SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Neuroscience Area – PhD course in Molecular Biology

Twin-pred:
a method to distinguish monozygotic twins in forensic science application

Candidate: Giorgia Esposito

Advisor: Michele Morgante

Co-advisor: Davide Scaglione

Academic Year 2021 2022

MA PER SEGUIR VIRTUTE E CANOSCENZA
• SISSA •

Science may set limits to knowledge,
but should not set limits to
imagination.

Bertrand Russell (1872-1970)

Contents

Abstract

Monozygotic (MZ) twins discrimination in forensic science remains an unsolved point. Nowadays, conventional DNA profiling techniques use the analysis of the Short tandem repeats (STR) to distinguish between suspects. However, since MZ twins share the same DNA sequences, their discrimination using STR analysis presents several limitations.

To overcome these limitations, scientists focused their attention on the study of DNA epigenetic modifications, and in particular on DNA methylation. DNA methylation is an epigenetic DNA modification that occurs at the 5′ positions of cytosine in CpG dinucleotides. So far, many works have identified epigenetics as a possible suitable solution for identical twins discrimination (Marqueta-Gracia et al., 2018; Vidaki et al., 2017b).

In our study, we set up a suitable protocol to distinguish between identical twins in forensic cases. We identified an end-to-end approach, which ranged from DNA extraction to final statistical analysis. To do that, we first set up the experiments as an analogy of a forensic case experiment. As starting material, we used the buccal swabs, often used in forensic science research. In detail, we collected from two couple of MZ volunteers' twins buccal swab samples, and we then extracted the total DNA. We then prepared NGS libraries using bisulfited conversion strategies, considered the gold standard in methylation study, to potentially target every single methylated cytosine state present in the twins genomes. To better asses a forensic case situation, for each couple we generated a set of Reference experiment libraries and four different Test libraries, two per individual.

In detail, Reference experiment libraries were created from sufficient quantities of DNA, simulating a standard buccal swab sampling while the Test libraries were created from a small quantity of starting DNA, simulating the limiting DNA quantity that may be available from a crime scene. Both reference and test data were analyzed using bioinformatic tools specific for DNA methylated samples *(Bismark Alignment software* and *Methylkit).* Once all cytosine methylation states were decided, we set up a statistical approach to infer the probability of one test sample being either one of the reference samples. Briefly, the binomial probability of each informative sequenced CpG was used as a part to derive a discriminatory estimate for each twin based on the known level of methylation at that site in the Reference data. The

aggregation of such probability components for all sites supplied the final association to the twin. We permuted several analysis parameters to find the best set and we assessed the reliability of the prediction by performing bootstrap analysis on the sets of parameters that gave back the best accuracy in the calling.

Lastly, we studied the trend of prediction accuracy of every single parameter passed into the function, to screen how and if they could affect the prediction.

List of abbreviation and Acronyms

5caC 5-carboxy-cytosine

5hmC 5-hydroxymethyl-cytosine

5hmU 5-hydroxymethyl-uracil

5fC 5-formyl-cytosine

5mC 5- methylated cytosines

AID Activation-induced cytidine deaminase

ANN Artificial neural network

APOBEC Apolipoprotein B mRNA-editing enzyme

BAH Bromo-adjacent homology

BER Base excision repair

CREB1 Cyclic AMP-responsive element-binding protein 1

dsDNA Double strand DNA

DNMT DNA methyltransferases

FBI Federal Bureau of Investigation

HDAC Histone deacetylase

HS High sensitivity

IAP Intracisternal A particle

MAD Mean absolute deviation

MAE Mean absolute error

MBD Methylation binding proteins

MeCP2 Methyl CpG binding protein 2

MeDIP Methyl-DNA immunoprecipitation

MR Million reads

MZ Monozygotic twins

NGS Next generation sequencing

PCR Polymerase chain reaction

PCR-HRM High-Resolution Melting PCR

qPCR Quantitative PCR

RRBS Reduced representation Bisulfite sequencing

RT Room temperature

SAH S-adenosyl homocysteine

SAM S-adenosyl-l-methionine

SMUG1 Single-strand selective monofunctional uracil DNA glycosylase

sjTREC Signal joint T-cell receptor excision circles STR Dhort tandem repeats

TDG Thymine DNA glycosylase

TE Transposable element

Tet Ten-eleven translocation enzyme

TRD Transcriptional repression domain

UHFR Ubiquitin-like, having PHD and RING finger domains

UMI Unique Molecular Identifiers

WGBS Whole genome Bisulfite sequencing

ZBTB38 Zinc Finger and BTB Domain Containing 38

ZBTB4 Zinc finger and BTB domain-containing protein 4

1 The term epigenetics

The term epigenetics was first used in 1942 by Conrad Waddington, an embryologist and developmental biologist working at the Institute of Edinburgh, in Scotland. Thereafter, both the usage of the term and the study of the field increased significantly through the years.

Just in 2010, there were over 13,000 epigenetics publications, while in 2013 the available publications were more than 17,000 (Haig, 2012). From the very beginning, epigenetics referred to the mechanisms involved in cellular differentiation, and, at the time of Conrad Waddington, there were two main views explaining the process of differentiation and development.

The first way of thinking was known as preformation, while the second point of view was referred to as epigenesis. In the preformation theory, it was asserted that all adult characters were already present in the embryo and these characters just simply grow up with the embryo itself. On the other hand, the epigenesis theory posited that new tissues were created from successive interactions between the constituents of the embryo (Linda Van Speybroeck, 2002; Noble, 2015). Conrad Waddington believed that both preformation and epigenesis could be complementary and, taking into account both the hypothesis, he coined the term epigenetics, which refers to the biology that studies the causal interaction that occurs between genes and their products (Dupont et al., 2009; Waddington, 2012).

After this first definition by Conrad Waddington other scientists tried to give their explanations for epigenetics. In 1994 Robin Holliday gave a more detailed definition of epigenetics, summarizing the explanation in two different postulates. In the first one, he was stating that epigenetics was the study of gene expression changes, which occur in cells of adult organisms. In the second explanation, he was also admitting that epigenetics was depending on nuclear inheritance not based on differences in the DNA sequence (Holliday, R.,1994).

This explanation was the first one considering the heritability of expression state and, moving through the years, the term epigenetics bought a more generalized meaning.

Today, the term epigenetics refers to Wu and Morris definition (Wu and Morris, 2001), which is related to the study of changes in the gene function that are considered heritable and that do not entail a change in the entire DNA sequence.

2 Epigenetic modifications

The cells of an organism have the same DNA sequence. Nevertheless, cell types and functions are quite different and well specialized. This perfect differentiation in function and specialization is possible because of qualitative and quantitative differences in gene expression. During the early stage of embryo differentiation, the pattern of gene expression is set up and kept during every cell's division throughout life. In addition to inheriting genetic information, cells acquire information through epigenetics, which encloses different modifications (Gibney and Nolan, 2010). **(Figure 1).**

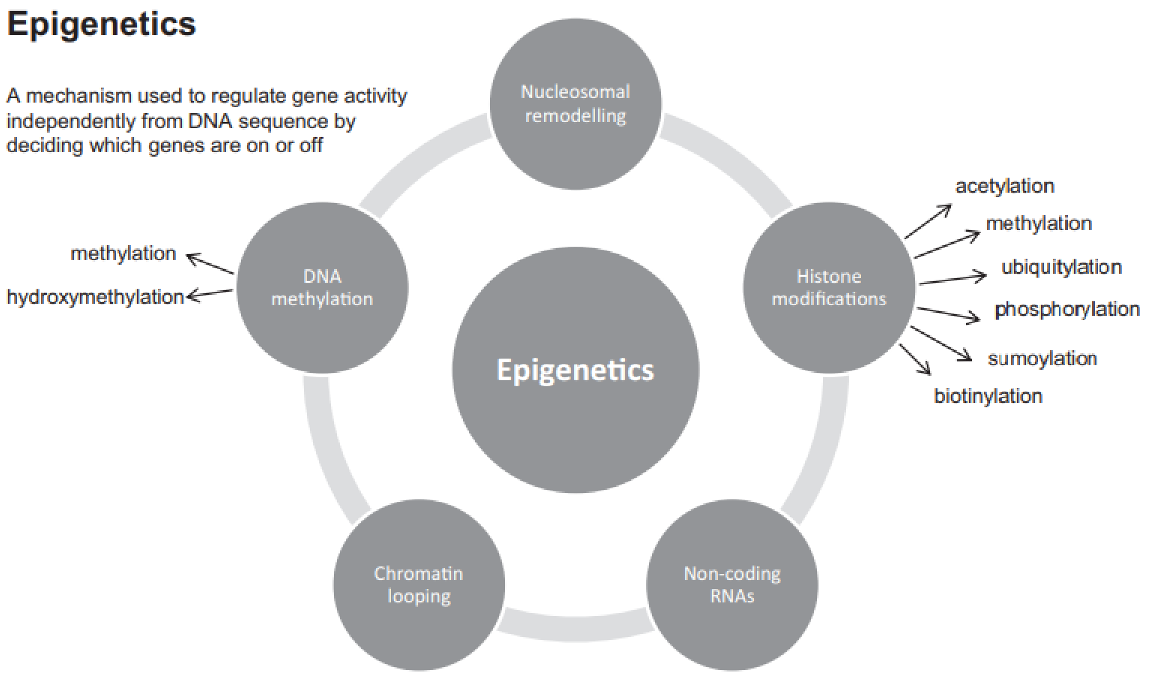


Figure 1 Different mechanisms involved in gene regulation. Epigenetic mechanisms refer to DNA methylation, histone modifications, and non-coding RNAs. One of these different mechanisms can further involve other chemical alterations (Vidaki et al., 2013).

The observed effect after each epigenetic modification is a reversible alteration of the chromatin fiber's structure that results in the transition from an open to close, or vice versa, state. All these epigenetic modifications have been observed to occur in response to environmental exposure and can be affected by various factors such as diet or smoking (Rando and Verstrepen, 2007).

Direct consequences of epigenetic processes are for example gene silencing, the X chromosome inactivation, formation of imprinted genes, and cell reprogramming. One of the most important cell functions regulated by epigenetic mechanisms in mammals is cell differentiation, where stem cells become fully differentiated cells during development (Rando and Verstrepen, 2007).

2.1 DNA Methylation

The experiment that discovered the chemical modification of the methylation of the DNA was performed in 1948 by Rollin Hotchickiss. More in detail, the scientist was working on the preparation of the calf thymus, and, using paper chromatography, he found some modified cytosine in the DNA sequence. He hypothesized that this fraction of modified cytosine was 5-methylcytosine and he suggested that this kind of modification existed naturally in DNA (Hotchkiss, 1948).

Since its first discovery, the biological importance of DNA methylation remains little investigated until 1980, when several studies demonstrated that DNA methylation was involved in gene regulation and cell differentiation (Holliday and Pugh, 1975).

From a chemical point of view, DNA methylation is a biochemical modification involving the addition of methyl group (-CH_3) at the 5' position of cytosine in CpG dinucleotides.

Compared to all the other cells in our body, embryonic stem cells are the only observed cells that do not present methylation in humans (Dodge et al., 2002).

Around 60% - 90% of all the CpGs nucleotides are present in areas of the genome known as CpG islands. These areas are close to promoter genes and can be dynamically methylated directly depending on the active or inactive state of the downstream gene.

On the other hand, the stable methylated cytosines are stored in repetitive areas of the genome, including DNA satellites, LINE-1, and LINE-2 families, Alu and Mir families, and parasitic elements (as DNA transposons and endogenous retroviruses). These areas of the genome are usually located around regulatory regions (5' end) of many human genes and are characterized as 300-3000 bp long (Espada and Esteller, 2010).

The family of enzymes responsible for DNA methylation is known as DNA methyltransferases (DNMTs) that transfer a methyl group from S-adenyl methionine (SAM) to the fifth carbon of a cytosine residue to form 5mC (**Figure 2).**

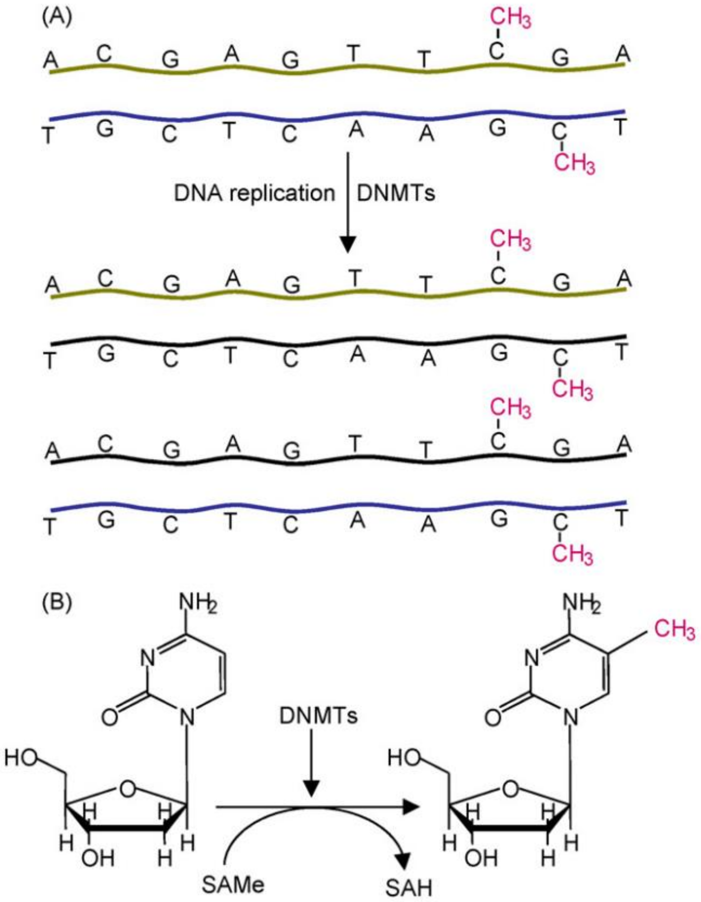


Figure 2 The chemistry behind DNA methylation. **(A)** During DNA replication the DNA methyltransferases (DNMTs) keep a stable pattern of methylation. **(B)** The DNMTs use S-adenosyl-l-methionine (SAMe) as a source of methyl groups, creating the S-adenosyl homocysteine (SAH). DNMTs enzymes can catalyze the addition of methyl groups to the 5 -position of the pyrimidine ring of cytosine (Espada and Esteller, 2010).

2.2 Mechanism of DNA methylation: the DNMTs

The chemical modification of DNA methylation is catalyzed by different classes of enzymes. These families of enzymes manage distinct functions depending on whether they write, read, or erase methylated DNA.

Eraser enzymes handle removing or modifying the methyl group, while readers can recognize and bind methylated DNA to influence gene expression. Writer enzymes refer to a particular family that can catalyze the addition of the methyl group to the cytosine residue. The writer's family are known as DNA methyltransferase (DNMTs), they catalyze the addition of the methyl group to the DNA and include three family members: Dnmt1, Dnmt3a, and Dnmt3b. All three enzymes share a similar structure with a large N-terminal domain that has a regulatory function and a C-terminal with a catalytic domain (Xie and He, 1999; Yen et al., 1992).

Within the DNMTs family, Dnmt1 is the best-studied and well characterized.

The Dnmt1 enzyme is highly expressed in mammalian tissues including the brain(Goto et al., 1994). Dnmt1, differently from the other DNMTs members, preferentially methylates the hemi methylated DNA (Ramsahoye et al., 2000). During DNA replication the Dnmt1 enzyme perfectly replicates on the newly synthesized filament the pattern of methylation present on the original strand of DNA (Hermann et al., 2004). In 2005 Mortusewicz and collaborators have also shown that Dnmt1 also can repair DNA methylation (Mortusewicz et al., 2005). Dnmt1 critical role in cellular differentiation and in dividing cells was demonstrated in knockout mice of Dnmt1 which results in embryonic lethality between E8.0 and E10.5 (Li et al., 1992).

Other members of the DNMTs family are Dnmt3a and Dnmt3b. These two enzymes are extremely similar in structure to Dnmt1, but they show expression patterns quite different from Dnmt1. These two members introduce the methylated group to naked cytosine, and, for this reason, they are known as *de novo* DNMTs.

While Dnmt3a is expressed relatively ubiquitously, Dnmt3b is poorly expressed by the majority of differentiated tissues except for the thyroid, testes, and bone marrow (Xie and He, 1999). Dnmt3a knockout mice can survive up to 4 weeks after their birth, while the Dnmt3b knockout mice, as happens for Dnmt1, are embryonically lethal (Okano et al., 1999).

In DNMTs family, the last characterized member is the Dnmt3L enzyme. This member lacks the catalytic domain present in the other Dnmts enzymes (Aapola et al., 2000).

The Dnmt3L enzyme is always associated with the other Dnmts members because of the lacking of the catalytic domain and it has been reported to stimulate their methyltransferase activity (He et al., 2011).

The Dnmt3L expression is found in early development. Its role seems to be fundamental to establishing maternal and paternal imprinting (Bourc’his and Bestor, 2004; Webster et al., 2005; Zamudio et al., 2011). After the first steps of development, in adulthood, Dnmt3L expression is restricted to the germ cells and in the thymus (Aapola et al., 2000) **(Figure 3)**.

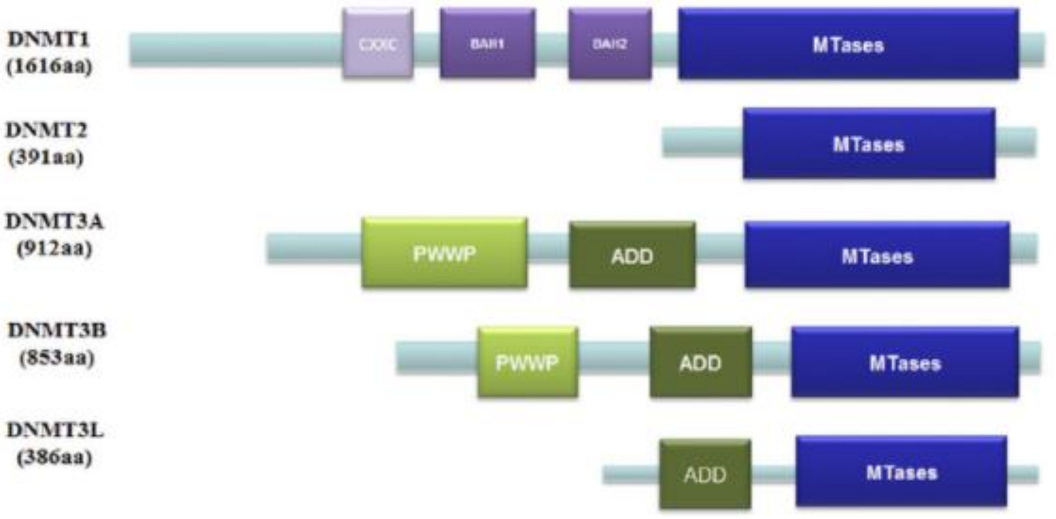


Figure 3 Members of the DNMTs family. The Dnmt1 has three major domains: one zinc finger domain (Cys-X-X-Cys), which can recognize unmethylated CpG, and two BAH domains (Bromo-adjacent homology), and one catalytic domain. The Dnmt2 catalyzes are responsible for the tRNA methylation, it has just the C-terminal catalytic domain, able to bind the cofactor SAM. Dnmt3 shows two domains, one is localized at the N-terminal (PWWP domain), able to bind methyl lysine histones, and one can recognize the non-methylated H3K4, which is the ADD domain (Li et al., 2013).

2.2.1 Reading the DNA methylation

DNA methylation itself can reduce gene expression by preventing transcriptional activator factors, but there are other groups of proteins responsible for inhibiting transcription factor binding; the methylation binding proteins (MBD), the ubiquitin-like-containing PHD, RING finger domain (UHFR) proteins, and the zinc-finger proteins (Moore et al., 2013).

The MBD proteins include different members such as MBD1, MBD2, MBD3, MBD4, and the best characterized, MeCP2. All these proteins contain a conserved methyl-CpG-binding domain that gives the members of the family a higher affinity for single methylated CpG sites (Nan et al., 1993).

The best-studied member, as mentioned above, is MeCP2. This protein has the unique role of directly binding Dnmt1 via the transcriptional repression domain (TRD), and this recruitment is required to maintain the DNA methylation state (Kimura and Shiota, 2003). **(Figure 4)**

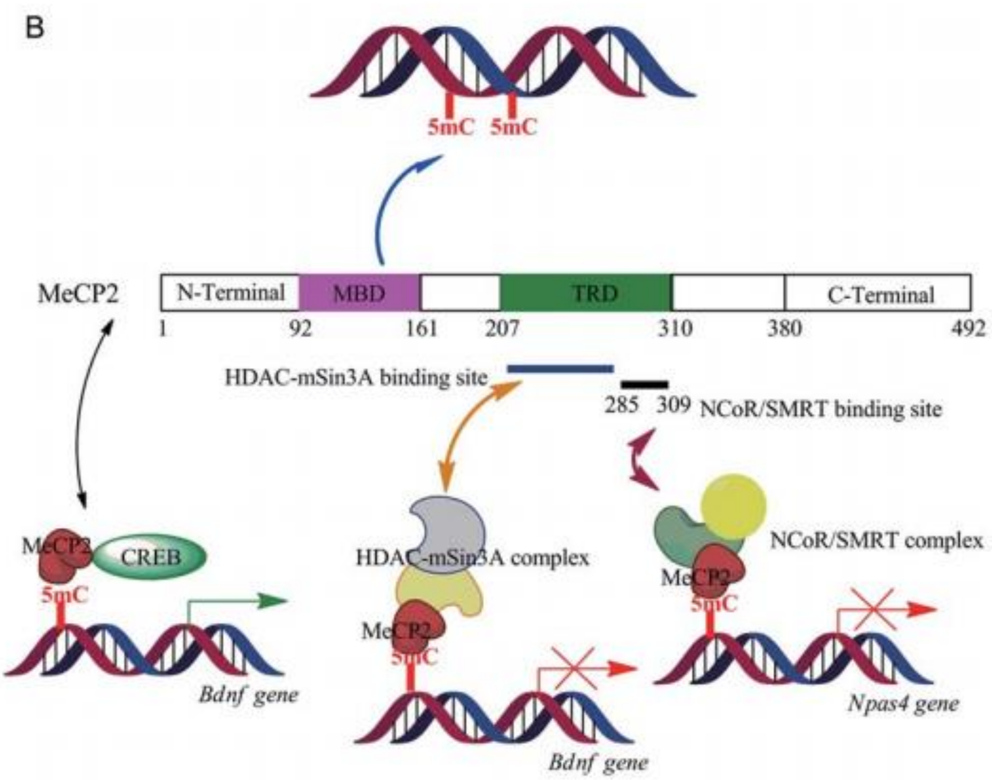


Figure 4 MeCP2 has two functional domains, a methyl DNA-binding domain (MBD) and a transcriptional repression domain (TRD). MeCP2 binding to a methylated DNA site is mediated by

the MBD domain. The physical interaction between MeCP2 and co-repressor complexes (such as HDAC-mSin3A and NcoR-SMRT) depends on the TRD. In addition to co-repressor complexes, MeCP2 has been shown to bind the transcriptional activator CREB1 (Cheng and Qiu, 2014).

The second family of reading proteins is the UHFR family. This family includes UHRF1 and UHRF2 members. These two proteins are multidomain that flip out and bind methylated cytosines via SET- and RING-associated DNA-binding domains (Hashimoto et al., 2009).

The UHRF proteins interact directly with the Dnmt1 to target the hemi-methylated DNA and maintain this state, particularly during the replication (Achour et al., 2008). Because of this particular role, deletion of this protein, as for Dnmt1, leads to embryonic lethality (Muto et al., 2002).

The last zinc-finger family include different proteins, such as Kaiso, ZBTB4, and ZBTB38.

Kaiso protein is shown to preferentially bind two consecutively methylated CpG sites (Daniel, 2002). Similar to the MBD family, the zinc-finger proteins repress transcription in a DNA methylation-dependent manner (Lopes et al., 2008; Prokhortchouk, 2001; Yoon et al., 2003).

2.2.2 Erasing DNA methylation

Active DNA demethylation occurs in multiple ways and involves different enzymes. Up to now, in mammals, unknown mechanisms can cleave the covalent carbon to carbon bound present in cytosines that have a methyl group. However, demethylation can also occur through a series of chemical reactions that turns methylated cytosines by deamination and/or oxidation.

The deamination and oxidation can be recognized by the base excision repair pathway (BER) which will replace the deaminated or oxidated cytosine with a naked cytosine (Moore et al., 2013).

Deamination of the amine of the 5mC into the carbonyl group can be catalyzed by the activation-induced cytidine deaminase/apolipoprotein B mRNA-editing enzyme complex (AID/APOBEC), that converts 5mC into thymine, creating a G/T mismatch recognized after by the BER pathway. Knockout mice for AID results to be viable and fertile (Wu and Zhang, 2014), suggesting that there are different mechanisms for DNA demethylation.

Another mechanism of active demethylations can be mediated by the ten-eleven translocation (Tet) enzymes (Tet1, Tet2, and Tet3). These enzymes add a hydroxyl group to the methyl group of 5mC to produce a 5hmC (Ito et al., 2012). Once this intermediate is formed, two different mechanisms convert 5hmC into naked cytosine. One possible mechanism could be the oxidation by Tet enzymes, forming first from 5hmC the 5-formyl-cytosine and then to 5-carboxy-cytosine (Ito et al., 2011). The second possible mechanism provides that the 5hmC is deaminated by AID/APOBEC to form 5-hydroxymethyl-uracil (Guo et al., 2011).

The BER pathways act after both the mentioned mechanisms. In order to cleave off the modified residue of thymine (5-hydroxymethyl-uracil, 5-formyl-cytosine, and 5-carboxy-cytosine), the BER pathway uses the thymine DNA glycosylase (TDG) which replaces the modified thymine with a naked cytosine (Cortellino et al., 2011; He et al., 2011). **(Figure 5)**

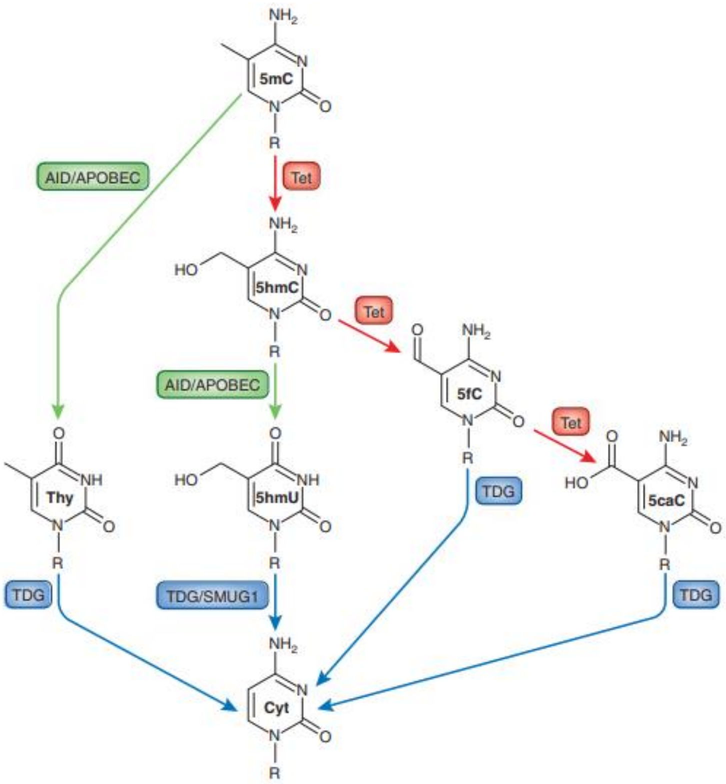


Figure 5 Possible DNA demethylation pathways. The green path describes the deamination by the AID/APOBEC mechanisms, which handle the conversion of 5mC into thymine (Thy). Another pathway (in red) is the addition mediated by Tet enzymes of a hydroxyl group to obtain the 5-hydroxymethyl-cytosine (5hmC) from a methyl group of 5mC. After its formation, the 5hmC can also be changed chemically in the amine group and the hydroxymethyl group. The AID/APOBEC can also deaminate 5hmC to produce 5-hydroxymethyl-uracil (5hmU) (in small green arrow). The last possible pathway refers to the possible action of Tet proteins. They can further oxidize 5hmC to form 5-formyl-cytosine (5fC) and then to 5-carboxy-cytosine (5caC). Eventually, the products of each pathway can be cleaved and replaced with a naked cytosine mediated by TDG and/or SMUG1, both components of the BER. Image modified from (Moore et al., 2013).

2.3 DNA methylation regions

In mammals, cytosine methylation occurs in a different type of cells at various stages of differentiation. In human, the total amount of 5mCs is around ~1% (Ehrlich et al., 1982).

In humans, the most highly methylated DNAs can be found in the thymus and the brain, while the two least methylated DNAs can be found in the placenta and sperm (Ehrlich et al., 1982). 5mCs can be easily deaminated into thymine and this results in the depletion of most of the CpG sites in the mammalian genome. 5mC that avoids deamination into thymine remains highly methylated all across the genome, except for the CpGs islands (Bird et al., 1985). DNA methylation in different genomic regions is proven to exert different influences on gene activity and regulate tissue-specific gene expression, including X chromosome inactivation, and genomic imprinting during the first moment of embryo formation.

2.3.1 Intergenic Regions

In the mammalian genome, almost half (around 45% of the genome), consists of transposable elements (TEs). TEs present in the genome is usually silenced by methylation (Schulz et al., 2006). The silencing of the TEs in the intergenic region is necessary to avoid a possible harmful role in the human genome. TEs are inactivated either by DNA methylation or by different mutations acquired over time, including 5mC deamination (Aran et al., 2011; Hellman and Chess, 2007). One of the most studied cases of transposable element silencing is the intracisternal A particle (IAP). The IAP particle is methylated throughout life in gametogenesis, development, and also adulthood (Gaudet et al., 2004; Walsh et al., 1998). During embryo development, the Dnmt1 maintains the IAP elements strongly methylated, even if most of the embryo genome is still hypomethylated (Gaudet et al., 2004). If Dnmt1 is depleted by genetic mutations, the IAP elements result be expressed, according to the extensive state of hypomethylation (Hutnick et al., 2010; Walsh et al., 1998). Considering these earlier scientific findings, within the intergenic regions, one of the possible roles of DNA methylation could be the active repression of the expression of potentially harmful genetic elements.

2.3.2 Gene bodies

The term gene body refers to the region of a gene found between the start and the stop site codon. This part of the genome refers to areas that will be translated into proteins (exons) and parts that will be cut out from the mRNA and do not translate into proteins (introns). Some studies suggest that the DNA methylation of a gene body is associated with a higher level of gene expression in dividing cells (Aran et al., 2011; Hellman and Chess, 2007). In 2011, Aran and collaborators tried to find a direct correlation between gene-body methylation and gene expression. Their work revealed efficient maintenance of high methylation levels along active gene bodies of the tissues and cell lines studied. Moreover, they also suggested a tissue-specific gene-body methylation pattern that directly reflects the different cell specification history, this hypothesis stemming from the observation that active and inactive genes are equally methylated during early development stages.

However, in non-dividing cells such as the brain, gene body methylation seems not to be associated with increased gene expression (Aran et al., 2011; Xie and He, 1999).

2.3.3 CpG Islands

CpG islands are pieces of DNA of around 1000 base pairs that show a higher CpG density than the rest of the genome. The human genome holds around 30,000 CpG islands found mostly at the promoters level of a gene (roughly 70% of CpGs), and approximately 3% of these human cytosines are methylated (Bird et al., 1985).

CpG islands, especially those associated with promoters, are conserved among species, e.g. mice and humans (Illingworth et al., 2010) and this probably led to the fact that CpG areas are preserved during evolution because of their functional importance.

Interestingly, in mammals, the presence of multiple methylated Cs sites in CpG islands of promoters genes can cause the silencing of genes, i.e., leading to the active or inactive state of a gene **(Figure 6).**

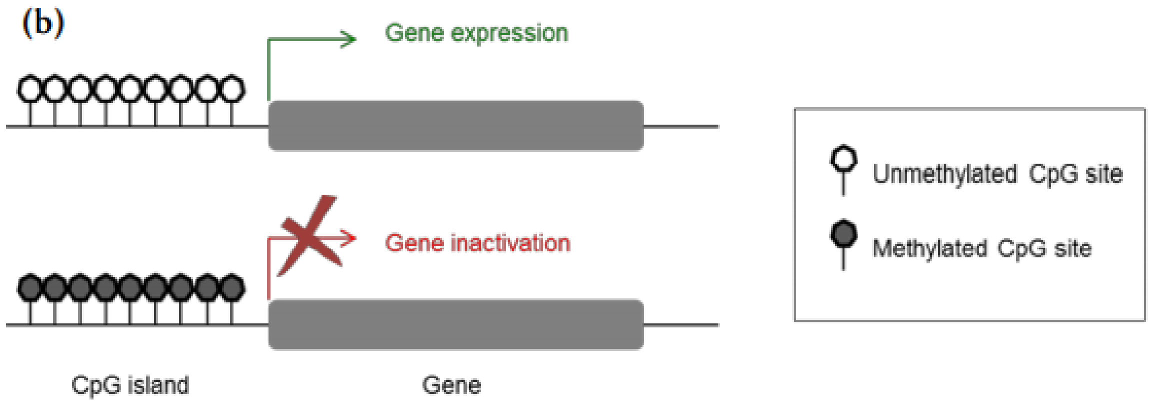


Figure 6 Representation of a gene that has a CpG island. In all cases in which a CpG island is unmethylated the downstream gene is expressed. In contrast, strong methylation of CpG islands results in silencing inactivation of the downstream gene (Vidaki et al., 2013)

While in physiological conditions CpG islands are methylated, it has been shown that in many cancers, this pattern of stable methylation of the promoters' region can be altered leading to cancer development and progression **(Figure 7)**

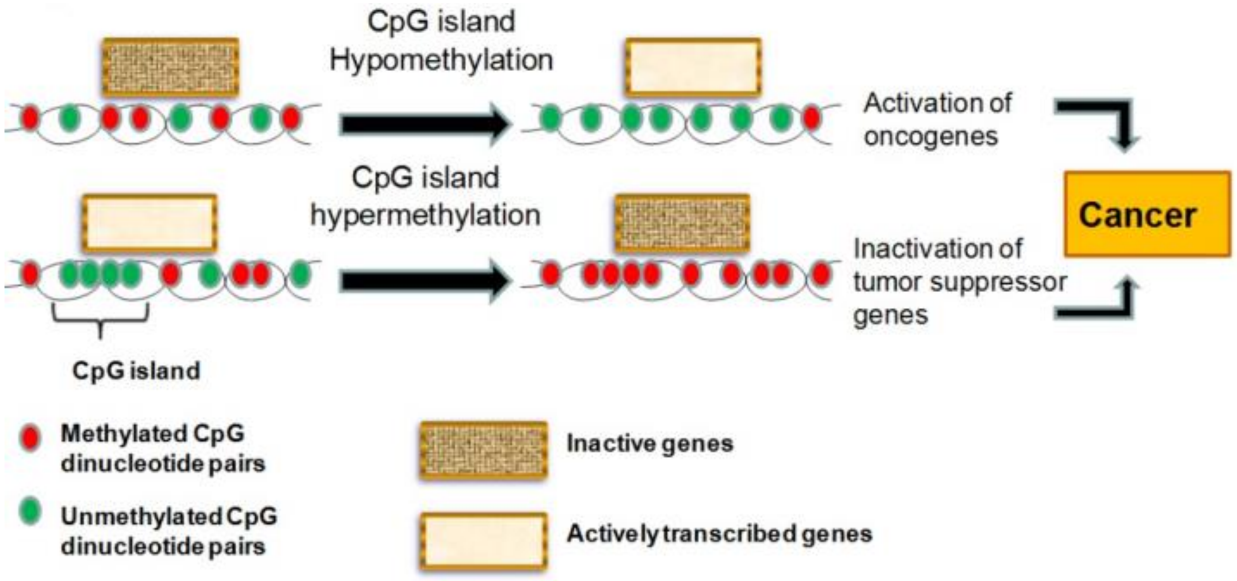

CpG island Hypomethylation
Activation of oncogenes
CpG island hypermethylation
Cancer
Inactivation of tumor suppressor genes
CpG island
Methylated CpG dinucleotide pairs
Unmethylated CpG dinucleotide pairs
Inactive genes
Actively transcribed genes

Figure 7 Aberrant DNA methylation takes part in the development of cancer. More in detail CpGs hypomethylation leads to an active gene state. Thus, when the activation occurs in oncogenes this can directly influence cancer development. On the other hand, hypermethylation of CpG island at the promoter region of tumor suppressor genes can directly lead to the inactivation of protective mechanisms against cancer development (Dricu et al., 2012).

During gametogenesis and embryonic development, CpG islands change their pattern of methylation leading to the moment of development (Kantor et al., 2004). For example, imprinted genes, i.e., that class of genes that are expressed in only one of the two inherited parental chromosomes, are under methylation control. In this case, DNA methylation of CpG islands regulates gene expression during development and differentiation (Meissner et al., 2008; Mohn et al., 2008). Besides being associated with stable silencing of gene expression, CpG island silencing is linked to tissue specification. In intergenic regions and gene bodies, CpG islands show a tissue-specific pattern of methylation, while CpG islands in promoter regions do not (Illingworth et al., 2010; Maunakea et al., 2010, p. 1; Rakyan et al., 2004).

However, further studies are needed to understand the degree of DNA methylation in CpG islands that regulate gene expression.

3 Forensic genetic and Human Identification

Human identification of stays can occur for example after natural disasters or terroristic attacks in which a large number of people are involved. In these particular situations, forensic scientists need to find a way to attribute identity and legal names to unknown individuals. Likewise, circumstances such as incidents or homicides also require scientists' intervention to correctly find the involved people.

Human identification is also needed in case of criminal investigations in which two different suspects are involved.

In all these different situations the available forensic approaches are different and can be distinguished in genetic and physical approaches, depending on the starting case situation.

In particular, physical evidence refers to fingerprints or dentition. These features can be used to find an unrecognizable individual as long as the material composition is not compromised.

In other cases, it is not possible to attribute the identity through physical features, for instance when decomposition or trauma from high-impact events take place. In these situations, the genetic approach using DNA evidence can be the only remaining source of setting up a reliably human identity.

3.1 Conventional DNA-based methods for human identification: The Short tandem repeats.

Between individuals, just 0.3% of the human genome can be considered different and unique for every person. Thus, forensic scientists have been focusing on this 0.3%, trying to set up a common pattern of discriminatory sites for every individual.

Nowadays, the main set of markers on which forensic scientists rely is the Short Tandem Repeats (STRs). STRs are regions of the DNA known to be very high in mutation rates and variability between human individuals (Amorim and Pereira, 2005).

Polymerase chain reaction (PCR) typing of STRs is the preferred method for DNA human identification in forensic science. During forensic human identification, the STR analysis results in the generation of a profile that can be directly compared against an already created database that has all the STR profiles obtained from crime scene samples, suspected people, missing people, and their family members. After all the available comparisons, the proper match between the two profiles is giving a level of statistical power in assessing the profile in a population database that estimates the allele frequencies in a representative population.

Since the beginning of the use of STRs, in 1990, the Federal Bureau of Investigation (FBI) in the United States has created a database of DNA profiles called the Combined DNA Index System, CODIS (Norrgard, K. (2008) Forensics, DNA fingerprinting, and CODIS (*Nature Education* 1(1):35) **(Figure 8).**

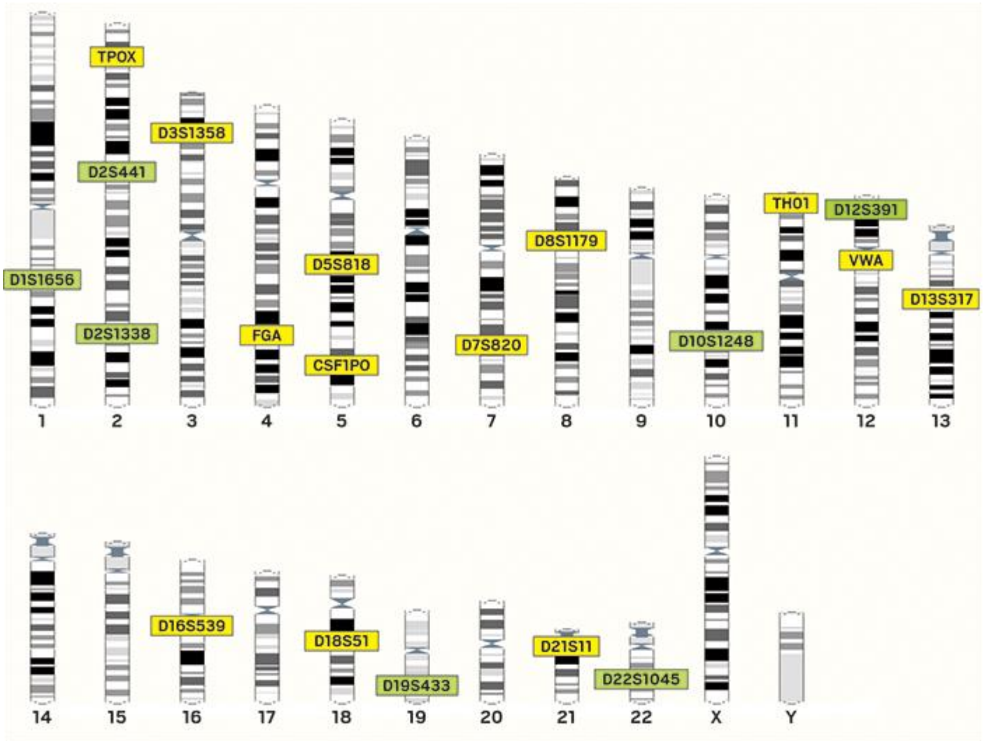


Figure 8 22 plus X and Y chromosomes and corresponding STR loci used in the Combined DNA Index System (CODIS). In the yellow boxes, the original 13 STR sites were used, in the green boxes, the newest seven were added in January 2017 (Ensemble (map)/NIST (loci locations).

3.2 Influence of low quality and poor quantity of DNA

In crime scenes, the biological samples available for forensic studies are not enough both in terms of quantity and quality, due to the degradation of the DNA molecules.

The biological DNA decomposition is mainly caused by two major factors: environmental conditions and post-mortem interval (Burger et al., 1999).

Environmental conditions refer to external factors such as pH, temperature, and humidity of the air. All these factors can greatly influence the intensity and advancement of the degradative processes (Rohland et al., 2018). Moreover, conditions such as warm or wet habitats could also promote microbial infestation, resulting in altered DNA base composition (Alaeddini et al., 2010). It is also known that a more humid environment can strongly accelerate the chemical breakdown of the sugar-phosphate backbone in DNA, resulting in

DNA fragmentation. Covalent linkages between C or T bases could be induced by photochemical exposure. This modification results in pyrimidine dimers that can cause DNA polymerases to get stuck during PCR replication.

The post-mortem DNA damage causes a strong DNA fragmentation that affects directly the STR amplification. In detail, the amplification of STR loci targets DNA sequences between 100 bp-500 bp and requires 80 intact cells to obtain the right amount of DNA for successful typing (Kline et al. 2005).

3.3 Forensic epigenetic

Forensic epigenetics refers to the field of study that uses epigenetic techniques to address questions of interest to the court of law and criminal investigators.

During the last years, forensic epigenetics has become significantly important in forensic investigations because of the possibility to obtain not only the molecular fingerprint or the STRs profile of a suspect but also information about the physical traits or lifestyle of the suspect.

In contrast to forensic epigenetics, current standard DNA profiling techniques are not able to predict the possible appearance traits and remain completely comparative, i.e., these technologies are just able to match DNA profiles from crime scene traces with that of known suspects.

Today, epigenetics in forensics is used for the determination of body fluids, the estimation of a person's age, and the discrimination between identical individuals, such as in the case of monozygotic twins (Forat et al., 2016; Vidaki et al., 2018, 2017a).

From 2016 to now, in the field of forensic epigenetics, more than 20 papers have been published in which DNA methylation was used to estimate a person's age and tissue identification.

In the future, the use of epigenetics could help even more. As already discussed, epigenetic modifications are directly correlated with external environmental factors. This aspect will help forensic scientists for example in predicting the appearance traits of a suspect or giving details about his or her habits, e.g., if they could be a smoker, their alcohol intake, or drug consumption **(Figure 9).**

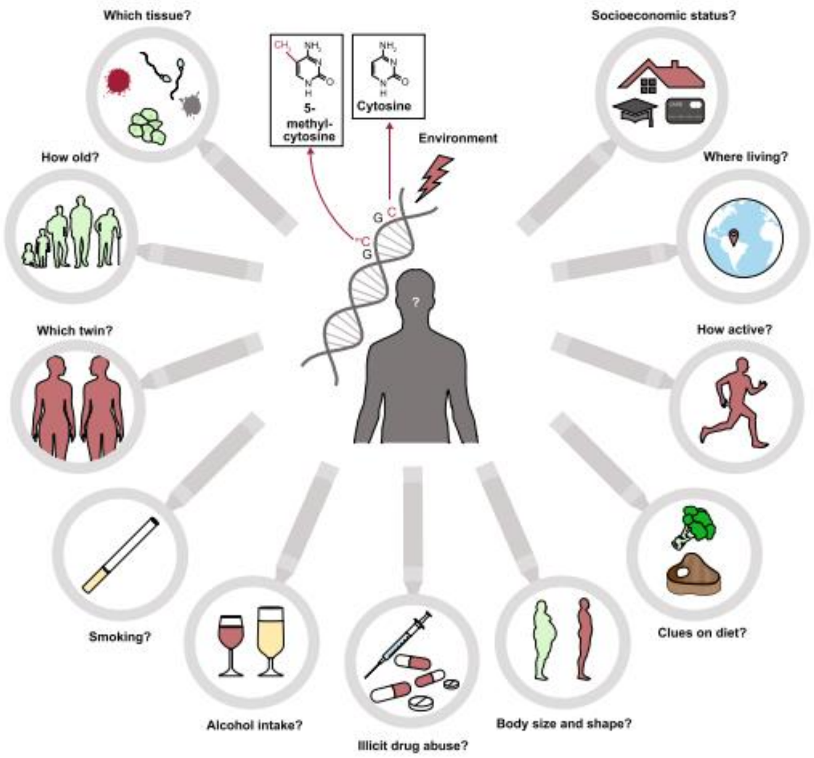


Figure 9 Possible questions forensic epigenetics could answer in the future (Vidaki and Kayser, 2017).

3.3.1 DNA methylation for tissue type identification

Finding tissue sources of biological material on a crime scene can be powerful information in many forensic cases. Nowadays, suitable tissue-specific CpG markers have been found and confirmed among different tissue types. At the same time, different technical approaches have been confirmed to evaluate CpGs markers of discrimination.

In this area of study in 2016, Lin et al. were able to find a set of eight tissue-specific CpG markers. They were able to identify the new tissue-specific associated markers using the Illumina Human Methylation450 Beadchip microarray, which taken together with two control markers formed a 10-plex assay based on the methylation-specific restriction enzyme (MSRE)-PCR system (Lin et al., 2016).

In the same year, Forat and collaborators found other 150 candidates recognized as possible tissue-specific markers for the identification of saliva, semen, blood, vaginal fluid, and menstrual blood, using the Illumina Human Methylation450 Beadchip microarray. Between these 150 candidates, 9 markers were found to be promising for tissue discriminations. During the discovery of the markers, they also checked the genomic stability of these 9 selected markers studying the potential influence of 12 relatively common tumors on the selected sites, concluding that just in the case of cervix carcinoma the vaginal samples could results as affected and different from the normal state (Forat et al., 2016).

In 2016, Vidaki and collaborators found 11 body fluid-specific CpG sites from blood, semen, and buccal cell samples, showing tissue-specific methylated levels (Vidaki et al., 2016).

During sexual attacks, a crucial point is the possibility to discriminate between whole blood and menstrual blood. Until now, the identification of menstrual fluid in trace evidence was challenging. In 2018 Holtkötter and collaborators evaluated a total of 11 reported CpG sites for their potential to differentiate between whole and menstrual blood, identifying BLU2 as the most suitable and reproducible markers of discrimination in these two fluids (Holtkötter et al., 2018) **(Figure 10-11).**

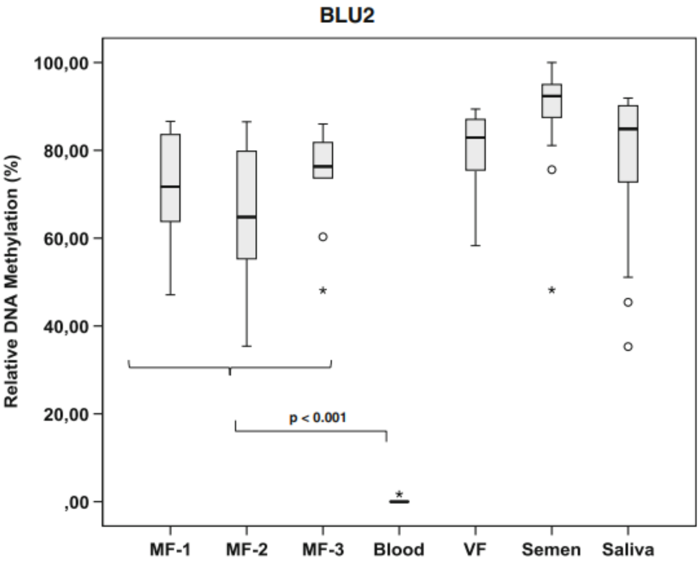


Figure 10 % of DNA in methylation for the blood-specific marker BLU2. This marker was identified as the most discriminant between blood and menstrual fluid. Moreover, it shows the lowest variation in methylation levels within body fluid sets (MF menstrual fluid, 1 day 1 of the menses, 2 days 2 of the menses, 3 days 3 of the menses, VF vaginal fluid) (Holtkötter et al., 2018).

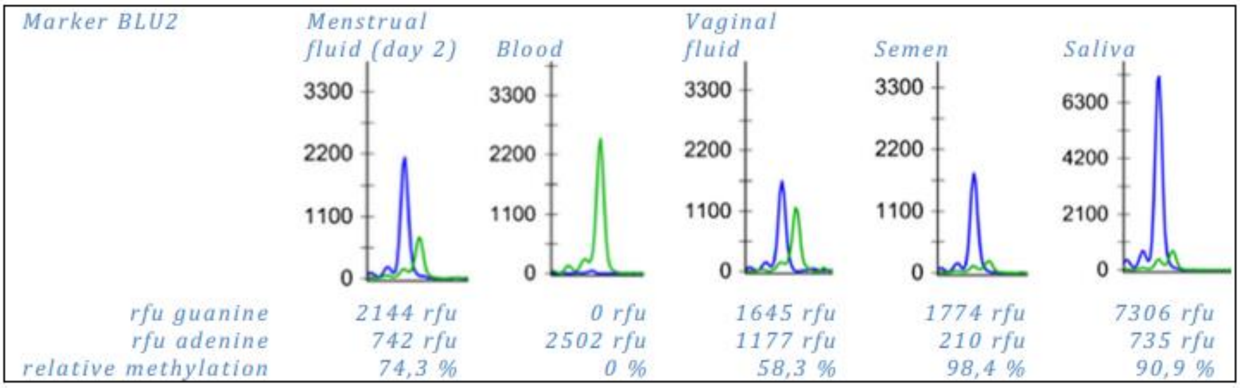
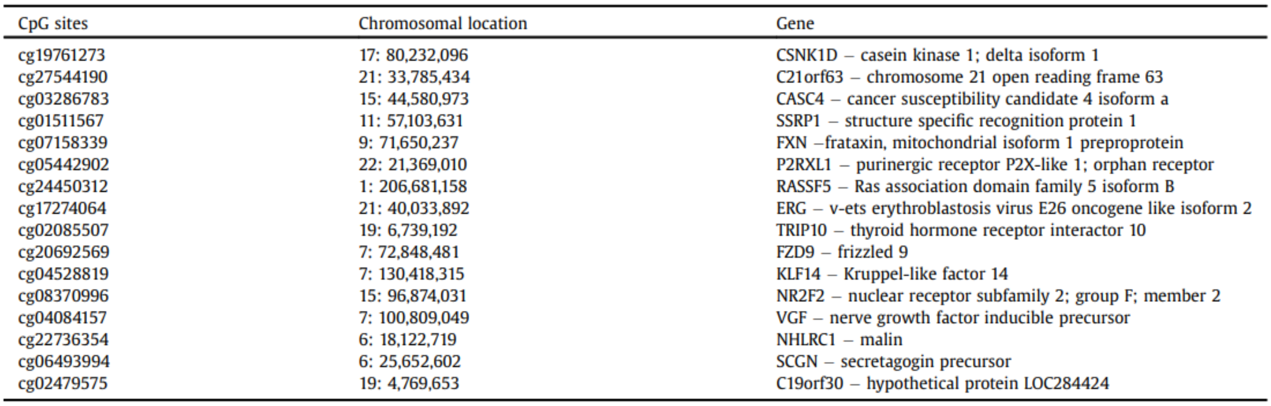


Figure 11 Use of the marker BLU2 found by Holtkötter and collaborators. Using BLU2 it is possible to distinguish blood from the menstrual fluid as well as from vaginal fluid, semen, and saliva. While blood shows complete unmethylation, all other fluids express hypermethylation. The blue peak (guanine) represents the methylated and the green peak (adenine) the unmethylated cytosines (Holtkötter et al., 2018).

3.3.2 DNA methylation for age-person evaluation

During forensic cases, estimating the age of the suspects from a biological sample may supply the essential information. Age prediction in forensic cases is relevant not only on its own but also in combination with features that can be strongly related to age-dependent traits (such as hair loss or hair greying). As proposed strategies for age person estimation, scientists can use different molecular evaluations, such as (i) telomere length, which is known to decrease with increasing age (Weidner et al., 2014), (ii) mutations that occur in the mRNA molecules, which result in the accumulation with the increasing of age person, the rearrangement of T-cell DNA (sjTREC) (Zubakov et al., 2016), and (iii) alterations that are associated to proteins, such as the racemization of aspartic acid and advanced glycation end-product (Wochna et al., 2018). Among all the proposed methods of identification and all the studied strategies, nowadays, many scientists are focusing their attention on DNA methylation, which gives the best accuracy in terms of the degree of errors.

In 2017, Hong and collaborators were able to generate individual genome-wide DNA methylation profiles from 54 individuals. Among the DNA-methylation profiles of all participants, they were able to find CpG markers that showed a high correlation between methylation and age. Six of these key age-dependent sites were present in saliva. Moreover, the 6 recognized sites were combined with a cell type-specific marker for both blood and buccal cells allowing them to create a novel 7-plex methylation SnaPshot® system that provides an age prediction showing a mean absolute deviation (MAD) of 3.2 years (Hong et al., 2017). During the same year, Alghamin and collaborators, using bisulfite pyrosequencing, found a new set of methylated markers to estimate human age. In their work, they used a large court of blood and saliva samples (72 blood samples and 91 saliva samples) from people aged 5 to 73, to find the linear correlation between some genetic loci and chronological age. The scientists examined 27 different CpG sites at three previously reported genetic loci (SCGN, DLX5, and KLF14), proposing single- and dual-locus age models resulting in MAD=8 years and MAD=7.1 years, respectively for single and dual locus (Alghanim et al., 2017). Finally, in the same year, Vidaki and collaborators proposed another approach to create a set of discriminatory age-related markers. In their work, they create not only a potential age-associated marker set but also a novel method for prediction analysis, namely machine learning by artificial neural network analysis (ANN). By doing that, Vidaki and collaborators used both machine learning and NGS-based DNA methylation detection. More in detail, they studied 45 age-associated CpG sites selected from available methylation data obtained from 1156 whole blood samples (aged 2 to 90 years) and analyzed them with genome-wide methylation platforms. From this set of markers, they then applied stepwise regression for variable selection, resulting in the identification of 23 CpG sites. They then performed a regression analysis of the 23 CpG sites and found that these markers supply a correct prediction of age (R2 = 0.92, mean absolute error (MAE) = 4.6 years). After applying a generalized regression neural network model, the age prediction improved significantly (R2 = 0.96) with an MAE = 3.3 years. In total, they found 16 CpG sites as reliable age-related markers **(Figure 12)**. The advantage of using machine learning, and in particular the approach that Vidaki and collaborators used, is given by the fact that the use of the ANN approach proved to be a successful strategy to find underlying trends in complex datasets (for example the age-estimation of a person).



CpG sites	Chromosomal location	Gene
cg19761273	17: 80,232,096	CSNK1D – casein kinase 1; delta isoform 1
cg27544190	21: 33,785,434	C21orf63 – chromosome 21 open reading frame 63
cg03286783	15: 44,580,973	CASC4 – cancer susceptibility candidate 4 isoform a
cg01511567	11: 57,103,631	SSRP1 – structure specific recognition protein 1
cg07158339	9: 71,650,237	FXN –frataxin, mitochondrial isoform 1 preproprotein
cg05442902	22: 21,369,010	P2RXL1 – purinergic receptor P2X-like 1; orphan receptor
cg24450312	1: 206,681,158	RASSF5 – Ras association domain family 5 isoform B
cg17274064	21: 40,033,892	ERG – v-ets erythroblastosis virus E26 oncogene like isoform 2
cg02085507	19: 6,739,192	TRIP10 – thyroid hormone receptor interactor 10
cg20692569	7: 72,848,481	FZD9 – frizzled 9
cg04528819	7: 130,418,315	KLF14 – Kruppel-like factor 14
cg08370996	15: 96,874,031	NR2F2 – nuclear receptor subfamily 2; group F; member 2
cg04084157	7: 100,809,049	VGF – nerve growth factor inducible precursor
cg22736354	6: 18,122,719	NHLRC1 – malin
cg06493994	6: 25,652,602	SCGN – secretagogin precursor
cg02479575	19: 4,769,653	C19orf30 – hypothetical protein LOC284424

Figure 12 The 16 CpG sites were identified as age-related markers by the study by Vidaki et al., 2017a.

Moreover, to hypothesize an accurate age-related test, they quantified through NGS the methylation status of the selected 16 CpG sites using the Illumina MiSeq platform (Vidaki et al., 2017a).To further validate the accuracy of the data set, they also checked the identified markers in an independent cohort of 53 monozygotic twins and a cohort of 1011 disease state individuals.

Altogether, these studies highlight that the introduction of NGS technology in the discovery of the markers for age prediction supplies a more correct screening of the samples, due to its high sensitivity. Moreover, NGS technology can easily be used in combination with other DNA marker analyses, such as mRNA mutations or telomere length setup.

4 Biology of monozygotic twinning

Current available forensic techniques used to distinguish between two different people are always useful in ambiguous two-suspects discrimination. As discussed above, one of the biases for forensic scientists might be sample availability and composition. However, there is one particular condition in which forensic scientists have enough material and of excellent quality but still, this is not sufficient to find the guilty ones in a forensic case. This particular case occurs when the two suspects are completely identical as it is in the case of monozygotic twins.

Unless the two identical suspects leave their unique fingerprints, they cannot be distinguished using the standard available forensic techniques. For this reason, monozygotic twins, nowadays, represent a strong limit for the application of markers and analytical methods that are routinely used in forensic science, meaning that a reliable and efficient method of discrimination still needs to be found.

Dizygotic twins (also known as fraternal) and monozygotic twins development is different. Fraternal twins develop from two eggs that have been fertilized simultaneously by two different sperm. In this case, each zygote develops in its own amniotic fluid-filled inner sac and has its placenta. In contrast, monozygotic twins arise from one single fertilized egg (zygote) and on average they represent about a third of all spontaneous twins births (Hall, 2003).

Furthermore, while dizygotic twins always have different embryonic adnexa, monozygotic twins, depending on their moment of splitting, can share or not the placenta (monochorionic) and/or the amniotic sac (monoamniotic) **(Figure 13).**

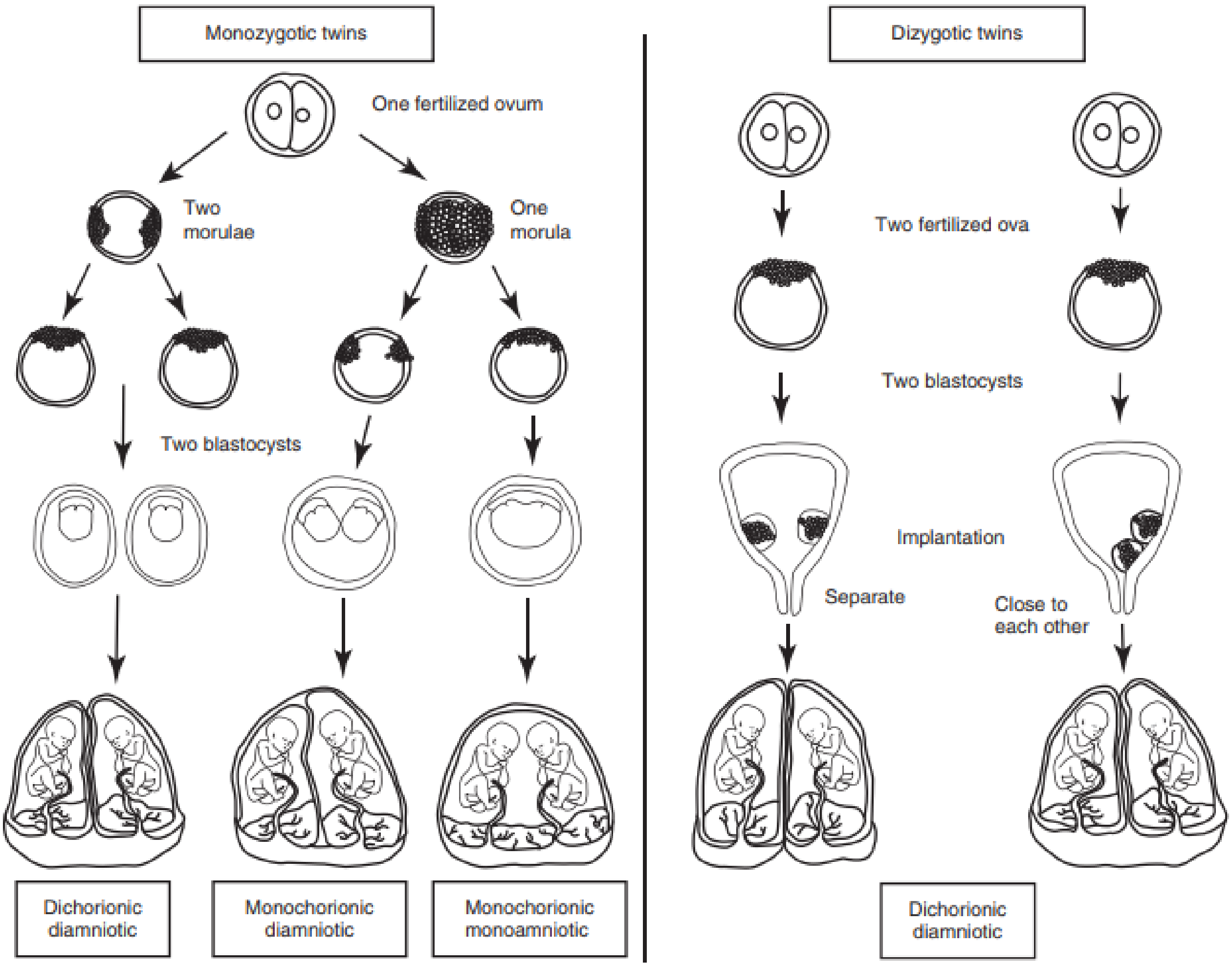


Figure 13 Development of various types of twin pregnancies (Kurt Benirscheke, M.D, and Chung K. Kim, M.D, 1973).

These early developmental differences could, in terms of genetics, results in the fact that their genetic makeup would be expected to be almost identical, characteristics that gave to them the appellative of clones.

However, after their splitting, i.e., during the early developmental processes, somatic mutations within their cells start to appear. Apart from stochastic mutations, other differences can be found in adult identical twins due to external and environmental factors that directly interact with the genome of the twins (Hall, 2003).

4.1 Epigenetic variations in monozygotic twins

Despite their almost identical genetic sequence, monozygotic twins can show phenotypic discordance traits at the very beginning of their life, another possible cause of the epigenetic discrepancy between monozygotic twins may be due to complex diseases (Boomsma et al., 2002). In the case of complex diseases, the aberrant phenotype is the result of shared mechanisms between different epigenetic modulations. As discussed above, epigenetic modifications refer not only to DNA methylation but also to other epigenetic mechanisms, such as histone modifications. All epigenetic modifications are strongly related and act together at multiple levels, from the chromatin structure to the DNA sequence, to modulate gene transcription, also in the case of disease development **(Figure 14).**

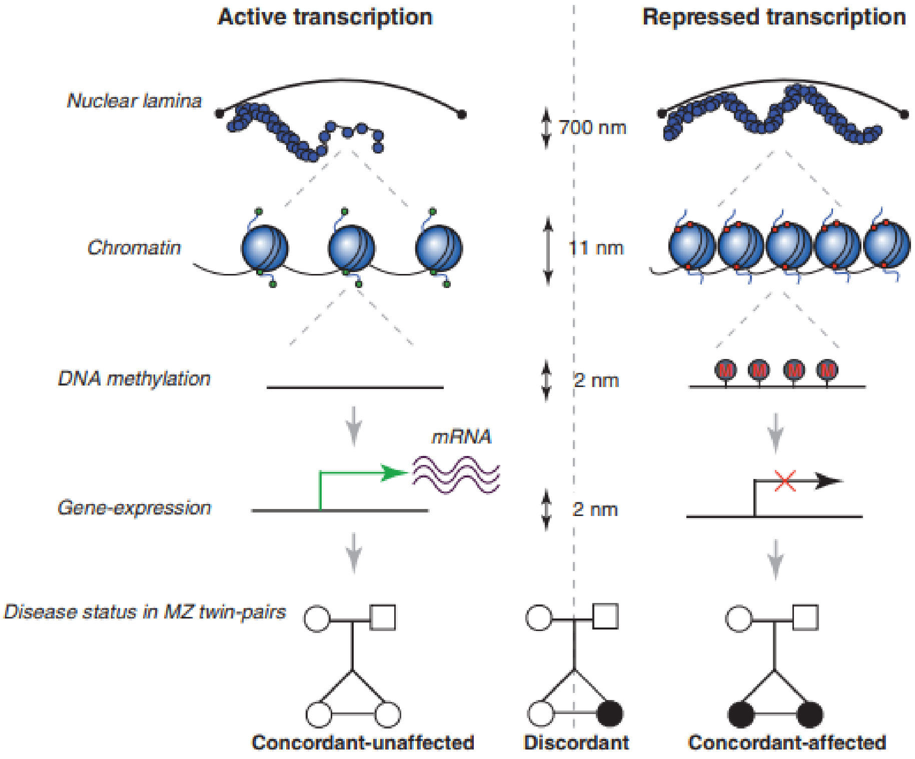


Figure 14 Epigenetic regulation of chromatin structure, associated with gene expression and disease status in a sample of MZ twins. The active or inactive transcription is regulated at various levels. From the top: higher chromatin loop configurations in case of active transcription or inactive transcription, the attachment to the nuclear lamina is the first step of the chromatin

activation/inactivation state. The second step refers to the chromatin beads on a string configuration, loose chromatin organization correlates with the active state of the transcription, whereas packed chromatin results in inactive transcription. Histones can further be modulated: the green dots refer to modifications that lead to an active chromatin state, while the red dots refer to modifications that lead to a repressed state. At the bottom of the image, possible effects of these different changes at different levels in the disease status of MZ twins, particularly for unaffected-concordant, discordant, and disease-concordant MZ twins (Bell and Spector, 2011)

The use of MZ twins to study complex diseases is useful because it supports the findings that not only genetic variation is fundamental, but also the interplay between genes and the environment. Nowadays, several pathologies are studied using the monozygotic twins approach. Monozygotic twins are found to be discordant for complex diseases such as type 1 diabetes and type 2 diabetes (Condon et al., 2008), schizophrenia (Beckmann and Franzek, 2009), as well as different types of cancer. Differences in the methylation profile of MZ can be associated with the presence of epimutation arising during the DNA replication or can be the result of the interaction between the genome and external factors.

4.2 Early monozygotic twins discordance

During the first stages of embryo development, the process of establishment and maintenance of methylation marks could be one of the possible explanations that make MZ twins different. As discussed previously, DNA methyltransferases are the enzymes responsible for DNA methylation. During every cell division, the methylation profile is inherited with epimutations (differences in epigenetic marks) arising during this process. The Dnmt1 is the DNA methyltransferase responsible for the maintenance of the right pattern of methylation through each cell division. Dnmt1 can methylate newly synthesized DNA, which lacks methylation marks, using the hemi-methylated parental DNA molecule as a substrate to methylate the corresponding nucleotides on the complementary strand. During this process, epimutations can occur as probabilistic errors. In the case of the Dnmt1, this skips about 4 to 5% of methylation sites and shows *de novo* methylation activity near densely methylated regions

(Vilkaitis et al., 2005). The presence of these epimutations in the DNA directly correlates with a subset of early-stage pathologies known as imprinting disorders. Imprinting disorders refer to pathologies associated with parent-of-origin-specific gene expression, in which the imprinted gene is differently methylated in the maternal and paternal allele. Prader-Willi syndrome and Angelman syndrome are some of the most studied cases. Interestingly, the methylation of the Prader-Willi imprinting center in chromosome 15q on the paternal allele causes the Prader-Willi syndrome, while failure in the same location but on the maternal allele causes Angelman syndrome. Furthermore, loss of DNA methylation at the KCNQ1OT1 gene on the maternal allele causes Beckwith-Wiedemann syndrome. Other than that, differences that arise in the early stage of development could be associated with the way of twinning. Moreover, somatic point mutations are quite common with a frequency of 1.2 × 10−7 per base pair per twin pair (Elhamamsy, 2017).

Lastly, recent studies, showed that the DNA methylation profiles are more similar within pairs of monozygotic twins that shared placenta (monochorionic), compared to MZ twins that did not share the placenta (dichorionic) (van Dongen et al., 2021), further suggesting that differences *in utero* conditions could also influence the epigenetic twins profile.

4.3 Monozygotic twins discordance due to environmental factors

Differences due to environmental factors are related to a more complex phenotype in MZ twins. MZ twins and DZ twins can easily be used to detect the proportion of influence from genetic factors in a particular disease. MZ twins are assumed to share the exact 100% of their genetic profile, while the DZ twins share just 50% of their genetic profile. Considering this, scientists could easily measure how much is the proportion of the genetic and/or environmental influence. A greater phenotype concordance in MZ twins could suggest a higher contribution of genetics to the disease examinate. Measuring the genetic proportion, consequentially, measures also the non-genetic proportion, which refers to environmental influences and random error (**Figure 15).**

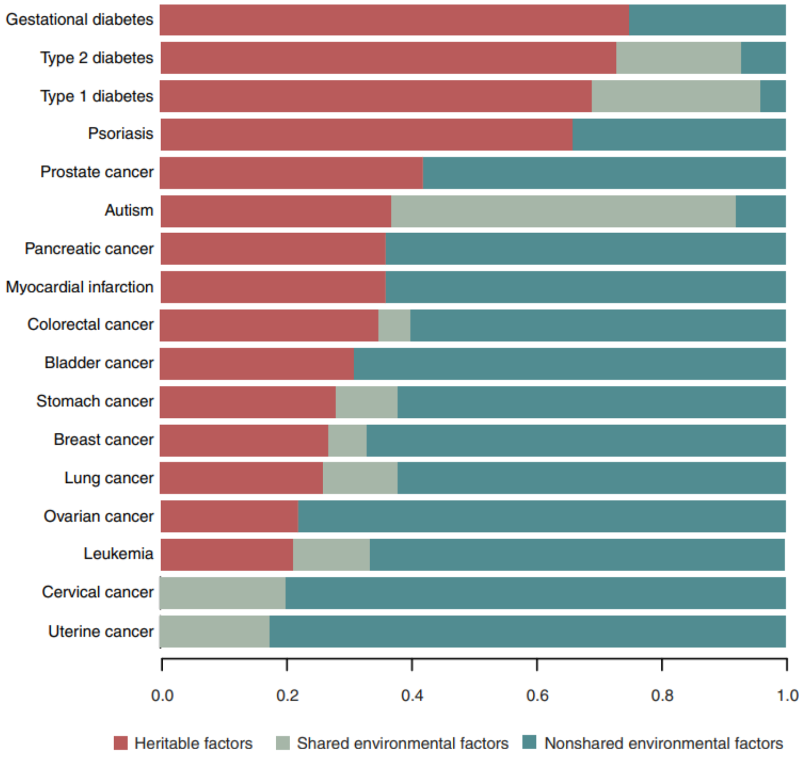


Figure 15 Portion of environmental and heritable factors that contribute to different diseases. At the top, the fraction of the phenotypic variance explained by heritable factors shared environmental factors, and non-shared environmental factors are shown (Castillo-Fernandez et al., 2014).

External and environmental factors, as already mentioned, directly interact with the genome not only in the case of pathologies but also during normal aging. Smoking habits, diet, and physical activity are just a few of the external factors that can have a long-term influence on the genome. MZ twin pairs that share a common environment for the majority of their life seem to have less degree of difference in terms of DNA methylation, while MZ twins with different life habits have more variability. Moreover, the methylation profile of young couples of MZ twins is much more similar compared to older ones, confirming that external factors and lifestyle can directly influence the methylation patterns of all individuals. Thus, MZ twins could help to better understand the external contribution to the human genome modifications **(Figure 16).**

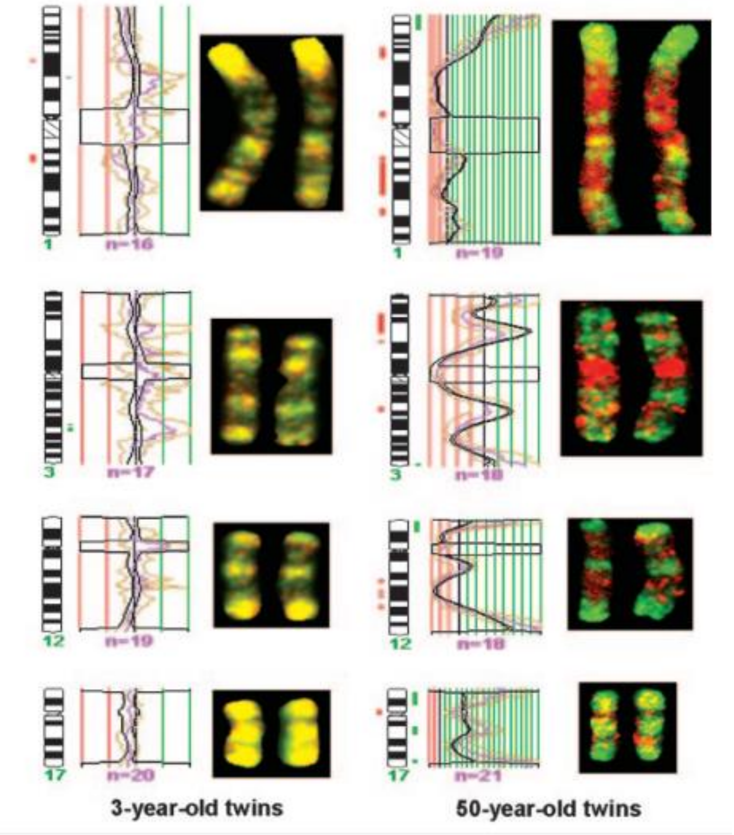


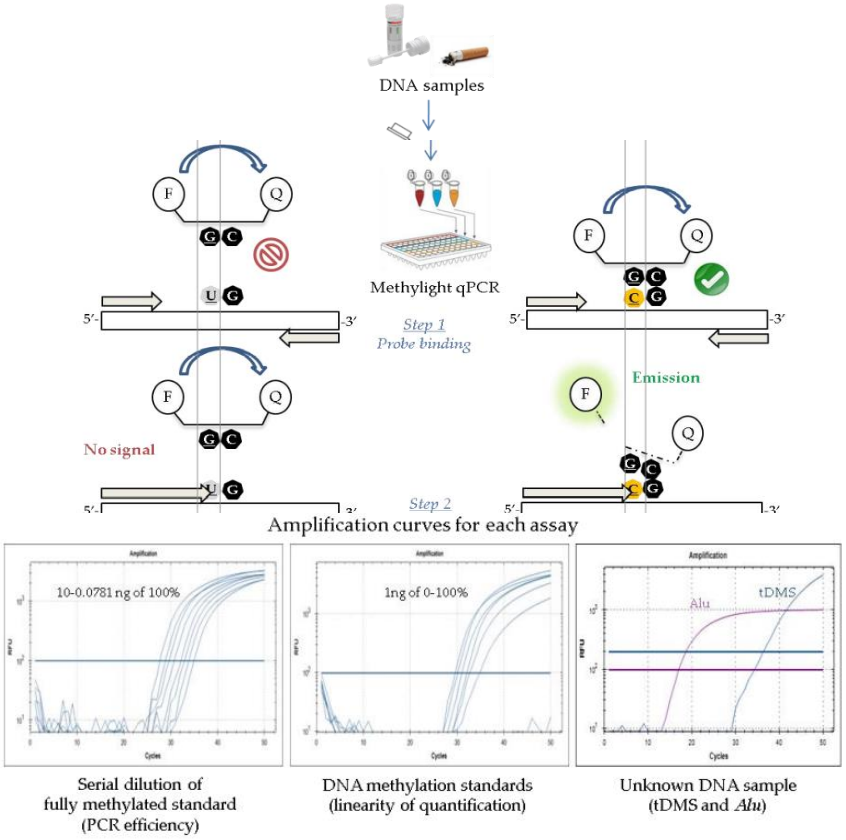
Figure 16 Comparison between a 3-year-old twins couple and a 50-year-old couple. Hybridization of regional chromosomes 1, 3, 12, and 17, mapped using comparative genomic hybridization for methylated DNA. The 50 years old-twin pair shows more changes in the methylation pattern compared to the 3 years old twin-pair. Different methylation patterns can be appreciated thanks to the green and red signals showing hypermethylation and hypomethylation events, respectively.

4.4 Monozygotic twins discrimination in forensic science

As already mentioned, MZ twins discrimination in forensic science is still an unsolved problem. Forensic cases in which identical twins are involved exist and, because of that, forensic scientists need to have a discrimination protocol. Recent works based on epigenetics, and in particular on DNA methylation, seem to open a way to solve this issue. Most of the already published and ongoing works are focusing their attention on the identification of different methylated sites in the genomes of MZ twins. Among all of these studies, scientists are using different approaches to find the most reliable answer to the question of "who is who" between the two twins. One proposed approach is the one used by Marqueta-Gracia and collaborators in 2018. In their work, they proposed a suitable solution, the use of High-Resolution Melting PCR (PCR-HRM) technology. The study was performed on saliva samples of 18 MZ twin pairs, selecting 6 different CpG regions found at ITGA2B, ASPA, PDE4C, ZIC5, USP11, and NOP14 loci, which have previously been shown to be variant during the human lifetime. They found regions showing significant within-pair differences located at ITGA2B, ASPA, and ZIC5 loci. Most of them were identified in the oldest couple of twins, ranging from 59 to 66 years old (Marqueta-Gracia et al., 2018).

The method proposed by Marqueta-Gracia and collaborators results is very cost-effective and rapid but unfortunately, more details need to be added in terms of other biological fluids and also different quantities of starting materials.

At the Department of Genetic Identification at the Erasmus University Medical Center in Rotterdam, Vidaki and collaborators tried to find another suitable protocol of discrimination. First, they tried to name sites of discrimination in blood samples, considering both reference-type DNA, which refers to the validated CpG candidates, and trace-type DNA, which instead refers to the DNA that could be found on a crime scene. In a work of 2017, they first tried to discover a set of candidates using a genome-wide methylation profile (using The Illumina Human Methylation 450K BeadChip array) of 10 MZ twin couples, then, after two steps of normalization (FUNNORM and SWAN methods), they validated the sites using a SYBR green-based PCR assay (Lo et al., 2009) and finally tried to identify the validated markers in the trace-type DNA (Vidaki et al., 2017b). In the second study of 2018, the same group of scientists used a similar experimental set-up, considering again reference-type DNA and trace-type DNA. In this second work, Vidaki and collaborators tried to use as reference type-DNA buccal swabs, and trace-type DNA saliva and cigarette butts (Vidaki et al., 2018, p. 29018). They analyzed one couple of female MZ twins, finding, using genome-wide DNA methylation microarray analysis, the suitable sites in reference-type DNA followed by candidates' sites analysis in trace-type DNA using the TaqMan-based quantitative PCR (qPCR) method (MethyLight). **(Figure 18).**



Serial dilution of fully methylated standard (PCR efficiency)

DNA methylation standards (linearity of quantification)

Unknown DNA sample (tDMS and *Alu*)

Figure 18 Scheme of the experimental setup for the DNA methylation analysis using the methylLight protocol (Vidaki et al., 2018)

The identified markers (22 CpG sites) were not consistent in reference-type DNA and trace-type DNA, attributable to the fact that reference and trace-type DNA came from different tissue types. Lastly, a recent paper from Planterose Jiménez and collaborators tried to move the attention to the site of discrimination that could differ because of stochastic methylation variation. In this study, they identified 333 CpG sites that displayed similarly large methylation variation between both monozygotic co-twins and also unrelated individuals highlighting the possible use of epigenetics as a unique human fingerprint (Planterose Jiménez et al., 2021).

5 Methodology to study DNA methylation

Nowadays, several approaches can be used in finding CpGs differentially methylated sites, and each of them proves its advantages and drawbacks.

In forensic cases, most of the approaches used based their technology on bisulfite conversion, an assay to find single nucleotide methylation variation. Many approaches to the study of DNA methylation combine bisulfite conversion and other techniques such as, for example, in the methyl-DNA immunoprecipitation (MeDIP) approach. This technique, after bisulfite conversion, performs a step of denaturation and precipitation of cleaved DNA using a 5mC antibody followed by sequencing (Weber et al., 2005). The shotgun type of sequencing allows single-base resolution of bisulfite sequencing. However, this approach cannot distinguish between 5-hydroxymethylcytosine and 5mC (Cokus et al., 2008). In another study Park and collaborators discriminated against individuals in a pair of monozygotic twins using The Infinium HumanMethylation450 BeadChip array (Illumina). This included the use of arrays that can identify more than 450,000 CpGs throughout the entire genome (Park et al., 2017). Notably, most of the techniques used to find CpG methylation includes the detection of areas of the genome that are rich in CG nucleotides rather than screening the entire genome. On the other hand, to get a better overview of the entire genome, other approaches, such as the whole-genome bisulfite sequencing are preferred.

In this chapter, a general overview of some of the available techniques used to study DNA methylation is provided.

5.1 Bisulfite conversion

Today, bisulfite conversion is considered the gold standard to study DNA methylation. This procedure was described for the first time in 1992 by Frommer et al. (Frommer et al., 1992). The chemistry beyond is based on the deamination of unmodified cytosines to uracil. The methylated cytosines (5mC) and (5hmC), differently, are resistant to deamination. This process allowed to then, followed by PCR amplification, resulting in the substitution of thymine for the not methylated cytosine, while 5-mC or 5-hmC residues get amplified as normal unmodified cytosines **(Figure 19)** (Li et al., 2019).

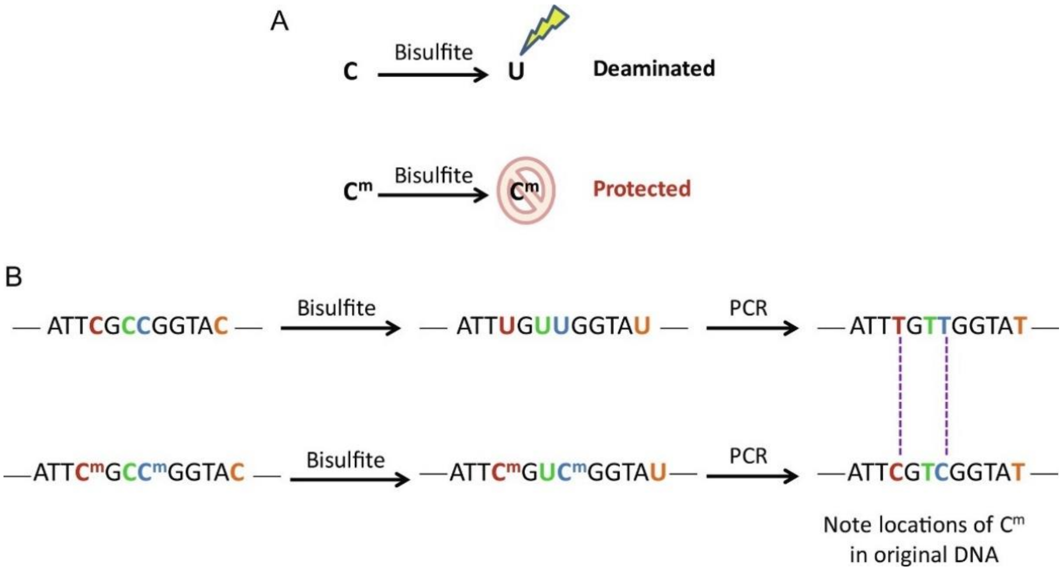


Figure 19 A) Cytosine treated with bisulfite is converted into uracil while methylated cytosine is protected from the conversion. **B)** During PCR amplification reaction, the DNA polymerase substitutes the uracil with thymine (Li et al., 2019).

Nowadays bisulfite conversion stays the most used technique also in forensic epigenetic studies, but, even with its wide use, this treatment has some disadvantages. Conversion with bisulfite causes high fragmentation of the DNA, resulting in small sequences, caused by the aggressive reaction condition of the conversion (pH 5 and temperatures up to 90°C). In forensic cases, as already discussed, most of the time starting DNA material is not enough in quantity and not even in quality resulting in a worsening of the material.

5.2 DNA methylation analysis by pyrosequencing

Pyrosequencing is one of the most used techniques to detect the degree of methylated and unmethylated cytosine residues after DNA bisulfite conversion. This technique is a sequencing-by-synthesis method that monitors the real-time incorporation of bases thanks to the enzymatic conversion of released pyrophosphate into a proportional light signal (Tost and Gut, 2007). To do that, the technique employs a cascade of coupled enzymatic reactions, that used different proteins such as DNA polymerase, ATP sulfurylase, and luciferase, used to check DNA synthesis, with a nucleotide-degrading enzyme in the cascade enabling iterative nucleotide dispensation to the reaction mixture. **(Figure 20).**

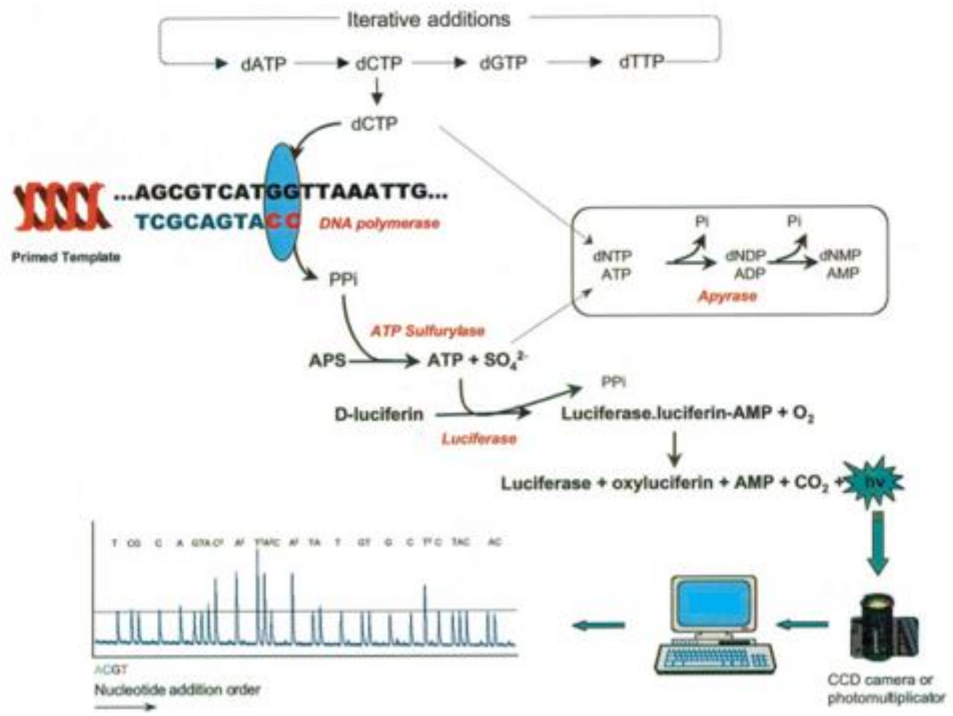


Figure 20 Schematic representation of Pyrosequencing workflow (Image modified from Gharizadeh et al., 2001).

5.3 Methylation analysis using arrays

Large-scale methylation screening is also performed using arrays technology. Some of the most used bead array technologies are the ones proposed by Illumina. Illumina developed different platforms for array-based screening of DNA methylation: the GoldenGate (which is currently out of production), the Infinium Human Methylation27, and the newest and popular Infinium HD 450K methylation array (Roessler et al., 2012).

The Methylation BeadChips developed by Illumina offers a combination of comprehensive, high-throughput large sample size screening for methylation studies in which individual CpG sites within a given DNA sample are interrogated simultaneously. The Infinium methylation assay interrogates the chemically differentiated loci using site-specific probes, designed for the methylated locus or the unmethylated locus. Then a single-base extension of the probes incorporates labeled ddNTP to the fragment and, the labeled ddNTP is then stained with a fluorescent reagent and detected. Different methylation levels for the interrogated locus can be deduced by computing the ratio of the fluorescent signals from the methylated vs. unmethylated sites. The detection of single CpGs range from 3,000 to 850,000 sites, interrogating a vast number of informative positions. Up to now, the technology was a hybrid of two different chemical assays, the Infinium I and Infinium II assays (**Figure 21)**.

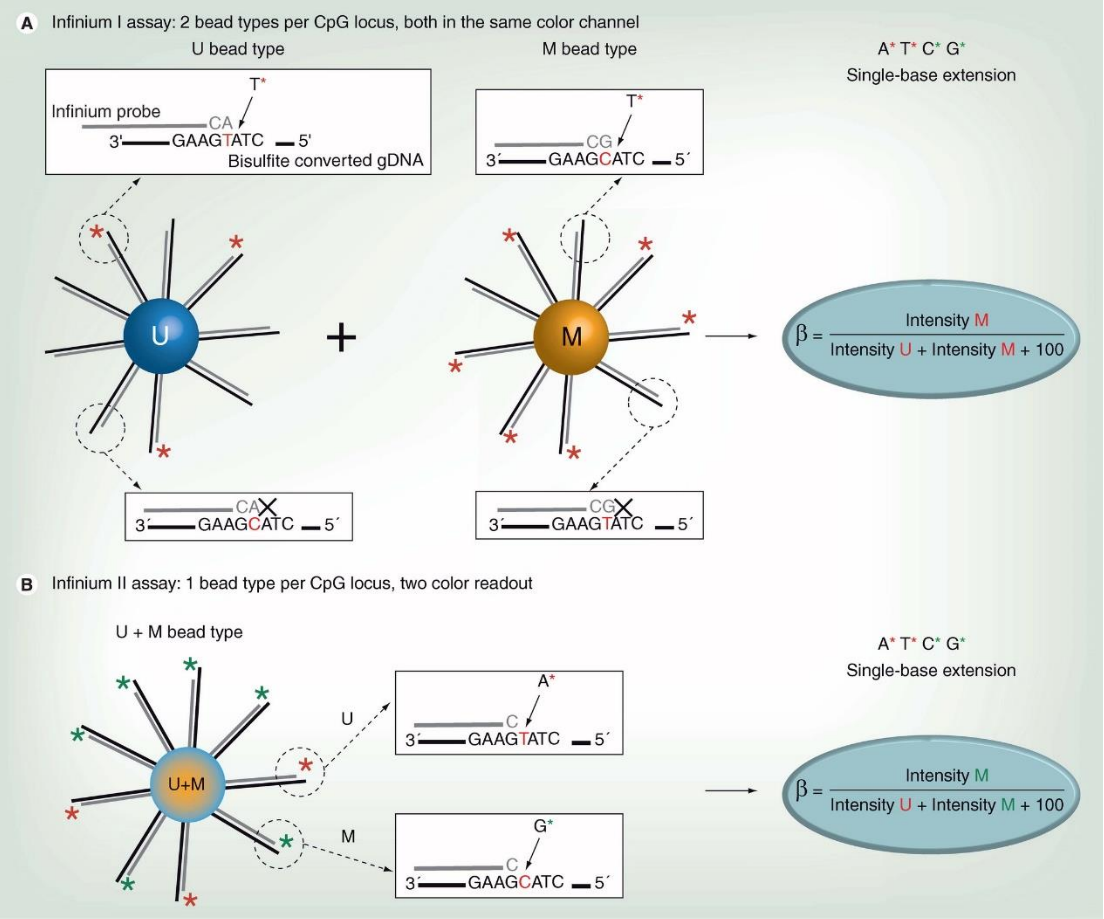


Figure 21 Overview of the Infinium I and Infinium II (Dedeurwaerder et al., 2011).

The main difference between the two assays is the design of the beads. In the Infinium I assay, two types of probes are used (one for the methylated allele and one for the unmethylated allele), while in the Infinium II assay a single probe is used for both alleles, and base extension depends on the methylation state of the hybridized genomic DNA molecule.

5.4 Reduced representation Bisulfite sequencing: Ovation® RRBS Methyl-Seq System

Other suitable techniques to study DNA methylation are sequencing-based techniques following the bisulfite conversion. The Ovation® Reduced representation Methyl-Seq system proposed by Nugen, is a cost-effective, fast, and correct approach. In the Ovation® RRBS Methyl-Seq System assay, differently from many other sequencing approaches, no prior fragmentation is needed, thanks to the MpsI restriction enzyme fragmentation. The MpsI restriction enzyme can cut symmetrically at the region CCGG of the genome, generating small fragments covering ∼1% of the human genome **(Figure 22).** (Gu et al., 2011)

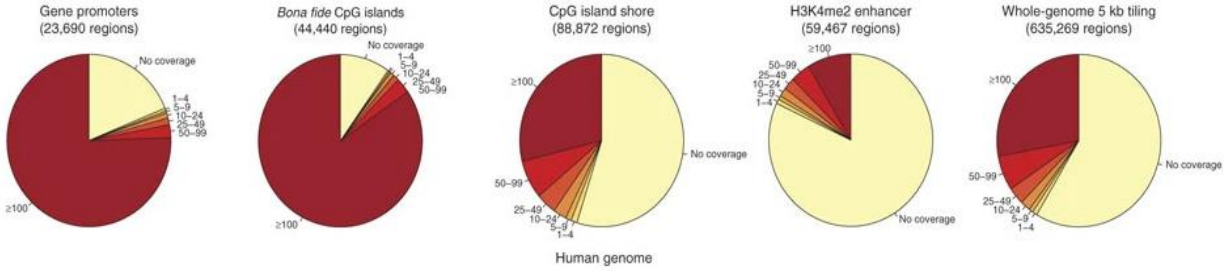


Figure 22 Pie charts illustrating RRBS library coverage for CpG methylation sites for five different genomic regions: gene promoters, CpG islands, CpG shores (identified as s 2-kb genomic regions close to CpG islands), enhancer elements from histone H3K4, and genomic tiling (Image modified from Gu et al., 2011).

The MpsI enzyme digestion is completely random, and it does not relate to the cytosine status (methylated or not methylated), moreover, considering the cutting region specificity of the enzyme (CCGG), it allows to obtain in each different fragment at least one informative cytosine.

The reduced representation bisulfite sequencing technique is a cost-effective way of obtaining single-base resolution DNA methylation information from a genomic sample. The term “reduced representation” refers to the fact that this technique examinates just a small part of the genome, resulting in a reduced amount of sequencing needed compared to a whole genome approach. The approach efficiently combines bisulfite sequencing with methylation insensitive MpsI enzyme to get the major areas of the genome that are covered by CpG islands.

Compared to other methods (some of them already described) the RRBS procedure efficiently covers all the important genomic regions that have at least one cytosine. In 2010, Laird and collaborators compared different techniques used to study DNA methylation and found RRBS as one of the most advantageous in terms of cytosine per sample analyzed **(Figure 23).**

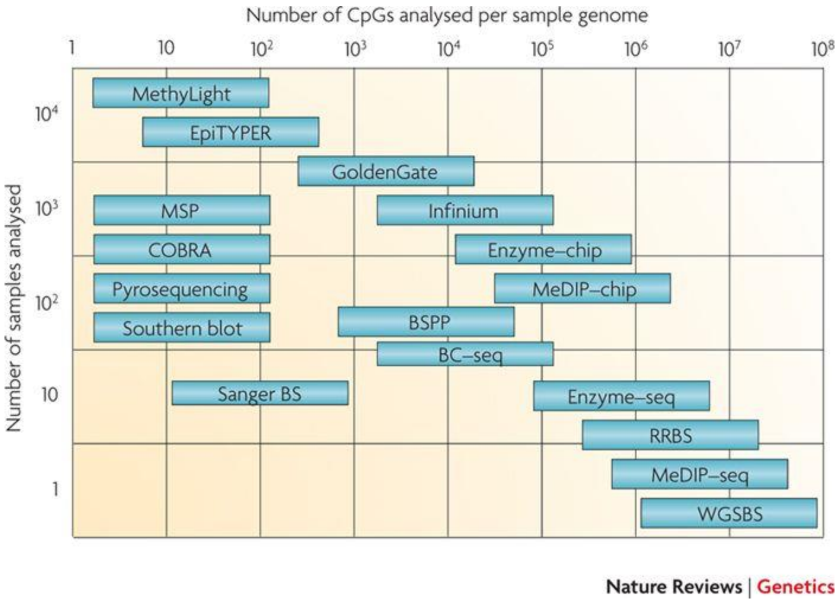


Figure 23 Different techniques used to study DNA methylation. The figure plots the sample throughput against genome coverage. Coverage is decided by the number of CpGs in the genome that could be analyzed for the experiment while the sample throughput refers to the number of a sample that can be analyzed in every experiment. RRBS results are one of the best methodologies to question a large number of cytosines per sample (Laird, 2010).

Also, one of the major advantages of using the RRBS technique is the low cost per experiment performed. As shown in the figure below, RRBS or capture-based techniques, are the most helpful in terms of cost compared to other techniques (**Figure 24).**

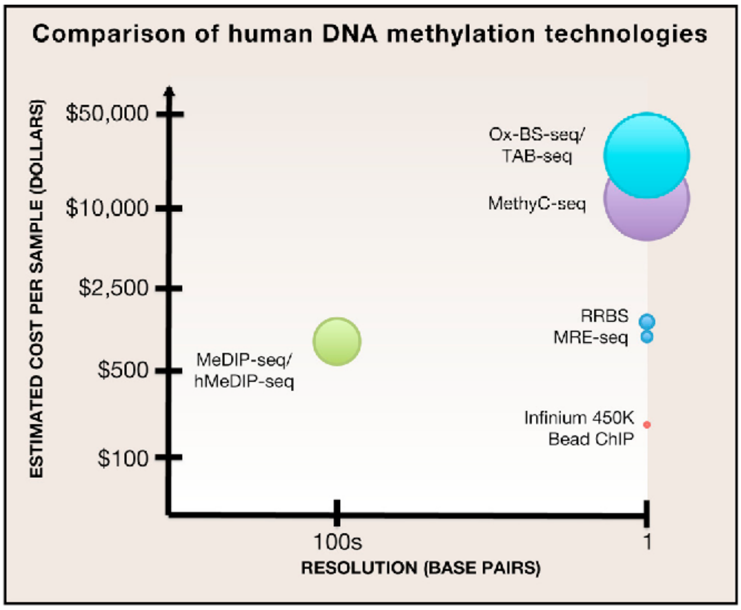
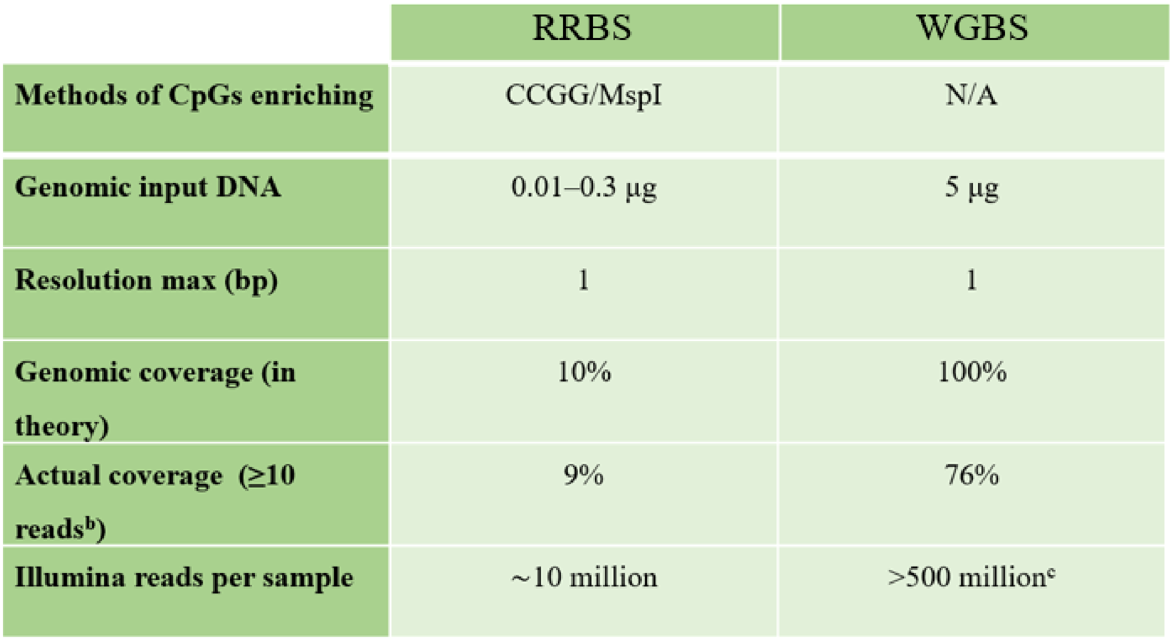


Figure 24 Comparison of different DNA methylation techniques used with the relative cost-effectiveness per sample in dollars (Rivera and Ren, 2013).

5.5 Whole-genome bisulfite sequencing: Ovation® Ultralow Methyl-Seq Library Systems

In contrast to the RRBS technique, WGBS does not interrogate a few areas of the genome but works on the entire DNA of the samples taken in the study. The steps in the protocol are similar to the RRBS except for the use of MpsI which in WGBS is substituted by DNA fragmentation.

The main differences between WGBS and RRBS approaches are described in **Table 1**.



	RRBS	WGBS
Methods of CpGs enriching	CCGG/MspI	N/A
Genomic input DNA	0.01–0.3 µg	5 µg
Resolution max (bp)	1	1
Genomic coverage (in theory)	10%	100%
Actual coverage (≥10 reads[b])	9%	76%
Illumina reads per sample	~10 million	>500 million[c]

Table 1 Major differences between RRBS and WGBS techniques (Table modified from Tost, 2018)

WGBS allows the screening of up to 500 million reads simultaneously compared to the 10 million screened by the RRBS assay. The screening performed by WGBS techniques examines the entire genome, focusing the attention not only on CpG located at relevant genomic areas (such as CpG islands) but also on sites randomly distributed and related to stochastic variation. In a recent paper from Jiménez et al, the importance of stochastic methylation variation has gained prominence. (Planterose Jiménez et al., 2021). In this study, 333 CpGs sites were found to display a large methylation variation between monozygotic co-twins and also unrelated individuals. The study highlights the importance of universal stochastic variation that occurs not only in unrelated individuals but also in identical twins. To better exclude any kind of epigenetic discordance related to epigenetic drift or environmental effects, the authors conducted the study on a cohort of early-age monozygotic twins. The study strongly suggested the existence of an epigenetic fingerprint relevant for the identification of MZ twins implicated in forensic cases.

6 Aims

Monozygotic twins discrimination stays one of the major pitfalls in forensic cases. The standard analysis of STRs is completely useless when applied in the case of identical twins discrimination. Nowadays, there are no available techniques able to distinguish between identical twins.

Together with the inability to use standard forensic techniques, other major problems noticed during in-loco inspections in forensic cases are represented by biological samples' availability and composition. In most cases, only a small amount of DNA is recovered from a crime scene and, most of the time, this DNA is also degraded.

One of the practical solutions that can be used in forensics to distinguish between identical twins, especially when dealing with degraded and low input material is the use of epigenetics, and in particular the study of DNA methylation. DNA methylation has been gaining increased attention in the past years, especially in forensics to distinguish identical twins at a crime scene. Therefore, the aims of this thesis are to:

- Mimic the unlucky condition of poor quantity and quality DNA on the crime scene.

  To achieve this goal, from two different couples of monozygotic twins 1 ng of starting DNA was used to mimic trace type DNA (DNA found on crime scene).

  Furthermore, from the same couple of twins, two reference trace DNA tables were created to simulate the available DNA required during the investigations.

- Create a reliable statistical approach able to lead back each unknown trace type table to the proper reference table.

7 Materials and Methods

7.1 Sample collection

In this study, the biological samples were collected after all the participants supplied signed informed consent. In total, one female MZ couple and one male MZ couple were included in the study, aged 25 and 40 years, respectively. Participants collected their body fluids samples (buccal swabs) on their own, according to the laboratory's instructions, using sterile and disposable buccal swabs (Omni Swabs from Whatman Bioscience), provided by our laboratory. After the collection, each sample was anonymized soon by the participants, as to not link any biological information to the individuals. For our study, each twin provides us with two different buccal swabs, to perform the entire protocol in duplicates.

Once the collection was done, all the samples were stored in our laboratory at 4° C until the DNA extraction was performed after 1 week after the collection.

7.2 DNA extraction (QIAamp® DNA Mini and Blood Mini Handbook Qiagen)

DNA extraction from stored buccal swabs of the two MZ couples was performed using the QIAamp® DNA Mini and Blood Mini Handbook, which allowed scientists to obtain biological materials from various kinds of samples, such as whole blood, plasma, serum, swabs, and tissue.

The DNA purification from the buccal swab could be performed both using spin-column or using vacuum protocol. For this study, the spin protocol was preferred. Using this procedure, the DNA is adsorbed on the QIAmp silica membrane while all the other materials, thanks to the salt composition and pH conditions, are not kept on the membrane and could be accumulated in the filtrate below the spin column. Extraction using the spin protocol requires no phenol-chloroform or alcohol precipitation and involves little handling, moreover, the protocol results are quite easy and quick and can be adjusted depending on the buccal swab type used in the study. Any other information about the protocol can be found in the QIAamp® DNA Mini and Blood Mini Handbook.

In this study, the head of the buccal swab was separated from the stick using sterile scissors. After obtaining the swab's head, each of them was placed in the 2 ml microcentrifuge tube and 600 µl of PBS was added to each sample. To perform the lysis 20 µl of QIAGEN proteinase K stock solution (already prepared in the kit) is added simultaneously with the lysis buffer (labeled as AL) to all the samples. To ensure efficient lysis after adding the Buffer and the proteinase, samples need to be at once mixed. The solutions were then incubated on a heated shaker at 56 C for 15 minutes and then briefly centrifugated to remove drops present inside the lid.

Following the incubation, 600 µl of ethanol 96% - 100% fresh prepared was added to each sample, mixed, vortexed, and briefly centrifugated to rescue drops in the lid. After the lysis and cleaning steps, the obtained mixture is ready to be added to the spin column. Of the mixture obtained 700 µl was added to the QIAmp spin column and placed in a 2 ml collection tube, taking care of not wetting the rim. The spin column in the collection tube is then centrifugated at 6000 x g (8000 rpm) for 1 minute and then transferred into a clean collection tube to discard the filtrate. This step of filtration was then repeated a second time with another 700 µl of the mixture.

Once these steps were done, to remove residual contaminants, 500 µl of Buffer AW1 was added to the samples, followed by 6000 x g (8000 rpm) for 1-minute centrifugation. After this first step of DNA cleaning, 500 µl of Buffer AW2 was then applied to the spin column, following 20,000 x g (14,000 rpm) for 3 minutes. These two-washing buffer steps could help significantly improve the purity of the DNA and the maximum speed centrifugation was then needed to remove all the possible ethanol carryover that may interfere with downstream DNA analysis. Purified DNA was then eluted in 150 µl of elution Buffer (AE) Nuclease-free Water. For our experiments, we used 150 µl of AE buffer and we then incubate at room temperature (15-25 C) for 5 minutes before centrifugate at 6000 x g (8000 rpm) for 1 minute. A second elution step with another 150 µl of AE buffer was performed to increase the yield significantly. We obtained for all the samples a total volume of 140 µl and we then stored the samples at -20 C before use.

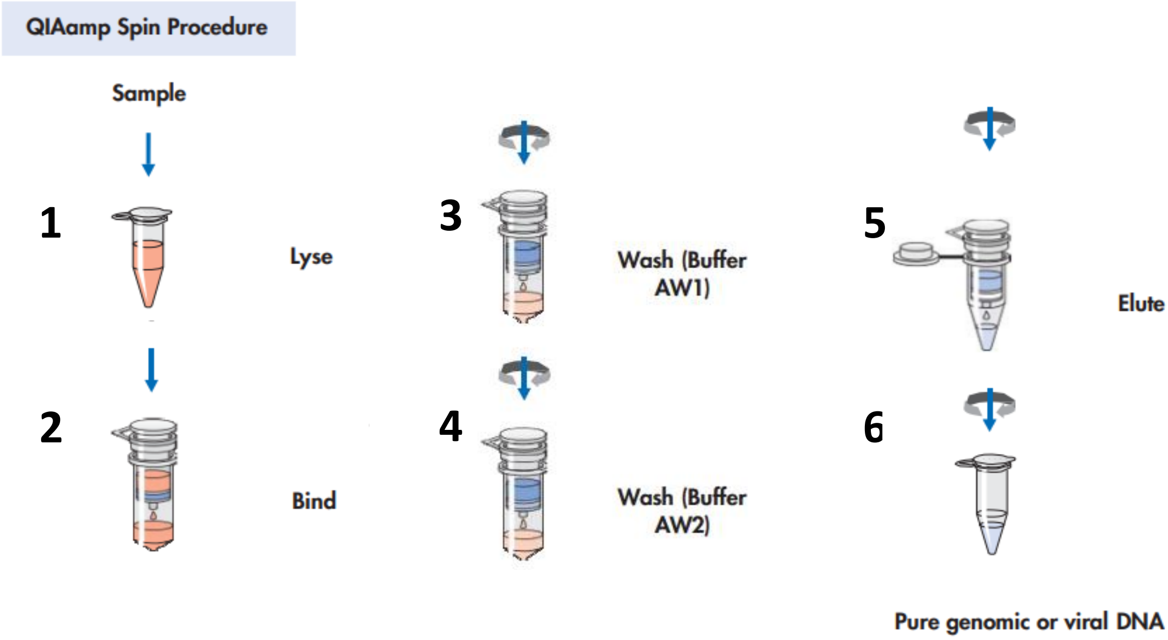


Figure 25 Procedures of DNA extraction using DNA Purification from Buccal Swabs (Spin Protocol). (Images changed from QIAamp® DNA Mini and Blood Mini Handbook).

7.3 DNA quantification; Qubit® dsDNA HS Assay Kits Life Technologies

To check the concentration of the samples the Qubit® dsDNA HS (High Sensitivity) Assay Kits were used, which give the possibility to, thanks to target-selective dyes emission bound to the DNA, measure the concentration of the sample.

The Qubit fluorometric quantification is a fast and simple assay that allowed us to obtain the correct quantification for our two samples. For these samples, the Qubit dsDNA High Sensitivity Assay was used, which has an assay range from 0.2–100 ng, while the starting concentration needed for the samples should be in the range of 10 pg/ µl to 100 ng/ µl.

Reagents used for the Qubit™ measurement were prepared as described in the image below **Figure 26**

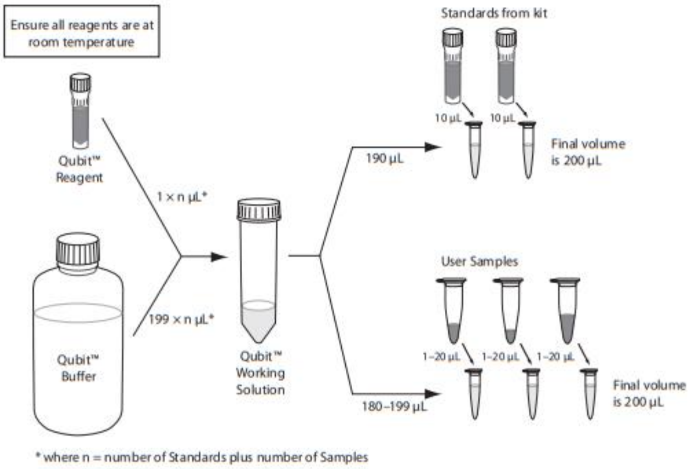


Figure 26 Qbit protocol: master mix prepared using 1 ng of reagent into 199 µl of buffer in a Qbit tube recommended by the protocol.

7.4 DNA fragmentation (The Diagenode Biorupter® Sonicator)

To obtain DNA fragments needed in the following steps, the Diagenode Biorupter® Sonicator was used.

This system is based on a water bath that generates indirect sonication waves, which emanate from an ultrasound element below the water tank. The ultrasound used by the system creates mechanical stress able to lyse or shear DNA. When the ultrasound waves pass through each sample, they create a flux of expansion and contraction into the liquid. During the step of the expansion, a negative pressure pulls the molecules away from each other and creates a space, or a bubble, resulting in a process called cavitation. At a certain point, when the pressure on the bubble becomes higher, it can no longer sustain the energy on itself, and this results in implosion, producing a strong force that disperses molecules.

The mechanical force of the Bioruptor randomly generates DNA fragments that show a correct and precise size distribution depending on the time and force applied by the system to the samples.

The time and force to be applied to the molecules can be easily chosen from the control units.

For human samples, the protocol is described in the table below (Table 2), with also the specification for the protocol recommended for the bacteriophage Lambda, the control DNA that we used in our study.

Human DNA	4 cycles	30 seconds	90 seconds stop
Lambda DNA	3 cycles	15 seconds	90 seconds stop

To perform the fragmentation, the starting materials required by the instrument were 100 μl of samples and water, in proper Bioruptor tubes. After the mechanical fragmentation, the fragment size distribution can be analyzed using various methods such as agarose gel electrophoresis, chip-based electrophoresis, or also capillary electrophoresis. The fragment size distribution was approximately 200 base pairs (bp).

7.5 Agilent High Sensitivity DNA Kit

DNA fragments obtained from the Bioroptur fragmentation were then checked for the proper distribution of sizes. To do that the Agilent High Sensitivity DNA Kit was used that find fragments that show a size distribution from 50 to 7000 bp**,** with a typical sizing accuracy of ± 10 % CV. The Agilent High Sensitivity DNA Kit technology is based on the use of a chip priming station. Into the chip station, just 1 μl of starting material for each sample is required, for a total of 11 samples to be analyzed simultaneously. The protocol required the preparation of a Gel-dye mix as follows; the tubes holding the blue-capped High Sensitivity DNA dye concentrate and the red-capped High Sensitivity DNA gel matrix are taken from 4° C and equilibrate at RT for 30 minutes. Once these reagents were equilibrated for the estimated time, they could be vortexed and spin down. Then, 15 μl of the High Sensitivity DNA dye (blue capped) are pipetted in the High Sensitivity DNA gel matrix vial (red-capped). This freshly prepared solution needs to be stored in the dark to preserve its functionality. The gel-dye mix was then transferred on a spin filter and then placed in a microcentrifuge for 10 minutes at 2240 g, or 6000 rpm, at RT. Before loading into the cheap gel-dye mix, the freshly prepared mix needs to be equilibrated at RT for 30. A total volume of 9.0 µl was pipetted in the well-marked with a G on the chip, after that, through the use of the plunger, the gel-dye mix could be dispensed through the chip. Once the procedure of the gel-dye mix was complete, 5 µl of the High Sensitivity DNA marker (marked with the green cap), was pipetted into the well with the ladder specific symbol and in each of the 11 samples well. After adding the marker, 1 µl of the High Sensitivity DNA ladder (with the yellow cap), was pipetted in the specific well marked with the ladder symbol. Finally, in each of the 11 sample wells, 1 µl of samples was pipetted. After the complete preparation, the chip was vortexed for 60 seconds at 2400 rpm and, once ready, placed in the Agilent 2100 Bioanalyzer.

The Agilent High Sensitivity DNA Kit gave a final report from which the size distribution of the fragments could be easily visualized by checking the electropherogram. **Figure 27**

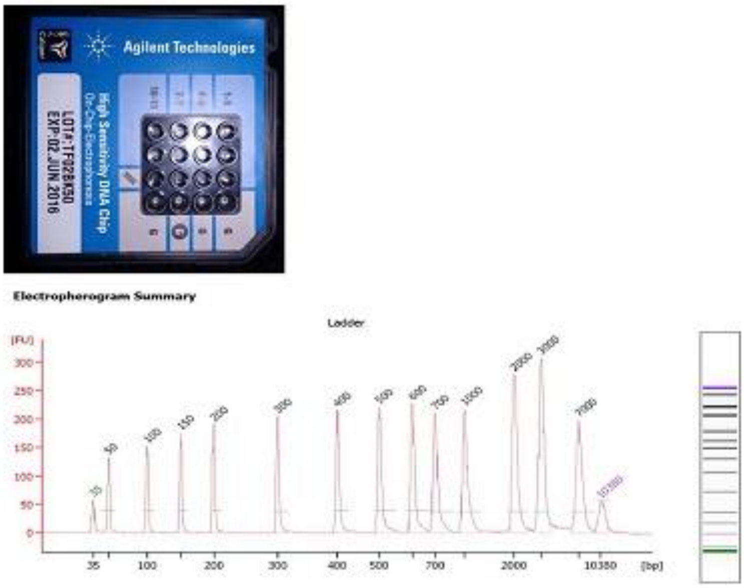


Figure 27 Representative image of a chip priming station and an Electropherogram that can show the abundance of any fragments (Performance characteristics of the High Sensitivity DNA kit for the Agilent 2100 Bioanalyzer System Tech note)

7.6 DNA purification using AMPure XP, Beckman Coulter.

Following the DNA fragmentation, fragments obtained were purified using magnetic beads (AMPure XP, Beckman Coulter). The purity of DNA is critical to generating the best libraries before sequencing. Contaminants present in the samples can directly interfere with downstream reactions. In the image below the AMPure XP, workflow is described. Figure 28

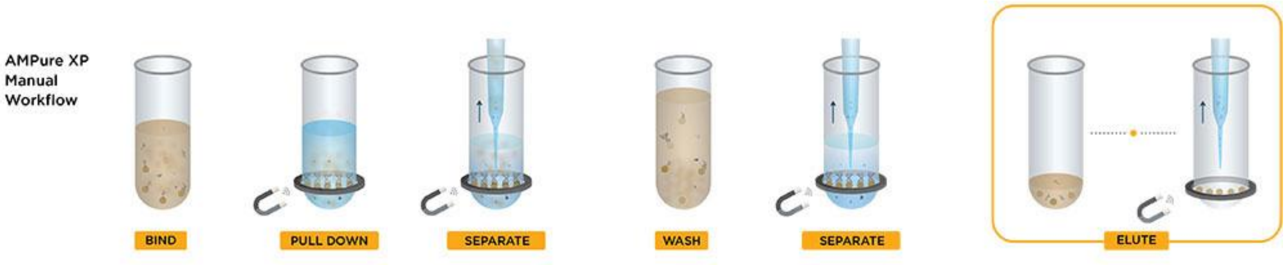


Figure 28 Magnetic beads procedure used to purify the fragmented DNA. (Imaged modified from AMPure XP, Beckman Coulter protocol).

The purification protocol requires as the first step to add 2X quantity of beads into the tube having the starting fragmented material. After adding the beads, the DNA needs to be completely resuspended into the solution to maximize the beads-DNA interaction. Once the mixing was done, the solution was left for 10 minutes at RT and then placed for 5 minutes onto the magnet. The liquid was then removed and two washes with freshly prepared 70% ethanol were performed. Once the ethanol results be completely dried 13 µl of Nuclease-free Water was added. Finally, the tubes were transferred onto the magnet for at least 5 minutes to completely clear the solution from beads. A final volume of 13 µl of eluate for each sample was then placed in a new PCR tube.

As already discussed in the introduction, two approaches were used for this study. In this section, each step of the two protocols, the RRBS technique, and the WGBS technique are step by step described.

7.8 First approach: The Reduced representation bisulfite sequencing (RRBS) technique: Ovation® RRBS Methyl-Seq System

Before moving to the Whole-genome sequencing preparation, we tried to use the Reduced representation approach as a possible experimental setup to prepare the reference tables.

The experimental workflow we use is described in **figure 30.**

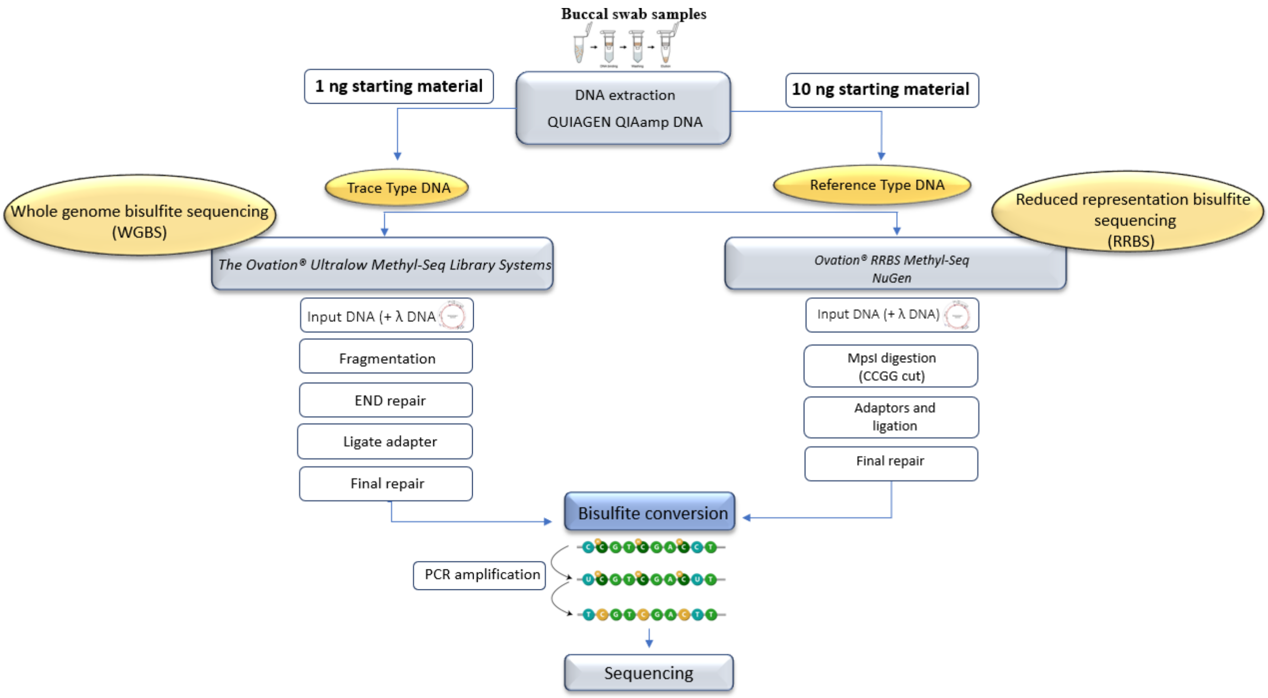


Figure 30 RRBS laboratory workflow.

In the paragraph below experiments performed using this technology are described.

RRBS protocol consists of five main steps: first MspI digestion, adaptor ligation, final repair, bisulfite conversion, and the PCR amplification of the obtained library. **Figure 31.**

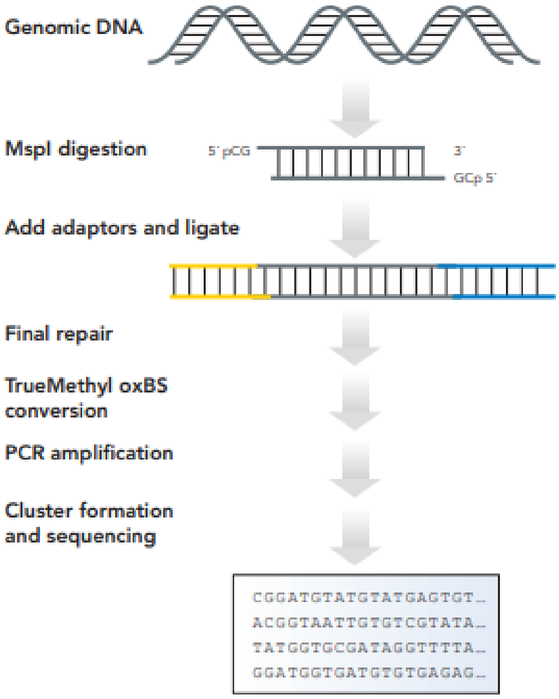


Figure 31 Schematic representation of the five main steps in RRBS workflow

MspI Digestion

The first step in the protocol consists of MpsI enzyme use. This step allows not to fragment the input DNA, as already described in the section *"DNA fragmentation (The Diagenode Biorupter® Sonicator)* ", because of the enzymatic digestion properties. During this first step, a MspI master mix was prepared to combine 1.0 µl of MpsI buffer mix and 0.5 µl of MpsI enzyme mix into a new tube. Once the mix was ready 1.5 µl of MspI Master Mix was added to each sample. Then the reaction was performed into a thermal cycler programmed as follows:

37 °C – 60 min, hold at 4 °C

Adapter Ligation

Ligation of adapters is needed to univocally mark the samples of the library. To perform this second step, a ligation adapter mix was prepared to add in a new tube 2 µl of nuclease-free water, 4 µl of ligation buffer mix, and 1.0 µl of ligation enzyme mix. Of this freshly prepared mix 7 µl were added to each sample and then 3 µl of specific adapter were added to each tube. The tubes were then placed in a heated thermal cycler with the programs specified:

25 °C – 30 min, 70 °C – 10 min, hold at 4 °C

Final Repair

For the final repair steps, we combined 6 µl of final repair buffer mix, 0.5 µl of final repair enzyme mix, and 13.5 µl of nuclease-free water. With a total amount of 20 µl of mix in each sample of the final repair master mix. The tubes were then placed into the thermal cycler:

60 °C – 10 min, 70 °C – 10 min, hold at 4 °C

Bisulfite Conversion

To perform the bisulfite conversion, 30 µl of Bisulfite Reagent Solution were then added to each sample, all the tubes were then incubated in a pre-warmed thermal cycler with the following program:

95 °C – 5 min, 60 °C – 20 min, 95 °C – 5 min, 60 °C – 40 min, 95 °C – 5 min, 60 °C – 45 min, hold at 20 °C

Bisulfite-Converted DNA Desulfonation and Purification

The magnetic bead binding solution is prepared by combining 200 µl of binding buffer and 2.4 µl of magnetic bead solution. Carefully 160 µl of the fresh prepared magnetic bead solution were added to each bisulfite converted sample for a total of 200 µl. After mixing properly, the samples were left 5 minutes at RT and put on to a magnet for 5 minutes. The supernatant was then discarded, and the DNA was then washed using 200 µl of 70% ethanol two times, paying attention to completely resuspending the beads in every single wash. After these two steps of washes, the DNA was then put onto the magnet and the ethanol was discarded. 200 µl of Desulfonation Buffer with EtOH was added directly onto the bead pellet, then the tubes were incubated for 5 minutes at RT. With the tubes still on the magnet the total amount of liquid, approximately 200 µl, was removed and the other two steps of ethanol washes were performed. Finally, the beads were completely air-dried and a 21 µl Elution Buffer was added to each sample. Samples were then put onto the magnet for 5 minutes at RT before the complete transfer of the final eluate.

7.9 Second approach: The whole-genome bisulfite Sequencing (WGBS): Ovation® Ultralow Methyl-Seq Library Systems Nugen

As well as in the case of RRBS experiments, the genomic DNA from the same two couple of MZ twins (one female and one male) have been sampled as shown in the figure below **figure 32**

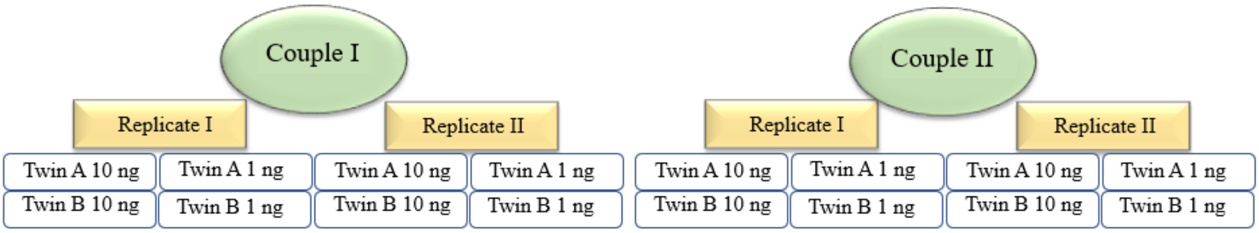


Figure 32 One male (40 years old) and one female (25 years old) couple were used in this study. Two replicates of each twin are produced.

A sampling of each individual occurs in duplicates, resulting in two different replicates for every twin sample. From the total starting material obtained once the mechanical fragmentation was ended, two different concentrations before starting library preparation were set up: 10 ng and 1 ng.

The 10 ng starting samples were considered as REFERENCE DNA, while the samples having 1 ng starting material were considered TEST or TRACE type DNA.

The entire lab procedures explaining for each type of DNA are explained in the workflow described in **Figure 33**

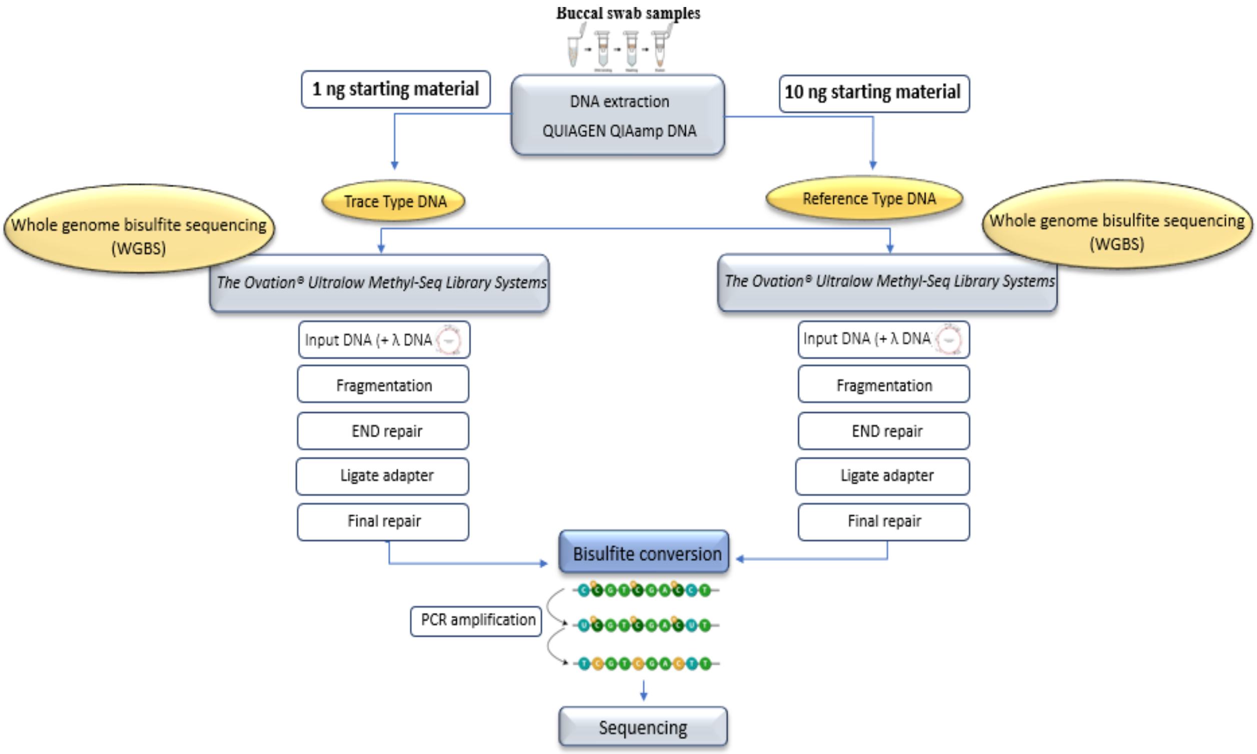


Figure 33 Experimental workflow performed. Both for trace type DNA and reference type DNA the method used to prepare the library was WGBS. The only difference between the two experimental sets up was the starting material. Starting DNA concentration for trace type DNA was 1 ng. Starting DNA concentration to prepare reference type DNA was 10 ng. λ genome was spiked in any samples with a percentage of 1%.

The WGBS method consists of five different steps; the fragmentation of genomic DNA through the use of the already mentioned technology (*"DNA fragmentation (The Diagenode Biorupter® Sonicator)* ", the steps of end repair to generate blunt ends, adaptor ligation, the bisulfite conversion, and the final PCR amplification to produce the library. **Figure 34.**

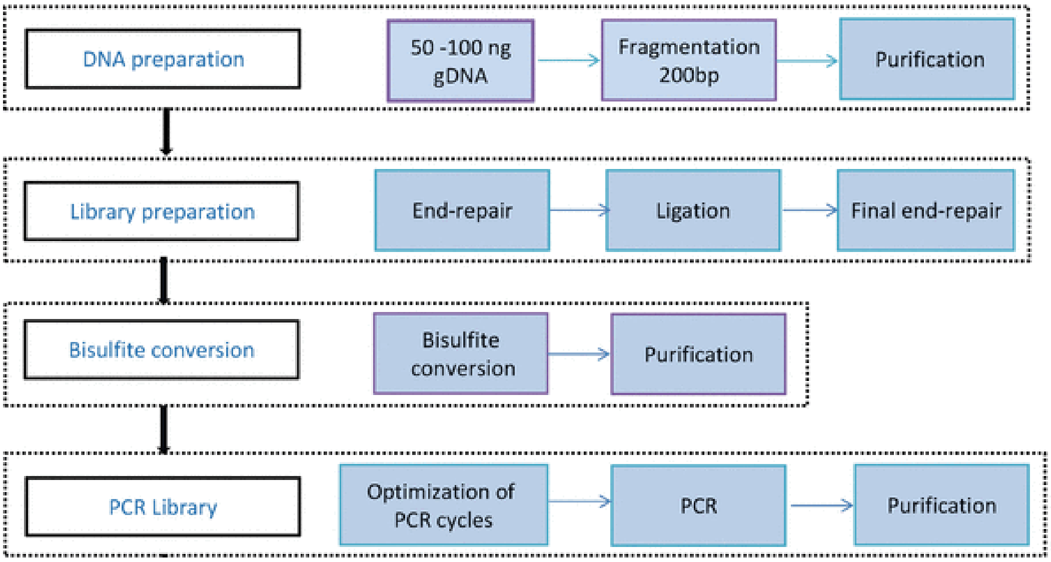


Figure 34 Schematic representation of the WGBS workflow (Image modified from Tost, 2018)

5’ repair ends

After the steps of fragmentation and purification of the input DNA, samples were then end-repaired. This step is needed to convert the 5’ overhangs into blunt ends and phosphorylates 5’ ends using the End-Repair mix. The step was performed according to the protocol (Ovation® Ultralow Methyl-Seq Library Systems Nugen), combining 2 µl of End Repair Buffer, 0.5 µl of End Repair Enzyme mix, and 0.5 µl of End Repair Enhancer. The freshly prepared mix was then vortexed, spin down, and placed on ice until use. A total of 3 µl of End repair mix was added to each sample, mixing and spinning down each sample. The tubes were then placed in a pre-warmed thermal cycler programmed as follows:

25 °C – 30 min, 70 °C – 10 min, hold at 4 °C

Ligation with different indexing adapters

After the end repair PCR was complete, samples were spin down and placed on ice until their use. Following the end repair step, it is necessary to ligate multiple adapters to the DNA fragments. For our samples we choose the adapters described in the table below:

The ligation master mix was prepared by combining 4.5 µl of nuclease-free water, 6 µl of Ligation Buffer Mix, and 1.5 µl of Ligation Enzyme Mix in a new tube. Then, 12 µl of Ligation Master Mix was added to each sample. After the addition of the mix 3 µl of each specific and unique adapter was pipetted into the specific sample. Once the ligation mix and the adapter were added, all of the samples were then placed in the preheated thermal cycler using the following program:

25 °C – 30 min, 70 °C – 10 min, hold at 4 °C

Following the step of ligation, a purification step is then required to decontaminate the samples from possible residues of adapters not ligated to the target DNA. Residues of adapters not bound to the target DNA can directly affect the run during the libraries sequencing, generating clusters on the sequencing flow cell that results be not informative in the study. To avoid this kind of contamination from adapters, 45 µl of beads were added to each sample and resuspended as already described, samples were then placed on a magnet at RT for 5 minutes and then 65 µl of supernatant were discarded. While the samples were still on the magnetic stand, 200 µl of freshly prepared 70% absolute ethanol was added to each tube, after 30 seconds at RT the supernatant was discarded, and this cleaning using EtOH 70% was repeated twice. Once the cleaning by EtOH was performed the tubes are left with the cap open to completely air-dry the beads. To finally eluate the DNA, 16 µl of Nuclease-free Water were added to each tube, and samples were then homogenized by pipetting. Tubes were then placed 5 minutes on the magnet and 15 µl of supernatant were transferred into a new clean tube.

Final Repair

After the purification of the samples, the final repair step is then needed. The final repair master mix was performed by adding 4.5 µl of Final Repair Buffer Mix and 0.5 µl of Final Repair Enzyme Mix in a new tube. Then 5 µl of Final Repair master mix was added to each sample. After the mixing of the samples and the final repair mix, the tubes were placed in preheated thermal cycler using the following program:

60 °C – 10 min, hold at 4 °C.

Bisulfite conversion

The bisulfite conversion required the addition of 30 µl of Bisulfite Reagent Solution, after proper mixing, to each sample. Then tubes are placed in a pre-warmed thermal cycler with the following program:

95 °C – 5 min, 60 °C – 20 min, 95 °C – 5 min, 60 °C – 40 min, 95 °C – 5 min, 60 °C – 45 min, hold at 20 °C

After the conversion step, DNA was purified using the magnetic beads solution master mix prepared as follows; 200 µl of Binding buffer are mixed with 2.4 µl of magnetic bead solution. 160 µl of this freshly prepared solution was added to each sample. After adding the mix samples were left for 5 minutes at RT, then 5 minutes at RT on the magnet. The supernatant was then carefully removed and discarded. After this step, 200 µl of fresh 70% Ethanol were added to each sample to resuspend the converted DNA. Once the DNA was incubated with the EtOH, 200 µl of desolfonation buffer was added to each sample. Samples were then placed on the magnet, cleaned from the desolfonation buffer, and again washed with fresh EtOH 70%. During the last step, the DNA was eluted in 23 µl of elution buffer before being amplificated during the step of library amplification.

Optimization of the library amplification cycle using qPCR

To assess the best number of cycles to yield a high diversity library with minimal levels of duplicates in each sample, a qPCR was performed.

After the detected of required cycles, we then performed the PCR amplification. This step was performed by combining 4 µl of Amplification Primer Mix with 20 µl of Amplification Enzyme Mix. Once the mix was pipetted properly and spun down, 24 µl of PCR master mix was pipetted to each sample tube. Samples were then placed in a preheated thermal cycler with the following program:

95 °C – 2 min, N (95 °C – 15 s, 60 °C – 1 min, 72 °C – 30 s), 72 °C – 5 min, hold at 10 °C

Where N shows the cycles of amplification obtained for each sample in the qPCR optimization. After the library amplification, one last step of purification with beads is needed. For this purification step, 50 µl of beads were resuspended in each reaction tube. After the two EtOH 70% washes, 20 µl of Nuclease-free Water was added to the samples. Finally, 18 µl of the eluate were removed from each sample and placed in new tubes.

8 Next-generation sequencing

Before sequencing, library concentrations were evaluated using Qbit High sensitivity assay and library fragment sizes were also evaluated, as described in section XXX. Fragments of the samples taken in this study should appear in the range of the lower marker (35 bp) and the upper marker (10380 bp), all the different libraries show a fragment size of 150 bp.

The sequencing step was performed using standard Illumina protocols in 1 × 150 (in the case of RRBS) and 2 × 150 (in the case of WGBS) base pairs. Sequencing was performed using the NovaSeq 6000 instrument at Area Science Park in Trieste.

9 Analysis Methods

9.1 Data analysis performed

The analysis methods described in this chapter refer to WGBS data.

RRBS performed experiments were analyzed using the same software described below (except for the -rrbs specific choice when needed and for the duplicates, removal using Unique Molecular Identifiers UMIs) but will not be described in this section. Despite the computed analysis of the RRBS data, we decide to not continue using this approach as it did not meet our experimental needs.

The sequencing step of WGBS data produced a total number of raw reads per sample that are shown in the bar plot below. **Figure 35**

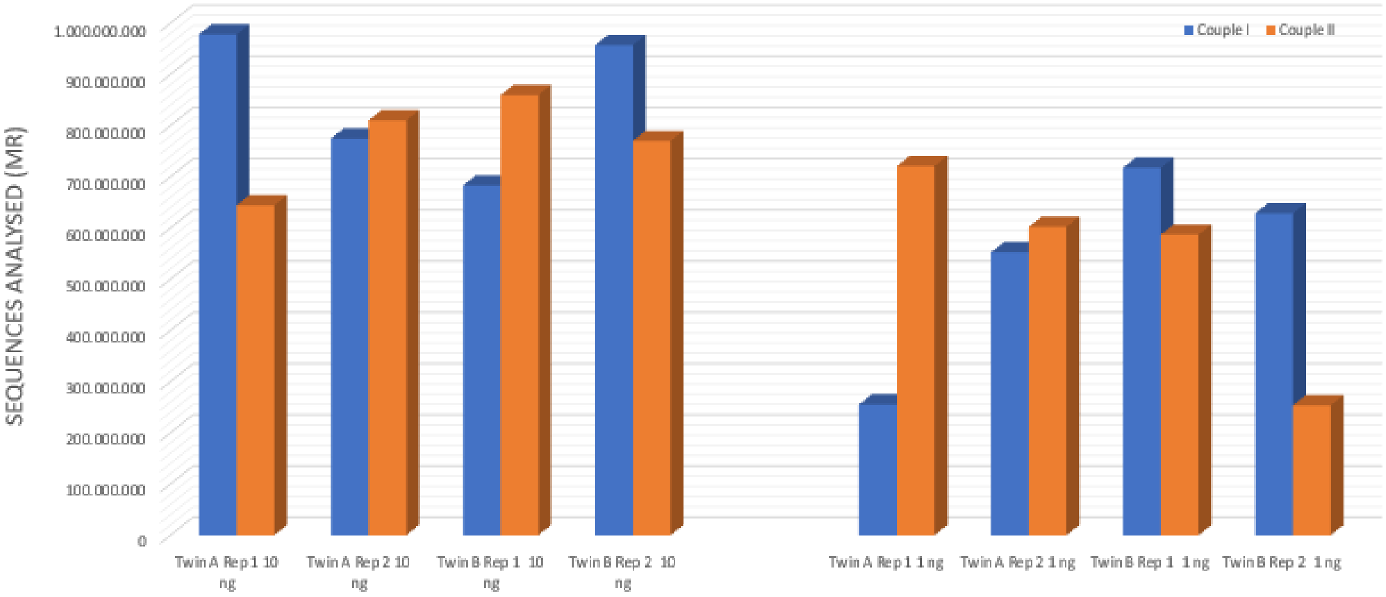


Figure 35 Total raw reads for each sample sequenced on NovaSeq 6000 system.

After sequenced the libraries, the bisulfite conversion efficiency was evaluated by mapping the λ reads to the bisulfite-converted genome of the phage. For all the samples taken in the study, the % of λ was in the range of 97-99%, as showed by the protocol

The entire data analysis workflow performed is summarised in **figure 36.**

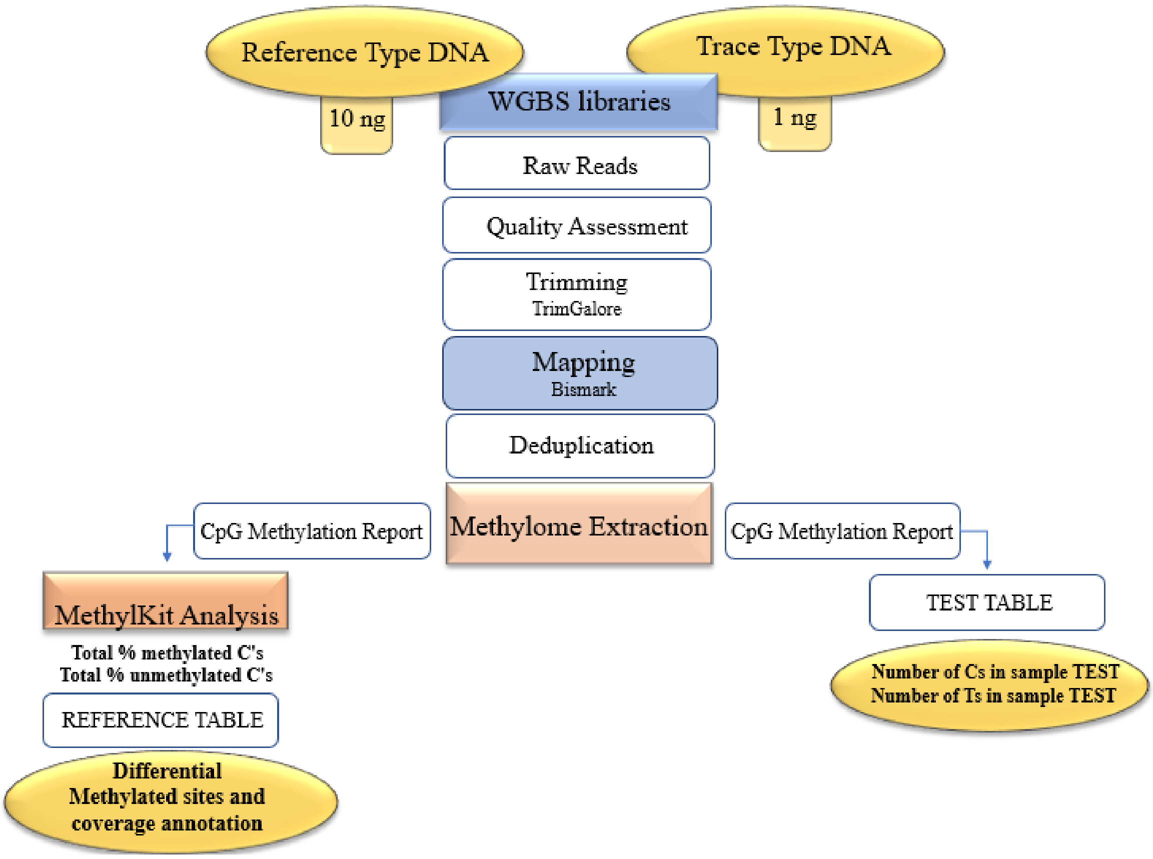


Figure 36 Analysis performed for REFERENCE type DNA and TRACE type DNA.

9.2 Quality control of sequenced libraries using FastQC

Information from the sequencing run for each library was stored in a text-based format file called FASTQ. FASTQ files were specific for all the 16 libraries obtained and have information about each read that was generated during the forward sequencing and the reverse sequencing. In the FASTQ files are stored not only the sequence for each read but also the quality obtained for each base in the read in ASCII code. Before performing further analysis firstly, quality control for all the obtained sequenced libraries was done. FastQ quality control was performed using software called FastQC. FastQC supplies a simple way to do quality control checks on raw sequence data coming from high throughput sequencing and gave the possibility to easily visualize quality metrics for the sequencing data. One of the most important features of the FastQC report is that it gave a plot of per base sequence content. In the case of the WGBS library, the displayed % of different bases present should be different if compared to normal sequenced libraries. F**igure 37** can be compared to the normal output of a balanced library not converted with bisulfite and one of the samples used in this study bisulfite-treated.

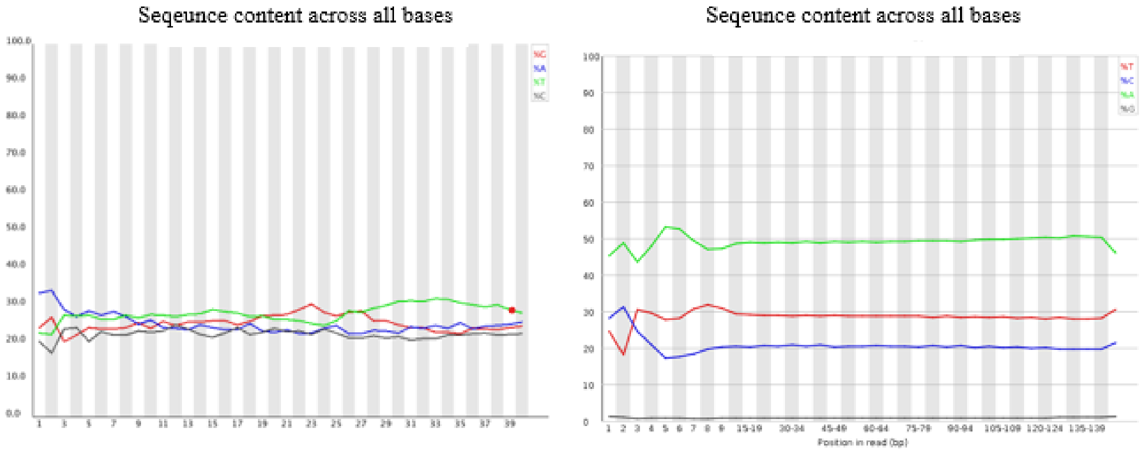


Figure 37 A) Per base content in normal sequenced libraries. **B)** per base content in one of our samples bisulfite-treated. Bases are imbalanced thanks to the chemical conversion of unmethylated cytosines into thymine.

9.3 Adapter trimming for WGBS

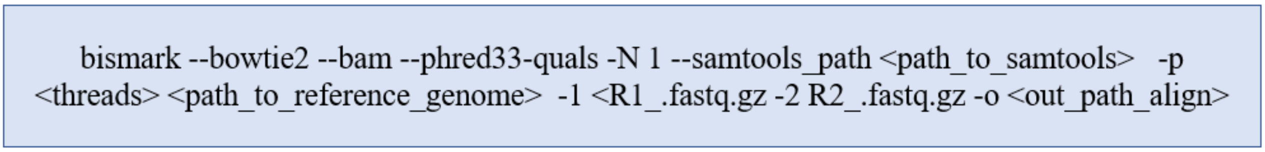
In this study, adapter sequences were removed from the WGBS data using the software TrimGalore. TrimGalore requested to specify the type of libraries to analyze. In the case of the libraries taken in the study, the commands specified were:

```
trim_galore --phred33 --illumina --paired --adapter2 <adapter_Sequence>  --output_dir  <output_directory>
<R1> <R2>
```

-- *paired* (specifying paired-end sequencing), the *adapter* option is the sequence of R2 adapters, *output_dir* the output directory and --*phred33*, to use the ASCII+33 quality scores as Phred scores.

9.4 Alignment to the reference Human genome using Bismark

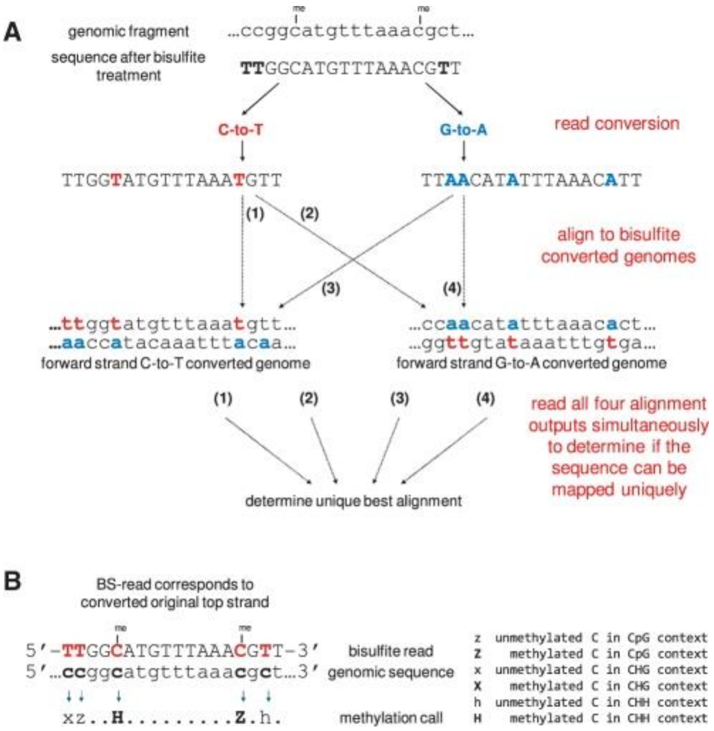
The second step of the data analysis required that reads obtained are mapped to a reference genome. WGBS reads were mapped to the human reference genome using software called Bismark. Bismark uses short read aligner Bowtie 1, or Bowtie 2 to map bisulfite converted sequence reads to the genome (Krueger and Andrews, 2011). Bismark aligner is created to univocally map bisulfite treated sequencing reads. For this study, the reference genome used was the Human Genome GRCh38 obtained from the NCBI websites (ftp://ftp.ncbi.nih.gov/genomes/). Bismark needed a converted reference genome before performing the alignment. To do that human reference genome downloaded was converted using the command *bismark_genome_preparation*. With this command, Bismark will produce two individual folders, one for a C->T converted genome and another one for the G->A converted genome. To align the sample reads to the converted human genome the command used was:



```
bismark --bowtie2 --bam --phred33-quals -N 1 --samtools_path <path_to_samtools>   -p
<threads> <path_to_reference_genome>  -1 <R1_.fastq.gz -2 R2_.fastq.gz -o <out_path_align>
```

Where *–phred33-quals* is the quality format, -*N 1* Sets the number of mismatches to be allowed in a seed alignment during multiseed alignment, -*p* is the number of cores used during the alignment, --*samtools-path* recognizes the samtools path directory, and -1 and -2 shows paired reads to be aligned.

One more choice available while running Bismark for WGBS converted libraries, is the possibility to create libraries in two different modes, directional or non-directional. All the libraries in this study were done in directional mode, meaning that the sequencing reads will correspond just to a bisulfite-converted version of the original forward or the reverse strand. This step is needed because Bismark software cannot know the strand identity a priori and, because of that, it looks for the unique alignment by running four alignment processes simultaneously. As the first step bisulfite reads are transformed into a C-to-T and G-to-A version (equivalent to a C-to-T conversion on the reverse strand). Each of the transformed reads is then aligned to equivalently pre-converted forms of the reference genome, using four parallel instances of the short read aligner Bowtie **(Figure 38)** This kind of mapping enables the alignment software to uniquely determine the strand origin of the bisulfite read (Krueger and Andrews, 2011)



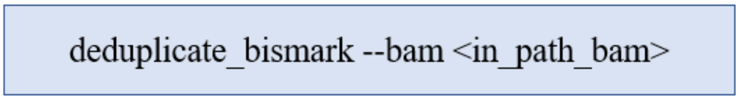
A
genomic fragment
...ccggcatgtttaaacgct...
me
me
sequence after bisulfite treatment
TTGGCATGTTTAAACGTT
C-to-T
G-to-A
read conversion
TTGGTATGTTTAAATGTT
TTAACATATTTAAACATT
(1)
(2)
align to bisulfite converted genomes
(3)
(4)
...ttggtatgtttaaatgtt...
...aaccatacaaatttacaa...
forward strand C-to-T converted genome
...ccaacatatttaaacact...
...ggttgtataaatttgtga...
forward strand G-to-A converted genome
(1)
(2)
(3)
(4)
read all four alignment outputs simultaneously to determine if the sequence can be mapped uniquely
determine unique best alignment
B
BS-read corresponds to converted original top strand
me
me
5'-TTGGCATGTTTAAACGTT-3'
bisulfite read
5'...ccggcatgtttaaacgct...3'
genomic sequence
xz..H.........Z.h.
methylation call
z unmethylated C in CpG context
Z methylated C in CpG context
x unmethylated C in CHG context
X methylated C in CHG context
h unmethylated C in CHH context
H methylated C in CHH context

Figure 38 A) Reads coming from the genomic fragments on the top of the figure are converted into a C-to-T and a G-to-A version. Reads are then aligned to equivalently converted versions of the reference human genome. Bismark aligner looks for the unique best hit, which is figured out from the four parallel alignment processes [in this particular case represented in the figure, for example, the best alignment has no mismatches and directly comes from the thread. (**B**) Every single methylation state of positions involving cytosines is revealed by comparing the read sequence with the corresponding reference genomic sequence. Depending on the strand the read mapped against the reference genome can involve looking for C-to-T (as shown here) or G-to-A substitutions (Krueger and Andrews, 2011).

All WBGS data were sequenced for 800/900 MR as already described in the paragraph *Quality control of sequenced libraries using FastQC*. From the total amount of reads sequenced and aligned to the reference human genome, Bismark looked for the unique best hit, meaning that the algorithm will exclude all the ambiguous alignments generated by mapping reads in several points and give back just all the reads mapped univocally to a position.

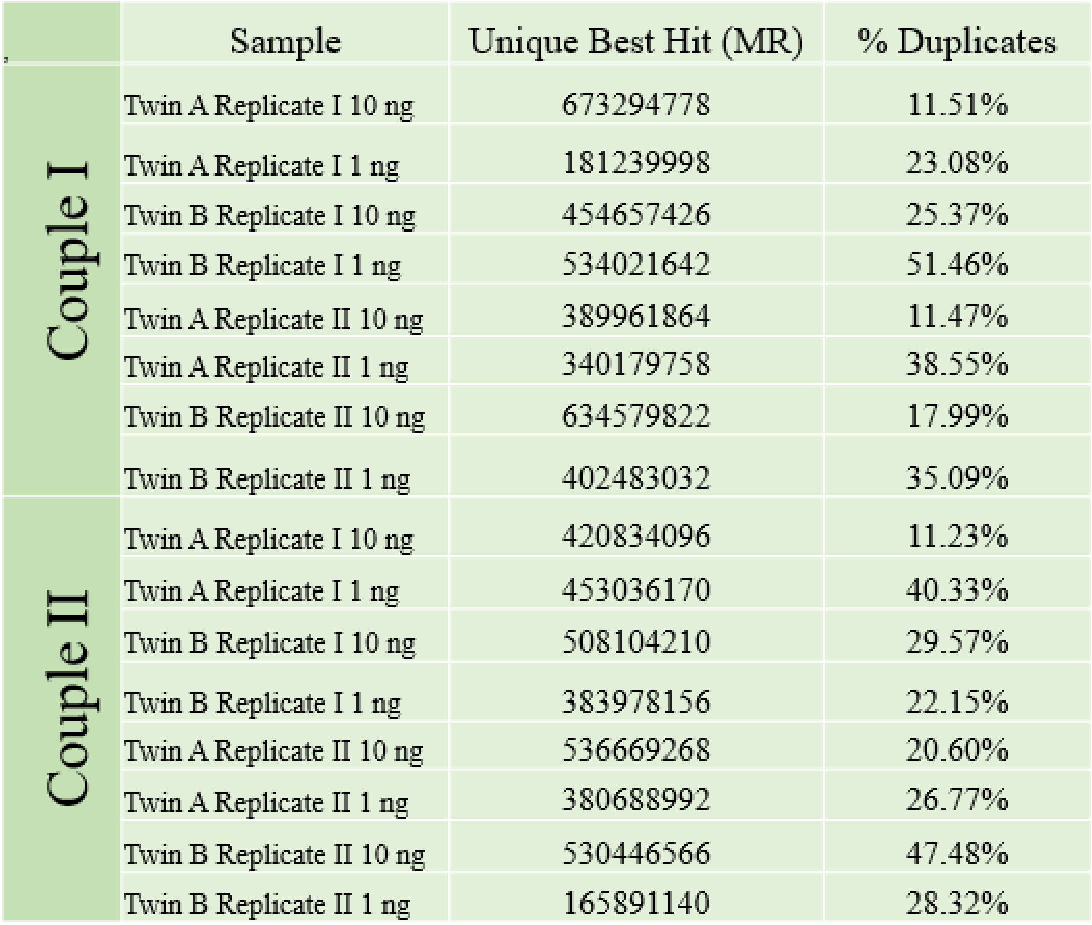
9.5 Duplicates removal

Reads that result from the alignment as identical, starting and ending at the same base pare and having the same sequence, are called duplicates. Duplicates need to be removed from further analysis because they could directly compromise the methylation calling, resulting in overcalling some different methylation bases. To remove duplicates from WGBS libraries the Bismark software was used. In this case, the *deduplicate_bismark* command was specified as shown below



```
deduplicate_bismark --bam <in_path_bam>
```

In the table below the total amount of unique best hit aligned for each library and the relative percentage of deduplicates removed are shown **(Table 3)**

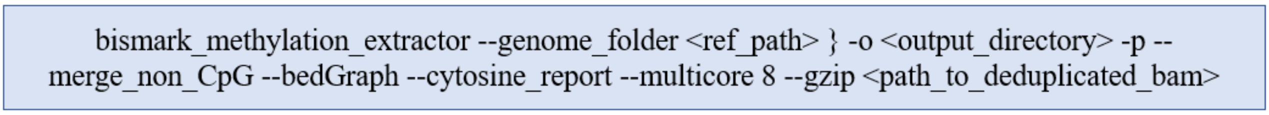


	Sample	Unique Best Hit (MR)	% Duplicates
Couple I	Twin A Replicate I 10 ng	673294778	11.51%
	Twin A Replicate I 1 ng	181239998	23.08%
	Twin B Replicate I 10 ng	454657426	25.37%
	Twin B Replicate I 1 ng	534021642	51.46%
	Twin A Replicate II 10 ng	389961864	11.47%
	Twin A Replicate II 1 ng	340179758	38.55%
	Twin B Replicate II 10 ng	634579822	17.99%
	Twin B Replicate II 1 ng	402483032	35.09%
Couple II	Twin A Replicate I 10 ng	420834096	11.23%
	Twin A Replicate I 1 ng	453036170	40.33%
	Twin B Replicate I 10 ng	508104210	29.57%
	Twin B Replicate I 1 ng	383978156	22.15%
	Twin A Replicate II 10 ng	536669268	20.60%
	Twin A Replicate II 1 ng	380688992	26.77%
	Twin B Replicate II 10 ng	530446566	47.48%
	Twin B Replicate II 1 ng	165891140	28.32%

TABLE 3 Table listed all the MR obtained from Bismark recognized as the best unique hit. On these total amounts of obtained reads, Bismark will remove the % of duplicates shown on the right of the table.

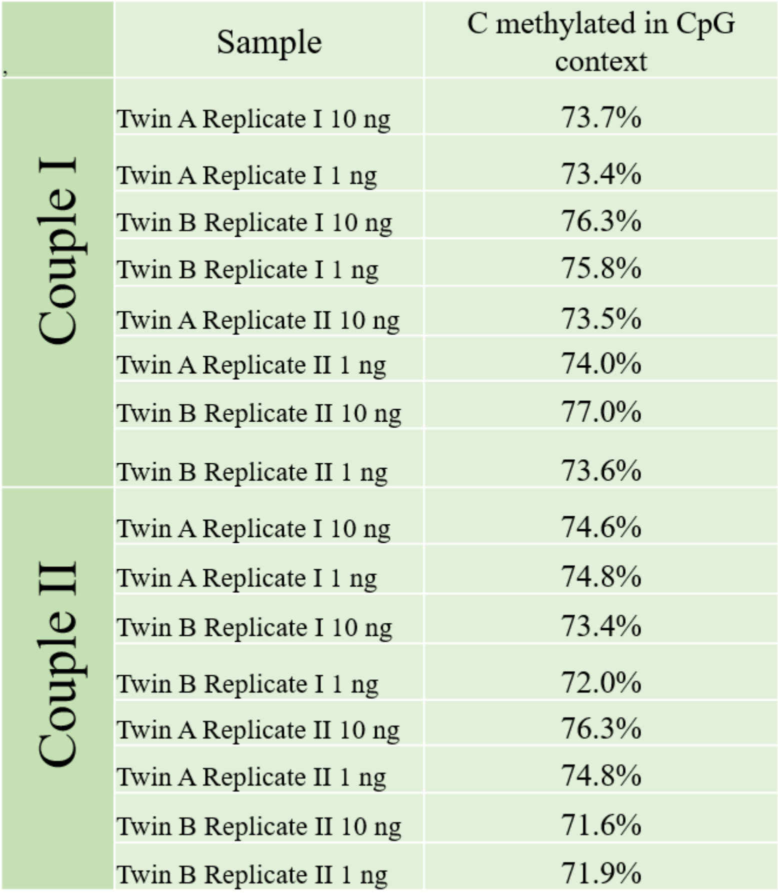
9.6 Methylome Extraction

For all the samples strand-specific methylation information and the total methylation coverage files were obtained from the Bismark alignment using the optional command of methylation extractor code. The *bismark_methylation_extractor* was used to run directly on Bismark deduplicate bam output to extract the methylation call for every single Cs analyzed. The position of every single C will be written out to a new output file in which the desired context is needed (in the case of libraries in this study the only screened context was CpG), whereby methylated Cs will be labelled as forward reads (+), non-methylated Cs as reverse reads (-). Bismark_methylation_extractor was run using the following specification



```
bismark_methylation_extractor --genome_folder <ref_path> } -o <output_directory> -p --
merge_non_CpG --bedGraph --cytosine_report --multicore 8 --gzip <path_to_deduplicated_bam>
```

were --*genome_folder* which specifies reference genome path, --*cytosine_report*, which gives a detailed report with several Cs detected, -*p*, to specify paired-end sequencing reads, and --*merge_non_CpG*, because of in this particular case just CpG context is needed. **Table 4** shows % of cytosines in Cpg context obtained from Bismark tool



,	Sample	C methylated in CpG context
Couple I	Twin A Replicate I 10 ng	73.7%
	Twin A Replicate I 1 ng	73.4%
	Twin B Replicate I 10 ng	76.3%
	Twin B Replicate I 1 ng	75.8%
	Twin A Replicate II 10 ng	73.5%
	Twin A Replicate II 1 ng	74.0%
	Twin B Replicate II 10 ng	77.0%
	Twin B Replicate II 1 ng	73.6%
Couple II	Twin A Replicate I 10 ng	74.6%
	Twin A Replicate I 1 ng	74.8%
	Twin B Replicate I 10 ng	73.4%
	Twin B Replicate I 1 ng	72.0%
	Twin A Replicate II 10 ng	76.3%
	Twin A Replicate II 1 ng	74.8%
	Twin B Replicate II 10 ng	71.6%
	Twin B Replicate II 1 ng	71.9%

Table 4 Methylation percentage result from Bismark alignment

10 Results

10.1 Setting of the REFERENCE table

Once all the cytosines were extracted from the command of Bismark methylation extraction two different analyses were performed to obtain the reference table and test table.

To obtain the reference table all the data derived from the 10 ng libraries from the couple I and couple II were analyzed as described in **Table 5**

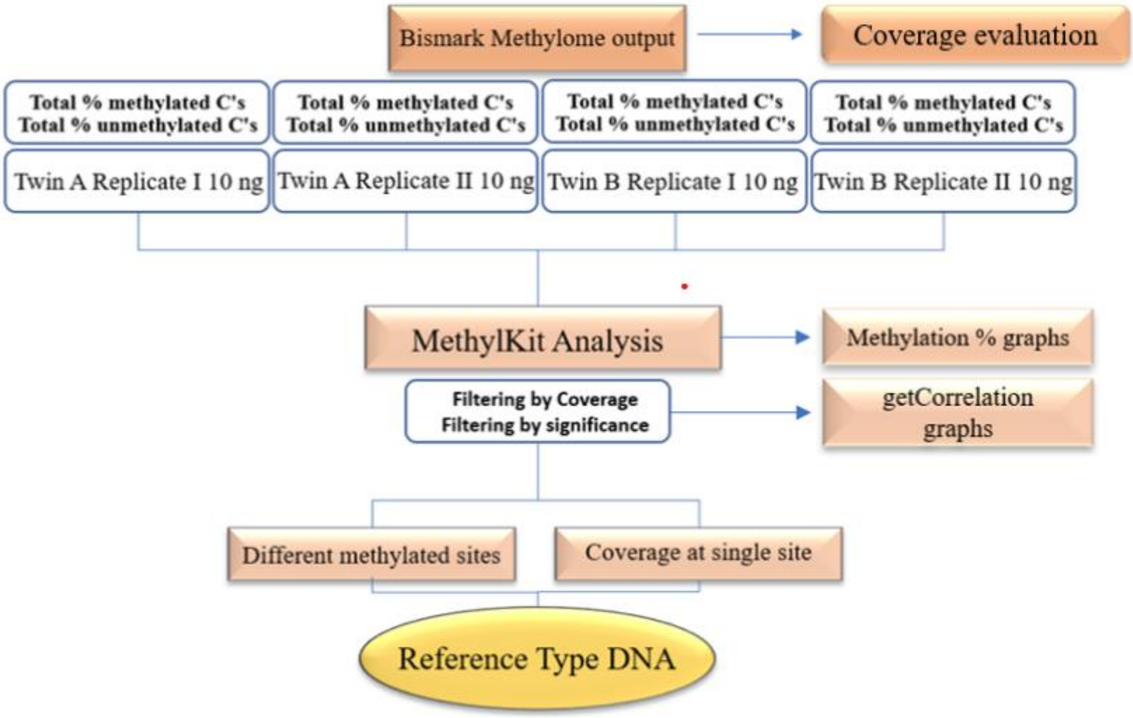


Table 5 Schematic workflow on how to reference tables was created from couple I and couple II.

Before starting the analysis using MethylKit, the reference sample coverage across all different bases was evaluated using the DeepTools plotCoverage function.

The plotCoverage function was performed to assess the sequencing depth of all the given samples. Plots obtained from Couple I and Couple II reference samples are shown in **figure 38**

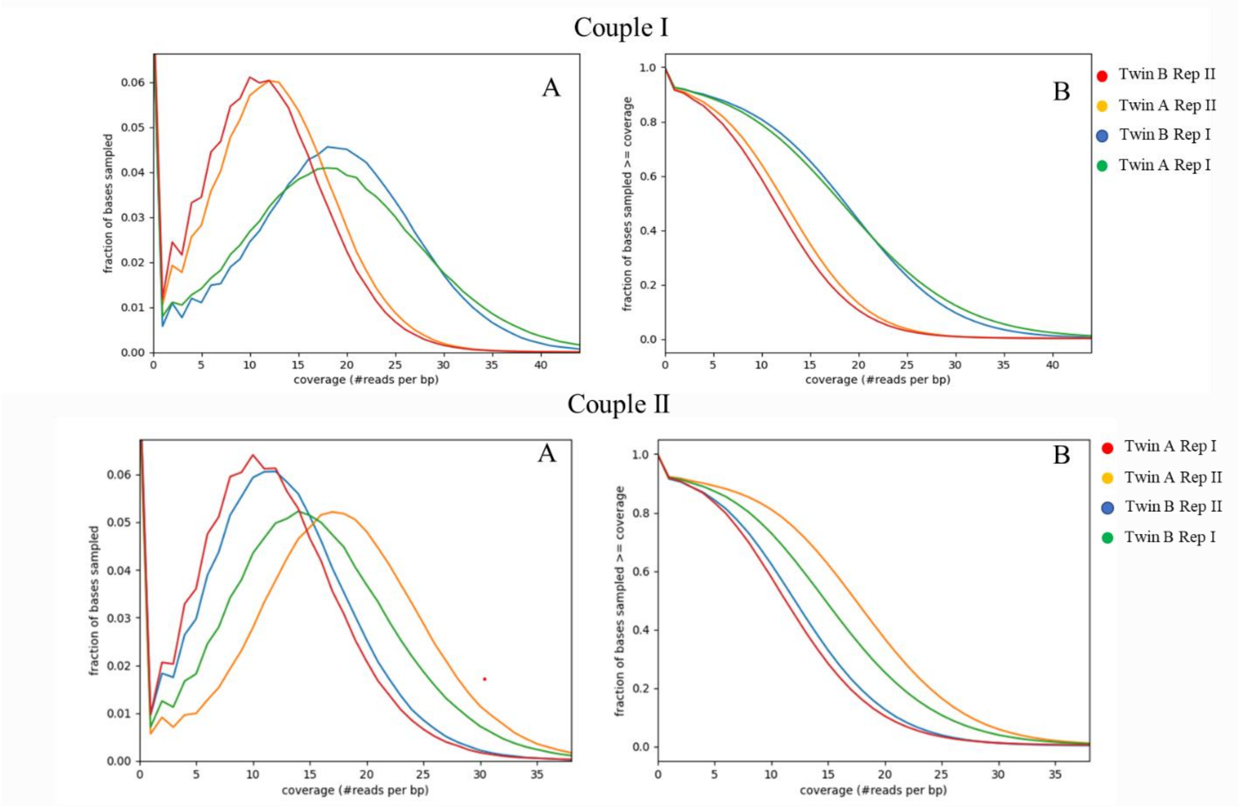


Figure 38 Coverage obtained from reference samples for Couple I and Couple II. For both couples, distribution of read coverage are reported in the A panels (Couple I: Twin B Replicate II mean coverage = 11.27, Twin A Replicate II = 12.10, Twin A Replicate I = 17.91, Twin B Replicate I = 18.10. Couple II: Twin A Replicate I = 11.88, Twin B Replicate II = 11.36, Twin B Replicate I = 14.66, Twin A Replicate II = 16.84). B panels supply the same information about read coverage being stratified by genome fraction.

10.1.1 Descriptive Results of Reference samples using MethylKit

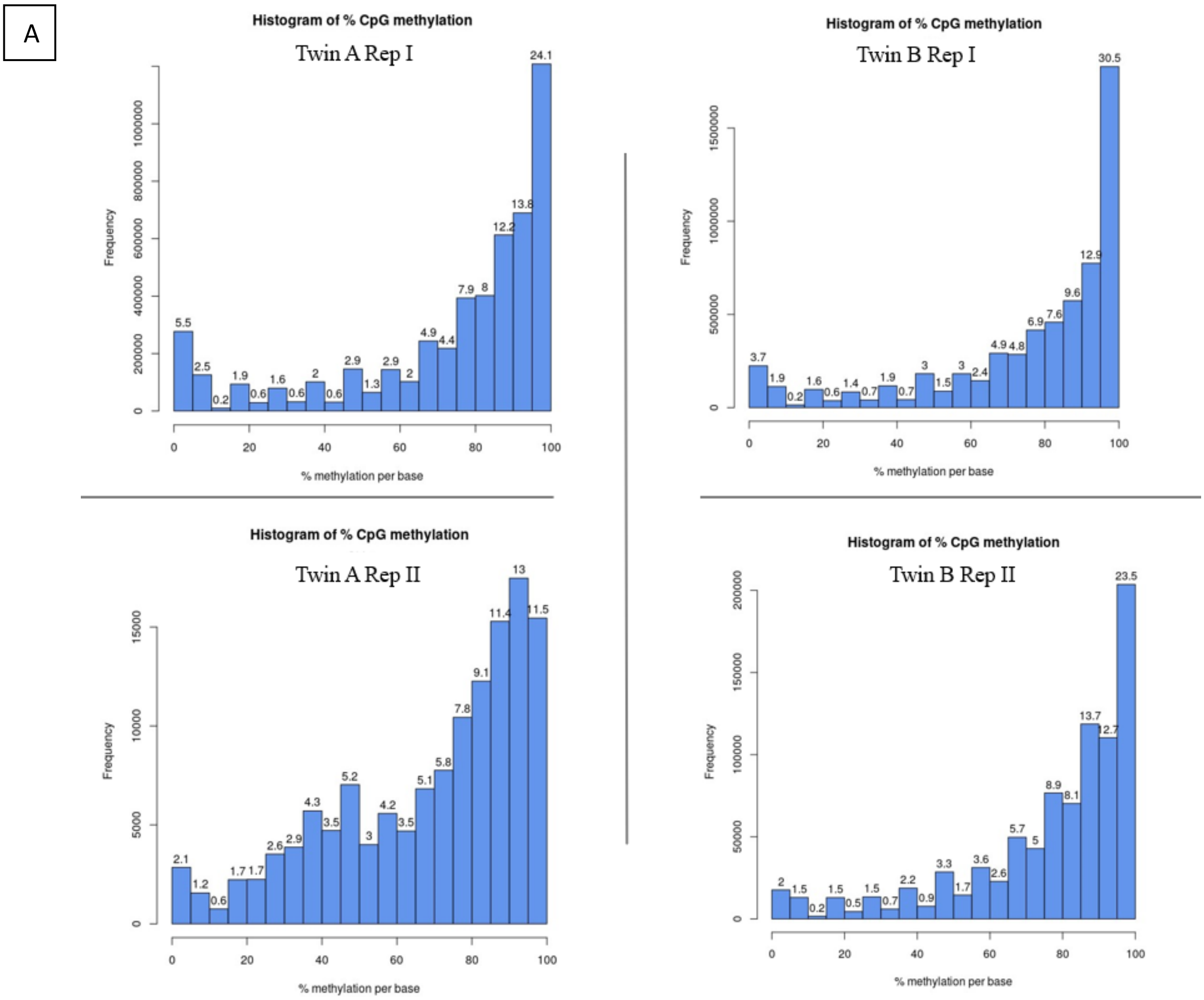
From Bismark methylome extraction output, for all the reference data sets, the % of methylated Cs and unmethylated Cs were analyzed using the Bioconductor R package called MethylKit. This package allows users to obtain a table storing information about chromosome position, strand, coverage, % of Cs, and % of Ts.

All four samples from each couple were converted into a MethylKit object to perform a preliminary descriptive statistics analysis.

From the MethylKit object we first obtained methylation information per sample using the command below:

```
getMethylationStats(myobj[[1]],plot=TRUE,both.strands=FALSE)
```

In the command. The number in square brackets identical samples in the list. This command plots histograms for percent methylation distribution for each sample **(Figure 39)**



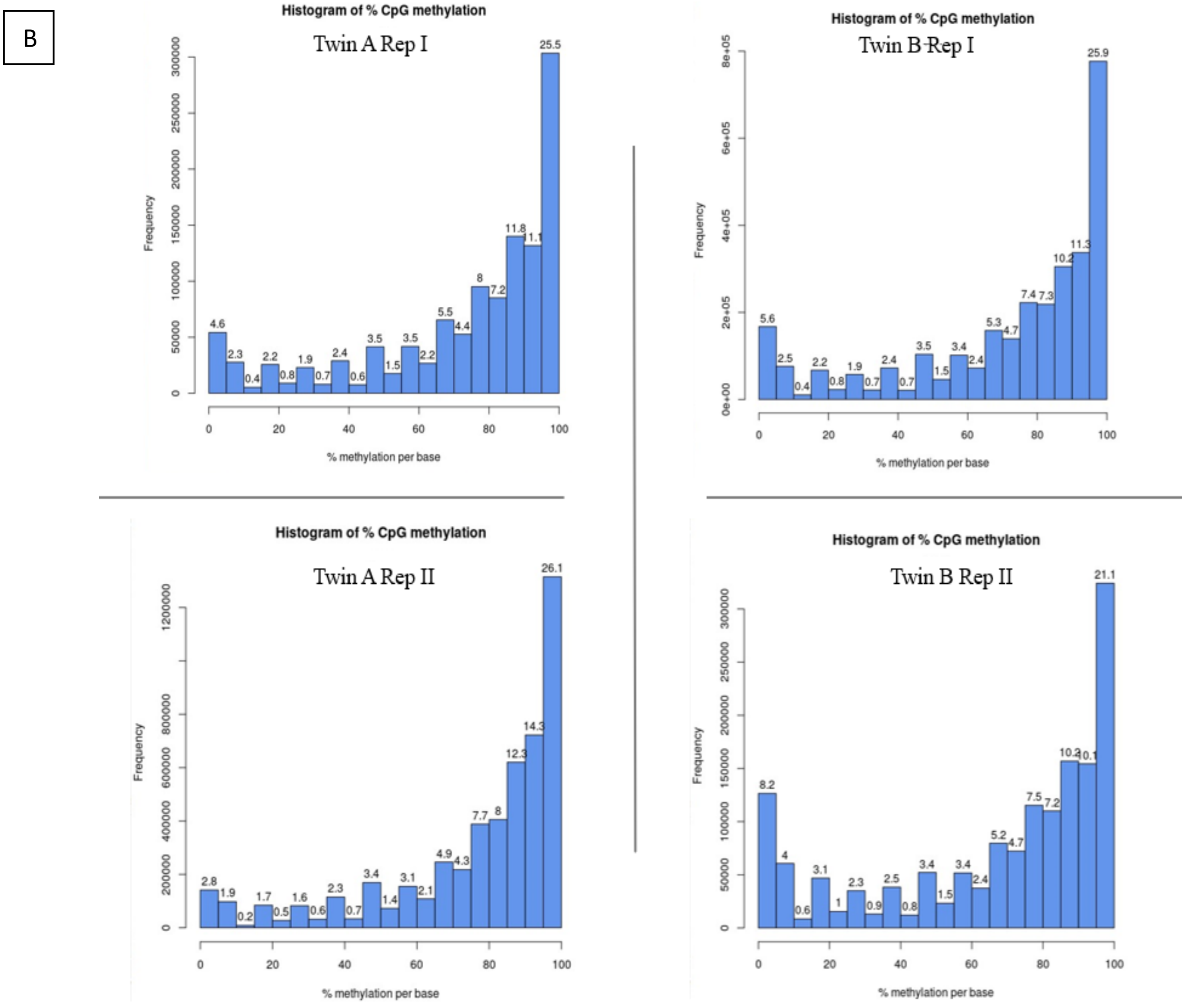
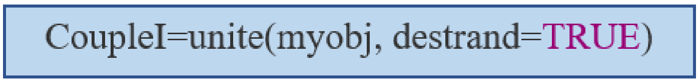


Figure 39 Histogram of the percentage of CpG methylation in Reference Samples in Couple I **(A)** and Couple II **(B)**. Numbers on the bars show what percentages of CpG sites are contained in each bar and refer to methylated or not methylated single sites.

Once the histogram of the percentage of CpG methylation for each sample was performed, to do further analysis, we first merge all samples in one object that includes all pair base locations covered in all samples for one couple.

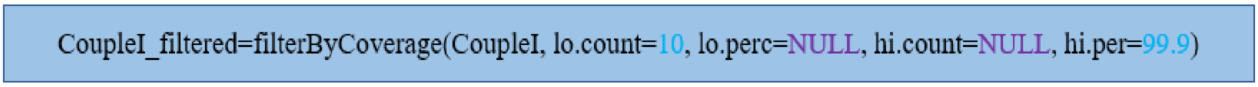
The command needed to merge the samples was:



```
CoupleI=unite(myobj, destrand=TRUE)
```

The destrand choice was set to TRUE, resulting in the merging of the reads from both strands of a CpG dinucleotide. This option allowed us to supply better coverage for every single CpG site given its mostly symmetric methylation mechanism. This option, as described in MethylKit (Akalin et al., 2012) should be preferred when operating on base-pair resolution, such as in our case.

Only sites with a minimum coverage of 10 across all replicates and samples within a twin pair were retained. The command we used to filter for the required coverage was:

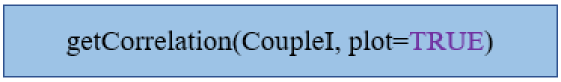


```
CoupleI_filtered=filterByCoverage(CoupleI, lo.count=10, lo.perc=NULL, hi.count=NULL, hi.per=99.9)
```

This command allowed us to discard bases that have low read coverage, less than 10 reads, and bases that have more than 99.9th percentile of coverage in each sample.

For all the samples we also check the correlation in each of the two couples in the study.

To look graphically at the correlation, we used the function below:



```
getCorrelation(CoupleI, plot=TRUE)
```

The command results in different heat plots from which it can be noticed that, for both couples, the correlation within (replicates) and across the two twins was not different **(Figure 40)**

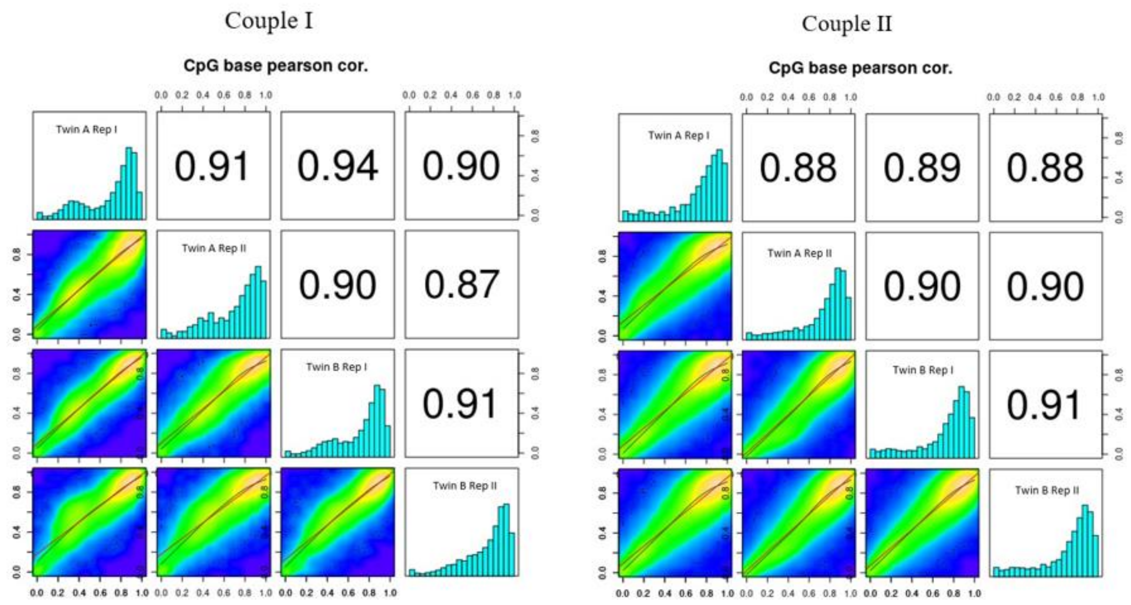


Figure 40 Scatter plot and correlation of CpG methylation between Twin A and Twin B for both twins couple. Heat plots on the bottom part show % methylation for pairwise comparisons of four samples. Numbers in the upper right corner denote the Pearson correlation coefficients. The histograms on the diagonal are the frequency of % methylation per cytosine for each Twin in each couple.

10.1.2 Finding differentially methylated sites in Reference samples using Methylkit

The next step in the reference table creation required the identification of differentially methylated sites specific for each couple. To achieve this step MethylKit needed the below function:

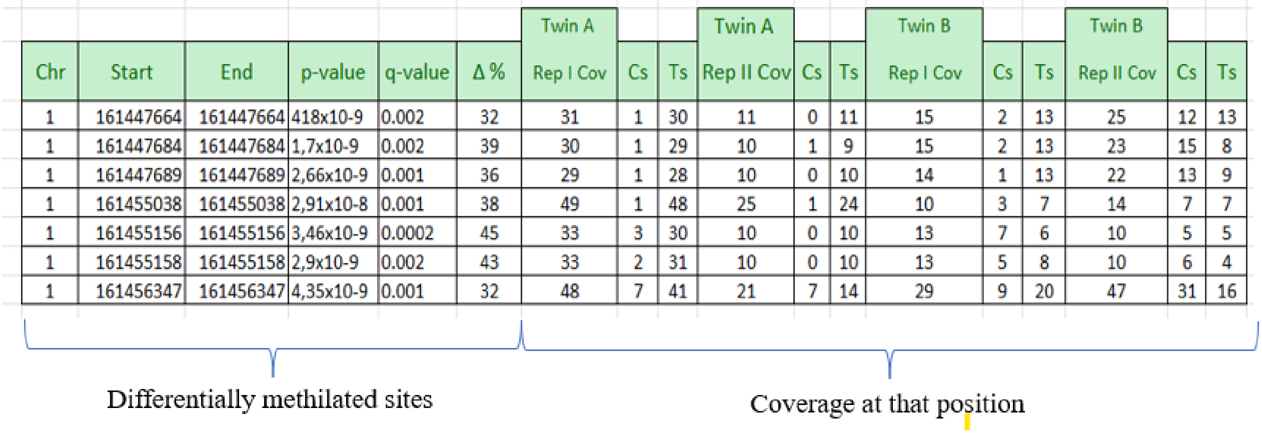
```
myDiff=calculateDiffMeth(CoupleI_filtered)
```

This function will give back a table having different methylated sites, the q-value, the p-value, and the percentage of methylation of the "treatment" sample. In our study, we decided to set as treatment always the twin B in each couple and as "control" the twin A. To calculate p-values, MethylKit will either use Fisher's exact or logistic regression depending on the sample size per set. In both our cases, the function will automatically use logistic regression. Using the logistic regression, the function will try to model the log odds ratio based on the methylation proportion of a CpG, (π in the formula), using the “treatment” vector (twin B for each couple of twins), which denotes the sample group membership for the CpGs in the model.

In the example below, the “treatment” variable is used to predict the log-odds ratio of methylation proportions:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Treatment_i$$

From the calculateDiffMeth function of MethylKit we obtain all the differentially methylated bases between the two samples of each couple. To create the complete reference table, we intersected the table holding the different positions recognized as differently methylated and the raw coverage data. This allowed us to obtain, for each single differentially methylated site, the total coverage in each replicate for each twin, also having the total number of Cs and Ts present at that position. This step results in the creation of our two reference tables that has the format of the table below (**Table 5)**



						Twin A			Twin A			Twin B			Twin B		
Chr	Start	End	p-value	q-value	Δ %	Rep I Cov	Cs	Ts	Rep II Cov	Cs	Ts	Rep I Cov	Cs	Ts	Rep II Cov	Cs	Ts
1	161447664	161447664	418x10-9	0.002	32	31	1	30	11	0	11	15	2	13	25	12	13
1	161447684	161447684	1,7x10-9	0.002	39	30	1	29	10	1	9	15	2	13	23	15	8
1	161447689	161447689	2,66x10-9	0.001	36	29	1	28	10	0	10	14	1	13	22	13	9
1	161455038	161455038	2,91x10-8	0.001	38	49	1	48	25	1	24	10	3	7	14	7	7
1	161455156	161455156	3,46x10-9	0.0002	45	33	3	30	10	0	10	13	7	6	10	5	5
1	161455158	161455158	2,9x10-9	0.002	43	33	2	31	10	0	10	13	5	8	10	6	4
1	161456347	161456347	4,35x10-9	0.001	32	48	7	41	21	7	14	29	9	20	47	31	16

Table 5 Example of the reference table created. For each identification of differentially methylated sites, coverage, resulting in the number of Cs and the number of Ts are reported. This table will be used to compute statistical tests.

10.1.3 Distribution of differentially methylated single sites in Reference samples.

Before computing the statistical analysis, we first inspected the location of differently methylated sites both within the chromosomes and within the genome.

To archive this goal, we filter all the positions that had at least 10X of coverage and that showed at least 25% of different methylation values. We found a total number of 390 differentially methylated sites for the couple I and 220 differentially methylated sites for couple II. For these identified sites we used a Bioconductor package called *genomation* to annotate them into the genome. The *genomation* package gave us the possibility to visualize and quantify genomic intervals over pre-formed functional regions, such as promoters, exons, introns, or no specific areas. First, we plot the percentage of differentially methylated bases overlapping with exon/intron/promoters areas in the genome. This function easily helped us to visualize if the identified sites of methylation in the reference samples match with the region in the human genome accounting for promoters, introns, or exons areas, as shown in Figure 41

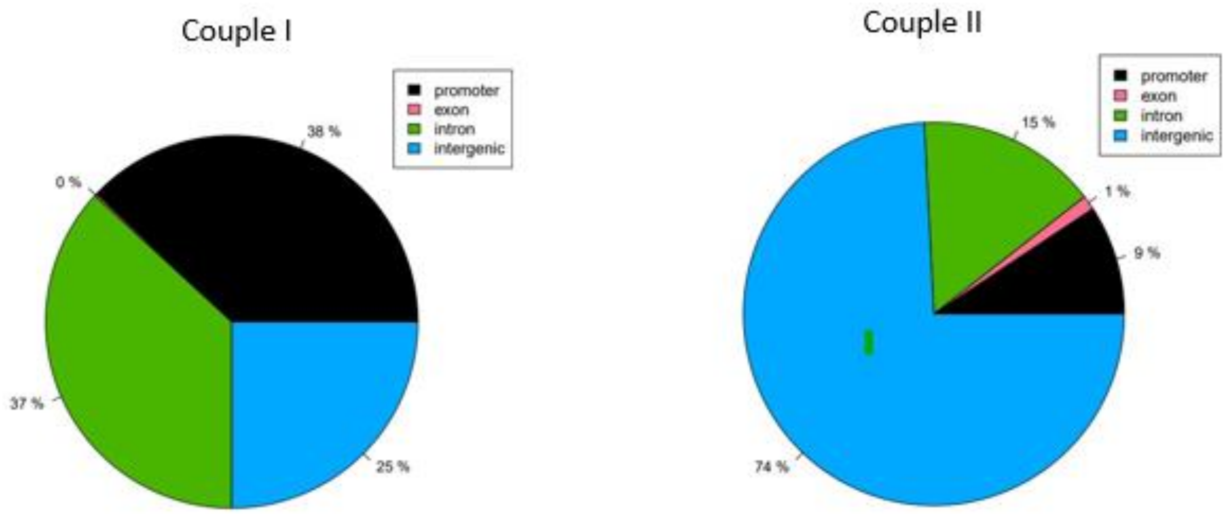


Figure 41 Pie charts for couple I and couple II representing the % of sites found in different areas of the human genome. Sites in the couple I and couple II are localized differently in the genome. In couple I we found equal distribution of sites in promoter regions and intronic regions (36% for promoters and 37% for intronic regions), and 25% of the identified locations are distributed in intergenic areas. In couple II most of the 220 sites are localized in intergenic regions (74%). While, differently from couple I, just 15% and 9% are sites localized at the intronic and promoter regions, respectively.

We then screened if the identified single sites are specifically in CpG islands, CpG island shores, or other regions **(figure 42).**

In this case, we found that at CpGs island, for couple I we had most of the sites localized at CpG island (58%), while fewer sites for couple II are found in such (23%). For couple I, the 26% of the sites results to be localized at CpG shores, and in couple II just 6% of the sites are in the same location. Most of the sites identified in couple II seem to be localized somewhere else (71%, mentioned as “other” locations).

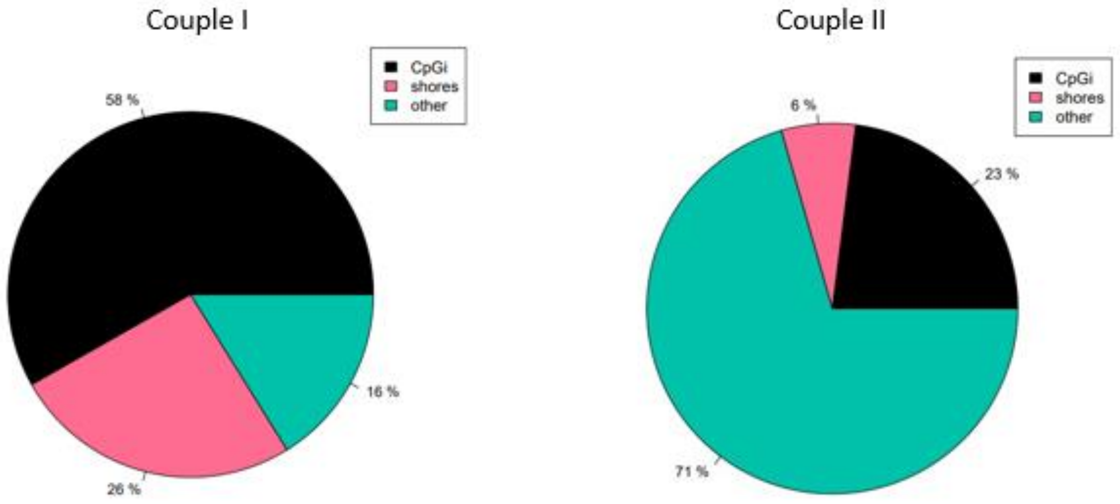
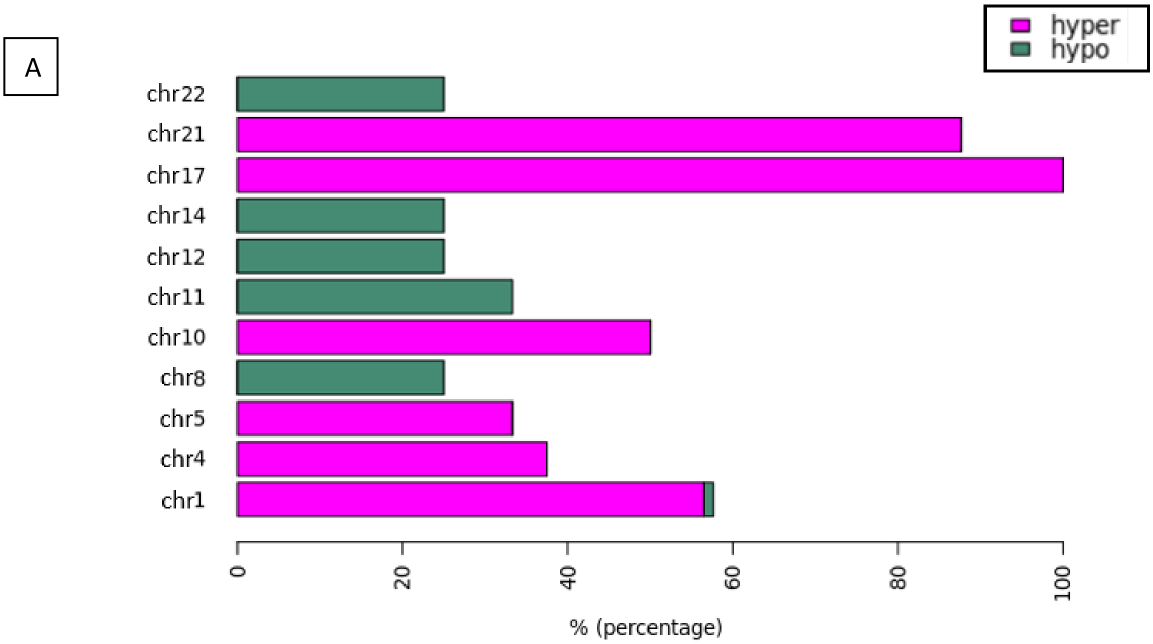


Figure 42 Pie charts for couple I and couple II representing % of sites found in CpG areas, on shores or other locations. In couple I 58% of fund sites result to be in CpGs, while in couple II 23% of the sites are in CpGs. The other sites in couple I are founded in shores 26% and other locations, 16%. Couple II had most of the sites in other locations (71%) and just the 6% at CpGs shores.

We finally screened where the 390 differentially methylated sites in couple I and the 220 methylated sites in couple II were distributed at the chromosomal level.



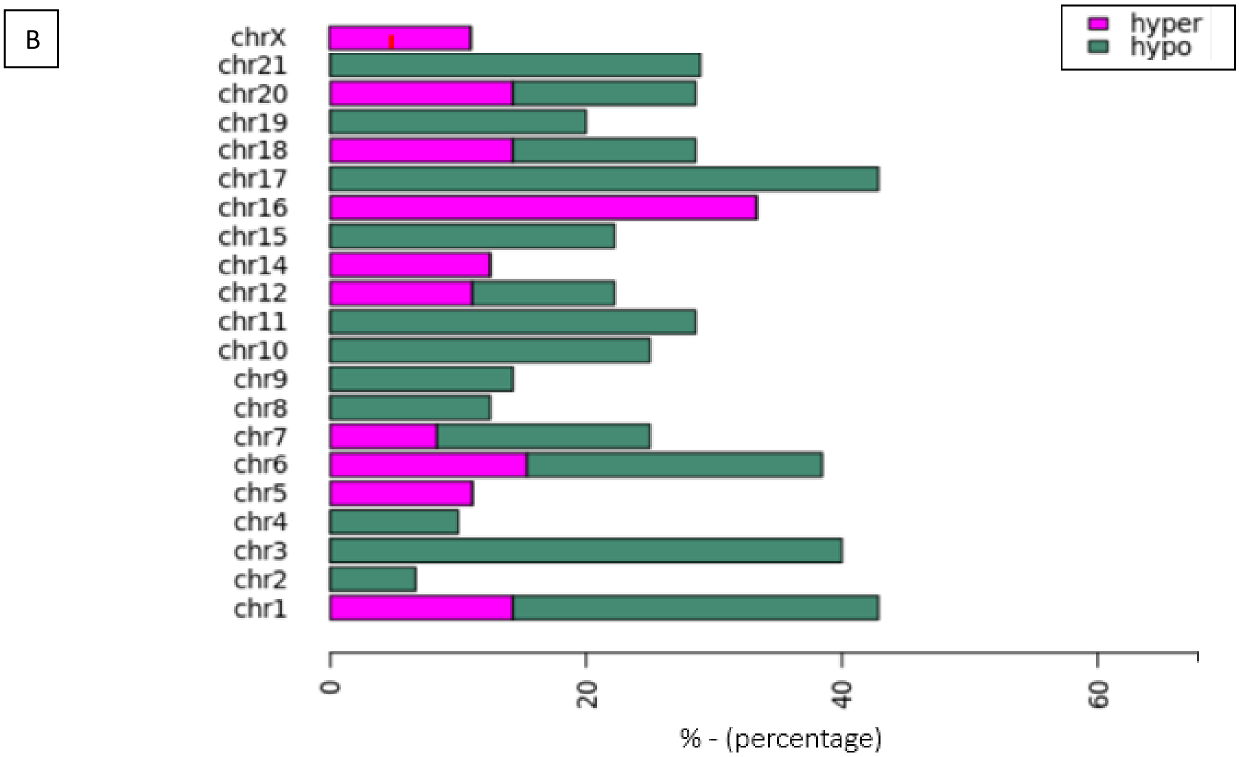


Figure 43 Chromosomal distribution between couple I and couple II differently methylated sites. Couple II shows the major distribution of the identified sites between the chromosomes, showing a maximum level of methylation difference of 40%, while in couple some sites showed a higher difference of methylation (around 60%).

We suppose that the distribution between the different levels of methylation could be associated with the aging of this couple. As reported in Fraga and collaborators work from 2005 (Fraga et al., 2005), older homozygous twins exhibited remarkable differences in their overall content of sites of methylation compared to younger twins couple. In our case, when we sampled couple I, they had 40 years old while couple II had 25 years old. The different ages of the two couple could be recognized as one of the possible explanations for why we saw a different distribution and different levels of the methylated sites in chromosomes.

11 Setting of the TEST table

Test table creation was performed following the below scheme **(Figure 44):**

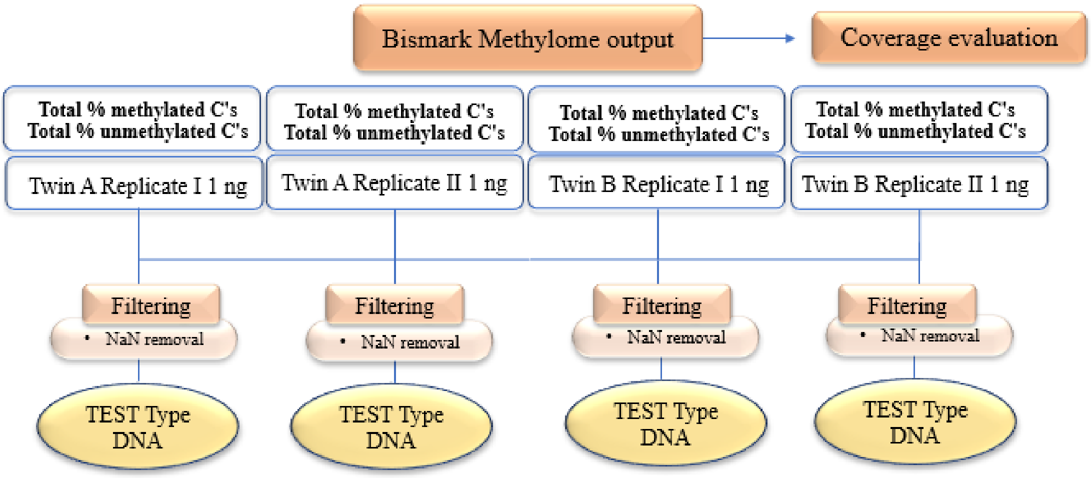


Figure 44 Scheme of TEST (trace) type DNA creation starting from 1ng prepared libraries.

Percentage of methylated Cs and percentage of unmethylated Cs (Ts) were extracted for a couple I and couple II for all the samples obtained from 1 ng starting material. Once the methylome was obtained from Bismark methylation extraction, as described in paragraph *Methylome Extraction*, coverage for all the replicates in each twin couple was analyzed as done previously with the reference type DNA using the DeepTools options *plotCoverage*. **(Figure 45)**.

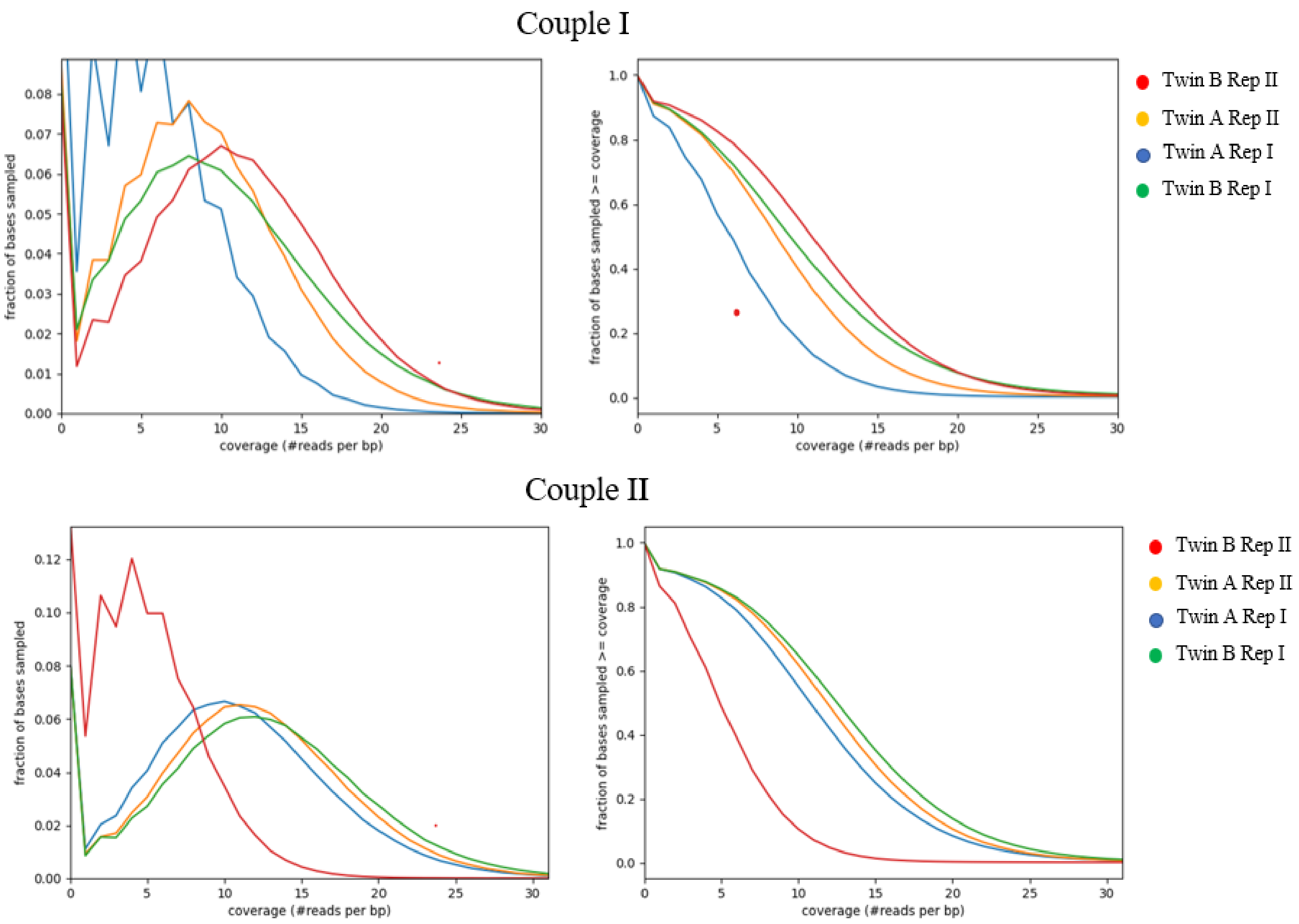


Figure 45 Coverage obtained from TEST samples for Couple I and Couple II. Both couples, I and couple II in panel A have represented the frequencies of reading coverage. (Couple I: Twin B Replicate II mean coverage = 10.73, Twin A Replicate II = 8.63, Twin A Replicate I = 5.83, Twin B Replicate I = 9.93. Couple II: Twin A Replicate I = 10.80, Twin B Replicate II = 4.92, Twin B Replicate I = 12.33, Twin A Replicate II = 11.65). Panel B gave the same information about reading coverage but focused on the genome region covered in the genome.

In the case of test table creation, no further software was needed to obtain the final shape of the table, since the final trace type DNA table was the output of Bismark methylation extraction. To just obtain informative bases we filter out all the positions not covered at all.

12 Design of the statistical approach: the Binomial probability

Once the reference table and test table were created, we tried to look for the most suitable statistical approach to be computed. We aimed to be able to lead back, ideally sampling blindly one of the four test tables created, to the right twin based on the sites found in the reference table.

The reference table created had all the differentially methylated single bases between the two twins in the two couples, while the test table has all the information about methylated cytosine and unmethylated cytosines for the given trace-type sample.

Based on the study of Jiménez et al (Planterose Jiménez et al., 2021), in which researcher hypothesized that the presence of universal stochastic epigenetic variation can be isolated from MZ twins, the statistical approach we were looking for should be able to distinguish between MZ twins pairs not taking into account any epigenetic events that could be related to, for example, pathological discordance or aging drift.

This concept of universal epigenetic variation directly reflects the assumption that, between identical twins, could exist a set of stochastic variations that is not related to any kind of biological meaning or not related to genetic influence.

The statistical test that we considered the most suitable to test the hypothesis of the existence of these random distributed methylated single sites is the calculation of the Binomial probability or Bernoulli statistical trials.

In statistics, a Bernoulli trial is considered an experiment that results in just two possible outcomes, *success*, and *failure*. The assumptions of this binomial distribution are that: there is only one outcome for each different trial and that, for each trial, there is the same probability of getting success, and that every single trial is independent of one another.

In statistics the formula which describes the binomial distribution is calculated as:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

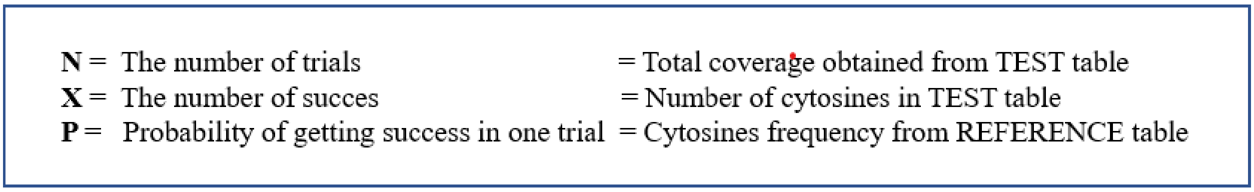
Where n is represented by the number of trials, x is the number of successes, p is the probability of getting success in one trial, and q=1-p, or the probability of getting a failure in a single trial. The idea of using the binomial probability on our genomic dataset was that, based on the different genomic base composition of each twin in a couple (the difference of the bases is expressed by the methylation level), we can statistically lead back the random dataset (what we called the test table), to the proper same dataset (what we called the reference table). One important assumption of the binomial probability is represented by the fact that each different action in the trials must be completely independent. If we consider our datasets, what makes us assume that each sampling stays independent from the others, is given by the fact that during the earlier analysis steps we make sure to remove all duplicates through the deduplication process discussed in paragraph (9.5 Duplicates removal). This allowed us to trust that the probabilities of having sampled one twin or the other are unique and independent in every single trial computed.

Moreover, as expected in a Bernoulli trial, because for each dataset we have a very large number of representative genomic sequences coming from the experimental design, the probability we found could be considered the same for every single trial, highlighting again that each event is completely separate and does not inherit any influence from earlier sampling. Last, the complement to the bisulfite conversion efficiency is the potential error in the trial results (3% at most since the conversion rate is always greater than 97%). With just 10 sampled cytosines the probability of having half of them wrongly evaluated by a lack of conversion is 2.43e-8.

12.1 Binomial probability customization based on our dataset

Based on the Bernoulli equation we adapt variables present in the Bernoulli formula to our sequencing data obtained. More in detail we set up the formula as follows:

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$



N =	The number of trials	= Total coverage obtained from TEST table
X =	The number of succes	= Number of cytosines in TEST table
P =	Probability of getting success in one trial	= Cytosines frequency from REFERENCE table

Our idea was to calculate the binomial probability, for every sequenced cytosine from a test data to be sampled from each twin, given its methylation state assessed in the reference data. One of the fundamental aspects of the setting of this model is given by the fact that we have tried, as much as possible, to set up an analysis that was able to attribute a probability to one rather than to another twin, also working on scarce sampling (i.e., very low coverage) that should also mime the usual forensic scenario as much as possible. So, the model described by us should therefore be able to attribute the belonging of the test table to its twin of sparse low-coverage sequenced CpG sites.

Referring to the Bernoulli formula described in the paragraph above, we match every single variable as described here; the trials are represented by the coverage obtained from the test table being an independent sampling of the same distribution of probability, and the number of successes is represented by the number of methylated cytosine present in the test table (we arbitrarily chose to use methylation as a sampling success) and the estimated probability of success is given by the percentage of bona fide methylation defined in the reference table for each twin. A binomial probability is therefore calculated for each twin in the analysis for each candidate site. The reference table created in MethylKit, moreover, has not been filtered for any parameters such as coverage or methylation difference between the two twins, so we delegated to a python script the intake of valuable sites for the probability estimation based on chosen parameters (i.e., all the technical parameters that we want to test during the binomial probability calculation).

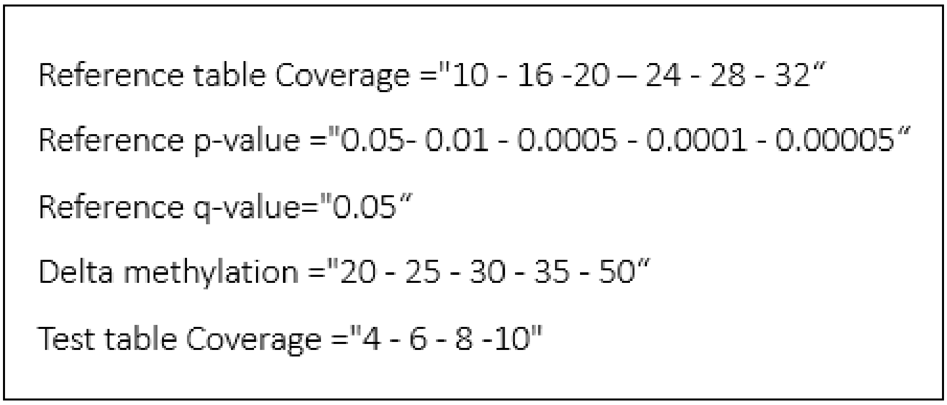
First, because our statistical test was based on the results obtained from the high-throughput sequencing of WGBS samples, we filter out from the reference table and the test table, a series of blacklisted regions. These regions could lead to anomalous mapping or result in high signals in next-generation sequencing experiments. The analysis of these regions could easily lead to inaccurate interpretation. Based on these assumptions we implement our script by filtering out the blacklisted region identified by Amemiya and collaborators (Amemiya et al., 2019) and not counting into the calculation of the single base binomial probability.

Moreover, as already described in the section *REFERENCE creation and TEST table creation*, our tables were not filtered for any parameters during their setups, such as coverage, significance, and methylation difference.

Regarding the attribution of significance used in our model, we have decided not to consider the possibility of correcting the obtained p-values. First of all, the exploratory nature of our study allows us to be able to work with p-value without correction for multiple testing. The study we carried out completely deviates from a confirmatory study, which would require an approach that considers the possibility of obtaining falsely significant results. The assumption is to potentially rely on low-coverage data, thus with a high number of sites tested for significance but with a lower statistical significance. Therefore, a conservative approach that would support the use of a correction for multiple tests, could result in the loss of too many discriminatory sites, to be summed in the cumulative probability.

These considerations led us to focus on the evaluation of several sets of parameter thresholds to skim the data ahead of the binomial probability calculation.

The parameters we permuted into the script are listed in the table below:



Reference table Coverage ="10 - 16 -20 – 24 - 28 - 32”

Reference p-value ="0.05- 0.01 - 0.0005 - 0.0001 - 0.00005”

Reference q-value="0.05”

Delta methylation ="20 - 25 - 30 - 35 - 50”

Test table Coverage ="4 - 6 - 8 -10"

All the above combination results in a total number of 4800 distinct sets of parameters, meaning a total number of 600 different groups of parameters for each test table used.

For every single set, we find the total number of sites used to compute the binomial probability, and the total number of sites discarded (resulting in the sites that were present in the blacklisted regions or in sites that are not included by the above parameters), one file holding all the parameters used and one file having the final predicted twins.

To assess the final predicted twins all the individual probabilities obtained for every single site were then transformed using the logarithmic function and added together to obtain the final call referred to as one twin rather than the other, called in our script *cumulative probability*. The final cumulative probability stands for the summation of every single binomial probability computer for all the cytosines found in common between the test table and one individual of the reference table. The final call was given according to one or the other twin that showed the higher likelihood call.

The binomial probability calculation algorithm is summarised below:

1. Read the reference test table

    1.1 Parse parameters for the reference table (reference coverage, reference p-value, reference q-value, and delta methylation)

2. Define the binomial probability variables.

2.1 Define the number of successes (cytosine at that position in the test table)

2.2 Define the number of trials (number of cytosines and thymine from test table)

2.3 Define the probability of getting success in one trial (cytosine frequency from the reference table.

This will return

probC = stats.binom.pmf (nC, trials, probability)

3. Load the test table and load the blacklisted region file

3.1 Filter out sites present in the blacklisted region. Return file only having valid sites.

4. Function used to store used and discarded sites in the prediction

4.1 Kept sites - write sites used to compute the cumulative probability

4.2 Discarded sites - write sites discarded from cumulative probability

5. Logarithmic transformation of single probability computed

cumulative_prob_log[sample] += abs(math.log10(site_prob[pos][sample]))

6. Prediction of the right twin according to the summation of all the single probability obtained

6.1 Predicted label = Summation of the log-p probability obtained

Optional

5. Bootstrap function (described in the section below).

The required argument for these functions is:

--reference-table – supply path to the reference table. Required=True.

--test-table – supply path to the test table. Required=True.

--ref-config - Definition of reference columns. Required=True

Optional parameters to be added:

--encode-blacklist – bed file having a problematic position to be filtered out.

--out - Output prefix

--ground-truth - The known test twin label

--reference-min-coverage - Minimum coverage in reference table. For any site, it must be covered by these values to be processed.

--reference-minp-value - Minimum p-value in reference table.

--reference-min-value - Minimum q-value in reference table.

--reference-delta-meth - Minimum methylation difference in reference table.

The bootstrap argument is described in the section below.

13 Calculation of the prediction accuracy and bootstrap analysis

For all the obtained sets of parameters, we then want to assess which combination could lead to the highest accuracy in the prediction. To do that we calculate for each set of parameters how many times the result gave a good prediction and how many times the prediction performed the result to be wrong. The idea of setting an accuracy threshold was chosen to be able to sample only a subset of parameters that would show, as much as possible, a correct reproducibility in the prediction for the greatest number of test samples.

The accuracy was calculated, once defined all the correct and wrong calling, as follows:

Accuracy = correct prediction/ (correct prediction + wrong prediction)

Sorting the obtained *Accuracy* for the highest value of correct prediction, we obtained that 408 sets of parameters were able to predict correctly in 7 cases out of 8.

Unfortunately, no set of parameters can correctly predict all the tests used in the analysis.

We decided to set up this level of accuracy in a completely arbitrary way, just to consider only the subset of parameters that have the highest level of correct prediction, in our case, considering that we do not find one set able to predict correctly in 8 cases out of 8, was represented by the 7 cases out of 8, corresponding in the accuracy of the 87.5 %.

Differently from what we would have expected, we found all the incorrect predictions coming only from a particular dataset, that is the replicate II of the twin B in the pair I (as shown in figure 45). Since only in this particular case for no set of parameters did the prediction occurs correctly, we began to hypothesize that, perhaps, the dataset was affected by some sampling error.

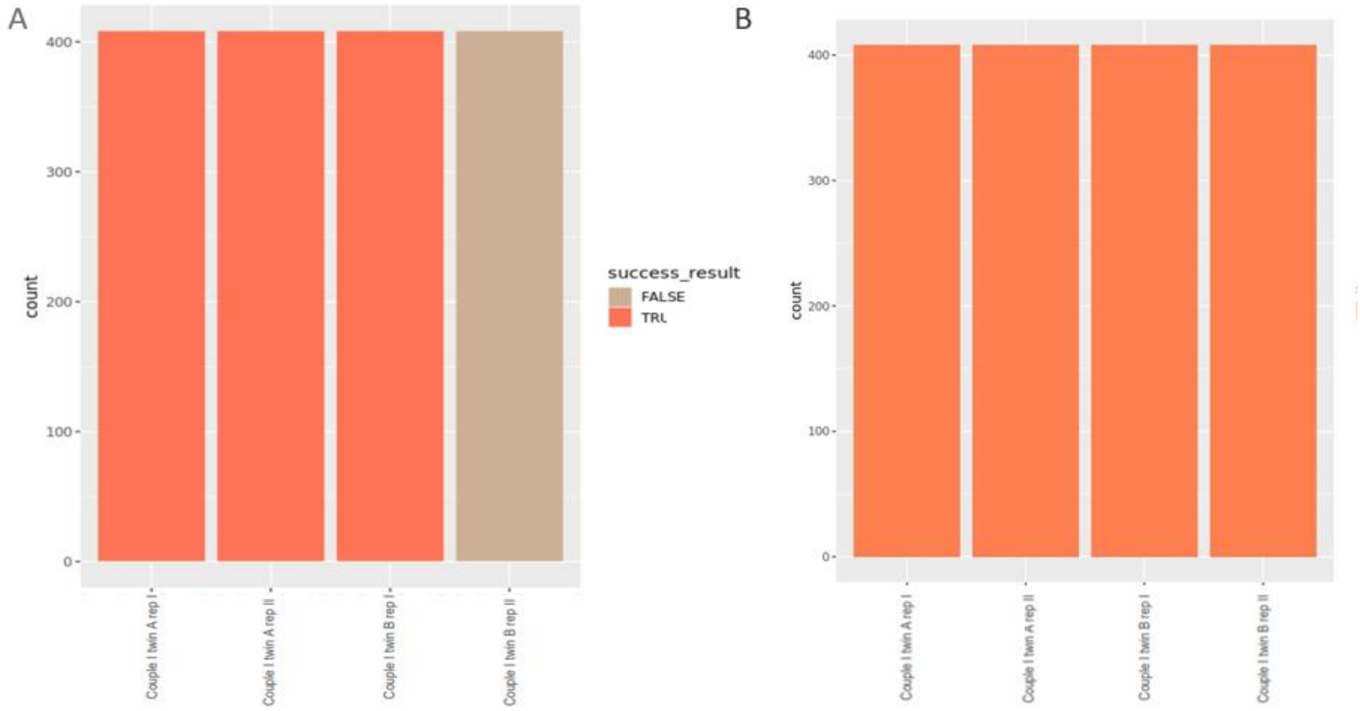


Figure 45 Prediction results were obtained for couple I (panel A) and couple II (panel B) using the 408 sets of parameters having the highest accuracy in the prediction. In couple I twin B replicate II was the only case in which the prediction result was completely incorrect.

Following the calculation of the probability, we then studied how many single sites had been used for every single dataset (test table) to predict the relative twin for the 408 sets of parameters having the highest value of accuracy. Between couple I and couple II we noticed a strong discrepancy, while in couple I the average of sites used is around 300-400 per dataset, for couple II we were able to sample an average of sites equivalent to a hundred. As is clear from panel B of figure 46, there would appear to be no correlation between the number of sites used and the outcome of the prediction. Even in the case of the pair II twin B replica II, although the probabilistic calculation only took place on a total of 55 sites, the prediction was made correctly.

As already discussed, we then noticed that in couple I twin B replicate II we obtain the only case in which an entire dataset shows a completely wrong prediction. This strongly discordant result in terms of prediction, as already hypnotized, led us to think that during the processing of this sample there may have been a problem of mis-assignment of the label. Unfortunately, however, we have not been able to verify and confirm this hypothesis and for this reason, we have preferred to continue to consider the dataset as belonging to the twin b of the couple I.

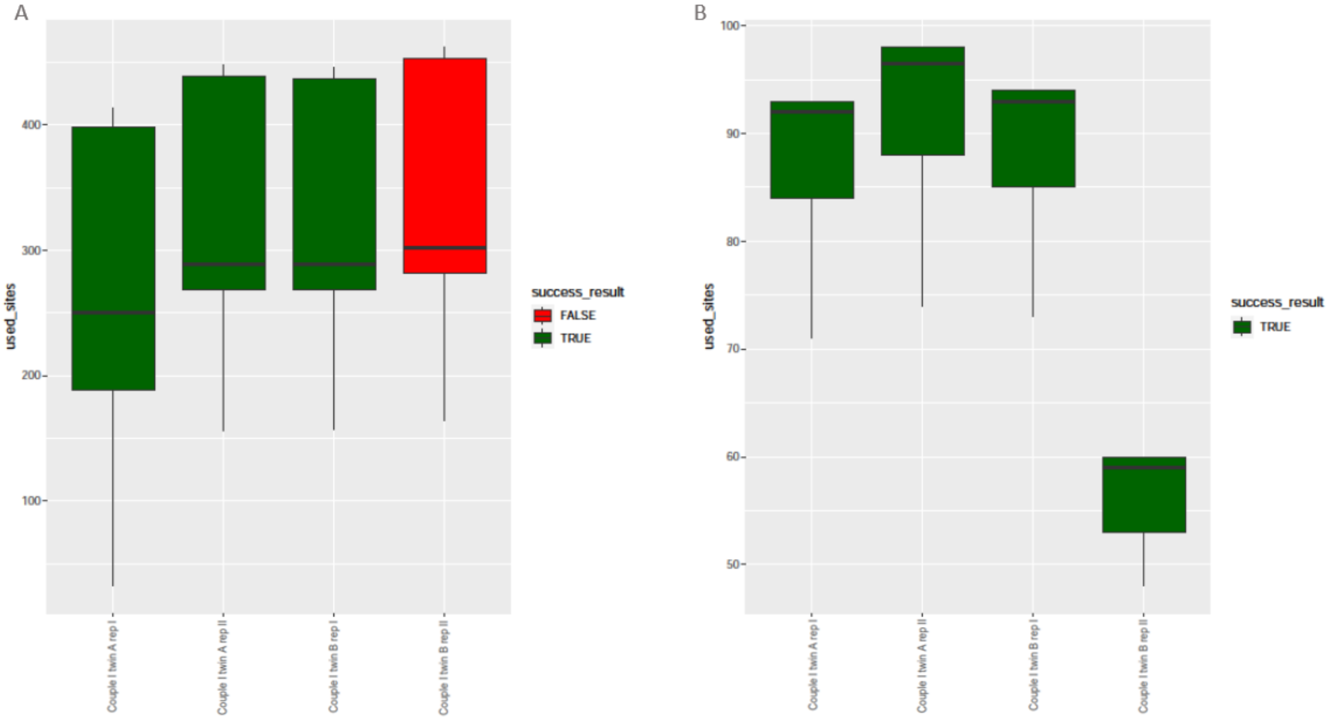


Figure 46 Result of the prediction compared to the number of sites used in the binomial probability calculation. Each different boxplot is represented in green if the obtained prediction was correct and in red if not. In couple I we obtain that twin B replicate II was the only case in which the prediction was completely incorrect. We were hypnotized that some sampling or analysis bias could have occurred during the experiments. For what that concern the number of sites used, we noticed a strong difference in couple I and couple II; despite the lower number of sites used for the computation for couple II the prediction was correct in all cases.

The second step of our analysis was conducted to evaluate the robustness of the prediction obtained by the 408 parameter sets. To test the reliability of the parameters used, we decided to see if these 408 sets of parameters recognized as the most accurately predicted parameter sets, were able to predict correctly even while performing a bootstrap analysis. The bootstrap analysis is a self-sustaining process that is based on the hypothesis that the sub-sample of size X is selected from the entire dataset and could be an estimation of the whole population.

These bootstrap methods help us to generate a set of random resampling subsets without replacement from the total amount of dataset that we previously used.

In a bootstrap validation, each group of size X is generated by a repeated random choice of a certain number of objects from the original dataset. To do that we implement into our script model the random module found in Python. This module helped us to randomly pick elements from our test table without repeating elements and returning a list of unique items chosen randomly from the tables.

For the bootstrap analysis computed in our mode, we decide to use 50% of the test table sites dataset for N=100 times. The bootstrap confidence obtained has been reported in the range from 0% to 100%.

The bootstrap function was integrated into the python script as summarised below:

import random (needed to perform the resampling)

```
1. Compute random sampling
    1.1 for perm in range(nperm)
2. Load randomly sampled subset
3. Calculate prediction
4. Count the number of times each twin was predicted
5. Calculate the most frequent prediction
6. Store in a file
```

The required argument for these functions is:

--bootstrap' – to activate the bootstrap resampling

--bootstrap-fraction'- to choose the fraction of resampling to use

--bootstrap-resample-n – to set the number of iterations to compute (default=100)

13.1 Parameters variation across the bootstrap analysis

As previously done during the calculation of the prediction, also during the bootstrap analysis we checked what was the trend of the parameters used.

As mentioned in the paragraph *Binomial probability customization based on our dataset*, we implement directly into the computation different values for reference table coverage, test table coverage, p-values, and different methylation levels. The trends of these parameters studied about the bootstrap confidence obtained are shown in **figure 47**.

To perform this study, we considered one single parameter at a time while all the rest of the parameters are fixed at a reference value, in order not to allow the parameter taken into consideration to change as the latter vary. The parameters that we have considered as fixed reference parameters have been chosen in a completely arbitrary way.

We have chosen the following as reference values (reference coverage = 20, test coverage =6, p-value = 0.01, methylation difference = 25).

In panel A of figure 47, the reference coverage was studied compared to the bootstrap % obtained. Boxplots for this first panel showed that each different value of reference coverage stays constant ranging from 75-90% of confidence. Panel B of reference coverage shows the density variation in all the values chosen. As already clear in the boxplot graph, there seems to be no difference among the proposed values, showing that in the case of the coverage of the reference tables, there would not seem to be a value for which the prediction becomes more correct and reliable in bootstrap analysis than the others.

During the reference table setting up (described in *Finding differentially methylated sites in Reference samples using Methylkit*), the function *CalculateDiffMeth* of MethylKit only sampled those cytosines that appear as differently methylated between the twins, excluding a priori those ambiguous single methylated sites. This could mean that the ratio between methylated and unmethylated cytosines for each identified site still is constant as coverage increases, always supplying the same information and not affecting either positively or negatively the result of the prediction.

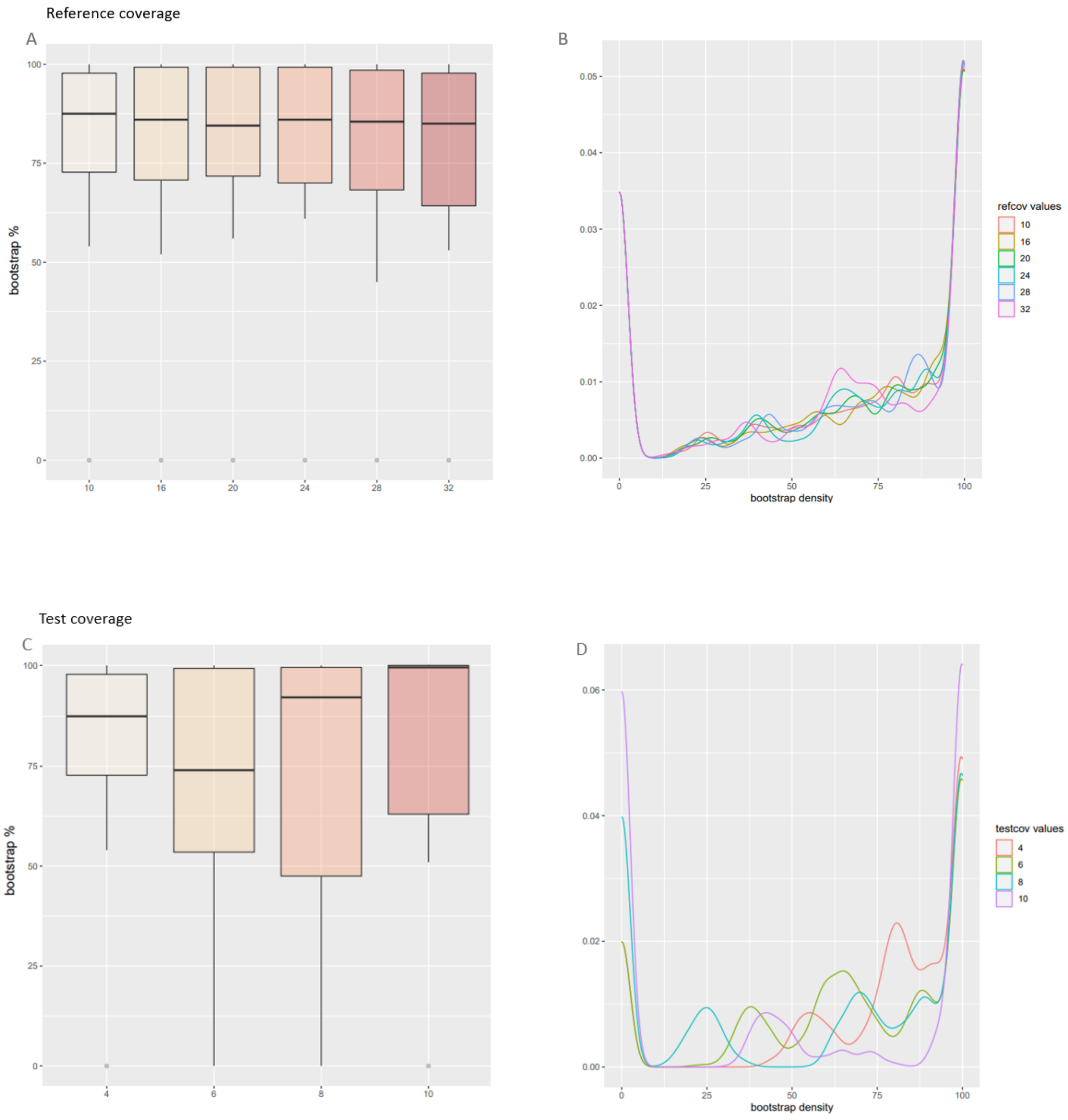
In panel C the test coverage variation across the bootstrap confidence evaluation is represented. We noticed that in this case, the bootstrap confidence had the highest confidence when the highest test coverage value was used (coverage equal to 10). This result is even more clear in panel C, where the bootstrap density is maximized when the parameter taken into consideration is the one equal to 10, which is the highest value we assigned to the coverage of the test tables.

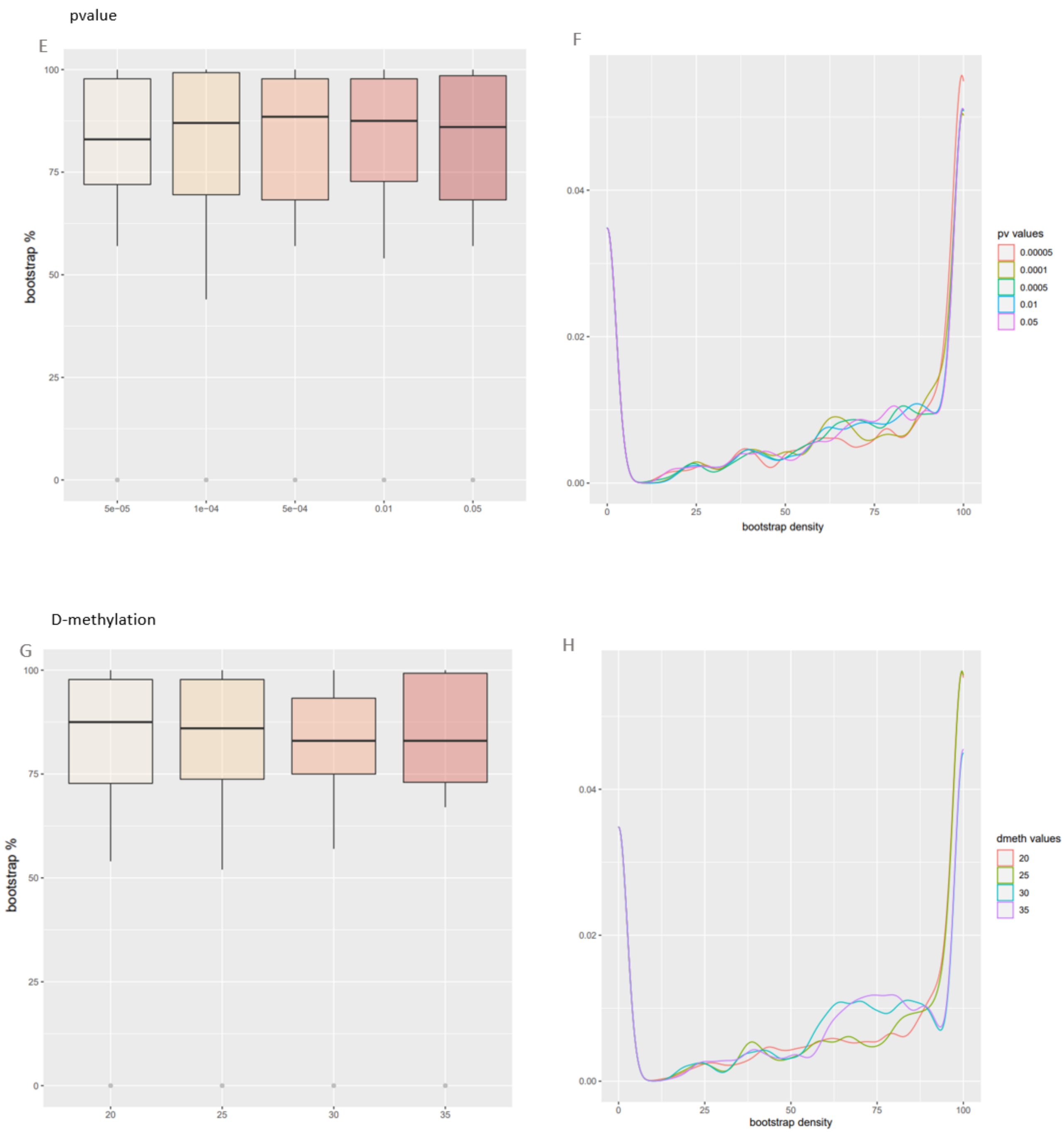
The fact that the robustness of the system (bootstrap analysis), turns out to be greater as the value of the test sampling increases (test coverage), shows us how one of the strongest limitations in this method was having a reduced sampling on the test tables. Increasing the coverage of the test tables would probably have drastically increased the robustness of the predictions obtained for each dataset under consideration.

We then studied the variation of the assigned values of significance. Panel E has displayed the trend of p-values used in the analysis. For p-values, the trend would seem to favour higher stringency values than looser parameters. In the boxplot of panel E, as well as in the graphical representation of the density of the parameters, the values that would seem to maximize the call are those of 0.0005, 0.0001, and 0.00005. Although these parameters allow us to consider those sites having a higher level of significance, it must always be considered that considering only and exclusively the sites having this high stringency, would entail a drastic reduction of the sites examined in the sampling.

Lastly, in panel G the difference in methylation variation is shown compared to the bootstrap confidence %. As expected, keeping the delta methylation parameter looser allows us to obtain higher bootstrap robustness values (for 20% and 25% close to 80%).

This result agrees with what was found during the descriptive analysis performed with MethylKit. In fact, during the descriptive analysis, keeping the delta methylation levels lower, allowed us to obtain a greater number of sites in the estimate. Keeping this parameter laxer is given by the fact that, in the case of studies of homozygous twins, it is very difficult to find cases in which there are strongly different methylation sites, if only in the case in which the individuals examined in the analysis are young and are not affected by particular pathologies.




pvalue

D-methylation

Figure 47 Panel A-H Parameters variation during bootstrap computing.

As previously discussed, one of the strongest limitations of this first preliminary analysis was given by the fact that the sampling of the test datasets was not extremely deep. Some parameters, such as the coverage test, showed how they drastically influenced the robustness of the prediction (bootstrap analysis), underlining that, as the value of this parameter increased, the validity of the prediction calls would also increase. For other parameters, such as the difference in methylation, it would seem the opposite is true. Considering higher methylation percentages (30%, 35%, 50%) leads to a drastic reduction in the robustness of the system since, as this parameter increases, the number of sites that can be used by the method drastically reduces.

13.2 Comparison between prediction and bootstrap calls

Following the analysis of the trends of the parameters, we want to verify if the bootstrap confirmed the prediction obtained from the calculation of the binomial probability and at what grade of confidence (represented by the bootstrap %).

In figure 48 each dataset is represented on the x-axis and the bootstrap % is reported on the y-axis. Each dot reported above the boxplot represented one of the 408 trials performed in the prediction using a specific set of parameters. These dots are represented in green if the result of the prediction obtained by these parameters was correct, and in red if the result of the prediction using the parameters was not correct. The bootstrap value reported on the y-axis stands for the confidence value obtained in the bootstrap analysis for that particular set of parameters (ranging from 0% to 100%).

As already discussed in the paragraph *Calculation of the prediction accuracy and bootstrap analysis*, in almost all the datasets the prediction through the calculation of the binomial probability seems to predict correctly except in one particular case. Unfortunately, in couple I, replicate II of twin B could have been affected by a mis-assignment problem, resulting in almost all of the calls being wrong with a bootstrap analysis toward the robust recall of the other twin.

For pair I in both replicates of twin A, we obtained the highest bootstrap values (ranging from 95% to 100%), confirming the maximum reliability of the prediction even while performing a subsampling of the dataset. In couple II, although the prediction is always able to correctly assign the call to the right twin, the bootstrap values tend to be lower in their distribution and do not have such robustness as to be able to confirm with absolute certainty the correctness of that particular prediction, although this was the case in one replicate of twin A for the couple II in certain parameter sets.

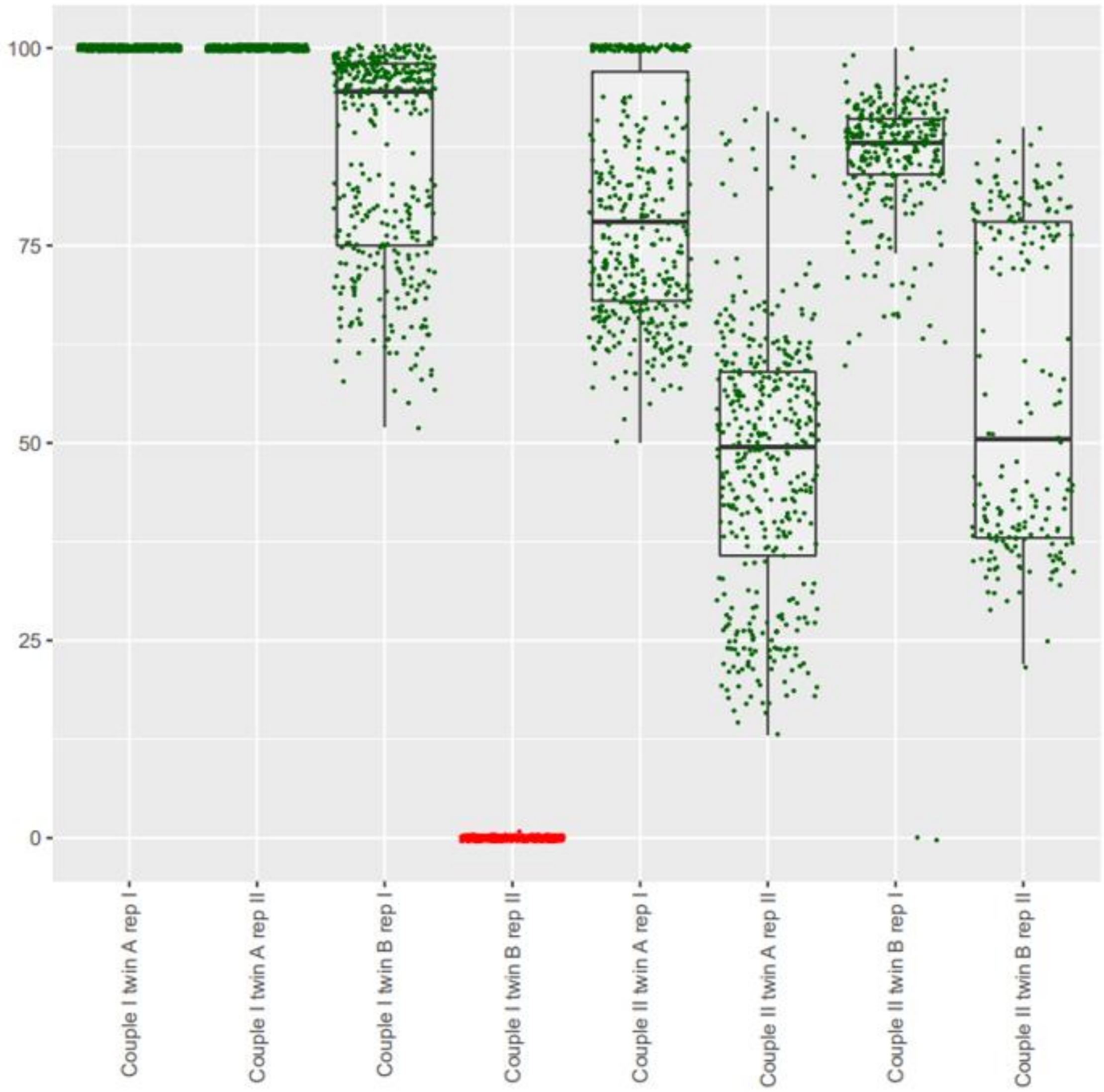


Figure 48 Summary graph of the totality of predictions made with the relative confidence values (bootstrap calls). Dots represented in green denote a good prediction call, meaning that the right twin was predicted, and red dots denote the wrong twin prediction. Every single dot is plotted on the y-axis according to the bootstrap value obtained for that particular set of parameters.

In couple I both replicates of twin A showed the highest robustness in the prediction, with bootstrap values of 100%. Twin B of couple I in replicate I, showed an average call confidence value of 88% (considering all the bootstrap confidence for every single set of parameters used). As previously discussed, twin B replicate II of couple I is probably the result of an error that occurred during sampling or mislabelling analysis.

In couple II replicate I of twin A we obtained almost 150 sets of parameters having 100% prediction reliability (bootstrap robustness values ranging from 95% to 100%), while the rest population of parameter sets ranged from 50% to 95% of bootstrap recall. In the other three dataset tests examined (couple II twin A replicate II, twin B replicate I, and replicate II), the prediction reliability of the parameter sets used reached values above 95% in a few cases, proving a lower level of reliability.

Taking into consideration all the results obtained, we then tried to find which were the sets of parameters in the totality of those named, which could be considered, strictly about the reliability obtained, those that gave a more correct and robust prediction.

Plotting every single set of parameters examined with their bootstrap value, we were able to narrow to some limited groups of parameters.

As already discussed, the subsets of found parameters that were shown to be the most exact in the prediction were the ones that had the highest value of test coverage compared to the less restricted values of methylation difference. In figure 49 are represented just 6 different sets are used that show the most correct calls. In the figure, green dots are the result of the prediction (green if it is correct and red if not), positioned at the relative bootstrap %.

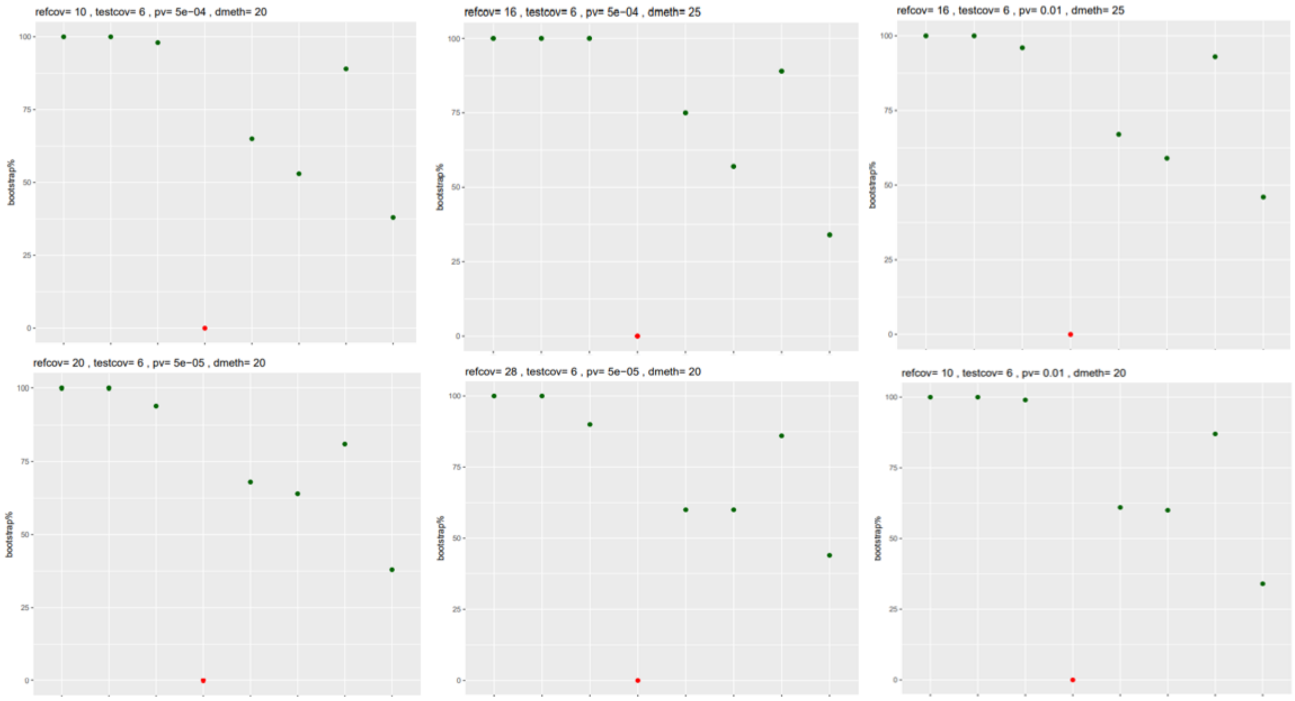


Figure 49 6 different sets of parameters show the best prediction results to the most exact calls (represented by the bootstrap % on the y-axis). Green dots stand for if the dataset is correctly reported to the right twin of membership, red dots are incorrect prediction and wrong twin membership. (In the figure are represented just a few examples of the best performing parameters used).

Unfortunately, as previously discussed, the accuracy of the prediction remained variable in couple II never reaching reliability equal to that found in couple I. Likely, this is strictly due to the number of sites used in the prediction for couple II (around a hundred, where couple I had a sampling of 300-400 sites used).

Even more in detail, the dataset showing the lowest % bootstrap value, and therefore the most variable called robustness, is the replicate B twin II of the second pair II. For this particular case, the number of sites on which the prediction was carried out was less than 100 (just 55 sites used).

It is therefore probable that during the bootstrap analysis, which involved a further subsampling of the dataset to evaluate the reliability, in many cases the number on which to perform the test was extremely small.

This result is strongly supported by the fact that, as shown in figure 50, if the number of test coverage increases, for some datasets the accuracy undergoes a drastic decrease which in some cases leads to a bootstrap value close to 0%, meaning that no value of robustness was assigned in not even one of the 100 iterations performed, or rather that it was not possible to assign any accuracy value. As can be seen from the image below, this situation was found in twin B of pair II in both replicas, in which in both datasets the number of sites named and used for the analysis was below 100.

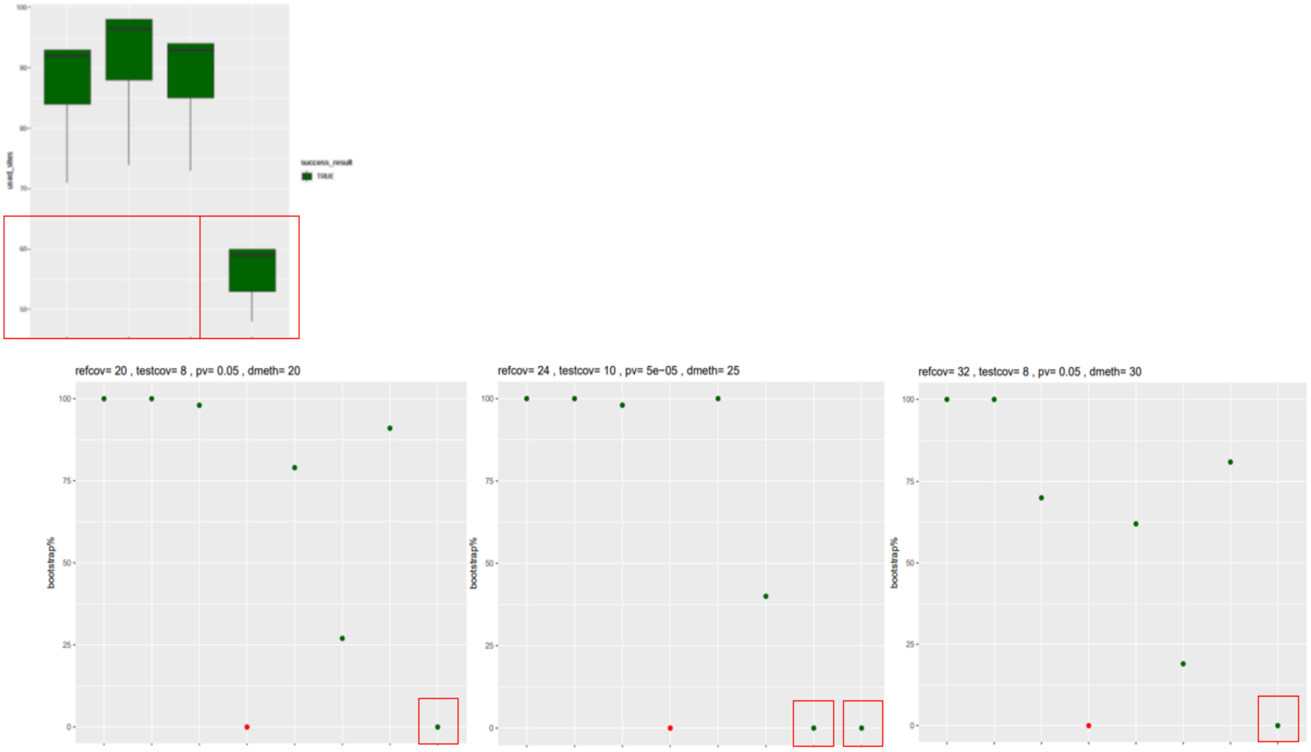


Figure 50 Boxplot in the upper right corner shows the couple II numbers of used sites in the prediction that was drastically lower than in couple I. Graphical representations of some sets of parameters show how increasing the test coverage threshold results in the loss of robustness calculation for some datasets. This is strongly correlated with the number of sites used in the prediction as clear in twin B replicate II.

14 Discussion

The goal of perfecting a protocol to distinguish monozygotic twins in forensic cases remains, up to now, an unsolved problem. Monozygotic twins are considered identical because they share the same DNA sequences and hence typically the same standard forensic DNA profiles. This result is a huge bias for forensic scientists that tried to find, based on the standard use of SNPs, the guilty twin in crime scenes.

Moreover, the low quality and -quantity of DNA recovered from crime scene traces is another limiting factor in the progress of forensic genetics and epigenetics, directly affecting not only the field of monozygotic twins' discrimination but also all the other forensic investigation types.

To overcome the problem of the identical genetic composition of monozygotic twins, nowadays, forensic scientists are trying to use epigenetics, and in particular DNA methylation, as one of the possible usable methods to create a standardized protocol of identical twins discrimination. During these last years, DNA methylation got attention on different forensic topics. Nowadays DNA methylation is used to perform sex determination, to find the tissue or the cell-type source of DNA evidence at crime scenes, and to estimate an individual's age. Apart from these topics, recently, DNA methylation is also used to differentiate between monozygotic twins.

For what that concern the sample quantity availability and quality in forensic cases that involved monozygotic twins, up to now the usable approaches are mainly divided into two, a more genome-wide screening approach or a targeted method. The two approaches change according to the scientific question and the context in which they are inserted.

Between the proposed and used approaches, in 2015, Stewart and collaborators firstly used the PCR-high resolution melting curve to observe differences in the methylation pattern of MZ twins in forensic cases (Stewart et al., 2015). This technique, which could be considered a targeted approach, was also used in 2018, by Marqueta-Gracia and collaborators. In this second work, scientists propose the use of the HRMA to discriminate between identical twins involved in crime scenes (Marqueta-Gracia et al., 2018). Scientists found three regions of the genes ITGA2B, ASPA, and ZIC5 that could enable discrimination in 44.4% of pairs, and, while this technique results to be one of the most cost-effective and easy to set up, it shows different limitations. The biggest problem of this technique was represented by the fact that in a relatively young MZ couple, the melting temperatures value, obtained from the HRMA, could not be discriminatory and so not be used in a court of law. Moreover, the methylation profile coming from different tissue types is one of the other biggest problems that forensic scientists need to overcome. Physiologically each tissue shows its pattern of methylation and could not be compared between different tissue sources. The tissue-to-tissue variations, as admitted by the authors themselves, were also one of the limitations encountered by Vidaki and collaborators in 2018, when they investigated the epigenetic discrimination of MZ twins using as sources of tissues, buccal swabs, saliva, and also cigarette butts (Vidaki et al., 2018).

Differently, observations derived from genome-wide approaches have identified potential candidate markers that could be capable of discriminating identical twins using different sources, such as blood (Saxonov et al., 2006) or buccal cells (Kaminsky et al., 2009) however, these studies still need implementations to become standardized and applicable in forensic cases investigation. In 2018 Vidaki and collaborators, used microarrays to find in reference-type buccal DNA 25 sites of discrimination showing >0.5 twin-to-twin differences. They also used the MethyLight quantitative PCR (qPCR) on 22 of the selected identified sites in the trace-type DNA, revealing in saliva DNA that six of the identified sites (27.3%) shows >0.1 twin-to-twin differences, seven (31.8%) shows smaller (<0.1) but robustly detected differences, whereas for nine (40.9%) the differences results to be in the opposite direction if compared to the microarray data; for cigarette butt DNA, results were 50%, 22.7%, and 27.3%, respectively. The identified discrepancy they obtained was properly justified and explained by the authors, which lead back to the different obtained results in the method-to-method variation and samples variation used between reference type DNA and trace-type DNA.

Recently, in 2021, Benjamin Planterose Jiménez and collaborators firstly used a different method to approach the MZ twins discrimination problem in forensic cases.

In this work, the authors tried to isolate stochastic inter-individual epigenetic variation, not related to epigenetic drift or genetic influence (Planterose Jiménez et al., 2021). To achieve this goal, they used MZ twins, unrelated individuals, with technical replicates obtained from whole blood, adipose tissue, and post-mortem tissues to find methylated sites that cannot be explained by epigenetic drift and/or measurement error. For what that concern the techniques, the authors used different online available dataset derived from two main sources; the Illumina Infinium HumanMethylation450K Beadchip array (450K), which covers > 450,000 CpG sites, and the whole genome bisulfite sequencing (WGBS). Scientists found 333 differentially methylated sites displayed similarly large methylation variation between monozygotic co-twins and unrelated individuals.

Considering all the earlier mentioned works, we tried to be as match as possible faithful to the reality of the crime scene, in which there are mainly two different types of sources, the reference sample, and the trace type sample. Reference samples referred to the source of samples that could directly be requested by the court of law during investigations while trace-type samples refer to DNA sources found on crime scenes, and as already discussed, are characterized by inferior quality and low quantity. Considering these assumptions, we decide to set up in all of our experiments two starting material concentrations; the 10 ng, which mimics the reference type DNA, and the 1 ng sample, which mimics the trace-type DNA.

We then tried to find which approach could be the most cost-effective and at the same time the most high-throughput screening to find the larger number of differentially methylated sites. The RRBS technique was firstly used for this purpose but, unfortunately, it was not effective in our screening because of the deprived areas of screening.

The RRBS technique perfectly screens all the CpG-rich DNA areas, such as CpGs island, but in our case does not fit properly with the goal we had. During these first sets of experiments with RRBS we noticed that many of the differentially methylated sites found between the two twins were localized at the end of the reads obtained by the sequencing.

When we then computed the statistical test all the sites found in these particular areas of the reads often led back to the wrong twin, unlike those cases in which the methylation site was found within the reads. We, therefore, wondered if this could not be a bias related to the conversion step with bisulfite, which perhaps lost performance when the conversion site was found in the terminal section of the read.

Furthermore, the screening areas we obtained from the analysis with the RRBS technology were very limited, and, taking in particular account the work of Planterose Jiménez et al., we aimed to investigate as much as possible the totality of the cytosine in the genome, trying to found stochastic single nucleotide variation that could be discriminatory between identical individuals.

Based on that, we decided to move to the whole genome approach to better screen most of the cytosines in the genome and, to avoid method-to-methods discrepancy, we prepared for both the reference type and trace-type DNA WGBS libraries, with the only difference of starting material concentration. We used the 10-ng starting material library to set up the reference type DNA tables and the 1-ng starting material library to set up the test table.

The main goal was to lead back the 8-test table (4 tests each couple) obtained to the right twin based on the calculation of the binomial probability and assigning a final cumulative probability. To do that, we parsed various parameters related to reference coverage, test coverage, p-value, and methylation difference within the script.

We did not find a unique set able to predict correctly for all the test tables we had. On the 600 sets of parameters used, 408 showed accuracy in the prediction of 87.5%, calculated as the correct predictions, out of the total predictions (wrong and correct), meaning that these sets were able to lead back to the right test table 7 times above 8. Unfortunately, in one particular case, in couple I twin B replicate II, the prediction result was strongly incorrect. Due to the particular bootstrap outcome that this dataset presented we assumed that a DNA sample exchange may have occurred during one of the steps described in the materials and methods, leading to the processing of Twin A instead of twin B material. In any case, we have decided to include this dataset as clear evidence of how the bootstrap analysis can be an agnostic indicator of the reliability of the analysis.

Despite the case of the twin B replicate II of the first pair, for all the other datasets we were able to correctly predict which twin they belonged to. We, therefore, decided to test how robust each computed set of parameters was in the prediction, calculating bootstrap intervals.

The bootstrap analysis showed us a high level of reliability in the case of twin A of the first couple, assigning robustness equal to 100% to all the parameter sets for replicate I and replicate II. For most of the parameters used in the dataset of the twin B replicate I of the first pair, the bootstrap values were shown to be above 95%, affirming a high level of robustness of predictions.

In particular, the highest reliability (a bootstrap level between 95% and 100%), was highlighted for those groups of parameters that had 6-8-10 test coverage values and methylation difference values equal to only 20% and 25%.

This result was completely following the analysis of the parameters we performed, discussed in the section *Parameters variation across the bootstrap analysis*. In this analysis, we saw that the test coverage values and the difference in methylation levels were the most significant in figuring out the outcome. The bootstrap density confidence results to be maximized for the test coverage equal to 10 and delta methylation equal to 25% (results shown in figure 47).

It should also be emphasized that unlike the couple II, couple I presented most of the sites localized at the level of CpG islands or in proximity to the latter, at the level of the CpG shores, as discussed in the *"Distribution of differentially methylated single sites in Reference samples"* paragraph.

Moreover, the same evidence discussed above was also found for couple II. The maximum robustness values were found for high coverage values predetermined in the tests and for less stringent values of methylation difference.

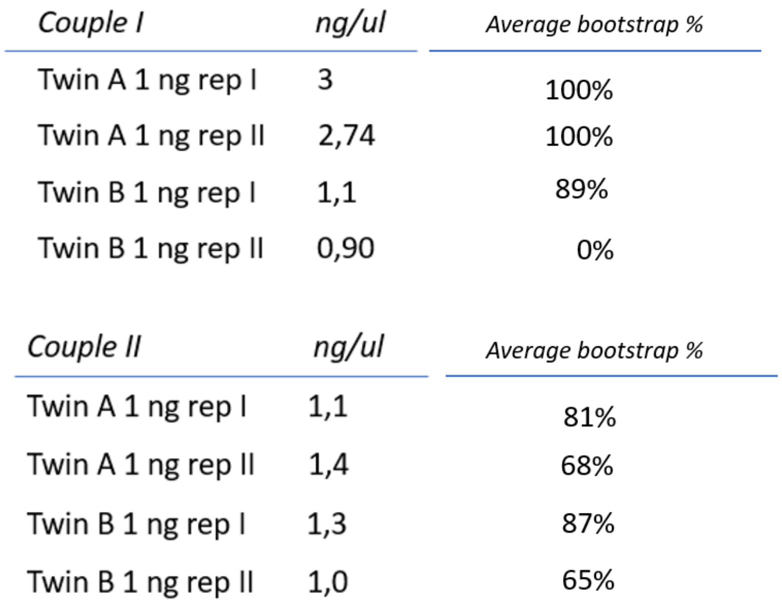
Unfortunately, in the case of the couple pair II the reliability of the system was more variable, obtaining a % of bootstrap lower than that seen for the couple pair I. More in detail, in about 150 sets of parameters for the twin A replicate I of the pair II the bootstrap defined robustness of the prediction equal to 100%, while both for the other parameters of the same dataset and for the other test tables considered, the robustness fell to reference values that fluctuated in a range that went from 95% down to even up to 60%, thus making the system drastically lose its robustness.

To understand what could have influenced so drastically the lowering of the confidence levels of the prediction in couple II, we realized that different from what happened in couple I, couple II was performing the prediction in each test table on a number of sites drastically inferior compared to couple I.

This observation has shown us how much, in this type of method, it is extremely necessary to obtain a sampling that could give a sufficient quantity of data that can become informative. The less data is produced by sampling the test, although this reduced sampling seems to be sufficient to correctly assign the call probability, the more difficult it is to assign a high robustness level since the individual sites on which to sample it are reduced.

Once we hypothesized this as one of the possible limitations present in the method, we, therefore, asked ourselves if the problem was upstream, and therefore attributable to the preparation of the samples taken into consideration.

We consider the data obtained from the buccal swab extraction and, in particular, what was clear to us is that for those datasets for which we were able to obtain a greater amount of post-fragmentation DNA, the downstream analysis was found to be better than the datasets for which the post-fragmentation amount was less (the values are shown in the table below).



Couple I	*ng/ul*	*Average bootstrap %*
Twin A 1 ng rep I	3	100%
Twin A 1 ng rep II	2,74	100%
Twin B 1 ng rep I	1,1	89%
Twin B 1 ng rep II	0,90	0%

Couple II	*ng/ul*	*Average bootstrap %*
Twin A 1 ng rep I	1,1	81%
Twin A 1 ng rep II	1,4	68%
Twin B 1 ng rep I	1,3	87%
Twin B 1 ng rep II	1,0	65%

Moreover, all the samples belonging to pair II, before reaching this final concentration, were concentrated by using a vacuum concentrator.

From these considerations, it is therefore fundamental for the development of the method we have introduced, to obtain a sufficient quantity of data right from the DNA extraction, since this sampling directly affects the downstream analysis.

15 Conclusion and future perspectives of the work

This thesis aimed to find one approach to use in forensic science to distinguish between identical twins. We obtain buccal swab samples from two different couple of twins and we then obtain libraries for NGS experiments. Considering the work from Vidaki and collaborators (Vidaki et al., 2018), we create for each couple test table, mimic the anonymous trace type DNA, and the reference table, corresponding to the DNA deposited by the suspects.

In the first trials of experiments, we use the RRBS techniques to try if through the CpG area screening alone it was possible to discriminate the identical couple. This technique proved to be non-informative in the case of methylation screening between identical individuals, due to the low percentage of areas of the genome studied by the methods.

Once we move to the WGBS approach, we realize that using a broad range of techniques gave us the possibility to focus not only on differently methylated areas previously described in the literature as involved in physiological or pathophysiological epigenetic phenomena but also on random methylation sites not dependent from these conditions.

We set up a statistical method called “Twin-pred” able to assess the binomial probability for every single site found by the WGBS and then compute a cumulative probability that will answer who is most likely to derive this test table from.

In a forensic analysis that has to assign guilt to one accused rather than another, the robustness and credibility of the final verdict must be as certain as possible. This is the reason behind our willingness to set, after performing the cumulative probability, a level of confidence and trust for the call. The bootstrap has provided us with the first approach towards a possible method that gives us a more or less high degree of security.

The Twin-pred predictor was designed to be able to help forensic scientists in those special cases in which two completely identical individuals are at the judge's bench to be indicted.

However, it is first necessary to confirm the robustness of the predictor system on a larger sampling dataset, as this has proved to be the main limitation of the method. If later and future analysis trials performed on larger datasets will show the same trend reported in our analysis on couple I, our forecasting method would lay the foundations for the development of a mathematical method that is correct and safe to use in those cases where not even the DNA sequence can help us.

16 References

Aapola U, Shibuya K, Scott HS, Ollila J, Vihinen M, Heino M, Shintani A, Kawasaki K, Minoshima S, Krohn K, Antonarakis SE, Shimizu N, Kudoh J, Peterson P. 2000. Isolation and Initial Characterization of a Novel Zinc Finger Gene, DNMT3L, on 21q22.3, Related to the Cytosine-5- Methyltransferase 3 Gene Family. *Genomics* **65**:293–298. doi:10.1006/geno.2000.6168

Achour M, Jacq X, Rondé P, Alhosin M, Charlot C, Chataigneau T, Jeanblanc M, Macaluso M, Giordano A, Hughes AD, Schini-Kerth VB, Bronner C. 2008. The interaction of the SRA domain of ICBP90 with a novel domain of DNMT1 is involved in the regulation of VEGF gene expression. *Oncogene* **27**:2187–2197. doi:10.1038/sj.onc.1210855

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**:R87. doi:10.1186/gb-2012-13-10-r87

Alaeddini R, Walsh SJ, Abbas A. 2010. Forensic implications of genetic analyses from degraded DNA—A review. *Forensic Science International: Genetics* **4**:148–157. doi:10.1016/j.fsigen.2009.09.007

Alghanim H, Antunes J, Silva DSBS, Alho CS, Balamurugan K, McCord B. 2017. Detection and evaluation of DNA methylation markers found at SCGN and KLF14 loci to estimate human age. *Forensic Science International: Genetics* **31**:81–88. doi:10.1016/j.fsigen.2017.07.011

Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**:9354. doi:10.1038/s41598-019-45839-z

Amorim A, Pereira L. 2005. Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Science International* **150**:17–21. doi:10.1016/j.forsciint.2004.06.018

Aran D, Toperoff G, Rosenberg M, Hellman A. 2011. Replication timing-related and gene body-specific methylation of active human genes. *Human Molecular Genetics* **20**:670–680. doi:10.1093/hmg/ddq513

Bell JT, Spector TD. 2011. A twin approach to unraveling epigenetics. *Trends in Genetics* **27**:116–125. doi:10.1016/j.tig.2010.12.005

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40**:91–99. doi:10.1016/0092-8674(85)90312-5

Boomsma D, Busjahn A, Peltonen L. 2002. Classical twin studies and beyond. *Nat Rev Genet* **3**:872–882. doi:10.1038/nrg932

Bourc’his D, Bestor TH. 2004. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**:96–99. doi:10.1038/nature02886

Burger J, Hummel S, Herrmann B, Henke W. 1999. DNA preservation: A microsatellite-DNA study on ancient skeletal remains. *Electrophoresis* **20**:1722–1728. doi:10.1002/(SICI)1522-2683(19990101)20:8<1722::AID-ELPS1722>3.0.CO;2-4

Castillo-Fernandez JE, Spector TD, Bell JT. 2014. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Med* **6**:60. doi:10.1186/s13073-014-0060-z

Cheng T-L, Qiu Z. 2014. MeCP2: multifaceted roles in gene regulation and neural development. *Neurosci Bull* **30**:601–609. doi:10.1007/s12264-014-1452-6

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE. 2008. Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning 12.

Condon J, Shaw JE, Luciano M, Kyvik KO, Martin NG, Duffy DL. 2008. A Study of Diabetes Mellitus Within a Large Sample of Australian Twins. *Twin Res Hum Genet* **11**:28–40. doi:10.1375/twin.11.1.28

Cortellino S, Xu J, Sannai M, Moore R, Caretti E, Cigliano A, Le Coz M, Devarajan K, Wessels A, Soprano D, Abramowitz LK, Bartolomei MS, Rambow F, Bassi MR, Bruno T, Fanciulli M, Renner C, Klein-Szanto AJ, Matsumoto Y, Kobi D, Davidson I, Alberti C, Larue L, Bellacosa A. 2011. Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. *Cell* **146**:67–79. doi:10.1016/j.cell.2011.06.020

Daniel JM. 2002. The p120ctn-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Research* **30**:2911–2919. doi:10.1093/nar/gkf398

Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. 2011. Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**:771–784. doi:10.2217/epi.11.105

Dodge JE, Ramsahoye BH, Wo ZG, Okano M, Li E. 2002. De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. *Gene* **289**:41–48. doi:10.1016/S0378-1119(02)00469-9

Dricu A, Oana S, Budiu R, Ola R, Elise D, Vlad A. 2012. Epigenetic Alteration of Receptor Tyrosine Kinases in Cancer In: Tatarinova T, editor. DNA Methylation - From Genomics to Technology. InTech. doi:10.5772/36012

Dupont C, Armant D, Brenner C. 2009. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Semin Reprod Med* **27**:351–357. doi:10.1055/s-0029-1237423

Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, Gehrke C. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucl Acids Res* **10**:2709–2721. doi:10.1093/nar/10.8.2709

Elhamamsy AR. 2017. Role of DNA methylation in imprinting disorders: an updated review. *J Assist Reprod Genet* **34**:549–562. doi:10.1007/s10815-017-0895-5

Espada J, Esteller M. 2010. DNA methylation and the functional organization of the nuclear compartment. *Seminars in Cell & Developmental Biology* **21**:238–246. doi:10.1016/j.semcdb.2009.10.006

Forat S, Huettel B, Reinhardt R, Fimmers R, Haidl G, Denschlag D, Olek K. 2016. Methylation Markers for the Identification of Body Fluids and Tissues from Forensic Trace Evidence. *PLoS ONE* **11**:e0147973. doi:10.1371/journal.pone.0147973

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML. n.d. Epigenetic differences arise during the lifetime of monozygotic twins. *MEDICAL SCIENCES* 6.

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL. 1992. A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences* **89**:1827–1831. doi:10.1073/pnas.89.5.1827

Gaudet F, Rideout WM, Meissner A, Dausman J, Leonhardt H, Jaenisch R. 2004. Dnmt1 Expression in Pre- and Postimplantation Embryogenesis and the Maintenance of IAP Silencing. *Mol Cell Biol* **24**:1640–1648. doi:10.1128/MCB.24.4.1640-1648.2004

Gharizadeh B, Kalantari M, Garcia CA, Johansson B, Nyrén P. 2001. Typing of Human Papillomavirus by Pyrosequencing. *Lab Invest* **81**:673–679. doi:10.1038/labinvest.3780276

Gibney ER, Nolan CM. 2010. Epigenetics and gene expression. *Heredity* **105**:4–13. doi:10.1038/hdy.2010.54

Goto K, Numata M, Komura J-I, Ono T, Bestor TH, Kondo H. 1994. Expression of DNA methyltransferase gene in mature and immature neurons as well as proliferating cells in mice. *Differentiation* **56**:39–44. doi:10.1046/j.1432-0436.1994.56120039.x

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc* **6**:468–481. doi:10.1038/nprot.2010.190

Guo JU, Su Y, Zhong C, Ming G, Song H. 2011. Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell* **145**:423–434. doi:10.1016/j.cell.2011.03.022

Haig D. 2012. Commentary: The epidemiology of epigenetics: Figure 1. *Int J Epidemiol* **41**:13–16. doi:10.1093/ije/dyr183

Hall JG. 2003. DEVELOPMENTAL BIOLOGY IV. *THE LANCET* **362**:9.

Hashimoto H, Horton JR, Zhang X, Cheng X. 2009. UHRF1, a modular multi-domain protein, regulates replication-coupled crosstalk between DNA methylation and histone modifications. *Epigenetics* **4**:8–14. doi:10.4161/epi.4.1.7370

He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L, Sun Y, Li X, Dai Q, Song C-X, Zhang K, He C, Xu G-L. 2011. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA **333**:6.

Hellman A, Chess A. 2007. Gene Body-Specific Methylation on the Active X Chromosome. *Science* **315**:1141–1143. doi:10.1126/science.1136352

Hermann A, Goyal R, Jeltsch A. 2004. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. *Journal of Biological Chemistry* **279**:48350–48359. doi:10.1074/jbc.M403427200

Holliday R, Pugh J. 1975. DNA modification mechanisms and gene activity during development. *Science* **187**:226–232. doi:10.1126/science.1111098

Holtkötter H, Schwender K, Wiegand P, Pfeiffer H, Vennemann M. 2018. Marker evaluation for differentiation of blood and menstrual fluid by methylation-sensitive SNaPshot analysis. *Int J Legal Med* **132**:387–395. doi:10.1007/s00414-018-1770-3

Hong SR, Jung S-E, Lee EH, Shin K-J, Yang WI, Lee HY. 2017. DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. *Forensic Science International: Genetics* **29**:118–125. doi:10.1016/j.fsigen.2017.04.006

Hotchkiss RD. 1948. THE QUANTITATIVE SEPARATION OF PURINES, PYRIMIDINES, AND NUCLEOSIDES BY PAPER CHROMATOGRAPHY. *Journal of Biological Chemistry* **175**:315–332. doi:10.1016/S0021-9258(18)57261-6

Hutnick LK, Huang X, Loo T-C, Ma Z, Fan G. 2010. Repression of Retrotransposal Elements in Mouse Embryonic Stem Cells Is Primarily Mediated by a DNA

Methylation-independent Mechanism. *Journal of Biological Chemistry* **285**:21082–21091. doi:10.1074/jbc.M110.125674

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *PLoS Genet* **6**:e1001134. doi:10.1371/journal.pgen.1001134

Ito S, D'Alessio AC, Taranova OV, Hong K, Zhang Y. 2012. Role of Tet proteins in 5mC to 5hmC conversion, ES cell self- renewal, and ICM specification 15.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333**:1300–1303. doi:10.1126/science.1210597

Kaminsky ZA, Tang T, Wang S-C, Ptak C, Oh GHT, Wong AHC, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, McRae AF, Visscher PM, Montgomery GW, Gottesman II, Martin NG, Petronis A. 2009. DNA methylation profiles in monozygotic and dizygotic twins. *Nat Genet* **41**:240–245. doi:10.1038/ng.286

Kantor B, Kaufman Y, Makedonski K, Razin A, Shemer R. 2004. Establishing the epigenetic status of the Prader–Willi/Angelman imprinting center in the gametes and embryo. *Human Molecular Genetics* **13**:2767–2779. doi:10.1093/hmg/ddh290

Kimura H, Shiota K. 2003. Methyl-CpG-binding Protein, MeCP2, Is a Target Molecule for Maintenance DNA Methyltransferase, Dnmt1. *Journal of Biological Chemistry* **278**:4806–4812. doi:10.1074/jbc.M209923200

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**:1571–1572. doi:10.1093/bioinformatics/btr167

Kurt Benirscheke, M.D, and Chung K. Kim, M.D. 1973. Multiple pregnancy.

Laird PW. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11**:191–203. doi:10.1038/nrg2732

Li E, Bestor TH, Jaenisch R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**:915–926. doi:10.1016/0092-8674(92)90611-F

Li KK, Luo L-F, Shen Y, Xu J, Chen Z, Chen S-J. 2013. DNA Methyltransferases in Hematologic Malignancies. *Seminars in Hematology* **50**:48–60. doi:10.1053/j.seminhematol.2013.01.005

Li LX, Agborbesong E, Zhang L, Li X. 2019. Investigation of epigenetics in kidney cell biologyMethods in Cell Biology. Elsevier. pp. 255–278. doi:10.1016/bs.mcb.2019.04.015

Lin Y-C, Tsai L-C, Lee JC-I, Su C-W, Tzen JT-C, Linacre A, Hsieh H-M. 2016. Novel identification of biofluids using a multiplex methylation sensitive restriction enzyme-PCR system. *Forensic Science International: Genetics* **25**:157–165. doi:10.1016/j.fsigen.2016.08.011

Linda Van Speybroeck. 2002. From Epigenesis to Epigenetics.

Lo P-K, Watanabe H, Cheng P-C, Teo WW, Liang X, Argani P, Lee JS, Sukumar S. 2009. MethySYBR, a Novel Quantitative PCR Assay for the Dual Analysis of DNA Methylation and CpG Methylation Density. *The Journal of Molecular Diagnostics* **11**:400–414. doi:10.2353/jmoldx.2009.080126

Lopes EC, Valls E, Figueroa ME, Mazur A, Meng F-G, Chiosis G, Laird PW, Schreiber-Agus N, Greally JM, Prokhortchouk E, Melnick A. 2008. Kaiso Contributes to DNA

Methylation-Dependent Silencing of Tumor Suppressor Genes in Colon Cancer Cell Lines. *Cancer Res* **68**:7258–7263. doi:10.1158/0008-5472.CAN-08-0344

Marqueta-Gracia JJ, Álvarez-Álvarez M, Baeta M, Palencia-Madrid L, Prieto-Fernández E, Ordoñana JR, de Pancorbo MM. 2018. Differentially methylated CpG regions analyzed by PCR-high resolution melting for monozygotic twin pair discrimination. *Forensic Science International: Genetics* **37**:e1–e5. doi:10.1016/j.fsigen.2018.08.013

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, Turecki G, Delaney A, Varhol R, Thiessen N, Shchors K, Heine VM, Rowitch DH, Xing X, Fiore C, Schillebeeckx M, Jones SJM, Haussler D, Marra MA, Hirst M, Wang T, Costello JF. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**:253–257. doi:10.1038/nature09165

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**:766–770. doi:10.1038/nature07107

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schübeler D. 2008. Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. *Molecular Cell* **30**:755–766. doi:10.1016/j.molcel.2008.05.007

Moore LD, Le T, Fan G. 2013. DNA Methylation and Its Basic Function. *Neuropsychopharmacol* **38**:23–38. doi:10.1038/npp.2012.112

Mortusewicz O, Schermelleh L, Walter J, Cardoso MC, Leonhardt H, Pardee AB. 2005. Recruitment of DNA Methyltransferase I to DNA Repair Sites. *Proceedings of the National Academy of Sciences of the United States of America* **102**:8905–8909.

Muto M, Kanari Y, Kubo E, Takabe T, Kurihara T, Fujimori A, Tatsumi K. 2002. Targeted Disruption of Np95 Gene Renders Murine Embryonic Stem Cells Hypersensitive to DNA Damaging Agents and DNA Replication Blocks. *Journal of Biological Chemistry* **277**:34549–34555. doi:10.1074/jbc.M205189200

Nan X, Meehan RR, Bird A. 1993. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucl Acids Res* **21**:4886–4892. doi:10.1093/nar/21.21.4886

Noble D. 2015. Conrad Waddington and the origin of epigenetics. *Journal of Experimental Biology* **218**:816–818. doi:10.1242/jeb.120071

Okano M, Bell DW, Haber DA, Li E. 1999. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99**:247–257. doi:10.1016/S0092-8674(00)81656-6

Park J-L, Woo K-M, Kim S-Y, Kim YS. 2017. Potential forensic application of DNA methylation to identify individuals in a pair of monozygotic twins. *Forensic Science International: Genetics Supplement Series* **6**:e456–e457. doi:10.1016/j.fsigss.2017.09.177

Planterose Jiménez B, Liu F, Caliebe A, Montiel González D, Bell JT, Kayser M, Vidaki A. 2021. Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. *Genome Biol* **22**:18. doi:10.1186/s13059-020-02223-9

Prokhortchouk A. 2001. The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes & Development* **15**:1613–1618. doi:10.1101/gad.198501

Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, Andrews TD, Howe KL, Otto T, Olek A, Fischer J, Gut IG, Berlin K, Beck S. 2004. DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project. *PLoS Biol* **2**:e405. doi:10.1371/journal.pbio.0020405

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences* **97**:5237–5242. doi:10.1073/pnas.97.10.5237

Rando OJ, Verstrepen KJ. 2007. Timescales of Genetic and Epigenetic Inheritance. *Cell* **128**:655–668. doi:10.1016/j.cell.2007.01.023

Rivera CM, Ren B. 2013. Mapping Human Epigenomes. *Cell* **155**:39–55. doi:10.1016/j.cell.2013.09.011

Roessler J, Ammerpohl O, Gutwein J, Hasemeier B, Anwar SL, Kreipe H, Lehmann U. 2012. Quantitative cross-validation and content analysis of the 450k DNA methylation array from Illumina, Inc. *BMC Res Notes* **5**:210. doi:10.1186/1756-0500-5-210

Rohland N, Glocke I, Aximu-Petri A, Meyer M. 2018. Extraction of highly degraded DNA from ancient bones, teeth and sediments for high-throughput sequencing. *Nat Protoc* **13**:2447–2461. doi:10.1038/s41596-018-0050-5

Saxonov S, Berg P, Brutlag DL. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences* **103**:1412–1417. doi:10.1073/pnas.0510310103

Schulz WA, Steinhoff C, Florl AR. 2006. Methylation of Endogenous Human Retroelements in Health and Disease In: Doerfler W, Böhm P, editors. DNA Methylation: Development, Genetic Disease and Cancer, Current Topics in Microbiology and Immunology. Springer Berlin Heidelberg. pp. 211–250. doi:10.1007/3-540-31181-5_11

Stewart L, Evans N, Bexon KJ, van der Meer DJ, Williams GA. 2015. Differentiating between monozygotic twins through DNA methylation-specific high-resolution melt curve analysis. *Analytical Biochemistry* **476**:36–39. doi:10.1016/j.ab.2015.02.001

Tost J, editor. 2018. DNA Methylation Protocols, Methods in Molecular Biology. New York, NY: Springer New York. doi:10.1007/978-1-4939-7481-8

Tost J, Gut IG. 2007. DNA methylation analysis by pyrosequencing. *Nat Protoc* **2**:2265–2275. doi:10.1038/nprot.2007.314

van Dongen J, Gordon SD, McRae AF, Odintsova VV, Mbarek H, Breeze CE, Sugden K, Lundgren S, Castillo-Fernandez JE, Hannon E, Moffitt TE, Hagenbeek FA, van Beijsterveldt CEM, Jan Hottenga J, Tsai P-C, BIOS Consortium, van Dongen J, Hottenga J-J, Genetics of DNA Methylation Consortium, McRae AF, Sugden K, Castillo-Fernandez JE, Hannon E, Moffitt TE, Hottenga J-J, de Geus EJC, Spector Timothy D., Min JL, Hemani G, Ehli EA, Paul F, Stern CD, Heijmans BT, Slagboom PE, Daxinger L, van der Maarel SM, de Geus EJC, Willemsen G, Montgomery GW, Reversade B, Ollikainen M, Kaprio J, Spector Tim D., Bell JT, Mill J, Caspi A, Martin NG, Boomsma DI. 2021. Identical twins carry a persistent epigenetic signature of early genome programming. *Nat Commun* **12**:5618. doi:10.1038/s41467-021-25583-7

Vidaki A, Ballard D, Aliferi A, Miller TH, Barron LP, Syndercombe Court D. 2017a. DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing. *Forensic Science International: Genetics* **28**:225–236. doi:10.1016/j.fsigen.2017.02.009

Vidaki A, Daniel B, Court DS. 2013. Forensic DNA methylation profiling—Potential opportunities and challenges. *Forensic Science International: Genetics* **7**:499–507. doi:10.1016/j.fsigen.2013.05.004

Vidaki A, Díez López C, Carnero-Montoro E, Ralf A, Ward K, Spector T, Bell JT, Kayser M. 2017b. Epigenetic discrimination of identical twins from blood under the forensic scenario. *Forensic Science International: Genetics* **31**:67–80. doi:10.1016/j.fsigen.2017.07.014

Vidaki A, Giangasparo F, Syndercombe Court D. 2016. Discovery of potential DNA methylation markers for forensic tissue identification using bisulphite pyrosequencing: Nucleic Acids. *ELECTROPHORESIS* **37**:2767–2779. doi:10.1002/elps.201600261

Vidaki A, Kalamara V, Carnero-Montoro E, Spector T, Bell J, Kayser M. 2018. Investigating the Epigenetic Discrimination of Identical Twins Using Buccal Swabs, Saliva, and Cigarette Butts in the Forensic Setting. *Genes* **9**:252. doi:10.3390/genes9050252

Vidaki A, Kayser M. 2017. From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence. *Genome Biol* **18**:238. doi:10.1186/s13059-017-1373-1

Vilkaitis G, Suetake I, Klimašauskas S, Tajima S. 2005. Processive Methylation of Hemimethylated CpG Sites by Mouse Dnmt1 DNA Methyltransferase. *Journal of Biological Chemistry* **280**:64–72. doi:10.1074/jbc.M411126200

Waddington CH. 2012. The Epigenotype. *Int J Epidemiol* **41**:10–13. doi:10.1093/ije/dyr184

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* **20**:116–117. doi:10.1038/2413

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schübeler D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* **37**:853–862. doi:10.1038/ng1598

Webster KE, O'Bryan MK, Fletcher S, Crewther PE, Aapola U, Craig J, Harrison DK, Aung H, Phutikanit N, Lyle R, Meachem SJ, Antonarakis SE, de Kretser DM, Hedger MP, Peterson P, Carroll BJ, Scott HS. 2005. Meiotic and epigenetic defects in Dnmt3L-knockout mouse spermatogenesis. *Proceedings of the National Academy of Sciences* **102**:4068–4073. doi:10.1073/pnas.0500702102

Weidner C, Lin Q, Koch C, Eisele L, Beier F, Ziegler P, Bauerschlag D, Jöckel K-H, Erbel R, Mühleisen T, Zenke M, Brümmendorf T, Wagner W. 2014. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biol* **15**:R24. doi:10.1186/gb-2014-15-2-r24

Wochna K, Bonikowski R, Śmigielski J, Berent J. 2018. Aspartic acid racemization of root dentin used for dental age estimation in a Polish population sample. *Forensic Sci Med Pathol* **14**:285–294. doi:10.1007/s12024-018-9984-8

Wu C -t., Morris JR. 2001. Genes, Genetics, and Epigenetics: A Correspondence. *Science, New Series* **293**:1103–1105.

Wu H, Zhang Y. 2014. Reversing DNA Methylation: Mechanisms, Genomics, and Biological Functions. *Cell* **156**:45–68. doi:10.1016/j.cell.2013.12.019

Xie S, He W-W. 1999. Cloning, expression and chromosome locations of the human DNMT3 gene family k 9.

Yen R-WC, Vertino PM, Nelkin BD, Yu JJ, El-Deiry W, Cumaraswamy A, Lennon GG, Trask BJ, Celano P, Baylin SB. 1992. Isolation and characterization of the cDNA encoding human DNA methyltransferase. *Nucl Acids Res* **20**:2287–2291. doi:10.1093/nar/20.9.2287

Yoon H-G, Chan DW, Reynolds AB, Qin J, Wong J. 2003. N-CoR Mediates DNA Methylation-Dependent Repression through a Methyl CpG Binding Protein Kaiso. *Molecular Cell* **12**:723–734. doi:10.1016/j.molcel.2003.08.008

Zamudio NM, Scott HS, Wolski K, Lo C-Y, Law C, Leong D, Kinkel SA, Chong S, Jolley D, Smyth GK, de Kretser D, Whitelaw E, O'Bryan MK. 2011. DNMT3L Is a Regulator of X Chromosome Compaction and Post-Meiotic Gene Transcription. *PLoS ONE* **6**:e18276. doi:10.1371/journal.pone.0018276

Zubakov D, Liu F, Kokmeijer I, Choi Y, van Meurs JBJ, van IJcken WFJ, Uitterlinden AG, Hofman A, Broer L, van Duijn CM, Lewin J, Kayser M. 2016. Human age estimation from blood using mRNA, DNA methylation, DNA rearrangement, and telomere length. *Forensic Science International: Genetics* **24**:33–43. doi:10.1016/j.fsigen.2016.05.014