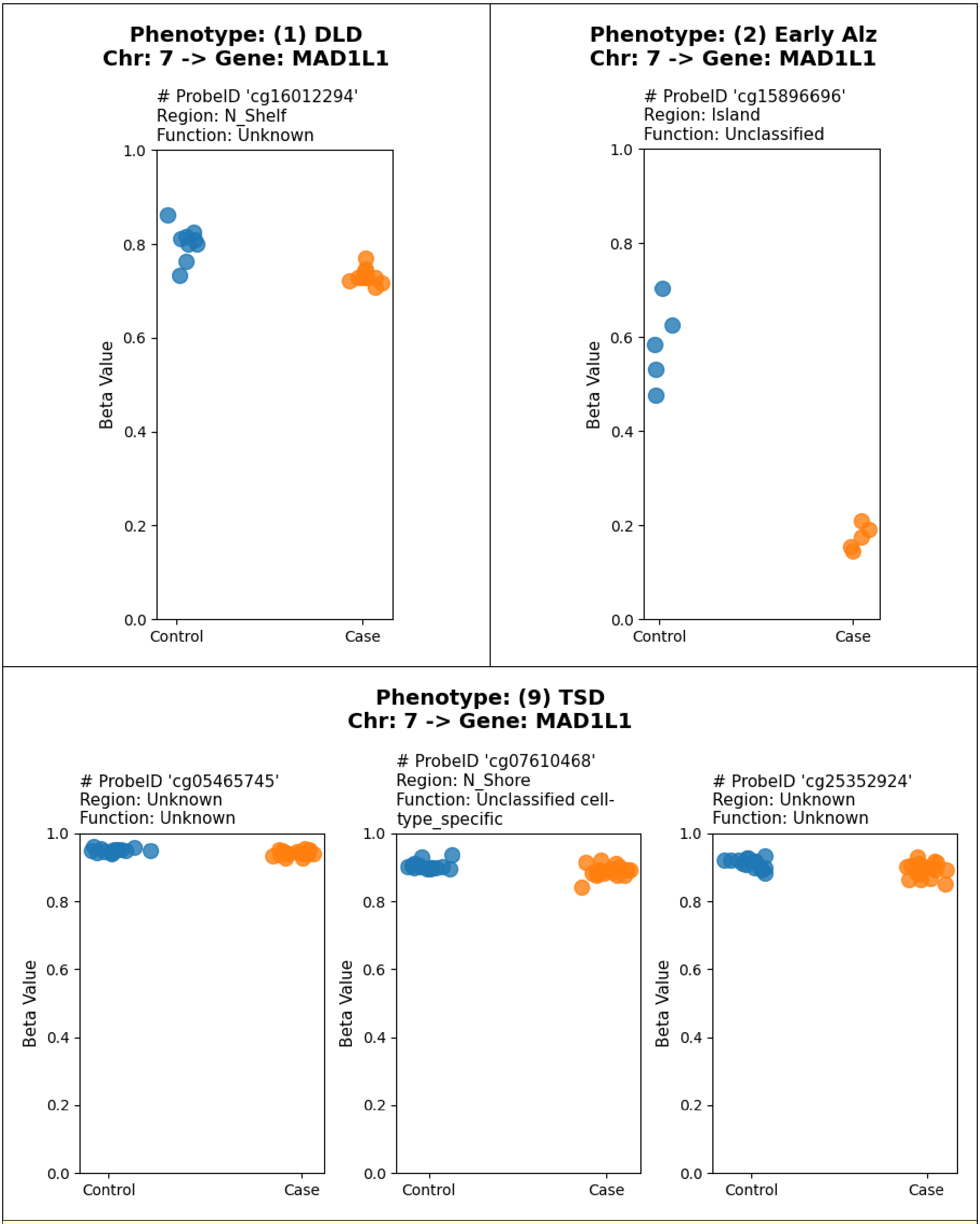


Figure 42. Gene: MAD1L1

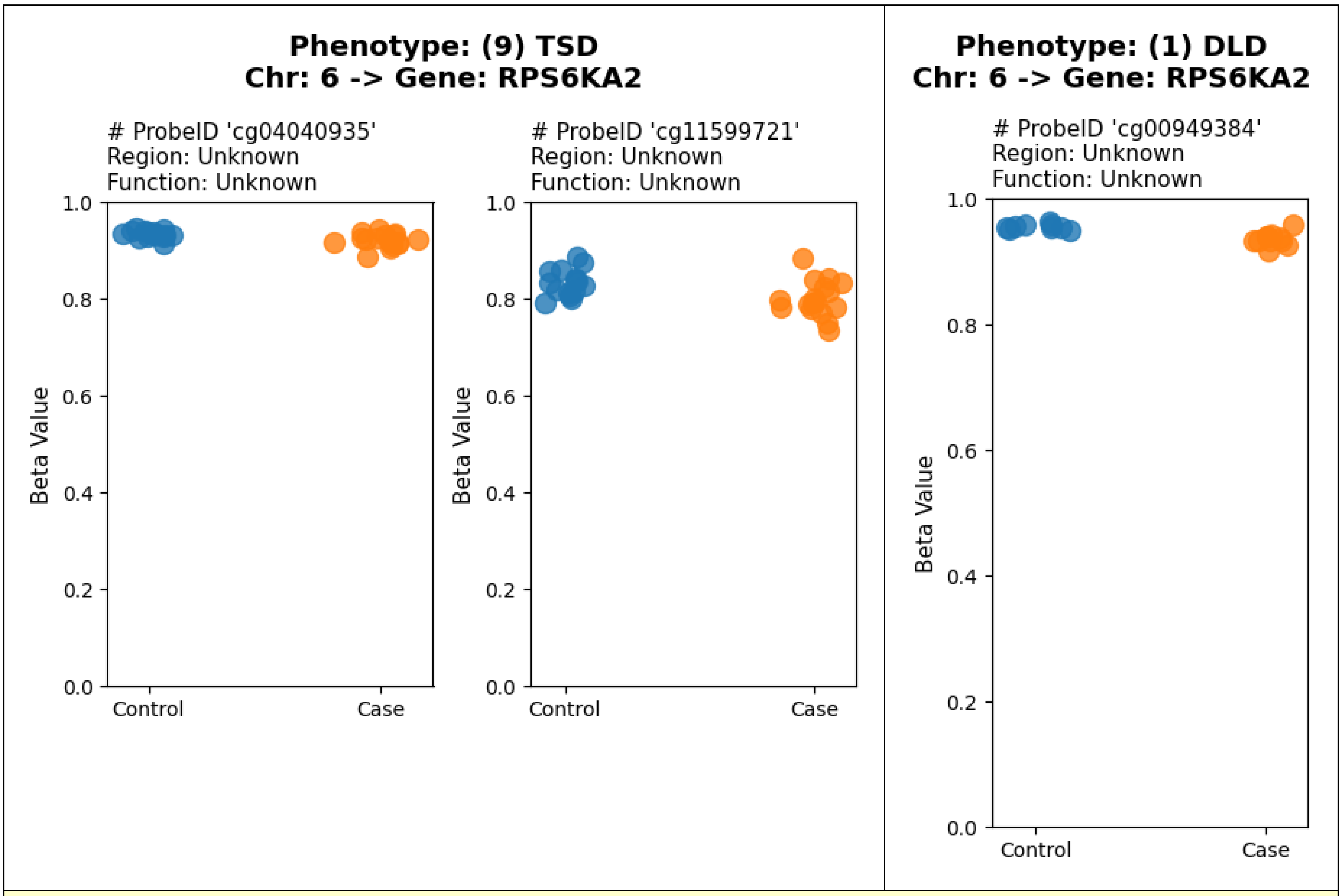


Figure 43. Gene: RPS6KA2

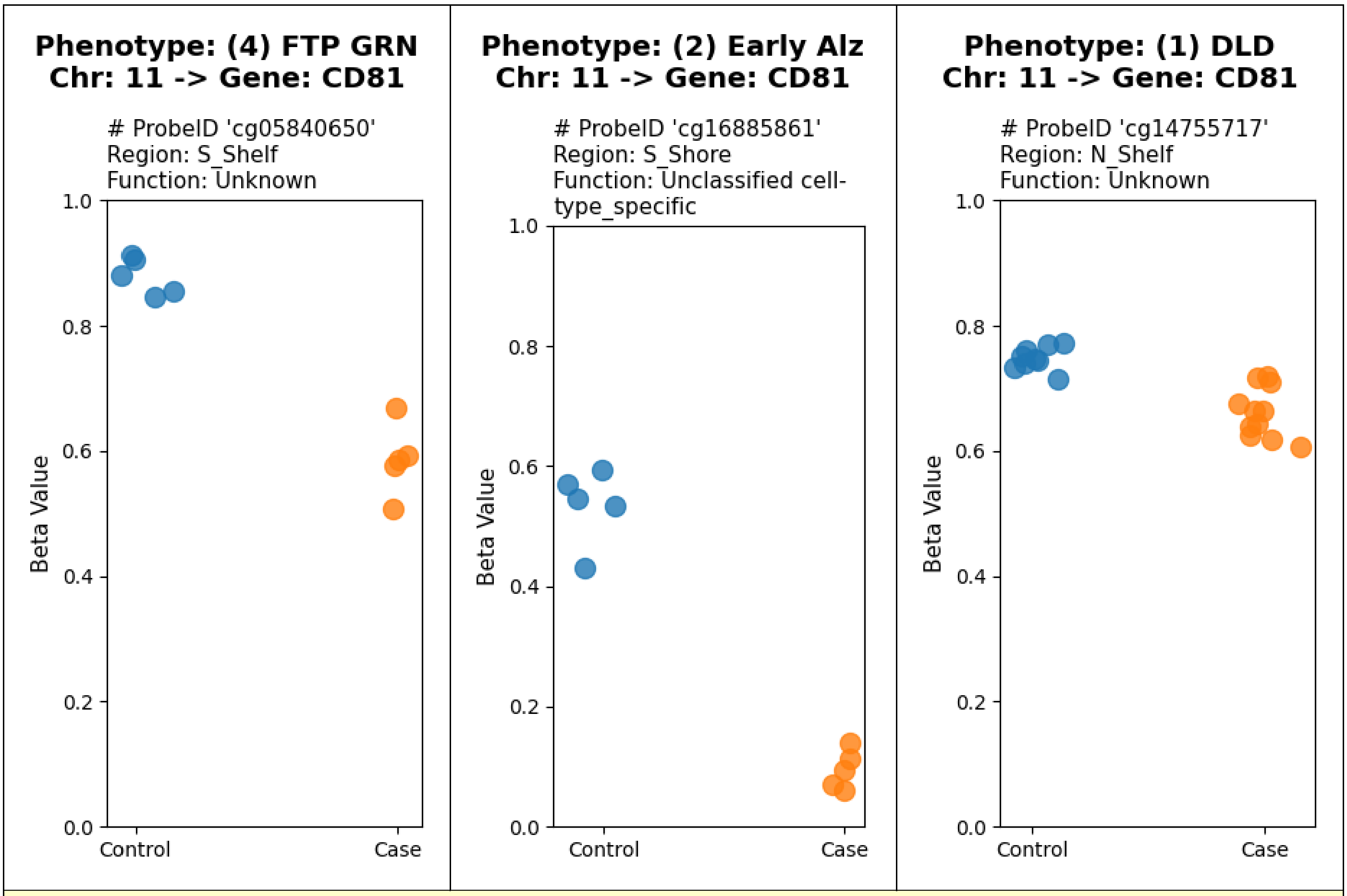


Figure 44. Gene: CD81

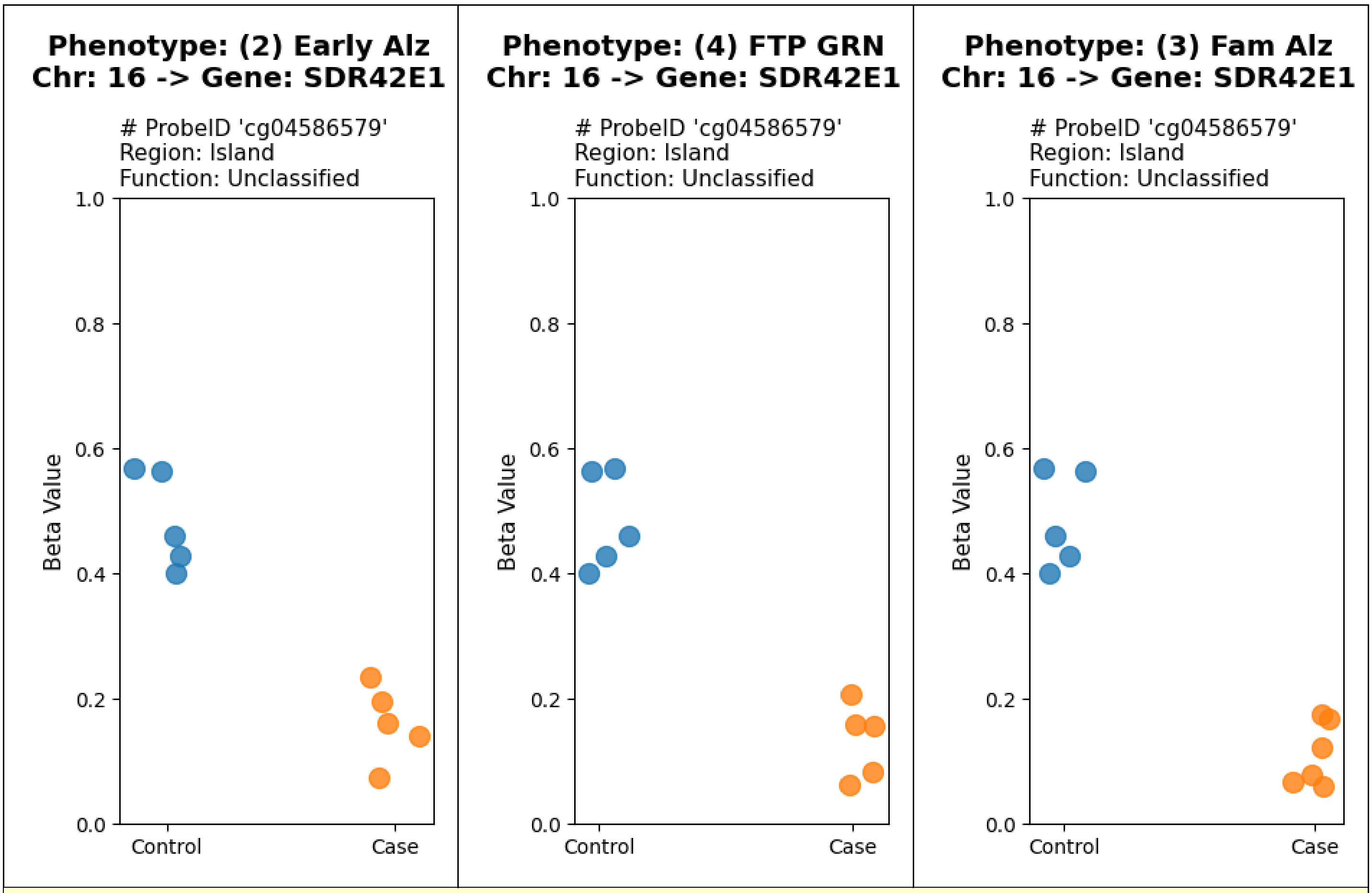


Figure 45. Gene: SDR42E1 (the same probes available in Figure 41)

1.6. Hypomethylated probes in nearby regions (< 100K bp) shared between two or more phenotypes:

None of the chromosomes expresses shared regions among the 9 phenotypes even with threshold of 100K bp.

Note: The source code used to plot all previous results is available in the supplementary material (*Shared probes, Shared Genes, Shared Regions*)

2. Discussion

The list of genes extracted for each dataset generally differed from the original results provided in each experiment. While a detailed comparison with the original results is beyond the scope of our study, the observed differences highlight the significant impact of varying normalization methods and pipelines on DMR results.

Based on logFC and the number of shared phenotypes, the most significant differentiations were selected. For hypermethylation, *SLC39A11*, *MAD1L1*, and *DENND2D* showed the highest differentiation. For hypomethylation, *SDR42E1*, *CD81*, and again *MAD1L1* were identified as top candidates. Tables 25 and 26 highlight the most significant differentiations for hypermethylated and hypomethylated genes, respectively. The maximum number of phenotypes that are detected to share certain loci is 3 for both hypermethylated and hypomethylated regions.

MAD1L1(Hypermethylation & Hypomethylation)

For hypermethylated genes, *MAD1L1* is the only to express availability in 3 phenotypes (DLD, Early Alz, Fam Alz). Interestingly, the same gene expresses hypomethylation in other locations (Figure 42) among also 3 phenotypes (DLD, Early Alz, and minimal differentiation in TSD). Upon reviewing available literature on the possible contribution of *MAD1L1* in behavior and psychiatric health, we found a recent study by Sokolov et al. (2023) that specifically identifies Three methylation loci (cg02825527, cg18302629, and cg19624444) as consistently hypomethylated in minor allele carriers of depression candidates. Su et al. (2015) and Levey et al. (2020), on the other hand, provide evidence for the association of *MAD1L1* variants with schizophrenia and anxiety, respectively. From an environmental perspective, Bozack et al. (2021) detected an association of the differentially methylated probe *cg26462130* (*MAD1L1*) in cord blood linked to prenatal metal exposure (specifically Mn), with their findings showing that the differentiation persisted when blood samples were collected during childhood. However, in our study, the environmental effect of sleep deprivation in the TSD group should be interpreted conservatively, as the DMRs for TSD did not pass FDR correction.

Other studies did not reveal a role for *MAD1L1* relative to our phenotypes. For example, Jansen et al. (2006) highlights *MAD1L1* as a potential target to improve survival in patients with ovarian cancer. The study demonstrates that *MAD1L1* overexpression delays cell proliferation, while its downregulation through hypermethylation contributes to disease progression. Similarly, Bandala-Jacques et al. (2020) shows that patients with the *MAD1L1* rs1801368 polymorphism are less likely to achieve optimal cytoreduction (a critical factor in improving overall survival in ovarian adenocarcinomas) compared to the non-polymorphic group. However, the study did not investigate an epigenetic contribution.

SLC39A11 Hypermethylation

Figure 31 shows significant hypermethylation of probe ID 'cg09945813', which corresponds to the *SLC39A11* gene. The hypermethylation was detected in the *Early Alz* and *Fam Alz-*

groups. Given that both phenotypes exhibit overall low quality of life and considering that hypermethylation is often associated with gene downregulation (reduced expression), we were particularly interested in the recent findings by Xia et al. (2024), which suggest a possible role of *SLC39A11* in overall longevity. The study found that a mutation in *SLC39A11*, leading to reduced expression, results in an accelerated aging phenotype in zebrafish. Additionally, the study reported that *SLC39A11* expression is significantly reduced in patients with Hutchinson-Gilford Progeria Syndrome (HGPS).

Chr	**Gene/**	**Loci**	**(1) DLD**	**(2) Early Alz**	**(3) Fam Alz**	**(4) FTP GRN**	**(5) FTP MAPT**	**(6) ELA**	**(7) SAD ELA**	**(8) SAD**	**(9) TSD**
17	SLC39A11			1		1					
13	NaN				1	1					
10	NaN			1		1					
7	MAD1L1		1	1	1						
1	DENND2D		2	1							
17	MSI2			2							2
10	FGFR2				1						2
3	P2RY14	~ 151M bp	1								1
3	MED12L	~ 151M bp	1								
3	P2RY13	~ 151M bp									1
10	TSPAN14	~ 82M bp	2								1
10	NaN	~ 82M bp									
16	NaN	~ 30M bp	1								
16	SNORA30	~ 30M bp	1								
16	PRR14	~ 30M bp									1
16	SRCAP	~ 30M bp									1

Table 25. Hypermethylated genes availability among the 9 phenotypes. Values (1,2) represents the number of probes).

DENND2D Hypermethylation

Figure 34 shows hypermethylation of 3 different probes that are mapped to DENND2D gene. Associated phenotypes were DLD and Early Alz groups. The differentiation was more obvious in Early Alz group. Although www.genecards.org stated that diseases associated with DENND2D include autism spectrum disorder, we could not actually identify studies with this information. On the other hand, the available literature shows that candidates of DENN family are poorly characterized (Yoshimura et al., 2010, Kumar et al., 2023). Kumar et al. (2023) suggests that DENND2B (another candidate from DENN damily) is involved in cancer and neurodevelopmental disorders. As per the study, loss-of-function mutation in DENND2B leads to severe mental retardation, seizures, neural hearing loss, unilateral cystic kidney dysplasia, frequent infections, and other congenital anomalies.

On the other hand, the only study that specifically discusses *DENND2D* hypermethylation is Kanda et al. (2013). The study highlights the frequent hypermethylation of *DENND2D* in hepatocellular carcinoma (HCC) tissues (75%) and its significant association with the downregulation of *DENND2D* mRNA expression. The study concluded that *DENND2D* plays an important role in hepatocarcinogenesis (Kanda et al., 2013).

TSPAN14 Hypermethylation in DLD & TSD groups

Our findings revealed two probes that are hypermethylated on chromosome 10 within the coordinate range (82247853.0 - 82265445.0) in the DLD group. Another two probes in a nearby region (82213490.0 - 82291886.0) were also detected to be hypermethylated in the TSD group. However, the differentiation in the DLD group was more significant than in the TSD group (Figure 38). According to www.genecards.org, *TSPAN14* is involved in the positive regulation of the Notch signaling pathway. Upon reviewing the available literature, Artavanis-Tsakonas and Muskavitch (2010) explain that Notch signalling plays a role in various developmental decisions in the nervous system. Salazar et al. (2020) also highlight the importance of Notch signalling in learning and memory across multiple species. The study further suggests that modulation of Notch activity may be effective in treating some symptoms associated with neurological disorders. These findings highlight the need for in-depth investigation of TSPAN14 in developmental delays, such as the DLD phenotype.

Chr	Gene	Probe/ Loci	(1) DLD	(2) Early Alz	(3) Fam Alz	(4) FTP GRN	(5) FTP MAPT	(6) ELA	(7) SAD ELA	(8) SAD	(9) TSD
16	SDR42E1			1	1	1					
7	MAD1L1		1	1							3
6	RPS6KA2		1								2
11	CD81		1	1		1					
10	NaN	cg15454820		1	1		1				

Table 26. Hypomethylated genes availability among the 9 phenotypes. Values (1,2,3) represents the number of probes).

SDR42E1 Hypomethylation

The most hypomethylation was observed in probe ID 'cg04586579' that maps to *SDR42E1* gene, shown in Figure 41. This probe was hypomethylated across three phenotypes (Early Alz, Fam Alz, FTP GRN), all belonging to the same experiment (E-MTAB-11975). We could not find relevant information in the available literature regarding *SDR42E1* role in nervous system symptomology. However, Bouhouche et al. (2021) which uses blood samples in their-

Study reveals an essential role of *SDR42E1* in maintaining connective tissue. Apart from relevant neurological symptoms, Meyer et al. (2021) on the other hand used TNBC samples (triple-negative breast cancer biopsy samples) and finds that *SDR42E1* is the only DMR that shows both altered methylation and expression in TNBC patients after Neoadjuvant chemotherapy (NAC).

CD81 Hypomethylation

Three hypomethylated probes were mapped to *CD81* which interestingly was associated among DLD, Early Alz, and FTP GRN groups (Fig. 44). Available literature demonstrates that *CD81* main role is to mediate signal transduction events (Hasterok et al, 2019). An analysis of single cell RNAseq of human Alzheimer's disease brains showed that *CD81* is upregulated in microglia module Mathys et al. (2019). The study uses samples from the prefrontal cortex of 48 individuals with varying degrees of Alzheimer's disease pathology across six major brain cell types (Figure 46). Together with our findings, this suggests a possible correlation between CD81 hypomethylation in blood cells and the upregulation of the same gene in microglia cells in Alzheimer's disease.

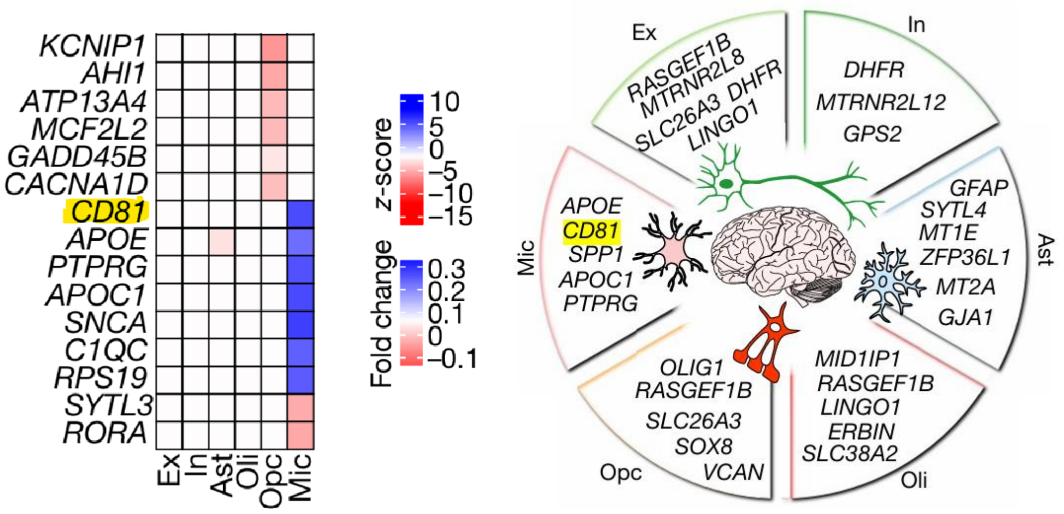


Figure 46. Diagram from Mathys et al. (2019) demonstrates differentially expressed genes in six cell types from prefrontal cortex tissue of Alzheimer's candidates.

Cell type (prefrontal cortex)	Abbreviation
Excitatory neurons	Ex
Inhibitory neurons	In
Astrocytes	Ast
Oligodendrocytes	Oli
Oligodendrocyte precursor cells	Opc
Microglia	Mic

Hypermethylated regions

MED12L has been reported to overlap with several genes, including *P2RY13* and *P2RY14* (Nizon et al., 2019). Notably, our approach for detecting probes with shared regions revealed hypermethylation within the genomic locus at chr3:151,016,761-151,077,307 (Figure 37) that includes *MED12L*, *P2RY13*, and *P2RY14*. *MED12L* is associated with Nizon-Isidor Syndrome, a neurodevelopmental disorder characterized by developmental delay, poor speech, and various symptoms, including sleep disturbances (Online Mendelian Inheritance in Man [OMIM], 2025). Given this context, our findings may offer promising insights in the field of epigenetics, potentially shedding light on the role of sleep in the development of such conditions or in explaining language delays in children without specific diagnoses. However, these interpretations should be approached with caution, specifically in terms of TSD group which did not pass the FDR correction.

3. Limitations

Further enrichment of the results, particularly for *MAD1L1*, *TSPAN14*, and *CD81*, could have provided deeper insights, especially when compared to results in other studies. This could be achieved using a tailored tool that accounts for multiple phenotypes. However, time constraints limited our ability to explore and test a sufficient number of tools. Additionally, data availability posed a limitation, as phenotype selection depended on publicly accessible datasets. On the other hand, our exclusion criteria in terms of platform compatibility needs to be re-evaluated, this is because our results mainly targeted probes that shares the same gene, rather than finding matched probes. Finally, the complete failure of the TSD dataset to pass FDR correction requires further investigation, particularly regarding the parameter settings of the *Limma* function for the 450K array.

4. Conclusion

Our research, alongside other studies, provides further evidence supporting the potential of peripheral blood biomarkers in reflecting neurological symptomatology. Phenotypes that exhibited patterns of alteration on identical probes were limited to Alzheimer's disorders groups from the same experiment (E-MTAB-11975). These probes need to be investigated in a separate study primarily dedicated to Alzheimer's disease. A key finding from our results is the hypomethylation of *CD81* in Alzheimer's samples (Early Alz and FTP GRN), which has also been reported as upregulated in the prefrontal cortex in another study. The alteration in methylation levels detected in this study was limited to a single probe differentiation. However, the methylation changes in *MAD1L1*-mapped probes in individuals with Developmental Language Disorder (DLD) and Early-onset Alzheimer's Disease (EOAD) are noteworthy, especially considering the existing literature discussing the methylation of the same gene in psychiatric and environmental conditions. These findings should serve as motivation for further investigation of *MAD1L1* to explore its potential contribution to neuro-system symptomology. From an analytical standpoint, it is clear that DNA methylation analysis would benefit from standardized methods tailored to specific cell types or phenotypes. Such standardization could improve result consistency and enhance the reliability of DNA methylation studies, especially for diseases lacking global methylation changes. On the other hand, our approach of performing QC procedures after a standard pipeline provides greater insights and easier sample-wise visibility into data quality but requires more processing time. Despite this, it remains valuable for large arrays with potential quality risks from specific samples. For smaller arrays, while less useful for removing low-performing samples, it offers a method to impute missing or masked beta values using weighted mean (WM).

References

1. Choi, S. H., Worswick, S., Byun, H., Shear, T., Soussa, J. C., Wolff, E. M., Douer, D., Garcia-Manero, G., Liang, G., & Yang, A. S. (2009). Changes in DNA methylation of tandem DNA repeats are different from interspersed repeats in cancer. International Journal of Cancer, 125(3), 723–729. https://doi.org/10.1002/ijc.24384

2. Takeshima, H., & Ushijima, T. (2018). Mechanisms of DNA methylation changes in cancer. In Elsevier eBooks. https://doi.org/10.1016/b978-0-12-801238-3.65058-4

3. Juo, Y., Johnston, F., Zhang, D., Juo, H., Wang, H., Pappou, E., Yu, T., Easwaran, H., Baylin, S., Van Engeland, M., & Ahuja, N. (2014). Prognostic value of CpG island methylator phenotype among colorectal cancer patients: a systematic review and meta-analysis. Annals of Oncology, 25(12), 2314–2327. https://doi.org/10.1093/annonc/mdu149

4. Elango, N., & Yi, S. V. (2011). Functional relevance of CPG island length for regulation of gene expression. Genetics, 187(4), 1077–1083. https://doi.org/10.1534/genetics.110.126094

5. Bibb, K., Arya, R., & Saldanha, S. N. (2017). Future challenges and prospects for the role of epigenetic mechanisms in cancer management. In Elsevier eBooks (pp. 345–372). https://doi.org/10.1016/b978-0-12-809552-2.00013-9

6. HumanMethylation27 product support files. (n.d.). https://emea.support.illumina.com/downloads/humanmethylation27_product_support_files.html

7. Infinium HumanMethylation450K V1.2 Product Files. (n.d.). https://sapac.support.illumina.com/downloads/infinium_humanmethylation450_product_files.html

8. Infinium MethylationEPIC v1.0 product Files. (n.d.). https://emea.support.illumina.com/downloads/infinium-methylationepic-v1-0-product-files.html

9. Infinium MethylationEPIC v2.0 Product Files. (n.d.). https://emea.support.illumina.com/downloads/infinium-methylationepic-v2-0-product-files.html

10. Introduction to DNA Methylation Analysis — methylprep 1.6.5 documentation. (n.d.). https://life-epigenetics-methylprep.readthedocs-hosted.com/en/latest/docs/introduction/introduction.html

11. Du, P., Zhang, X., Huang, C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11(1). https://doi.org/10.1186/1471-2105-11-587

12. Marabita, F., Almgren, M., Lindholm, M. E., Ruhrmann, S., Fagerström-Billai, F., Jagodic, M., Sundberg, C. J., Ekström, T. J., Teschendorff, A. E., Tegnér, J., & Gomez-Cabrero, D. (2013). An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics, 8(3), 333–346. https://doi.org/10.4161/epi.24008

13. Cassoff, J., Wiebe, S. T., & Gruber, R. (2012). Sleep patterns and the risk for ADHD: A review. Nature and Science of Sleep, 4, 73–80. https://doi.org/10.2147/NSS.S31269

14. O'Callaghan, F. V., Mamun, A. A., O'Callaghan, M., Clavarino, A., Williams, G. M., Bor, W., Heussler, H., & Najman, J. M. (2010). The link between sleep problems in infancy and early childhood and attention problems at 5 and 14years: Evidence from a birth cohort study. Early Human Development, 86(7), 419–424. https://doi.org/10.1016/j.earlhumdev.2010.05.020

15. Van Der Heijden, K. B., Smits, M. G., Van Someren, E. J., & Gunning, W. B. (2005). Idiopathic chronic sleep onset insomnia in Attention-Deficit/Hyperactivity disorder: a circadian rhythm sleep disorder. Chronobiology International, 22(3), 559–570. https://doi.org/10.1081/cbi-200062410

16. National Institute of General Medical Sciences (NIGMS). (Link: https://www.nigms.nih.gov/education/fact-sheets/Pages/circadian-rhythms.aspx#:~:text=Circadian%20rhythms%20are%20the%20physical,and%20temperature%20also%20affect%20them.)

17. Han, Y., Shon, J., Kwon, S. Y., & Park, Y. J. (2024). Effects of dietary protein intake levels on peripheral circadian rhythm in mice. International Journal of Molecular Sciences, 25(13), 7373. https://doi.org/10.3390/ijms25137373

18. Gurney, J. G., McPheeters, M. L., & Davis, M. M. (2006). Parental Report of health conditions and health care use among children with and without autism. Archives of Pediatrics and Adolescent Medicine, 160(8), 825. https://doi.org/10.1001/archpedi.160.8.825

19. Magnuson, K. M., & Constantino, J. N. (2011). Characterization of depression in children with autism spectrum disorders. Journal of Developmental & Behavioral Pediatrics, 32(4), 332–340. https://doi.org/10.1097/dbp.0b013e318213f56c

20. Vahratian, A., Blumberg, S. J., Terlizzi, E. P., & Schiller, J. S. (2021). Symptoms of anxiety or depressive disorder and use of mental health care among adults during the COVID-19 pandemic — United States, August 2020–February 2021. MMWR Morbidity and Mortality Weekly Report, 70(13), 490–494. https://doi.org/10.15585/mmwr.mm7013e2

21. Delpino, F. M., Da Silva, C. N., Jerônimo, J. S., Mulling, E. S., Da Cunha, L. L., Weymar, M. K., Alt, R., Caputo, E. L., & Feter, N. (2022). Prevalence of anxiety during the COVID-19 pandemic: A systematic review and meta-analysis of over 2 million people. Journal of Affective Disorders, 318, 272–282. https://doi.org/10.1016/j.jad.2022.09.003

22. Bourque, V., Poulain, C., Proulx, C., Moreau, C. A., Joober, R., D'Arc, B. F., Huguet, G., & Jacquemont, S. (2024). Genetic and phenotypic similarity across major psychiatric disorders: a systematic review and quantitative assessment. Translational Psychiatry, 14(1). https://doi.org/10.1038/s41398-024-02866-3

23. Alomari, N. A., Bedaiwi, S. K., Ghasib, A. M., Kabbarah, A. J., Alnefaie, S. A., Hariri, N., Altammar, M. A., Fadhel, A. M., & Altowairqi, F. M. (2022). Social anxiety Disorder: associated conditions and therapeutic approaches. Cureus. https://doi.org/10.7759/cureus.32687

24. Polderman, T. J. C., Benyamin, B., De Leeuw, C. A., Sullivan, P. F., Van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics, 47(7), 702–709. https://doi.org/10.1038/ng.3285

25. Anttila, V., Bulik-Sullivan, B., Finucane, H. K., Walters, R. K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G. J., Gormley, P., Malik, R., Patsopoulos, N. A., Ripke, S., Wei, Z., Yu, D., Lee, P. H., Turley, P., Grenier-Boley, B., Chouraki, V., Kamatani, Y., . . . Neale, B. M. (2018). Analysis of shared heritability in common disorders of the brain. Science, 360(6395). https://doi.org/10.1126/science.aap8757

26. Morbidity and Mortality Weekly Report (MMWR): Symptoms of Anxiety or Depressive Disorder and Use of Mental Health Care Among Adults During the COVID-19 Pandemic — United States, August 2020–February 2021

27. Sahoo, K., & Sundararajan, V. (2024). Methods in DNA methylation array dataset analysis: A review. Computational and Structural Biotechnology Journal, 23, 2304–2325. https://doi.org/10.1016/j.csbj.2024.05.015

28. Wang, Z., Wu, X., & Wang, Y. (2018). A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. BMC Bioinformatics, 19(S5). https://doi.org/10.1186/s12859-018-2096-3

29. Fortin, J., Labbe, A., Lemire, M., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M., & Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biology, 15(11). https://doi.org/10.1186/s13059-014-0503-2

30. Hansen, K. D., & Fortin, J. (2024, November 19). The minfi User's Guide. https://bioconductor.org/packages/devel/bioc/vignettes/minfi/inst/doc/minfi.html

31. Illumina. A Patient-Centric Methylation Pipeline (Methodology for detecting DNA methylation changes in microarray data.)

32. Welsh, H., Batalha, C. M. P. F., Li, W., Mpye, K. L., Souza-Pinto, N. C., Naslavsky, M. S., & Parra, E. J. (2023). A systematic evaluation of normalization methods and probe replicability using infinium EPIC methylation data. Clinical Epigenetics, 15(1). https://doi.org/10.1186/s13148-023-01459-z

33. Zhou, W., Triche, T. J., Laird, P. W., & Shen, H. (2018). SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. Nucleic Acids Research. https://doi.org/10.1093/nar/gky691

34. RNBeads. (n.d.). Bioconductor. https://www.bioconductor.org/packages/release/bioc/html/RnBeads.html

35. Xu, Z., Langie, S. a. S., De Boever, P., Taylor, J. A., & Niu, L. (2017). RELIC: a novel dye-bias correction method for Illumina Methylation BeadChip. BMC Genomics, 18(1). https://doi.org/10.1186/s12864-016-3426-3

36. SeSAME. (n.d). Bioconductor. https://doi.org/doi:10.18129/B9.bioc.sesame

37. Enmix. (n.d.). Bioconductor. https://doi.org/doi:10.18129/B9.bioc.ENmix

38. BioStudies. (n.d.). BioStudies < The European Bioinformatics Institute < EMBL-EBI. https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-13583?query=E-MTAB-13583

39. GEO Accession viewer. (n.d.). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL10878

40. Zhou, W., Hinoue, T., Barnes, B., Mitchell, O., Iqbal, W., Lee, S. M., Foy, K. K., Lee, K., Moyer, E. J., VanderArk, A., Koeman, J. M., Ding, W., Kalkat, M., Spix, N. J., Eagleson, B., Pospisilik, J. A., Szabó, P. E., Bartolomei, M. S., Schaaf, N. a. V., . . . Laird, P. W. (2022). DNA methylation dynamics and dysregulation delineated by high-throughput profiling in the mouse. Cell Genomics, 2(7), 100144. https://doi.org/10.1016/j.xgen.2022.100144

41. Wiegand, A., Kreifelts, B., Munk, M. H. J., Geiselhart, N., Ramadori, K. E., MacIsaac, J. L., Fallgatter, A. J., Kobor, M. S., & Nieratschker, V. (2021). DNA methylation differences associated with social anxiety disorder and early life adversity. Translational Psychiatry, 11(1). https://doi.org/10.1038/s41398-021-01225-w

42. Hop, P. J., Zwamborn, R. a. J., Hannon, E. J., Dekker, A. M., Van Eijk, K. R., Walker, E. M., Iacoangeli, A., Jones, A. R., Shatunov, A., Khleifat, A. A., Opie-Martin, S., Shaw, C. E., Morrison, K. E., Shaw, P. J., McLaughlin, R. L., Hardiman, O., Al-Chalabi, A., Van Den Berg, L. H., Mill, J., & Veldink, J. H. (2020). Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. NAR Genomics and Bioinformatics, 2(4). https://doi.org/10.1093/nargab/lqaa105

43. Ramos-Campoy, O., Comas-Albertí, A., Hervás, D., Borrego-Écija, S., Bosch, B., Sandoval, J., Fort-Aznar, L., Moreno-Izco, F., Fernández-Villullas, G., Molina-Porcel, L., Balasa, M., Lladó, A., Sánchez-Valle, R., & Antonell, A. (2024). Genome-Wide DNA methylation in Early-Onset-Dementia patients brain tissue and lymphoblastoid cell lines. International Journal of Molecular Sciences, 25(10), 5445. https://doi.org/10.3390/ijms25105445

44. Cabezón, M., Malinverni, R., Bargay, J., Xicoy, B., Marcé, S., Garrido, A., Tormo, M., Arenillas, L., Coll, R., Borras, J., Jiménez, M. J., Hoyos, M., Valcárcel, D., Escoda, L., Vall-Llovera, F., Garcia, A., Font, L. L., Rámila, E., Buschbeck, M., & Zamora, L. (2021). Different methylation signatures at diagnosis in patients with high-risk myelodysplastic syndromes and secondary acute myeloid leukemia predict azacitidine response and longer survival. Clinical Epigenetics, 13(1). https://doi.org/10.1186/s13148-021-01002-y

45. Jiang, R., Jones, M. J., Chen, E., Neumann, S. M., Fraser, H. B., Miller, G. E., & Kobor, M. S. (2015). Discordance of DNA Methylation Variance Between two Accessible Human Tissues. Scientific Reports, 5(1). https://doi.org/10.1038/srep08257

46. Van Doorn, R., Slieker, R. C., Boonk, S. E., Zoutman, W. H., Goeman, J. J., Bagot, M., Michel, L., Tensen, C. P., Willemze, R., Heijmans, B. T., & Vermeer, M. H. (2016). Epigenomic analysis of Sézary syndrome defines patterns of aberrant DNA methylation and identifies diagnostic markers. Journal of Investigative Dermatology, 136(9), 1876–1884. https://doi.org/10.1016/j.jid.2016.03.042

47. Illumina, Infinium Controls Training Guide PDF. https://support.illumina.com/content/dam/illumina-support/courses/eval-inf-controls/story_content/external_files/Infinium_Controls_Training_Guide.pdf

48. Illumina Website, What are Ch probes in Illumina Manifest. https://knowledge.illumina.com/microarray/general/microarray-general-troubleshooting-list/000005501

49. Bondhus, L., Wei, A., & Arboleda, V. A. (2022). DMRscaler: a scale-aware method to identify regions of differential DNA methylation spanning basepair to multi-megabase features. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-04899-1

50. Sokolov, A. V., Manu, D., Nordberg, D. O. T., Boström, A. D. E., Jokinen, J., & Schiöth, H. B. (2023). Methylation in MAD1L1 is associated with the severity of suicide attempt and phenotypes of depression. Clinical Epigenetics, 15(1). https://doi.org/10.1186/s13148-022-01394-5

51. Su, L., Shen, T., Huang, G., Long, J., Fan, J., Ling, W., & Jiang, J. (2015). Genetic association of GWAS-supported MAD1L1 gene polymorphism rs12666575 with schizophrenia susceptibility in a Chinese population. Neuroscience Letters, 610, 98–103. https://doi.org/10.1016/j.neulet.2015.10.061

52. Levey, D. F., Gelernter, J., Polimanti, R., Zhou, H., Cheng, Z., Aslan, M., Quaden, R., Concato, J., Radhakrishnan, K., Bryois, J., Sullivan, P. F., & Stein, M. B. (2020). Reproducible Genetic Risk LOCI for anxiety: results from ∼200,000 participants in the Million Veteran program. American Journal of Psychiatry, 177(3), 223–232. https://doi.org/10.1176/appi.ajp.2019.19030256

53. Bozack, A. K., Rifas-Shiman, S. L., Coull, B. A., Baccarelli, A. A., Wright, R. O., Amarasiriwardena, C., Gold, D. R., Oken, E., Hivert, M., & Cardenas, A. (2021). Prenatal metal exposure, cord blood DNA methylation and persistence in childhood: an epigenome-wide association study of 12 metals. Clinical Epigenetics, 13(1). https://doi.org/10.1186/s13148-021-01198-z

54. Jansen, R. A., Liu, J. C., Liyanarachchi, S., Crijns, A. P., Yan, P. S., Huang, T. H., Cohn, D. E., Fowler, J. M., Van Der Zee, A. G., & Brown, R. (2006). Prognostic impact of MAD1L1 promoter hypermethylation in advanced ovarian cancer. Journal of Clinical Oncology, 24(18_suppl), 5021. https://doi.org/10.1200/jco.2006.24.18_suppl.5021

55. Bandala-Jacques, A., Hernández-Cruz, I. A., Castro-Hernández, C., Díaz-Chávez, J., Arriaga-Canon, C., Barquet-Muñoz, S. A., Prada-Ortega, D. G., Cantú-De-León, D., & Herrera, L. A. (2020). Prognostic significance of the MAD1L1 1673 G:A polymorphism in ovarian adenocarcinomas. Revista De Investigaci□N Cl□Nica, 72(6). https://doi.org/10.24875/ric.19003280

56. Xia, Z., Tang, B., Li, X., Li, X., Jia, Y., Jiang, J., Chen, J., Song, J., Liu, S., Min, J., & Wang, F. (2024). A novel role for the Longevity-Associated protein SLC39A11 as a manganese transporter. Research, 7. https://doi.org/10.34133/research.0440

57. Yoshimura, S., Gerondopoulos, A., Linford, A., Rigden, D. J., & Barr, F. A. (2010). Family-wide characterization of the DENN domain Rab GDP-GTP exchange factors. The Journal of Cell Biology, 191(2), 367–381. https://doi.org/10.1083/jcb.201008051

58. Kumar, R., Francis, V., Ioannou, M. S., Aguila, A., Khan, M., Banks, E., Kulasekaran, G., & McPherson, P. S. (2023). DENND2B activates Rab35 at the intercellular bridge, regulating cytokinetic abscission and tetraploidy. Cell Reports, 42(7), 112795. https://doi.org/10.1016/j.celrep.2023.112795

59. Kanda, M., Nomoto, S., Oya, H., Takami, H., Hibino, S., Hishida, M., Suenaga, M., Yamada, S., Inokawa, Y., Nishikawa, Y., Asai, M., Fujii, T., Sugimoto, H., & Kodera, Y. (2013). Downregulation of DENND2D by promoter hypermethylation is associated with early recurrence of hepatocellular carcinoma. International Journal of Oncology, 44(1), 44–52. https://doi.org/10.3892/ijo.2013.2165

60. www.genecards.org (query: https://www.genecards.org/cgi-bin/carddisp.pl?gene=TSPAN14#:~:text=TSPAN14%20Gene%20%2D%20Tetraspanin%2014&text=Enables%20enzyme%20binding%20activity.,plasma%20membrane%3B%20and%20protein%20maturation.)

61. Artavanis-Tsakonas, S., & Muskavitch, M. A. (2010). Notch: The past, the present, and the future. Current Topics in Developmental Biology/Current Topics in Developmental Biology, 1–29. https://doi.org/10.1016/s0070-2153(10)92001-2

62. Salazar, J. L., Yang, S., & Yamamoto, S. (2020). Post-Developmental roles of notch signaling in the nervous system. Biomolecules, 10(7), 985. https://doi.org/10.3390/biom10070985

63. Bouhouche, A., Albaroudi, N., Alaoui, M. a. E., Askander, O., Habbadi, Z., Hassani, A. E., Iraqi, H., Fahime, E. E., & Belmekki, M. (2021). Identification of the novel SDR42E1 gene that affects steroid biosynthesis associated with the oculocutaneous genital syndrome. Experimental Eye Research, 209, 108671. https://doi.org/10.1016/j.exer.2021.108671

64. Meyer, B., Clifton, S., Locke, W., Luu, P., Du, Q., Lam, D., Armstrong, N. J., Kumar, B., Deng, N., Harvey, K., Swarbrick, A., Ganju, V., Clark, S. J., Pidsley, R., & Stirzaker, C. (2021). Identification of DNA methylation biomarkers with potential to predict response to neoadjuvant chemotherapy in triple-negative breast cancer. Clinical Epigenetics, 13(1). https://doi.org/10.1186/s13148-021-01210-6

65. Hasterok, S., Nyesiga, B., & Gjörloff-Wingren, A. (2019, September 1). CD81 (Cluster of Differentiation 81). Atlas of Genetics and Cytogenetics in Oncology and Haematology. Retrieved from https://atlasgeneticsoncology.org/gene/991/cd81-(cluster-of-differentiation-81)

66. Erratum to: Alzheimer's disease brain-derived extracellular vesicles spread tau pathology in interneurons. (2021). Brain, 144(4), e42. https://doi.org/10.1093/brain/awaa452

67. Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J. Z., Menon, M., He, L., Abdurrob, F., Jiang, X., Martorell, A. J., Ransohoff, R. M., Hafler, B. P., Bennett, D. A., Kellis, M., & Tsai, L. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature, 570(7761), 332–337. https://doi.org/10.1038/s41586-019-1195-2

68. Lambert, J., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M. J., Tavernier, B., Letenneur, L., Bettens, K., Berr, C., Pasquier, F., Fiévet, N., Barberger-Gateau, P., Engelborghs, S., De Deyn, P., Mateo, I., . . . Amouyel, P. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nature Genetics, 41(10), 1094–1099. https://doi.org/10.1038/ng.439

69. Nizon, M., Laugel, V., Flanigan, K. M., Pastore, M., Waldrop, M. A., Rosenfeld, J. A., Marom, R., Xiao, R., Gerard, A., Pichon, O., Caignec, C. L., Gérard, M., Dieterich, K., Cho, M. T., McWalter, K., Hiatt, S., Thompson, M. L., Bézieau, S., Wadley, A., . . . Isidor, B. (2019). Variants in MED12L, encoding a subunit of the

mediator kinase module, are responsible for intellectual disability associated with transcriptional defect. Genetics in Medicine, 21(12), 2713–2722. https://doi.org/10.1038/s41436-019-0557-3

70. Online Mendelian Inheritance in Man (OMIM). (2025). Nizon-Isidor Syndrome. Retrieved from https://www.omim.org/entry/618872

Appendices

Section 1 Pipelines Comparison

Comparison (1) between:	
Raw Data	Raw reads extracted using minfi package from the idat files provided on ArrayExpress.
Processed data	SWAN-Processed Data: Processed beta values provided by the publisher on ArrayExpress.

Fig 1. Shows the beta distribution curve for both outputs. From the provided metadata file (*E-MTAB-13583.txt.idf)* available with the experiment E-MTAB-13583 on ArrayExpress, the quality control protocol includes the following information:

- R packages *minfi*, ChAMP and RnBeads were used.
- The *limma* package was used by RnBeads to compute the p-values for all the covered CpGs.
- Samples and CpG islands (CpGs) that contained a substantial fraction of low technical quality measurements were discarded.
- Normalization done using SWAN method available in minfi package.

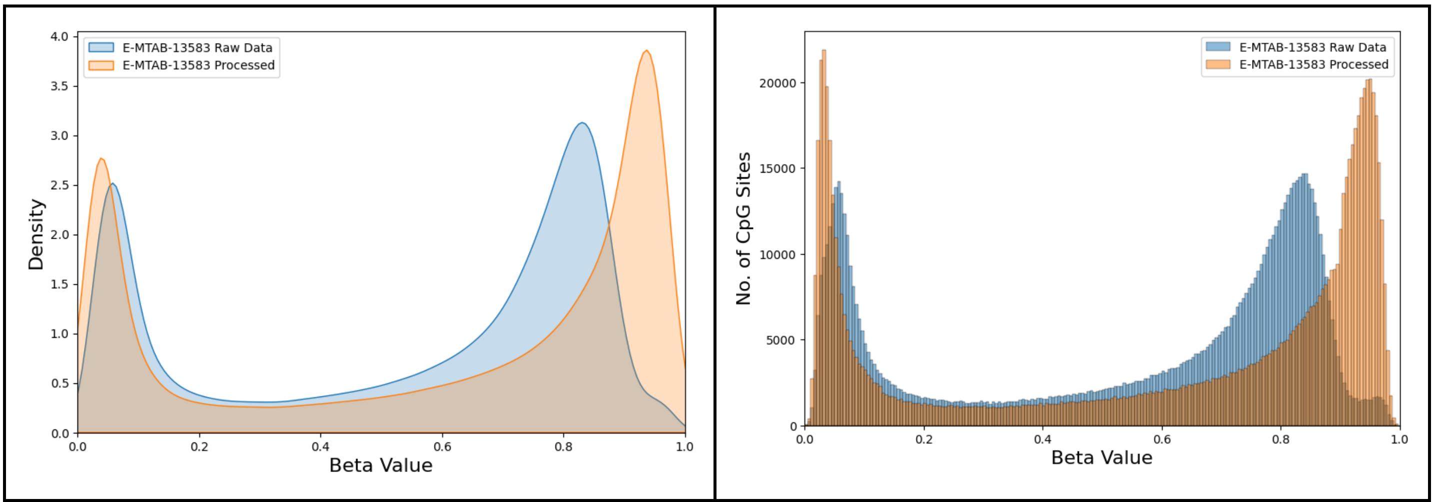


Fig 1. A comparison between processed and raw data publicly available on ArrayExpress for the experiment E-MTAB-13583

Section 1 Pipelines Comparison

Comparison (2) between:	
Processed data	SWAN-Processed Data: Processed beta values provided by the publisher on ArrayExpress.
Quantile pipeline (Minfi package)	*preprocessQuantile(data, fixOutliers = TRUE, removeBadSamples = TRUE, badSampleCutoff = 10.5, quantileNormalize = TRUE, stratified = TRUE).*

To compare the processed data with a complete pipeline, we have chosen *preprocessQuantile* from *Minfi* library as it provides full workflow including the removal of low-quality points in addition to fixing outlier (Fig. 2).

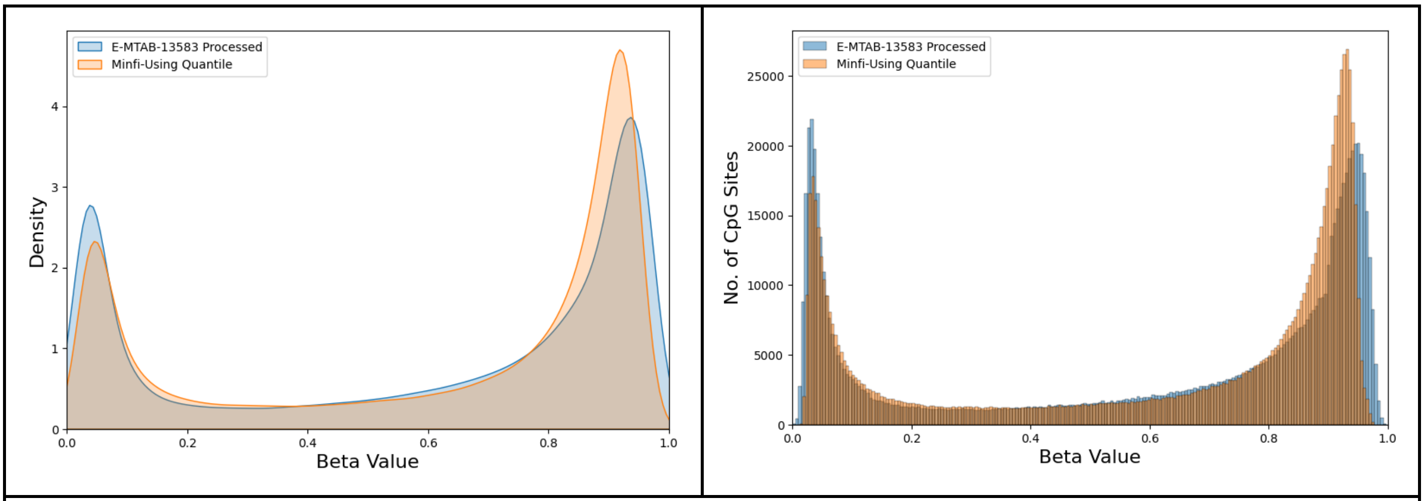


Fig 2. Comparing preprocessQuantile() from *Minfi* with the processed data from the experiment E-MTAB-13583 which used *SWAN* method.

The difference between the two methods is primarily due to difference in normalization technique in Quantile versus SWAN. However, the details of the SWAN-processed data published on ArrayExpress did not declare the specific arguments used in the function. To further confirm that different processing methods result in different distribution of beta values, we have tested one more pipeline used by Enmix, which is another popular library from R Bioconductor packages.

Section 1 Pipelines Comparison

Comparison (3) between:	
Quantile pipeline (Minfi package)	*preprocessQuantile(data, fixOutliers = TRUE, removeBadSamples = TRUE, badSampleCutoff = 10.5, quantileNormalize = TRUE, stratified = TRUE).*
mpreprocess() (Enmix standard pipeline)	*mpreprocess(data, nCores=2,bgParaEst="oob",dyeCorr="RELIC", qc=TRUE,qnorm=TRUE,qmethod="quantile1", fqcfilter=FALSE,rmcr=FALSE,impute=TRUE)*

One of the main differences observed in Enmix library, is the availability of *RELIC* method in adjusting dye bias. The corresponding method used in *Minfi* is the non-linear approach *(dyeCorr = NL)*. Despite that both functions use quantile and applies outlier imputation, The difference of distribution in beta values is still available (Fig. 3). Different factors that can contribute to such a difference. For example, one obvious factor is the application of different dye bias correction methods (RELIC vs Non-Linear).

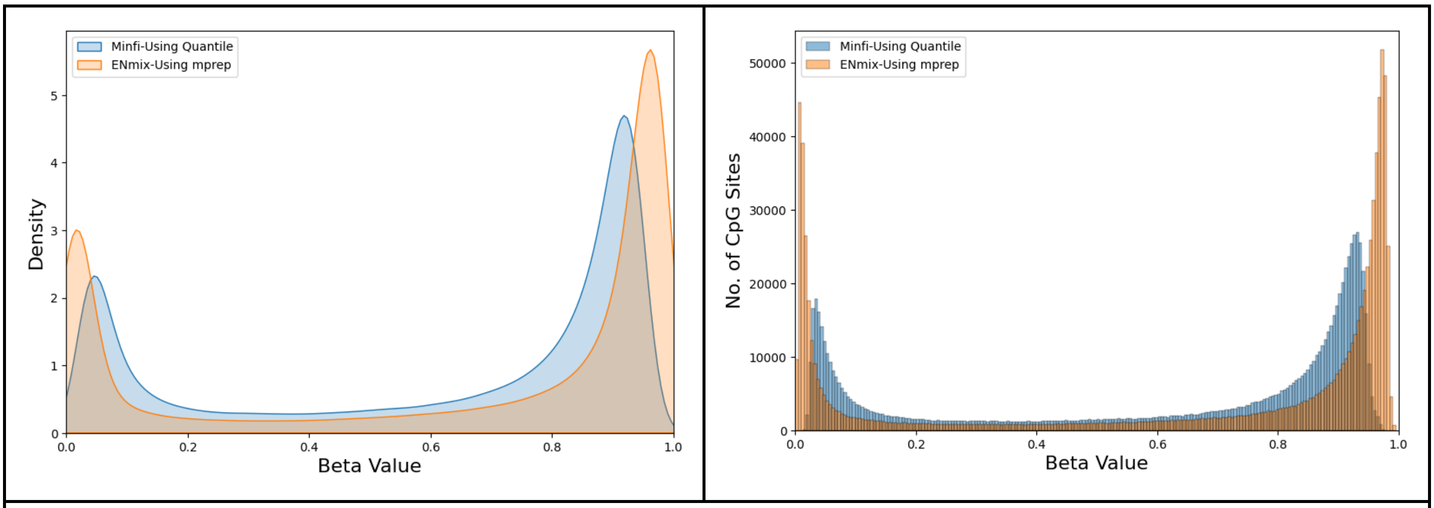


Fig 3. A comparison between mpreprocess() from Enmix and preprocessQuantile() from Minfi

Section 1 Pipelines Comparison

Comparison (4) between:	
Processed data	SWAN-Processed Data: Processed beta values provided by the publisher on ArrayExpress.
Standard preprocessSWAN() (Minfi package)	*if (require(minfiData)) { dat <- preprocessRaw(RGsetEx) preprocessMethod(dat) datSwan <- preprocessSWAN(RgsetEx, mSet = dat) datIlmn <- preprocessIllumina(RgsetEx) preprocessMethod(datIlmn) datIlmnSwan <- preprocessSWAN(RgsetEx, mSet = datIlmn) }*

Attempting to mimic SWAN of the processed data, we used preprocessSWAN() expecting that this attempt will give the same result. The code used for SWAN is as per Minfi reference manual (K. D. Hansen & Fortin, [Minfi User Guide]). Although the same normalization algorithm is used in pipelines, there is still a difference in the beta distribution curve that mostly resulted from quality control steps and thresholds used in each pipeline (Fig. 4).

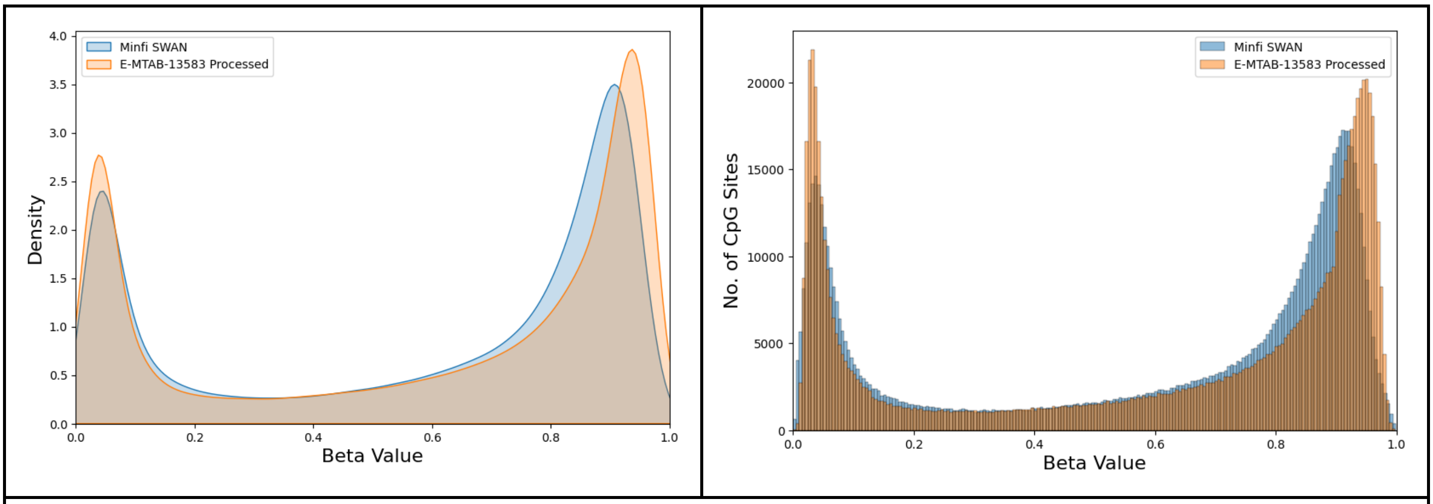


Fig 4. A comparison between preprocessSWAN() standard approach as per Minfi reference manual versus the preprocessSWAN() pipeline used in E-MTAB-13583 experiment.

Section 2 P-value Methods Comparison

```
# P value detection methods comparison
for (path in c(
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/001_HealthyChildren/E-MTAB-
12728",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/01_DLD/E-MTAB-13583",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/02_FTP/E-MTAB-11975",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/03_SAD/GSE164056",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/04_AcuteSleepDep/E-MTAB-4664",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/05_LowSleepImpact/E-GEOD-80559"
)) {

    print("calculating p values . . ")

    # Define input path
    setwd(path)
    input <- path

    # Initialize a data frame to store results
    result <- data.frame(row.names = c(
        "'pOOBAH' Method by SeSAME",
        "'detectionP(M+U)' Method by Minfi",
        "'oob' Method by ENmix",
        "'negative' Method by ENmix"
    ))

    # Sesame - pOOBAH
    idat_files <- searchIDATprefixes(input)
    sesame_counts <- c()

    for (prefix in idat_files) {
        sset <- readIDATpair(prefix) # Process each IDAT pair individually
        sset <- pOOBAH(sset) # Apply pOOBAH for detection p-values
        sesame_counts[basename(prefix)] <- sum(sset$mask) # Count failed probes
    }
    result["'pOOBAH' Method by SeSAME", names(sesame_counts)] <- sesame_counts

    # Minfi - detection P values
    rgSet <- read.metharray.exp(input)
    p_val_minfi <- detectionP(rgSet)
    count_minfi <- colSums(p_val_minfi > 0.05)
    result["'detectionP(M+U)' Method by Minfi", names(count_minfi)] <- count_minfi

    # ENmix - oob method
    rgSetEX <- read.metharray.exp(input, extended = TRUE)
    p_val_oob <- calcdetP(rgSetEX, detPtype = "oob")
    count_oob <- colSums(p_val_oob > 0.05)
    result["'oob' Method by ENmix", names(count_oob)] <- count_oob

    # ENmix - negative method
    p_val_neg <- calcdetP(rgSetEX, detPtype = "negative")
    count_neg <- colSums(p_val_neg > 0.05)
    result["'negative' Method by ENmix", names(count_neg)] <- count_neg

    # Add a column for averages
    result$Average <- rowMeans(result, na.rm = TRUE)

    # Write the result to a CSV file
    write.csv(result, file = "[1] Pval_Detec_Methods.csv", row.names = TRUE)

    print("Check csv comparison table")
}
```

Section 3 Preprocessing IDATs

```
print("[4] Processing Group of Idats..")
# Loop over each path and perform the tasks
for (path in c(
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/01_DLD/E-MTAB-13583",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/02_FTP/E-MTAB-11975",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/03_SAD/GSE164056",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/04_AcuteSleepDep/E-MTAB-4664",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/05_LowSleepImpact/E-GEOD-80559"
)) {
    print("Processing Starts. . . ")
    # Set Working Directory as same as the input
    setwd(path)
    input = path

    # Load the Required Libraries
    library(parallel) # in order to use mclapply()
    library(sesame)

    ### ------------------------------------
    ### Function to process each IDAT pair
    process_idat <- function(px) {
        # Step 1: Apply qualityMask
        masked_data <- qualityMask(readIDATpair(px))

        # Step 2: Apply dyeBiasNL and extract p-values
        corrected_data <- dyeBiasNL(masked_data, mask = TRUE) # Equal to standard
dyBiasNL()
        pvalues <- pOOBAH(corrected_data, return.pval = TRUE) # Extract p-values

        # Step 3: Apply pOOBAH (using corrected_data from step 2)
        p_value_data <- pOOBAH(corrected_data, combine.neg = TRUE, pval.threshold =
0.05) # Equal to standard pOOBAH()

        # Step 4: Apply noob
        noob_data <- noob(p_value_data, combine.neg = TRUE, offset = 15) # qual to
standard noob()

        # Step 5: Get betas
        betas <- getBetas(noob_data)

        return(list(betas = betas, pvalues = pvalues)) # Return both betas and p-
values
    }

    # Define input directory (Assign Multipe Inputs For Multi Experiments).
    # Each Input Needs to Be Different Path In Order Not To Mix Up The Outputs.

    ### ------------------------------------
    # Locate IDAT files and process using above function

    idat_prefixes <- searchIDATprefixes(input)
```

```r
    results <- mclapply(
        idat_prefixes,
        process_idat
    )

    # Combine betas and p-values
    betas <- do.call(cbind, lapply(results, `[[`, "betas"))
    pvalues <- do.call(cbind, lapply(results, `[[`, "pvalues"))

    # Convert betas and pvalues to data frames
    betas <- as.data.frame(betas)
    pvalues <- as.data.frame(pvalues)

    # Rename columns to end with _Betas and _Pval
    colnames(betas) <- paste0(colnames(betas), "_Betas")
    colnames(pvalues) <- paste0(colnames(pvalues), "_Pval")

    # Merge betas and p-values
    merged_data <- data.frame(ProbeID = rownames(betas)) # Start with ProbeID
    rownames(betas) <- NULL # Remove row names for binding
    rownames(pvalues) <- NULL

    # Interleave columns from betas and pvalues
    for (i in 1:ncol(betas)) {
        merged_data <- cbind(merged_data, betas[, i], pvalues[, i])
        colnames(merged_data)[(2 * i)] <- colnames(betas)[i] # Rename to beta column
        colnames(merged_data)[(2 * i + 1)] <- colnames(pvalues)[i] # Rename to p-
value column
    }

    # Write merged data to CSV file
    write.csv(merged_data, "[4] Betas_Pval.csv", row.names = FALSE)

    # Optional: frees up memory
    rm(list = ls())
    gc()

    # Check Betas QC in Python
    print("Check Betas in Python")
}
```

Section 4 Masking Summary

```
print("[2-3] Masking summary")

# Loop over each path and perform the tasks
for (path in c(
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/001_HealthyChildren/E-MTAB-
12728",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/01_DLD/E-MTAB-13583",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/02_FTP/E-MTAB-11975",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/03_SAD/GSE164056",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/04_AcuteSleepDep/E-MTAB-4664",
    "C:/Users/Saeed.LAPTOP-0UBK4QVG/Documents/Target/05_LowSleepImpact/E-GEOD-80559"
)) {
    # Set Working Directory as same as the input
    setwd(path)
    input <- path

    # Locate IDAT file prefixes
    idat_files <- searchIDATprefixes(input)

    # Initialize list to store results
    results <- list()
    results_1 <- list()

    # Iterate over each IDAT pair
    for (file in idat_files) {
        s <- readIDATpair(file)

        # Check if SeSAME Identifies the platform:
        platform <- sdfPlatform(s)

        # Step 1: Check initial masking
        initial_mask <- sum(s$mask)

        # How Many Missing Betas Before QC:
        qcstat <- sesameQC_calcStats(s)
        missing_betas <- sesameQC_getStats(qcstat, "frac_na_cg")

        # Modification 1: Calculate Probe Success rate using:
        rate_05_1 <- probeSuccessRate(s, mask = TRUE, max_pval = 0.05)
        rate_01_1 <- probeSuccessRate(s, mask = TRUE, max_pval = 0.01)

        # Step 2: Apply Quality Mask
        s1 <- qualityMask(s)
        quality_mask <- sum(s1$mask) # Masks the un unique mapping or influenced by
SNPs (SeSAME tutorial)

        # How Many Missing Betas After QC:
        qcstat_1 <- sesameQC_calcStats(s1)
        missing_betas_1 <- sesameQC_getStats(qcstat_1, "frac_na_cg")

        # Mddification 2: Calculate Probe Success rate using:
```

```
        rate_05_2 <- probeSuccessRate(s1, mask = TRUE, max_pval = 0.05)
        rate_01_2 <- probeSuccessRate(s1, mask = TRUE, max_pval = 0.01)

        # Step 3: Apply dyeBiasNL (mask not modified)
        s2 <- dyeBiasNL(s)
        dyeBiasNL_mask <- sum(s2$mask)

        # Step 4: Apply pOOBAH over the original data
        s3 <- pOOBAH(s)
        pOOBAH_mask <- sum(s3$mask)

        # Apply pOOBAH over the corrected Data
        s2_1 <- dyeBiasNL(s1) # s2_1 is the corrected Data
        s3_1 <- pOOBAH(s2_1)
        rate_05_3 <- probeSuccessRate(s3_1, mask = TRUE, max_pval = 0.05)
        rate_01_3 <- probeSuccessRate(s3_1, mask = TRUE, max_pval = 0.01)

        # How Many Missing Betas After pOOBAH for the original data:
        qcstat_2 <- sesameQC_calcStats(s3)
        missing_betas_2 <- sesameQC_getStats(qcstat_2, "frac_na_cg")

        # Step 5: Apply noob
        s4 <- noob(s)
        noob_mask <- sum(s4$mask)

        # How Many Missing Betas After noob for the original data:
        qcstat_3 <- sesameQC_calcStats(s4)
        missing_betas_3 <- sesameQC_getStats(qcstat_3, "frac_na_cg")

        # Mddification 3: Calculate Probe Success rate using:
        s4_1 <- noob(s3_1)
        rate_05_4 <- probeSuccessRate(s4_1, mask = TRUE, max_pval = 0.05)
        rate_01_4 <- probeSuccessRate(s4_1, mask = TRUE, max_pval = 0.01)

        # How many missing Betas After Complete Processing
        # QC Mask + Dye Corr + pOOBAH + noob:
        qcstat_4 <- sesameQC_calcStats(s4_1)
        missing_betas_4 <- sesameQC_getStats(qcstat_4, "frac_na_cg")

        # Calculate Total Masked
        Total_Masked_Probes <- sum(s1$mask) + sum(s2$mask) + sum(s3$mask) +
sum(s4$mask)
        Total_Missing_Betas <- missing_betas_4

        # Store results with base name of IDAT file
        results[[basename(file)]] <- c(platform, initial_mask, missing_betas,
quality_mask, missing_betas_1, dyeBiasNL_mask, pOOBAH_mask, missing_betas_2,
noob_mask, missing_betas_3, Total_Missing_Betas, Total_Masked_Probes)
        results_1[[basename(file)]] <- c(rate_05_1, rate_05_2, rate_05_3, rate_05_4,
rate_01_1, rate_01_2, rate_01_3, rate_01_4)
    }

    # Convert the results list to a data frame
    summary_df <- as.data.frame(do.call(cbind, results))
    summary_df_1 <- as.data.frame(do.call(cbind, results_1))
    # Ensure only numeric rows are used for averaging
    numeric_cols <- sapply(summary_df, is.numeric)

    # Calculate average for only the numeric columns
    summary_df$avg <- rowMeans(summary_df[, numeric_cols], na.rm = TRUE)

    # Assign row names for steps
    rownames(summary_df) <- c("Platform Recognized", "No. of Masked Probes in The
Raw Sample", "Perc. Of Missing Betas in The Raw Sample", "No. of Masked Probes After
qualityMask()", "Perc. Of Missing Betas After qualityMask()", "No. of Masked Probes
with dyeBiasNL()", "No. of Masked Probes with pOOBAH()", "Perc. Of Missing Betas As
a Result of pOOBAH", "No. of Masked Probes with noob()", "Perc. Of Missing Betas As
a result of noob()", "Perc. Of Missing Betas After (qualityMask() + dyeBiasNL() +
pOOBAH() + noob())", "Total Masked probes")
    rownames(summary_df_1) <- c("05 Rate for Raw Data", "05 Rate after
qualityMask()", "05 Rate after pOOBAH()", "05 Rate after noob()", "01 Rate for Raw
Data", "01 Rate after qualityMask()", "01 Rate after pOOBAH()", "01 Rate after
nnob()")

    # Write summary to CSV
    write.csv(summary_df, "[2] Masking_Summary.csv", row.names = TRUE)
    write.csv(summary_df_1, "[3] SuccessRate_Summary.csv", row.names = TRUE)


    # Optional: frees up memory
    rm(list = ls())
    gc()

    print("One iteration is done..")
}
```

Section 5 Limma Analysis

```
perform_limma_analysis <- function(case_path, control_path, filename) {
    # Load necessary library
    library(limma)

    # Load datasets
    case_data <- read.csv(case_path)
    control_data <- read.csv(control_path)

    # Ensure 'ProbeID' column exists in both datasets
    case_data <- case_data[, c("ProbeID", grep("_Betas$", names(case_data), value =
TRUE))]
    control_data <- control_data[, c("ProbeID", grep("_Betas$", names(control_data),
value = TRUE))]

    # Merge datasets on 'ProbeID'
    merged_data <- merge(case_data, control_data, by = "ProbeID")

    # Set ProbeID as row names and remove the column
    rownames(merged_data) <- merged_data$ProbeID
    merged_data <- merged_data[, -1]

    # Create group labels
    group <- factor(c(rep("Case", ncol(case_data) - 1), rep("Control",
ncol(control_data) - 1)))

    # Design matrix
    design <- model.matrix(~ group)

    # Fit the linear model
    fit <- lmFit(merged_data, design)

    # Apply empirical Bayes moderation
    fit <- eBayes(fit)

    # Get top differentially methylated probes
    results <- topTable(fit, coef = 2, number = Inf)  # coef=2 represents the
"group" variable

    # Save results to CSV
    write.csv(results, paste0(filename, " Results.csv"), row.names = TRUE)

    # Filter probes based on adjusted p-value
    significant_probes <- results[results$adj.P.Val < 0.05, ]

    # Save significant probes to CSV
    write.csv(significant_probes, paste0(filename, " Significant Probes.csv"),
row.names = TRUE)

    # Return significant probes for further analysis
    return(significant_probes)
}
```

```
################################################################
# Paths
> cases_path <- "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized"
>
    >
    > # Print the list of files
    > print(list.files(path = cases_path, full.names = TRUE))
[1] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/01_DLD(11)cases(NORM).csv"
[2] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/02_Early_Alz(5)cases(NORM).csv"
[3] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/02_Fam_Alz(6)cases(NORM).csv"
[4] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/02_FTP_GRN(5)cases(NORM).csv"
[5] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/02_FTP_MAPT(3)cases(NORM).csv"
[6] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/03_ELA(27)cases(NORM).csv"
[7] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/03_SAD(32)cases(NORM).csv"
[8] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/03_SAD_ELA(29)cases(NORM).csv"
[9] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/04_TSD(17)cases(NORM).csv"
[10] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/05_LSI(1)cases(NORM).csv"



> controls_path <- "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized"
>
    > # Print the list of files
    > print(list.files(path = controls_path, full.names = TRUE))
[1] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/01_DLD(9)controls(NORM).csv"
[2] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/02_FTP(5)controls(NORM).csv"
[3] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/03_SAD(42)controls(NORM).csv"
[4] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/04_TSD(15)controls(NORM).csv"
[5] "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/05_LSI(1)controls(NORM).csv"

################################################################
# Example usage

case_path <- "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CASES/Normalized/02_Early_Alz(5)cases(NORM).csv"
control_path <- "C:/Users/Saeed.LAPTOP-
0UBK4QVG/Documents/Target/0001_CONTROLS/Normalized/02_FTP(5)controls(NORM).csv"
filename <- "02 Early Alz"
```

```
significant_probes <- perform_limma_analysis(case_path, control_path, filename)
```

Supplementary Material

Note: project pipeline will be uploaded on github after official publishing.
Github link: https://github.com/users/saeed-svu/projects/1.

Access using google drive (link). List of contents available below:

#	Folder Name	Content
1	Limma DMPs	Differentially Methylated Probes for 9 phenotypes
2	Probe Info	Mapped DMPs infomration from Illumina manifest (HG19)
3	QC2	.ipynb notes (9 cases and 4 controls)*
4	Quality Score	Masked percentage per subject (9 tables)
5	Shared Genes	Genes list per phenotypes + combined
6	Shared Probes	Probes that are shared among phenotypes
7	Shared Regions	Detected coordinates for 9 phenotypes + a table for the shared regionsphenotypes
8	Subject Performance	A table that include subject quality score + outlier counts
9	DNA_Meth_Module_Limma.ipynb	A module that includes all functions and libraries**

* Pipeline example available on https://github.com/saeed-svu/DNAmeth_QC2_Pipeline.
**To be used in all steps except QC2.