UNIVERSITÀ POLITECNICA DELLE MARCHE

DIPARTIMENTO SCIENZE DELLA VITA E DELL'AMBIENTE

Corso di Laurea Magistrale-Biologia Molecolare e Applicata

ANALISI MUTAZIONALE DI UNA VASTA COLLEZIONE DI PAZIENTI CON CROMATINOPATIE HA PERMESSO DI RIDEFINIRE LA PREVALENZA DELLA SINDROME KABUKI

MUTATIONAL SCREENING OF A LARGE CHROMATINOPATHIES PATIENTS COLLECTION ALLOWS TO REDEFINE THE PREVALENCE OF KABUKI SYNDROME

Tesi di Laurea Magistrale di:

Matteo Russo

Relatore
Chiar.ma Prof.ssa

Anna La Teana

Correlatore:

Giuseppe Merla

Sessione ESTIVA

Anno Accademico 2019/2020

SUMMARY

ITALIAN SUMMARY **2**
ABSTRACT **5**
OBJECTIVES **7**

CHAPTER ONE: CHROMATINOPATHIES AND CROMATIN FUNCTIONING **8**
1.1 Chromatin regulation and functioning 8
1.2 The epigenetic machinery 9
1.3 DNA Methylation and Histone modifications 12
1.4 Histone Methylation 13
1.5 Kabuki Syndrome 15

CHAPTER TWO: MATERIALS AND METHODS **23**
2.1 Cohort of Samples analysed-NGS 23
2.2 Resequencing of the target and NGS 28
2.3 Bioinformatic analysis-NGS 29
2.4Kabuki Syndrome cohort of samples analysed-Sanger 30
2.4a PCR protocol *31*
2.4b Exosap protocol *31*
2.4c Sanger Sequencing Protocol *32*
2.5 Kabuki Prevalence Study 33
2.5a Epidemiological Part *33*
2.5b Genetic Part *36*

CHAPTER THREE: RESULTS **43**
3.1 Chromatinopathy and Kabuki syndrome cohort sequence analysis 43
3.2 Kabuki Syndrome: Meta-analysis of epidemiological studies 45
3.3 Kabuki Syndrome: Prevalence estimates from allele frequency 47
3.3a Known Disease Variants *47*
3.3b Predicted Disease Variants *53*

CHAPTER FOUR: DISCUSSION **56**

GLOSSARY **59**
BIBLIOGRAPHY **61**

ITALIAN SUMMARY

Le Cromatinopatie comprendono un gruppo di malattie genetiche rare causate da mutazioni nei geni essenziali per la regolazione della struttura e funzione della cromatina. Questo tipo di regolazione avviene attraverso meccanismi epigenetici che consentono il corretto funzionamento della cellula principalmente mediante il controllo dell'espressione genica. I geni che causano le Cromatinopatie si suddividono in 4 gruppi: writers, erasers, readers e remodelers. Generalmente, queste malattie si presentano con caratteristiche cliniche comuni quali la disabilità intellettiva, ritardo della crescita, cardiopatie e dismorfismi facciali tipici Una delle Cromatinopatie più note è la sindrome Kabuki, causata da mutazioni nei geni *KMT2D* e *KDM6A*, caratterizzata da anomalie congenite multiple con caratteristiche facciali tipiche tra cui: sopracciglia arcuate, ciglia lunghe, lunghe aperture delle palpebre (lunghe ragadi palpebrali) con le palpebre inferiori rivolte (ribaltate) ai bordi esterni, punta del naso allargata, grandi lobi delle orecchie sporgenti. Altri segni clinici: anomalie scheletriche, disabilità cognitiva lieve-moderata e deficit di crescita post-natale. Obiettivo principale del mio studio è stato quello di fare una nuova stima della prevalenza della sindrome Kabuki, attraverso una serie di analisi bioinformatiche combinate ad analisi mutazionali. Per la parte di laboratorio, ho eseguito l'analisi mutazionale

mediante Sanger del gene *KMT2D* su 172 pazienti diagnosticati con sindrome Kabuki ma nessuna mutazione è stata riscontrata. Infine, è stata eseguita prima un'analisi NGS e poi un'analisi di Sanger per un numero di pazienti con sospetta Cromatinopatia identificandone i geni causativi. Per il raggiungimento dell'obiettivo sono stati utilizzati due approcci: **epidemiologico** e **genetico**. Entrambi prevedevano l'utilizzo di parole chiave per la ricerca nei database PUBMED e EMBASE, di lavori scientifici che riportano la prevalenza della sindrome (PUBMED) e mutazioni nei geni *KMT2D* o *KDM6A* (EMBASE). Nel metodo epidemiologico i lavori raccolti sono stati poi sottoposti a meta-analisi, ottenendo una stima di prevalenza della malattia di 1/45000. L'approccio genetico ha tenuto conto di due classi di Varianti: Known Disease Variants (varianti note della malattia) e Predicted Disease Variants (varianti predette per la malattia). Per le Known Disease Variants, oltre i due database PUBMED e EMBASE, è stato utilizzato il dataset Clinvar per raccogliere tutte le varianti note di *KMT2D* e *KDM6A*. Solo le varianti per le quali è nota la frequenza allelica su GnomAD sono state analizzate. Usando il tool "https://www.cardiodb.org/allelefrequencyapp/", è stata ottenuta una stima della prevalenza dalle Known Disease Variants di 1:28000. Invece i dati per le Predicted Disease Variants sono stati raccolti da Clinvar, PUBMED, EMBASE, Ensembl e GnomAD. Solo le varianti

classificate come frameshift, stop codon, splice site, sono state analizzate con i tool “www.cardiodb.org “e “loftool”, ottenendo una stima di prevalenza di 1:18000. In conclusione, combinando i dati, la nuova stima sulla prevalenza globale della sindrome Kabuki è tra 3.6 e 9.2 su 100000 [tra 1:27933 e 1:10891].

ABSTRACT

Chromatinopathies are a group of disease caused by mutations in genes that are essential in the regulation of the structure and the function of chromatin. The chromatin is regulated through epigenetic mechanisms (such as the methylation of DNA, the genomic imprinting and the modification of the condensation of chromatin) which are important in order to manage the correct functioning of a cell, and thus regulate the gene expression, development and many other cell functions. Chromatinopathies are caused by mutations in one of four components of epigenetic machinery: writers, erasers, readers and remodelers. Generally, these diseases have common features such as: neurological dysfunction (among these the most frequent is the intellectual disability), autism, congenital heart disease or cancer. Learn more about functioning of epigenetics can be fundamental in order to understand how it regulates chromatin and related diseases.

One of the more studied Chromatinopathies is the Kabuki syndrome. The Kabuki Syndrome is a genetic disease that leads to multiple congenital anomalies with typical facial features including arched eyebrows, long eyelashes, long openings of the eyelids (long palpebral fissures) with the lower lids turned out (everted) at the outside edges, broadened tip of the nose, large protruding earlobes but also other clinical signs like skeletal anomalies,

mild-moderate cognitive disability and postnatal growth deficit. The KS syndrome is caused by mutations in *KMT2D* and *KDM6A* genes, the first gene provides instructions for making a lysine-specific methyltransferase 2D, while *KDM6A* encodes a lysine-specific demethylase 6A.

The prevalence of Kabuki Syndrome is estimated to 1:32.000 in Japan (Niikawa et al., 1988). In my thesis the main goal is to estimate the prevalence of Kabuki Syndrome, through a series of bioinformatics analysis, in order to update the prevalence of this rare genetic disease, which can turn be fundamental for the diagnosis. Moreover, I performed a genetic screening of patients clinically diagnosticated with KS. Finally, a Sanger analysis was also performed for patients with suspected Chromatinopathies, previously analysed with NGS, in order to confirm the mutations identified with the New Generation Sequencing methodology, and therefore to confirm the presence of the suspected pathology.

OBJECTIVES

In my thesis the main goal was to estimate the prevalence of Kabuki Syndrome, through a series of bioinformatics analysis, in order to update the prevalence of this rare genetic disease, which can turn be fundamental for the diagnosis. A second aim was to carry out a mutational screening in a number of patient's samples with a clinical diagnosis of Kabuki syndrome and other Chromatinopathies.

Chapter one

CHROMATINOPATHIES AND CHROMATIN FUNCTIONING

1.1 Chromatin regulation and functioning

Chromatin is a nucleoprotein complex in which nucleic acids are organized in eukaryotic organisms. This complex is fundamental in order to allow the compaction of DNA, its presence inside the nucleus, reducing the size of DNA, and also for its protection. The basic unit of chromatin is the nucleosome, that is formed by DNA, basic proteins called histones and other non-istonic proteins (acid proteins). It is made up of 147 base pairs of DNAs wrapped around a complex of eight histones (paired two by two) that are: H2A, H2B, H3 and H4. In addition, most chromatin condensates by winding in a polynucleosome fibre, that can be stabilized by the connection of histone H1 with all the nucleosomes by means a DNA linker. The regulation of the chromatin fibre allows the control of gene expression because it determines the accessibility and thus the recruitment of regulatory factors to the DNA. The structure of nucleosome may vary according to the different transcriptional stage. Indeed, the changes can affect the variation of the presence/absence of the chromatin constituents and/or the covalent modification of its components. The first types of changes concern the action

of repressors, activators, chromatin remodelling complexes, incorporation of histone variants. The second involves, instead, the methylation at cytosine residues and post-translational modifications of histone tails.

The Chromatin presents two state of condensations: euchromatin and heterochromatin. Euchromatin is the genome region where genes are actively transcribed or can be potentially transcribed, and is decondensed during interphase. The regulatory sequences in these regions are accessible to nucleases, have unmethylated CpG islands and the core histones H3 and H4 are hyperacetylated on their N-terminal lysine residues. In general, euchromatic domains replicate early in S phase. Conversely, the heterochromatin is correlated to the areas of the genome that are transcriptionally inactive and highly condensed (Arney & Fisher, 2004).

1.2 The epigenetic machinery

The Chromatinopaties (also known by the term Mendelian disorders of the epigenetic machinery or MDEMs) are rare genetic disorders that are caused by heterozygous loss of function mutations in one of the components of the epigenetic machinery: writers, erasers, remodelers and readers (Fahrner & Bjornsson, 2019). In these diseases, the state of chromatin may result compromised, the transcriptional state of the target genes altered, and changes can occur at DNA methylation level. **Writers** are proteins, enrolled by

transcription factors in characteristics regions of genome, that marks the genes which must be activated or inactivated by methylation/acetylation of histones or methylation of DNA; however, these modifications can be removed by **Erasers**. **Readers** allow the identification, interpretation of the epigenetic marks and also the connection with other component of epigenetic machinery. **Remodelers** deal with accessibility of DNA sites through the movement of nucleosomes on the DNA or by means the exchange of histone variants (**Figure 1a**). (Aygun & Bjornsson, 2020).

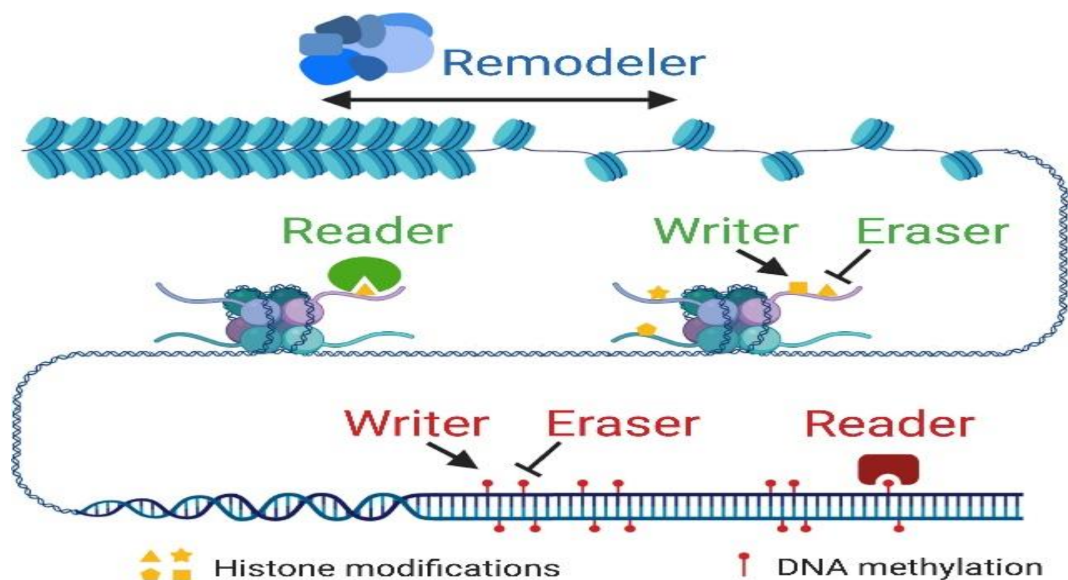


Figure 1a: Components of the epigenetic machinery. Writers, erasers, remodelers and readers are proteins that control gene expression, through specific modifications on DNA and/or histones.

Eight-two are the Chromatinopathies identified so far, that resulting from mutations in 70 epigenetic genes (Fahrner & Bjornsson, 2019). These disorders are mainly inherited in an autosomal dominant mode (**Figure 1b).**

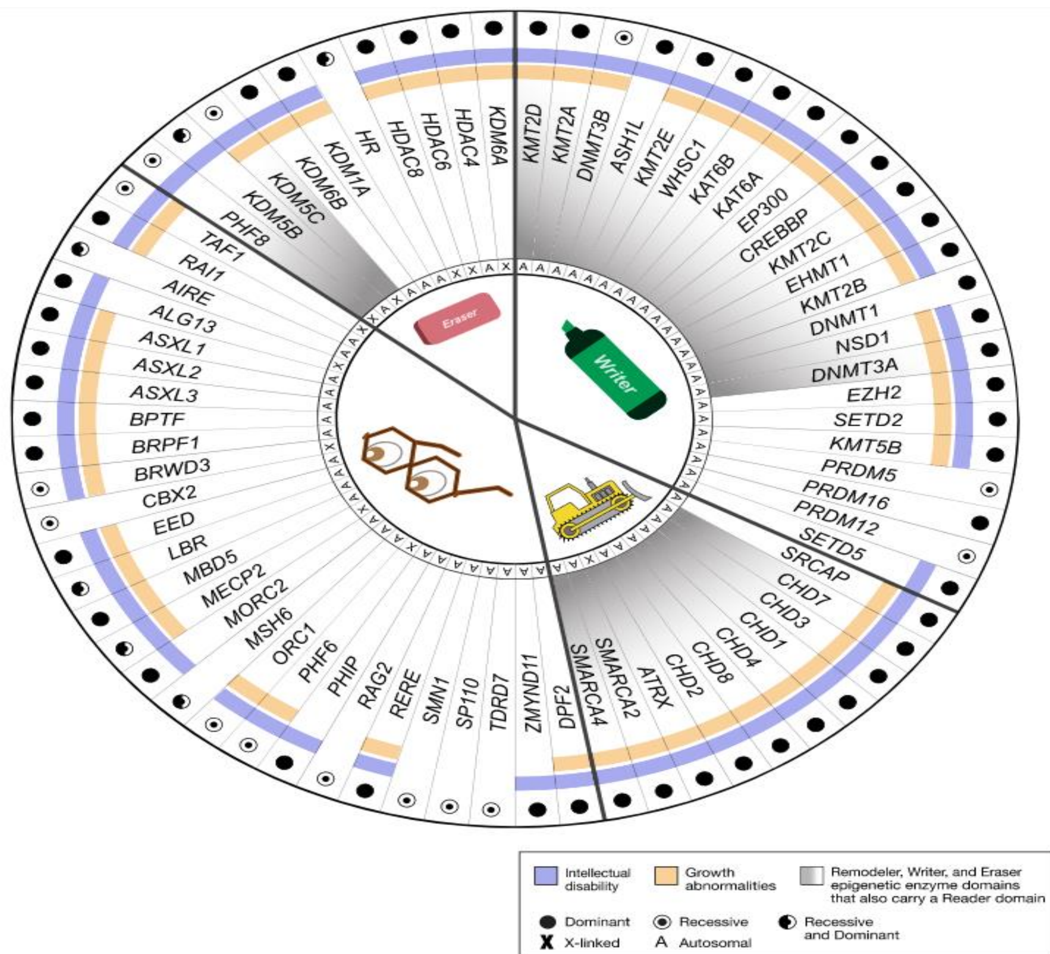


Figure 1b. List of Chromatinopathies so far identified.

1.3 DNA Methylation and Histone modifications

DNA methylation is an inherited epigenetic mark that provides the transfer of a methyl group at the C-5 position of cytosine ring of DNA, through activity of DNMTs (DNA methyltransferases) (Jin, Li, & Robertson, 2011). In mammals, DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis. (Gopalakrishnan, Van Emburgh, & Robertson, 2008; Jin et al., 2008; Jin et al., 2009; Robertson, 2005). The DNTM protein family are involved in regulation of DNA methylation. There are several members of this family: DNMT1, DNMT2, DNMT3A, DNMT3B, DNMT3L. (T. Bestor, Laudano, Mattaliano, & Ingram, 1988; T. H. Bestor, 2000; Okano, Xie, & Li, 1998; Yoder & Bestor, 1998). DNMT1 manages the methylation of hemi-methylated DNA, is positioned in the replication foci in S-phase and it allows the duplication on methylation patterns to daughter strands (Probst, Dunleavy, & Almouzni, 2009); DNMT2 methylates the cytosine-38 of tRNA (Goll et al., 2006); DNMT3A and DNMT3B adds the methylation mark at level of unmethylated CpG dinucleotides And finally DNMT3L increases the skill of methyltransferases in binding of SAM (S-adenosyl-L-methionine, a methyl group donor) (Kareta, Botello, Ennis, Chou, & Chedin, 2006). Histone

proteins can undergo different covalent modifications, which are essential in order to regulate gene expression, including acetylation, methylation, phosphorylation, ubiquitination and sumoylation (**Figure 2**) (Cedar & Bergman, 2009; Chi, Allis, & Wang, 2010).

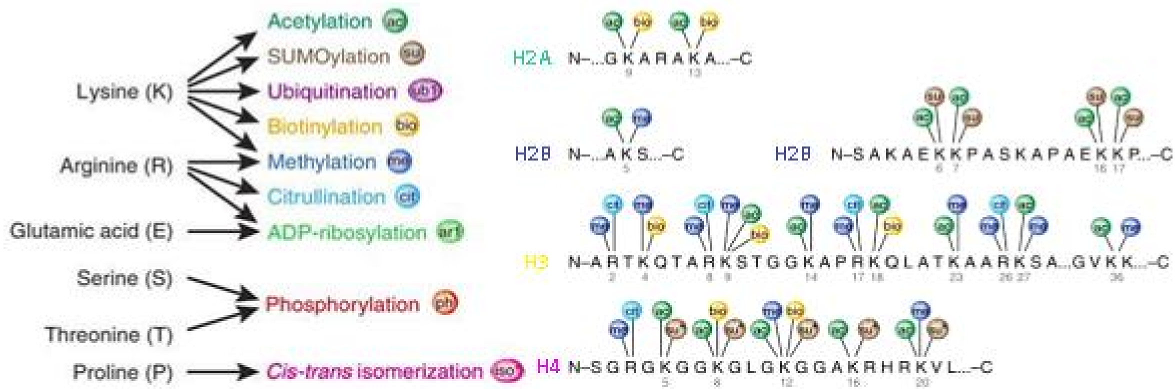


Figure 2: The histone post-translational modifications (PTMs). Each histone has a different profile of modifications, which occur on N-terminal tails of specific aminoacidic residues.

1.4 Histone Methylation

Histone methylation can take place on arginine, lysine, histidine. The Lysine residue can be mono(me1), di(me2) or tri(me3) methylated at the level of ε amine group (Hempel, Lange, & Birkofer, 1968; Murray, 1964; Paik & Kim, 1967). The arginine residue can be mono-methylated or demethylated at the level of guanidinyl group (Borun, Pearson, & Paik, 1972; Byvoet, Shepherd, Hardin, & Noland, 1972); finally, the histidine residue is monomethylated (Borun et al., 1972; Gershey, Haslett, Vidali, & Allfrey, 1969). The addition of methyl group, derived from SAM (s-adenosyl methionine) (Murray, 1964), to

histones is mediated by three families of enzymes that are: proteins that containing SET domain, Dot1 like proteins, and PRMT family. The first two families of proteins methylate lysine (Feng et al., 2002; Rea et al., 2000), instead the PRMT family methylate arginine (Bannister & Kouzarides, 2011). Furthermore, there are two highly conserved families of enzymes that demethylate the lysine residues methylated: amine oxidases (Shi et al., 2004) and Jumonji C domain containing (JmjC) (Cloos et al., 2006; Tsukada et al., 2006; Whetstine et al., 2006). The gene status depends by the position of the methyl lysine residue on the histone tail and on the number of methyl groups added. Indeed, H3K4me3 is associated with active transcription, while H3K27me3 is associated with repressed chromatin (Bernstein et al., 2002; Santos-Rosa et al., 2002). Moreover, H3K4me1 is associated with enhancer function (Heintzman et al., 2007) while H3K4me3 results in promoter activity. Finally, the simultaneous action of methyl-modifying enzymes is context specific. In fact, the H3K27 demethylase UTX is associated with H3K4 methyltransferase complex MLL2/MLL3, through the addition of "activating" mark H3K4m3 and by removing the inhibitory mark H3K27me3, in order to get an optimal transcriptional activation.

1.5 Kabuki Syndrome

Kabuki make-up syndrome (KMS) or Kabuki Syndrome is a multiple congenital anomalies syndrome caused by mutations in *KMT2D* and *KDM6A* genes first described in 1981 (Niikawa et al., 1988; Niikawa, Matsuura, Fukushima, Ohsawa, & Kajii, 1981). The KS syndrome is characterized by many facial features like: elongated palpebral fissures with eversion of the lateral third of the lower eyelid, arched and broad eyebrows, short columella with depressed nasal tip, and large, prominent, or cupped ears (Micale & Merla, 2012). The patients affected by this rare syndrome show also other common manifestations like skeletal anomalies (where there is a spinal column deformity, with or without sagittal cleft vertebrae, and brachydactyly), dermatoglyphic abnormalities (which comprises finger-tip pads, absence of digital triradius, increased digital ulnar loop and hypothenar loop patterns), mental retardation and finally postnatal growth deficiency. Other findings may include heart defects, genitourinary anomalies, cleft lip and/or palate, gastrointestinal abnormalities, ophthalmic defects, ptosis and strabismus, dental anomalies, with widely spaced teeth and hypodontia, and ear pits (Micale & Merla, 2012). Other less frequent signs may be visceral abnormalities and premature breast in females, delays in speech and language acquisition (Kuroki, Suzuki, Chyo, Hata, & Matsui, 1981; Niikawa et al.,

1981), autism or autistic-like behaviour, with difficult in communication and peer interactions (Ho & Eaves, 1997; Oksanen, Arvio, Peippo, Valanne, & Sainio, 2004). Furthermore, KS is distinguished by functional differences such as increased susceptibility to infections and auto-immune disorders, seizures, endocrinologic abnormalities, feeding problems, and hearing loss. KS is full penetrant syndrome and a child of a parent affected by KS have 50% of possibilities to have the syndrome (Micale & Merla, 2012). Whole-exome sequencing led to discovery heterozygous loss of function mutations in the *KMT2D* gene (MIM #602113, NM_003482.3, also known as *MLL2*) as the main cause of KS. The gene *KMT2D* encodes a member of the SET1 family of histone lysine methyltransferases, which catalyse the methylation of lysine 4 on histone H3 (H3K4), a modification associated with active transcription. In a minority of KS individuals has been identified, with either point mutations or microdeletions encompassing, the X-linked gene, *KDM6A* (MIM #300128, NM_021140.3, also known as *UTX*), which encodes for a Histone H3 lysine-27 demethylase. *KDM6A* plays a crucial role in chromatin remodelling and interacts with *KMT2D* in a conserved SET1-like complex (Cocciadiferro et al., 2018). *KMT2D* is localized at chromosome region 12q13, is a homologue of Trithorax group of proteins in Drosophila and manage the epigenetic control and chromatin activity. This gene is formed by

54 exons that lead to a protein with 5537 amino acid residues, with seven PHD (plant homeodomains), one HMG (high-mobility group domain), one LXXLL (Leu-X-X-Leu-Leu) and Coiled-coil motifs, one FYRC and FYRN domains and a SET domain. (**Figure 3**).

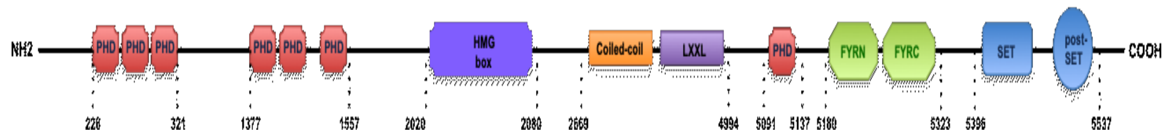


Figure 3. *KMT2D* protein domains. (Micale & Merla, 2012).

Since mutations in *KMT2D* lead to protein haploinsufficiency, this can cause the loss of its role in histone methylation pathway, which is reflected in an incorrect activation of genes in various organs and tissues, as demonstrated by the abnormalities in postnatal development of Kabuki patients. Also, variations in the *KMT2D* gene can lead to skeletal muscle and developmental anomalies, because it can activate specific genes involved in myogenesis and HOX genes (Micale & Merla, 2012). Moreover, patients who have mutations in *KMT2D* genes exhibit renal anomalies and short stature (Hannibal et al., 2011; Li et al., 2011; Paulussen et al., 2011). Taking into account all the pathogenic and likely pathogenic variants of the *KMT2D* gene (collected until March 2020), they are mostly frameshift (44%), then there are nonsense variants (35%), missense (11%), splice-site (9%), indel (1 %). (**figure 4**).

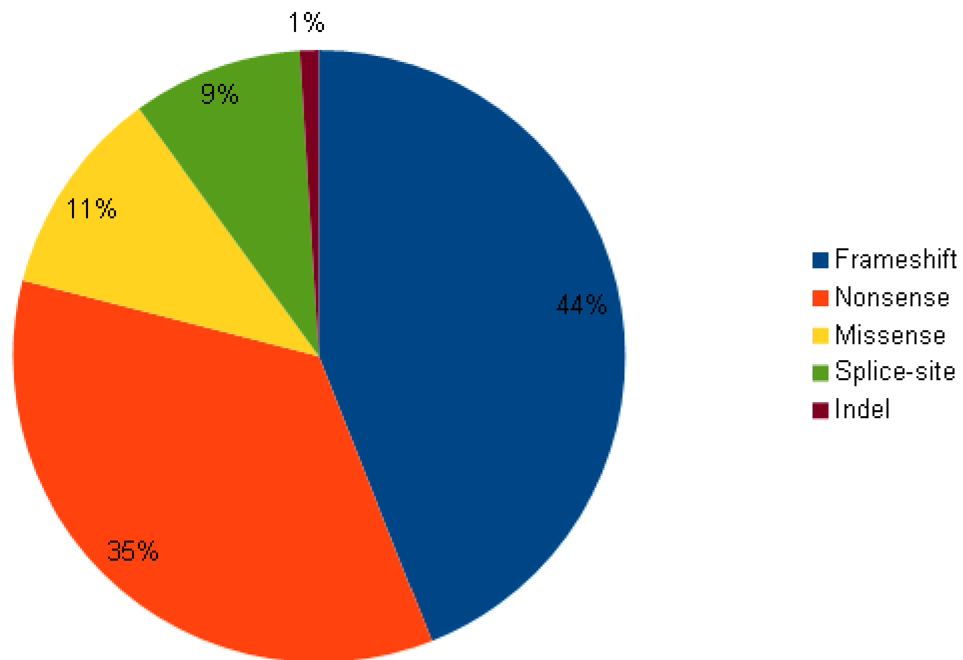


Figure 4. Percentage of various types of mutations for the *KMT2D* gene (considering only the pathogenic and likely pathogenic variants).

The majority of variants are localized in exon 39 and 48 (**figure 5**), and in particular in the LXXLL domain that is important in the binding between *KMT2D* and estrogen-receptor alpha. This brings to immunological defects and cardiac anomalies, but also early breast development (Micale & Merla, 2012).

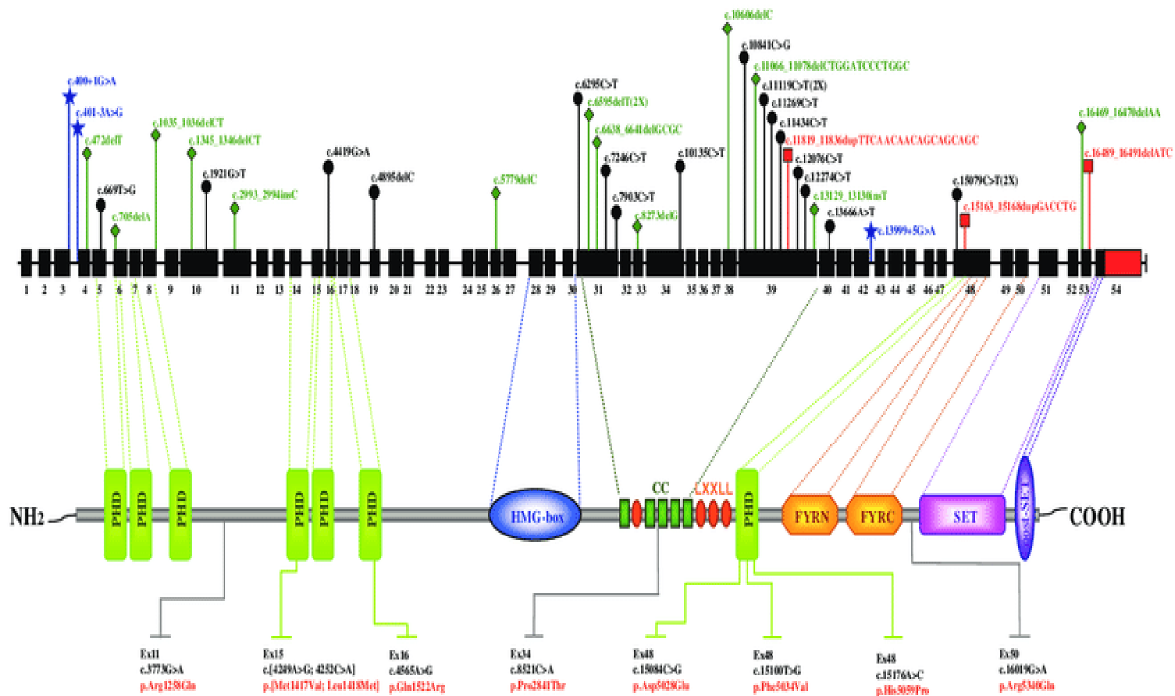


Figure 5. Mutation spectrum of *KMT2D*. Upper panel: genomic structure of the *KMT2D* gene including 54 coding exons (black rectangles), the 3'untranslated region (red rectangle) and introns (black horizontal line). Mutations are represented in black (nonsense), red (indel), blue (splice sites), and green (frameshift). Middle figure: *KMT2D* protein domain structures, PHD, plant homeodomain finger; HMG-box, high mobility group; CC, Coiled Coil; LXXLL domain, FYRN, FY-rich domain, N-terminal region; FYRC, FY-rich domain, C-terminal region; SET, (Su(var)3-9, Enhancer-of-zeste, Trithorax) domain; PostSET: PostSET domain. (L. Micale et al., 2011).

Regarding the other domains of the *KMT2D* gene, the PHD domain is a C4HC3 zinc-finger-like motif, involved in protein-protein interactions and controls the epigenetics and chromatin-mediated transcriptional regulation. HMG domain is a DNA-binding motif that can be found in in non-histone components of chromatin and in regulators of transcription a cell differentiation. Coiled coil domain deal with homo-oligomerisation (Micale & Merla, 2012). FYRN and FYRC motifs are two phenylalanine/tyrosine-rich

regions of 50 and 100 bases respectively that can be found in chromatin-associated proteins (Garcia-Alai, 2010) and that have the function of heterodimerisation between MLLN and MLLC terminal (Micale & Merla, 2012). Lastly, the SET domain is a histone methyltransferase, that are present in chromatin associated proteins with transcriptional activities (Jenuwein, Laible, Dorn, & Reuter, 1998) which deals with transferring a methyl group from S-adenosyl-L-methionine (SAM) to a lysine residue on histone or other proteins. Furthermore, *KMT2D* is part of SET1 family, that are present in conserved multi-subunit complexes that regulate the H3K4 metilation levels (Malik & Bhaumik, 2010). The gene *KDM6A* (also known as *UTX*) it's part of proteins family containing the domain JumonjiC (JmjC) that includes UTY (ubiquitously transcribed tetratricopeptide repeat, Y chromosome) and JMJD3 (Tsukada et al., 2006). The protein UTY is the Y-linked homologue of UTX. Instead, JMJD3, that don't have the TPR domains (tetratricopeptide repeat), allow the protein-protein interactions, has a sequence shared with UTX and UTY, both inside that outside JmjC domain. *KDM6A* gene has 29 exons that encode a 1401 residues protein which contains six TPR domains in the N-terminal half and the JmjC domain in the C-terminal half. **(Figure 6**).

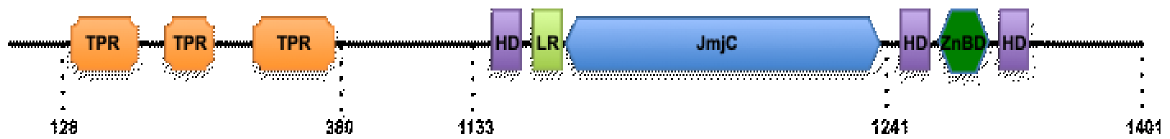


Figure 6. *KDM6A* protein domains (Micale & Merla, 2012).

KDM6A demethylates di- and trimethyl-lysine 27 (H3K27me2/3) on histone H3 (Hong et al., 2007; Lan et al., 2007). The Trimethylation of H3K27 (H3K27me3) is an important epigenetic mark that allows the maintaining of embryonic stem cell pluripotency and plasticity during the embryonic development, X chromosome inactivation and Polycomb-mediated gene silencing (Schuettengruber, Chourrout, Vervoort, Leblanc, & Cavalli, 2007). *KDM6A* binds with the complex MLL histone H3K4 methyltransferase, indeed this gene removes the histone H3K27 methyl mark when the MLL complex methylates histone H3K4, that are the antagonistic methyl marks for gene expression (Lee et al., 2007). Moreover, *KDM6A* and *KMT2D* works together for the epigenetic control of transcriptionally active chromatin by Polycomb-group proteins (PcG), which deal with tri- and demethylation of H3K27 (Aziz, Liu, & Dilworth, 2010). Finally, like *KMT2D*, the gene *KDM6A* it's involved in embryogenesis and development (Micale & Merla, 2012). Analysing all the *KDM6A* variants (collected until March 2020), considering only the pathogenetic and likely pathogenetic ones, they are

mainly splice-site (32%), then there are frameshift and nonsense (both 27%), missense (11%) and indel (3%). (**figure 7**).

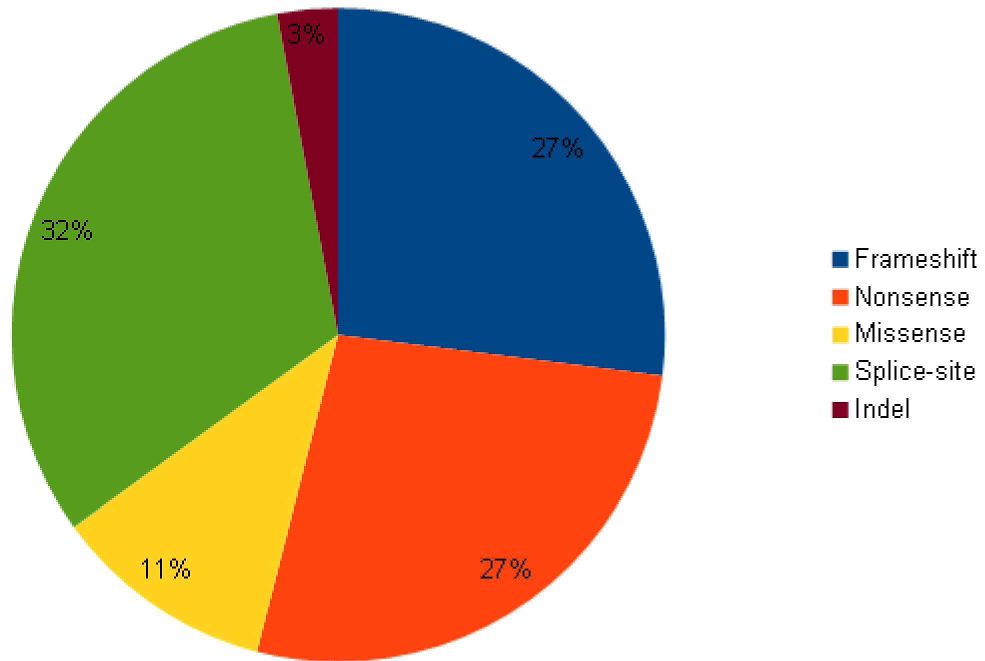


Figure 7. Percentage of various types of mutations for the *KDM6A* gene (considering only the pathogenic and likely pathogenic variants.

Chapter two

MATERIALS AND METHODS

2.1 Cohort of samples analysed-NGS

From January 2017 to January 2020, 307 patients with a clinical suspect of Chromatinopathies have been analysed (**Table 1, Table 2).**

The patients present in this cohort were included after the informed consent by the physicians and endorsement by local ethics committees. The Genomic DNA was extracted from fresh or frozen peripheral blood leukocytes, with the automated DNA extractor (EZ1 Advanced XL Qiagen, Germany).

Syndrome	Patient Analyzed
ATR-X syndrome	1
BCOR	1
Borjeson-Forssman-Lehmann syndrome	1
CHARGE syndrome	29
Coffin-Lowry syndrome	7
Coffin-Siris syndrome	6
Cornelia de Lange syndrome	17
Floating-Harbor syndrome	20
Kabuki Syndrome	135
Koolen-De Vries syndrome	3
Lujan-Fryns syndrome	2
Luscan-Lumish syndrome	1
MED 12	3
Nicolaides-Baraitser syndrome	1
Rubinstein-Taybi syndrome	48
Say-Barber-Biesecker-Young-Simpson syndrome	1
Smith-Magenis syndrome	2
Sotos syndrome	9
Undefined phenotype	18
Weaver syndrome	1

Table 1. Patients analysed with NGS, with suspect for Chromatinopathy, between January 2017-January 2020

Disorder	OMIM	Gene	OMIM Gene	NCBI Ref Seq	Inheritance	Disease Prevalence
Say-Barber-Biesecker-Young-Simpson syndrome	603736	**KAT6B**	605880	NM_012330.3	AD	<1/1000000 *
Genitopatellar Syndrome	606170	**KAT6B**	605880	NM_012330.3	AD	<1/1000000 *
Cerebellar ataxia, deafness, and narcolepsy	604121	**DNMT1**	126375	NM_001379.3	AD	<1/1000000 *
Neuropathy, hereditary sensory, type IE	614116	**DNMT1**	126375	NM_001379.3	AD	<1/1000000 *
Lujan-Fryns Syndrome	309520	**MED12**	300188	NM_005120.2	XLR	<1/1000000 *
Opitz-Kaveggia Syndrome	305450	**MED12**	300188	NM_005120.2	XLR	<1/1000000 *
Ohdo syndrome	300895	**MED12**	300188	NM_005120.2	XLR	<1/1000000 *
Alpha-thalassemia myelodysplasia syndrome	300448	**ATRX**	300032	NM_000489.4	NA	<1/1000000 *
Alpha-thalassemia/mental retardation syndrome	301040	**ATRX**	300032	NM_000489.4	XLD	<1/1000000 *
Mental retardation-hypotonic facies syndrome	309580	**ATRX**	300032	NM_000489.4	XLR	<1/1000000 *
Floating-Harbor Syndrome	136140	**SRCAP**	611421	NM_006662	AD	NA
Nicolaides-Baraitser Syndrome	601358	**SMARCA2**	600014	NM_003070.4	AD	<1/1000000 *
Dystonia 28, Childhood-onset	617284	**KMT2B**	606834	NM_014727.2	AD	NA
Mental retardation, X-linked 3	309541	**HCFC1**	300019	NM_005334.2	XLR	<1/1000000 *
Autism, Susceptibility To, 18	615032	**CHD8**	610528	NM_020920.3	AD	NA
Shprintzen-Goldberg Syndrome	182212	**SKI**	164780	NM_003036.3	AD	<1/1000000 *
Epileptic encephalopathy, childhood-onset	615369	**CHD2**	602119	NM_001271.3	AD	NA
Growth retardation, developmental delay, facial dysmorphism	612938	**FTO**	610966	NM_001080432.2	AR	<1/1000000 *
Holocarboxylase synthetase deficiency (biotin-responsive MCD)	253270	**HLCS**	609018	NM_000411.6	AR	NA
Pontocerebellar hypoplasia, type 8	614961	**CHMP1A**	164010	NM_002768.4	AR	<1/1000000 *
Microcephaly, postnatal progressive, with seizures and brain atrophy	613668	**MED17**	603810	NM_004268.4	AR	<1/1000000 *
Bohring-Opitz Syndrome	605039	**ASXL1**	612990	NM_015338.5	AD	<1/1000000 *
OFCD Syndrome	300166	**BCOR**	300485	NM_017745.5	XLD	NA

Table 2a

Disorder	OMIM	Gene	OMIM Gene	NCBI Ref Seq	Inheritance	Disease Prevalence
Tatton-Brown-Rahman Syndrome	615879	**DNMT3A**	602769	NM_022552.4	AD	<1/1000000 *
Townes-Brocks Syndrome	107480	**SALL1**	602218	NM_002968.2	AD	1-9/1000000 *
Borjeson-Forssman-Lehmann Syndrome	301900	**PHF6**	300414	NM_032458.2	XLR	<1/1000000 *
Smith-Magenis Syndrome	182290	**RAI1**	607642	NM_030665.3	AD	1-9 /100000 *
Kleefstra Syndrome	610253	**EHMT1**	607001	NM_024757.4	AD	<1/1000000 *
	617768	**KMT2C**	606833	NM_170606.3	AD	<1/1000000 *
Koolen-De Vries Syndrome	610443	**KANSL1**	612452	NM_015443.3	AD	NA
Wiedemann-Steiner Syndrome	605130	**KMT2A**	159555	NM_001197104.1	AD	<1/1000000 *
Luscan-Lumish Syndrome	616831	**SETD2**	612778	NM_014159.6	AD	NA
Claes-Jensen type Syndrome	300534	**KDM5C**	314690	NM_004187.3	XLR	<1/1000000 *
White-Sutton Syndrome	616364	**POGZ**	614787	NM_207171.2	AD	<1/1000000 *
Au-Kline Syndrome	616580	**HNRNPK**	600712	NM_002140.4	AD	<1/1000000 *
Weaver Syndrome	277590	**EZH2**	601573	NM_004456.4	AD	<1/1000000 *
		EED	605984	NM_003797.3	AD	NA
Siderius X-Linked mental retardation Syndrome	300263	**PHF8**	300560	NM_015107.2	XLR	<1/1000000 *
Coffin-Lowry Syndrome	303600	**RPS6KA3**	300075	NM_004586.2	XLD	1-9 /100000 *

Table 2b

Disorder	OMIM	Gene	OMIM Gene	NCBI Ref Seq	Inheritance	Disease Prevalence
Coffin-Siris Syndrome	135900	**ARID1B**	614556	NM_020732.3	AD	<1/1000000 *
	614607	**ARID1A**	603024	NM_006015.4	AD	<1/1000000 *
	614608	**SMARCB1**	601607	NM_003073.4	AD	<1/1000000 *
	614609	**SMARCA4**	603254	NM_003072.3	AD	<1/1000000 *
	616938	**SMARCE1**	603111	NM_003079.4	AD	<1/1000000 *
		HDAC4	605314	NM_006037.3	NA	<1/1000000 *
		MBD5	611472	NM_018328.4	AD	<1/1000000 *
		HUWE1	300697	NM_031407.6	XLD	<1/1000000 *
	618027	**DPF2**	601671	NM_006268.4	AD	NA
	617808	**ARID2**	609539	NM_152641.2	AD	NA
Cornelia de Lange Syndrome	122470	**NIPBL**	608667	NM_015384.4	AD	1-9 /100000 *
	300590	**SMC1A**	300040	NM_006306.3	XLD	1-9 /100000 *
	610759	**SMC3**	606062	NM_005445.3	AD	1-9 /100000 *
	614701	**RAD21**	606462	NM_006265.2	AD	1-9 /100000 *
	300882	**HDAC8**	300269	NM_018486.2	XLD	1-9 /100000 *
	615761	**SETD5**	615743	NM_001080517.2	AD	1-9 /100000 *
		BRD4	608749	NM_058243	NA	NA
Kabuki Syndrome	147920	**KMT2D**	602113	NM_003482.3	AD	1-9 /100000 *
	300867	**KDM6A**	300128	NM_021140.2	XLD	1-9 /100000 *
		RAP1A	179520	NM_001010935.2	NA	1-9 /100000 *
		RAP1B	179530	NM_015646.5	NA	1-9 /100000 *

Table 2c

Disorder	OMIM	Gene	OMIM Gene	NCBI Ref Seq	Inheritance	Disease Prevalence
Sotos Syndrome	117550	**NSD1**	606681	NM_022455.4	AD	1-9/100000 *
	614753	**NFIX**	164005	NM_001271043.2	AD	<1/1000000*
	617169	**APC2**	612034	NM_005883.2	AR	1-9/100000 *
Rubinstein-Taybi Syndrome	180849	**CREBBP**	600140	NM_004380.2	AD	1-9/100000 *
	613684	**EP300**	602700	NM_001429.3	AD	1-9/100000 *
Baraitser-Winter Syndrome	243310	**ACTB**	102630	NM_001101.3	AD	<1/1000000 *
	614583	**ACTG1**	102560	NM_001614	AD	<1/1000000 *
CHARGE Syndrome	214800	**CHD7**	608892	NM_017780.3	AD	NA
	214800	**SEMA3E**	608166	NM_012431	AD	NA
Beckwith-Wiedemann Syndrome	130650	**NSD1**	606681	NM_022455.4	AD	1-5/10000 *
	130650	**CDKN1C**	600856	NM_000076.2	AD	1-5/10000 *
ICF Syndrome	242860	**DNMT3B**	602900	NM_006892.3	AR	<1/1000000 *
	614069	**ZBTB24**	614064	NM_014797.2	AR	<1/1000000 *
Mental retardation, autosomal dominant 21	615502	**CTCF**	604167	NM_006565.4	AD	<1/1000000 *

Table 2d

Tables 2a, 2b, 2c, 2d. Chromatinopathy Panel (last update: June 2019 …). * Prevalence data were obtained from ORPHA.NET, NA= data not available

2.2 Resequencing of the target and NGS

The HaloPlex Target Enrichment panel (Agilent Technologies) was created with "www.agilent.com/genomics/suredesign" building a library that contains all the exons and the borders exon/intron of the genes **Table 2** following hg19 -GRCh37 reference genome. The generated probes cover about 384 kb of genome and the coverage wanders on 99.82%. To know the concentration of DNA, 50 ng of genomic DNA fluorimetry-based Qubit dsDNA BR Assay was used (Thermo Fisher Scientific). The samples were analysed only if the optical density ratio 260/280 was between 1.8-2.0 (NanoDrop2000c) and the distribution size is about 2.5 kb (0,8% gel electrophoresis). Using an indexing primer, the library of gDNA was hybridized to the probe and then the hybrids probe-fragment have been caught with magnetic beads. Later, the libraries fragments were amplified with a PCR to have suitable samples for sequencing. Both fragments of gDNA and final libraries were analysed through electrophoresis (Agilent Technologies) and a Fluorometer (Thermo Fisher Scientific). Lastly, the sequencing has been performed by Illumina MiSeq System, which provides the use of the protocol 151 bp paired-end and the kit MiSeq Reagent Kit V2—500 cycles (Illumina). This system can process from 11 to 22 samples per run to get reads depth bigger than 400X (for targeted nucleotide) and 20X in 98% of targeted region.

2.3 Bioinformatic analysis-NGS

Using FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc) the reads were analysed and trimmed with Trimmomatic (Bolger, Lohse, & Usadel, 2014) if the quality of the half of the reads was less than 10 (phred quality score). The adapter sequences that were not necessary were removed with the tool cutadapt (https://cutadapt.readthedocs.io/en/stable/). Then, the reads were aligned with the reference genome sequence (hg19) with the tool Bowtie2. The depth of coverage (for the target regions) was obtained with TEQC (https://doi.org/doi:10.18129/B9.bioc.TEQC) and also with other scripts. Then, using Haplotype Caller (GATK, version 3.8), the variants were identified, removed if the calling quality was lower than the phred value of 20 or not covered by at least 30 reads and finally annotated with ANNOVAR, through the use of RefSeq gene and transcript annotation. All the variants were getting from dbSNP (version 150), Exome Variant Server (http://evs.gs.washington.edu/EVS), ExAC (version 0.3), ClinVar, Kaviar and lastly GnomAD (version 2.1). After that, the artifacts which had previously been sequenced, were checked through coupling specific variants in the internal database of variants, then missense variants were annotated with dbNSFP (version 3.5), which allows to get pathogenicity predictions and evolutionary conservation measures. Lastly, the significance of variants was

determined by ACMG criteria, using tools like InterVar (http://wintervar.wglab.org/) and with Varsome (https://doi.org/10.1093/bioinformatics/bty897). The predicted pathogenic variants were then tested with PCR reaction and sequenced with Sanger protocol. Moreover, if parents' DNA were available, it was analysed the variant inheritance. The sequence of the various variants was reported with HGVS nomenclature available at "https://varnomen.hgvs.org/ ".

2.4 Kabuki syndrome cohort of samples analysed-Sanger

172 patients with suspected Kabuki syndrome, not analysed with NGS, were re-sequenced with Sanger method to search for mutations located in exon 40 and exon 41 of the *KMT2D* gene, two exons where most pathogenetic variants are located. This analysis was performed because of technical issues that hampered the first correct analysis of such cohort.

2.4a PCR Protocol

The protocol used for the PCR reaction involved:

- Initialization at 95°C for 12 minutes;
- 35 cycle, that included: denaturation a 95°C for 15 seconds, Annealing at 60°C for 30 seconds and Elongation at 72°C for 45 seconds;
- Final Elongation at 72°C for 7 minutes.

Furthermore, all the quantities of the various reagents used in the PCR protocol can be viewed in **table 3.**

2.4b Exosap Protocol

After checking the PCR product by agarose gel electrophoresis for 20 minutes at 150V (concentrated at 1.5%, so as to be able to test fragments less than 800bp), the amplified PCR was subjected to Exosap purification. This method is used to enzymatically eliminate PCR products not needed for Sanger sequencing. In fact, through the use of exonucleases and phosphatases, the elimination of primers and excess nucleotides is carried out by hydrolysis. The Exosap method involves a first incubation at 37° C to remove the primers and nucleotides, the second incubation instead provides a temperature of 80° C to inactivate the enzymes used in the Exosap. In order to be able to use this

procedure, PCR products must have a size between 100 and 20000 bp. In this technique 1.5 µl of PCR amplified and 0.5 µl of Exosap was used.

2.4c Sanger Sequencing Protocol

The next step involved preparing the samples for Sanger sequencing. Among the reagents used in this protocol we find: H20, buffer, Big Dye (which contains Polymerase, dNTP and ddNTP, the latter block the addition of further nucleotides) (**table 3**). Two separate reactions have been prepared for each primer pair (one for the forward primer and one for the reverse primer). The quantity of Big Dye to be used depends on the length of the fragment, in fact if the fragment has dimensions of 700-800 bp then it will be necessary to use a greater quantity of reagent, while if the fragment to be sequenced has smaller dimensions it will be necessary less Big Dye. The protocol used to prepare the samples for sequencing included:

- 96°C for 1 minute;
- 27 cycles, where the following temperatures are set: 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 35 seconds;
- extension to 15°C.

After which, the samples were sequenced by Illumina MiSeq System and checked with the tool “Sequencher 4.10.1”.

Reagent	Quantity	Concentration
DNA	2 µl	25 ng/µl
Buffer	1,25µl	10X
dNTP	1 µl	2,5mM
For	0,5 µl	15pm/µl
Rev	0,5 µl	15pm/µl
TaqGold	0,1 µl	5U/µl
H20	7,15µl	-
Total Volume	*12,5 µl*	

Reagent	Quantity	Concentration
PCR product	0,5 µl	-
Buffer	1,5 µl	5X
Big Dye	0,8 µl	-
Primer (F o R)	0,3 µl	15pm/µl
H20	6,9 µl	-
Total volume	*10 µl*	

Table 3. Reagents used for the PCR protocol (left figure) and Sanger Sequencing (right figure), with their quantities and their concentrations.

2.5 Kabuki Prevalence Study

In order to update the prevalence of the Kabuki Syndrome I used two distinct approaches: epidemiological and genetic.

2.5a Epidemiological Part

This part involves making a selection (on EMBASE and PUBMED) by setting various keywords to get all the papers reporting the prevalence of Kabuki Syndrome (**figure 8**). The search ended on March 2020. The selected papers have been subjected to meta-analysis to obtain the prevalence of Kabuki Syndrome from epidemiological papers. The meta-analysis was carried out using the Forest Plot method. [This analysis was performed by Dr. Massimiliano Copetti, Unit of Biostatistics, IRCCS- "Casa Sollievo della Sofferenza" - Hosital Poliambulatorio Giovanni Paolo II, San Giovanni Rotondo (FG), Italy].

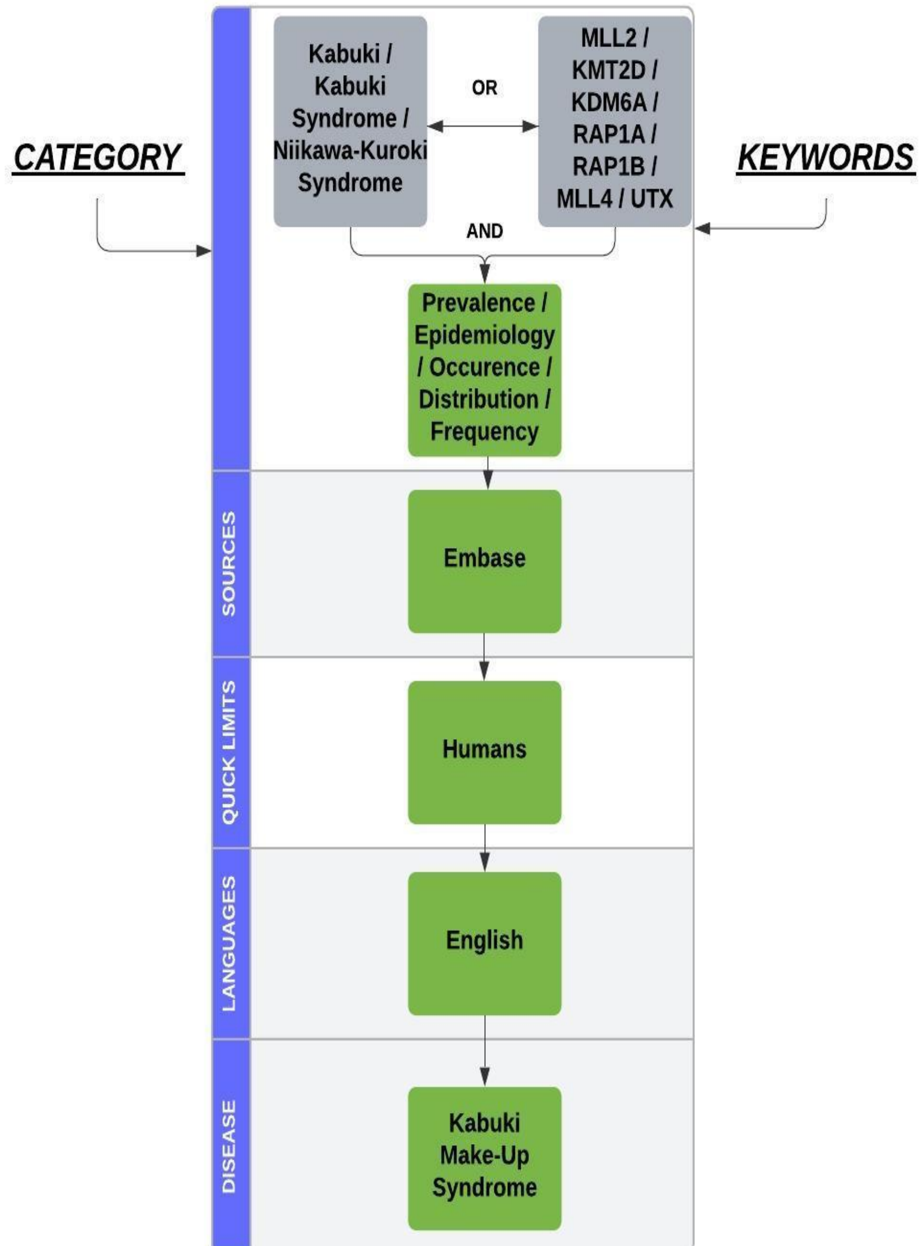


Figure 8a. Flowchart representing the epidemiological selection performed on the database EMBASE, indicating the keywords used in the various steps.

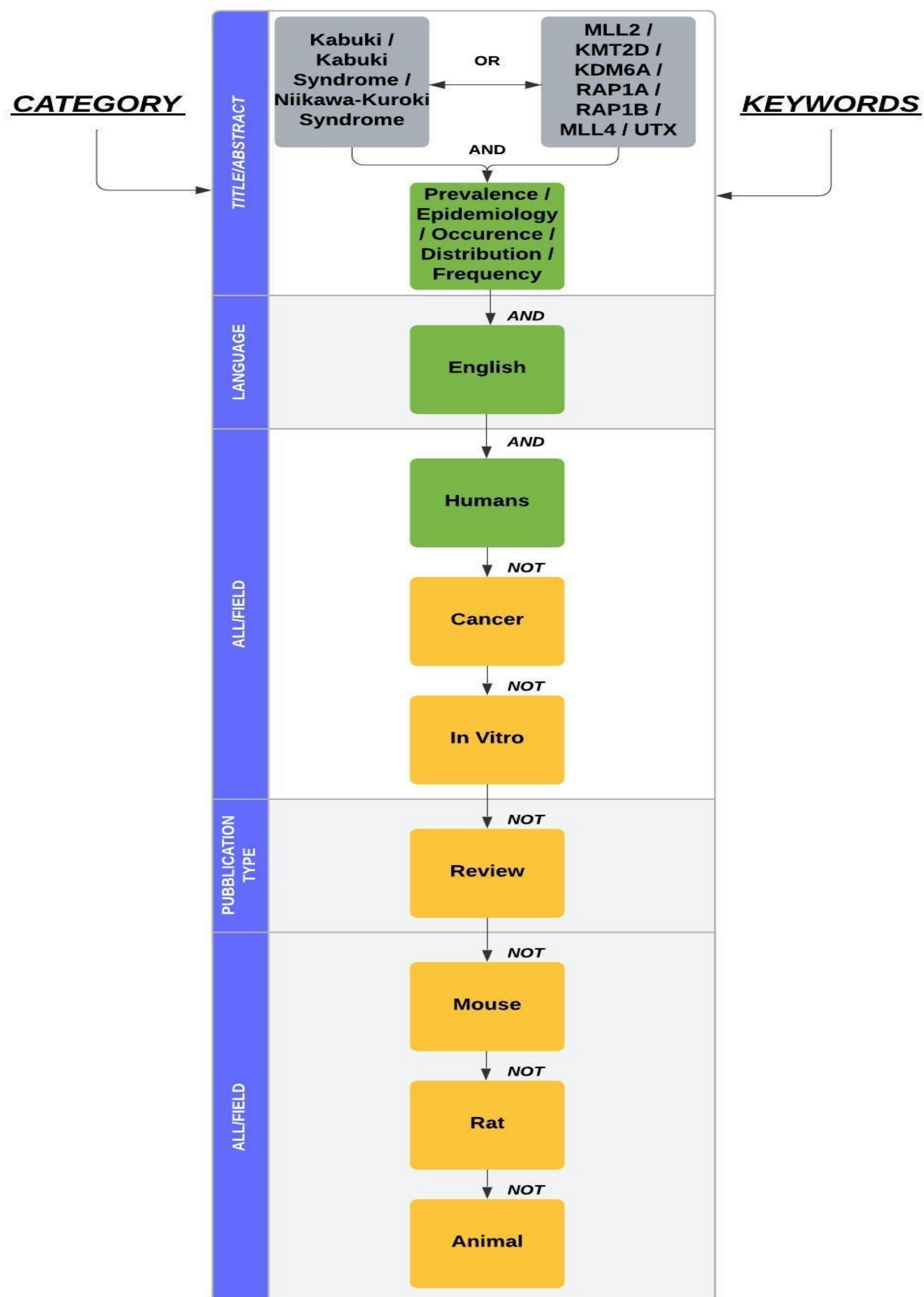


Figure 8b. Flowchart representing the epidemiological selection performed on the database PUBMED, indicating the keywords used in the various steps.

2.5b Genetic Part

The second part can be divided in "Known disease variants" and "Predicted disease variants". For the **Known disease variants**, was made a selection, using EMBASE and PUBMED as database, by setting keywords needed to get the papers concerning mutations in the *KMT2D* and *KDM6A* genes (**figure 9**). Then, all the *KMT2D* and *KDM6A* variants present on ClinVar were collected (https://www.ncbi.nlm.nih.gov/clinvar/) and recorded in our database indicating the consequence (cDNA, resulting protein), exon/intron concerned, type of mutation (frameshift, nonsense, missense, splice-site, indel), inheritance and reported pathogenicity (pathogenetic, likely pathogenetic, variant of unknown significance, benign). The tools MutationTaster (http://www.mutationtaster.org/) and VEP (https://www.ensembl.org/info/docs/tools/vep/index.html) were used to determine the consequence of frameshift, splice-site, and indel variants, where this consequence is not described. Indeed, Intervar (http://wintervar.wglab.org/) was used to determine the consequence of missense variants.

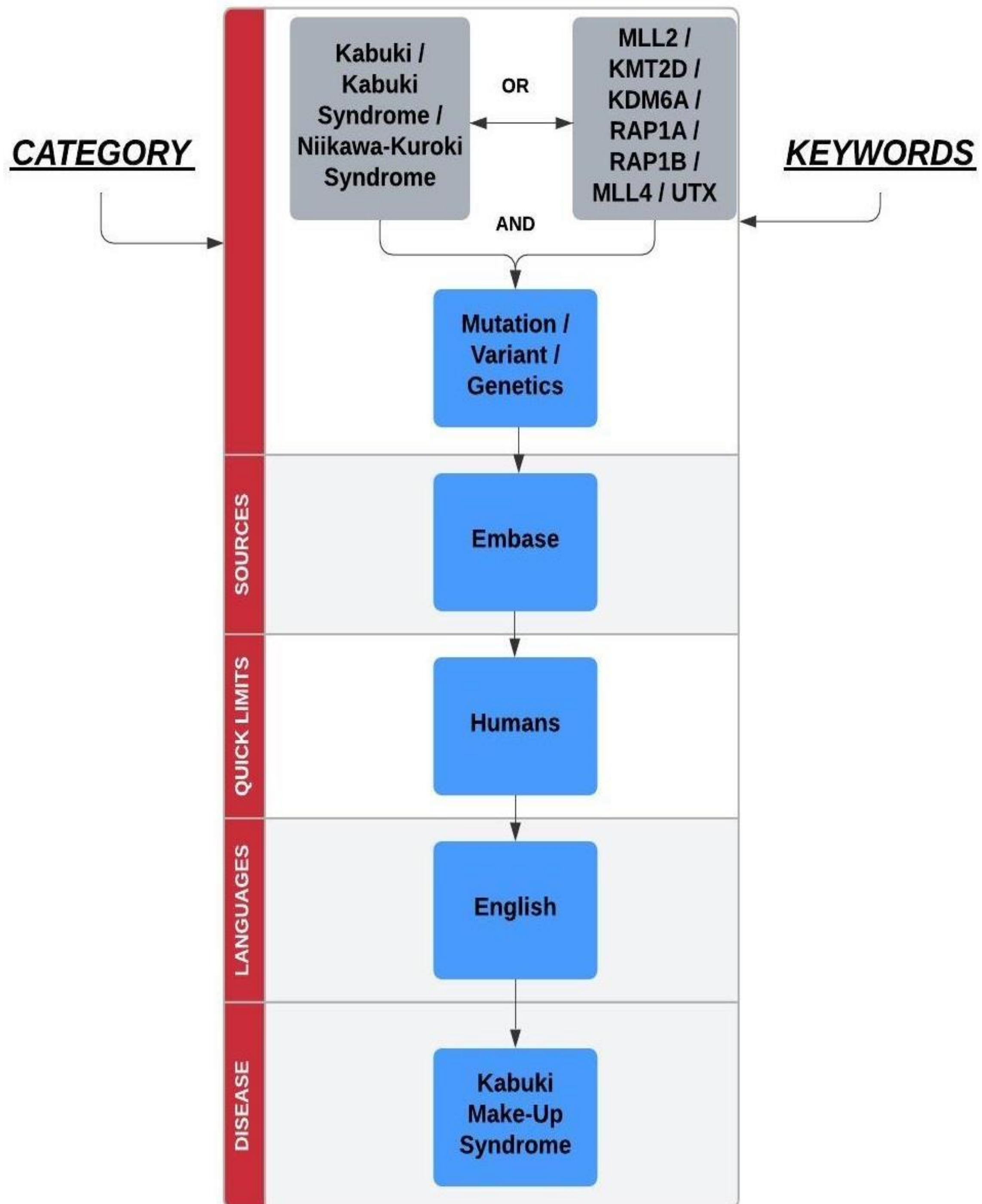


Figure 9a. Flowchart representing the genetic selection performed on the database EMBASE, indicating the keywords used in the various steps. The "/" symbol stands for *OR*.

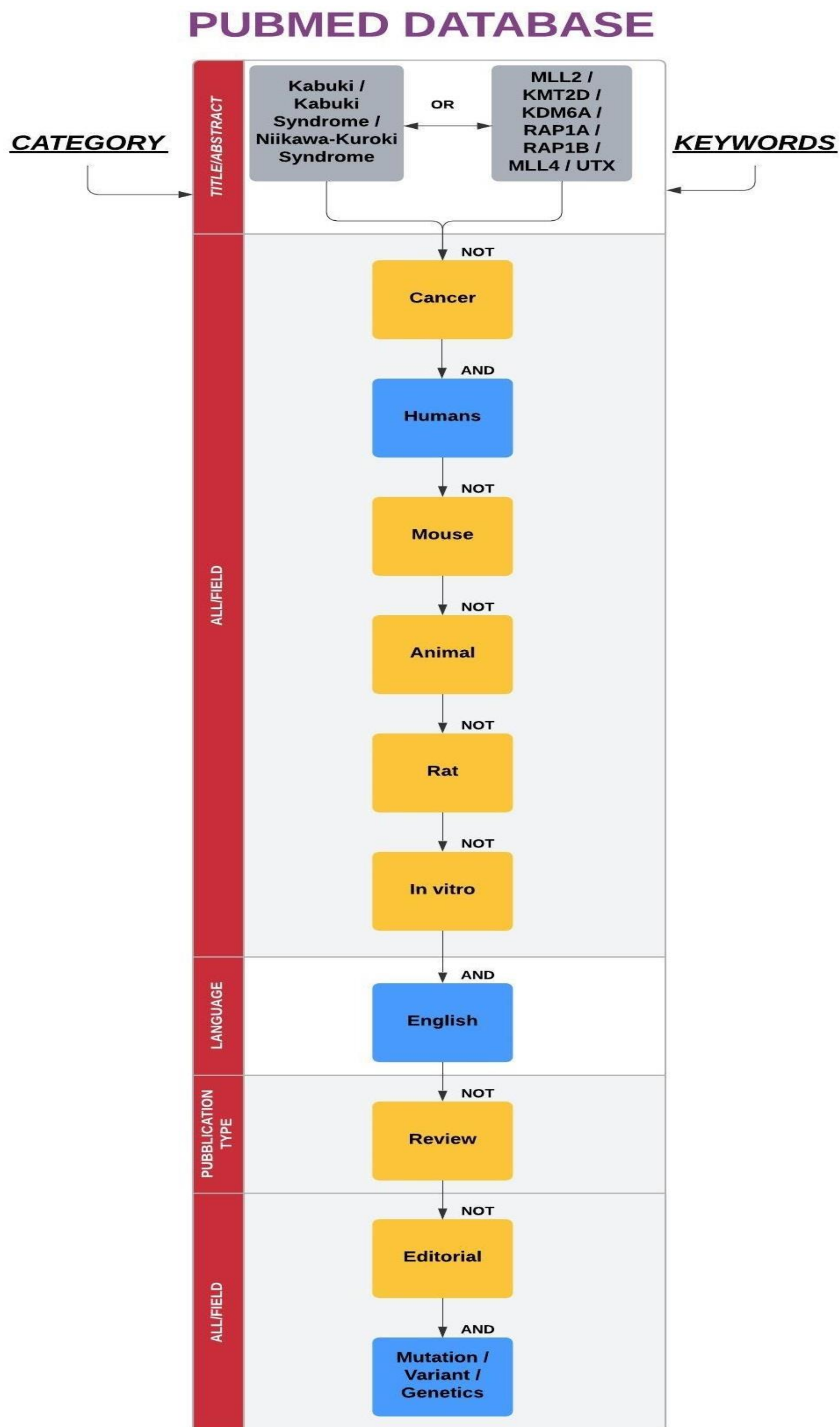


Figure 9b. Flowchart representing the genetic selection performed on the database PUBMED, indicating the keywords used in the various steps. The "/" symbol stands for *OR*.

The next step was to select the pathogenic and likely pathogenic variants that have allelic frequency data in GnomAD (https://GnomAD.broadinstitute.org/) Later, a filter was set using the software available at http://cardiodb.org/allelefrequencyapp/, that allows to calculate parameters like: Allele Frequency (AF), Maximum credible population AF, Maximum tolerated reference AC, penetrance and Filter allele frequency (all these terms can be found in section Glossary). So, with this tool, the Filtering AF was calculated for all the variants selected before, through use of Observed population AC" and "Observed population alleles sequenced (AN)" for each variant. The **Maximum credible population AF** and **Maximum tolerated reference AC** were calculated by setting all values required, which are: **inheritance** (monoallelic for Dominant condition or biallelic for Recessive condition), **prevalence**, **allelic heterogeneity**, **genetic heterogeneity**, **penetrance**, **confidence** and **Reference population size** (see glossary). Then for all the variants we compared the Filter AF with Maximum Credible population AF. Whether the **Filter allele frequency < Maximum credible population AF**, the variant is probably pathogenic, because it is uncommon in the healthy population, so it was included for further analysis. It's necessary discard also all the variants where the AC is bigger than Maximum tolerated reference AC (Gao, Brackley, & Mann, 2019). All the variants that

passed the filters were classified as "**Known disease variants**", and are used to calculate the prevalence of Kabuki syndrome. This can be calculated by sum all the Allelic Frequencies of these variants to get the **Global Mutant Allele Frequency.** The probability (P) that these variants cause the syndrome can be calculated multiplying by two the Global MAF. Then, multiplying the probability for 100.000, the prevalence of KS was calculated, from the Known Disease Variants. In **Figure 10a** there is a scheme of the procedure used to obtain prevalence from Known Disease Variants. Anyway, to estimate the true genetic prevalence of the Kabuki Syndrome it is necessary to include pathogenic variants that have not yet been identified in patients (Gao et al., 2019). For this purpose, these variants were called **Predicted disease variants.** To collect all the *KMT2D* and *KDM6A* predicted disease variants, we used Clinvar, Literature (EMBASE and PUBMED), Ensembl (https://www.ensembl.org/index.html), and GnomAD (https://GnomAD.broadinstitute.org/) datasets. Only the variants with a reported Allele Frequency were taken. Then we selected them according to the type of variants - frameshift, stop codon, splice site acceptor, splice site donor - in order to focus on the pathogenic variants only. After, the Filtering AF was calculated for each variant, and those in which Filtering AF was lower than Maximum Credible Population AF passed the selection. After that,

using LoFtool (https://omictools.com/loftool-tool) the variants were further filtered through a score given by the tool, that allow to identify LoF variants (loss of function). Finally, to obtain the prevalence of Kabuki syndrome from the *Predicted disease variants*, the same formulas used previously have been applied (**figure 10b**). For both classes of variants (Known and Predicted), the prevalence was calculated both globally and for the various ethnic groups (African, Latino, Ashkenazi Jewish, East Asian, European Finnish, European not Finnish, South Asian, Other).

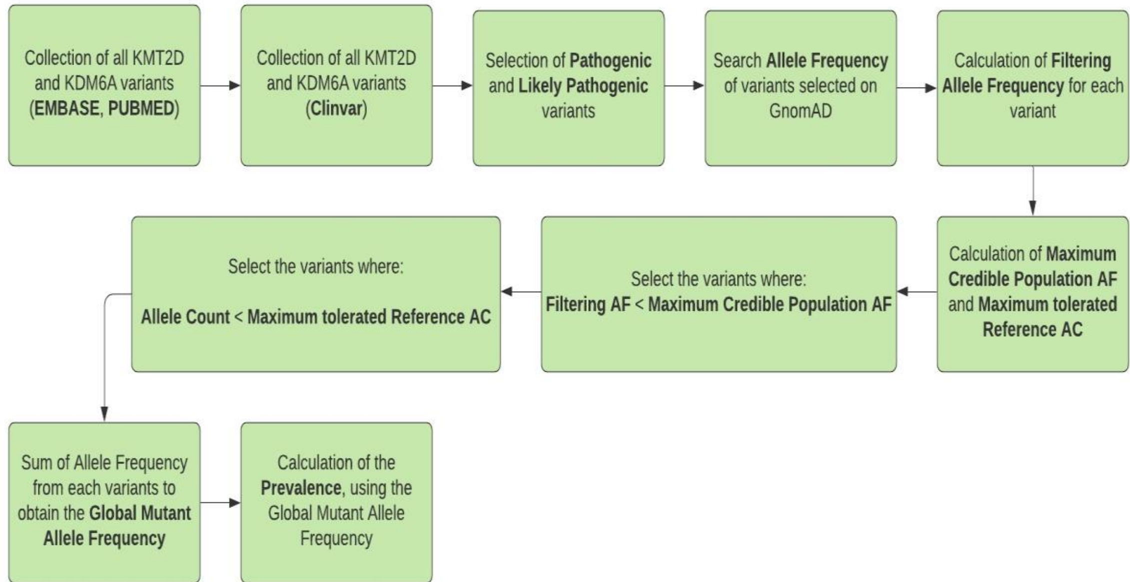


Figure 10a. Schematic representation of the procedure followed to calculate the prevalence starting from the Known Disease Variants.

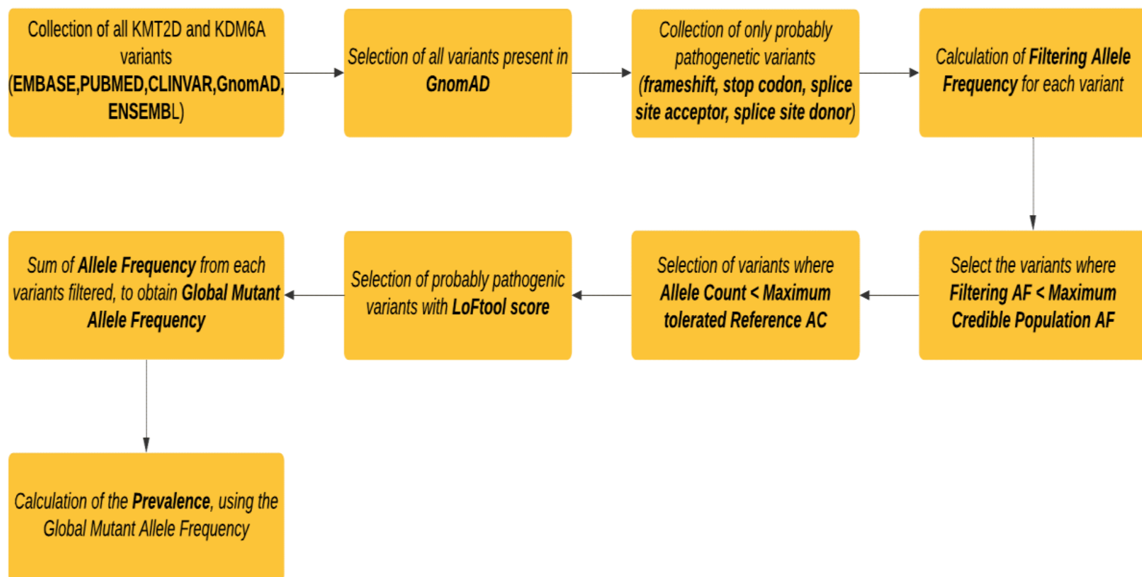


Figure 10b. Schematic representation of the procedure followed to calculate the prevalence starting from the Predicted Disease Variants.

Chapter three

RESULTS

3.1 Chromatinopathy and Kabuki syndrome cohort sequence analysis

Among the large cohort of Chromatinopathy's, analysed by NGS, and confirmed by Sanger sequencing, I analysed a small group of them and I found the results reported in (**Table 4**). Moreover, 172 patients with suspected Kabuki Syndrome, not included in the NGS study, were analysed by PCR, Exosap and Sanger sequencing, in order to find mutations in *KMT2D*. I did not find any causing mutations and, thus, these samples were not included in the further analysis.

sample	gene	Clinical Diagnosis	Phenotype observed	NM Gene
#1	SMARCA2	Kabuki Syndrome	Nicolaides-Baraitser syndrome	NM_003070:exon31:c.4364G>A
#2	MED12	Kabuki Syndrome	Lujan-Fryns S_Opitz-Kaveggia S_Ohdo syndrome	NM_005120.3(MED12_v001):c.1660G>A
#3	SMARCA4	Kabuki Syndrome	Coffin-Siris syndrome	NM_003072.4(SMARCA4_v001):c.3595G>A
#4	NSD1	CHARGE syndrome	Sotos syndrome	NM_022455.4(NSD1_v001):c.7799T>C
#4	DNMT1	CHARGE syndrome	Cerebellar ataxia, deafness, and narcolepsy; Neuropathy, hereditary sensory, type IE	NM_001379.4(DNMT1_v001):c.2983C>G
#4	ARID1B	CHARGE syndrome	Coffin-Siris syndrome	NM_020732.3(ARID1B_v001):c.356A>G
#4	KMT2B	CHARGE syndrome	Dystonia 28, Childhood-onset	NM_014727.2(KMT2B_v001):c.2053G>A
#5	SRCAP	Kabuki Syndrome	Floating-Harbor syndrome	NM_006662.3(SRCAP_v001):c.8647T>A
#5	SRCAP	Kabuki Syndrome	Floating-Harbor syndrome	NM_006662.3(SRCAP_v001):c.8648C>T
#5	KANSL1	Kabuki Syndrome	Koolen-De Vries syndrome	NM_015443.3(KANSL1_v001):c.680G>T
#6	NIPBL1	CHARGE syndrome	Cornelia de Lange syndrome	NM_015384.5(NIPBL_v001):c.2390G>A
#7	SETD5	CHARGE syndrome	Cornelia de Lange syndrome	NM_001080517.3(SETD5_v001):c.3929C>T
#7	HUWE1	CHARGE syndrome	Coffin-Siris syndrome	NM_031407.7(HUWE1_v001):c.9655C>T
#8	DNMT3B	CHARGE syndrome	ICF Syndrome	NM_006892.4(DNMT3B_v001):c.1144C>T
#9	KANSL1	Cornelia De Lange	Koolen-De Vries syndrome	NM_015443.3(KANSL1_v001):c.3028C>G
#9	CHD7	Cornelia De Lange	CHARGE syndrome	NM_017780.4(CHD7_v001):c.3949C>T

Table 4. Samples tested for generic Chromatinopathy. In this table there are information's like: the sample ID, the gene tested, Clinical Diagnosis, Phenotype observed, NM Gene, aminoacidic change, where is located the mutation (Exon or intron) and exonic function of the mutation.

3.2 Kabuki Syndrome: Meta-analysis of epidemiological studies

The first step was to find all the papers, using two databases (EMBASE, PUBMED), in which there were data concerned the prevalence of Kabuki Syndrome. The search ended on March 2020. The results obtained are present in **Figure 11**.

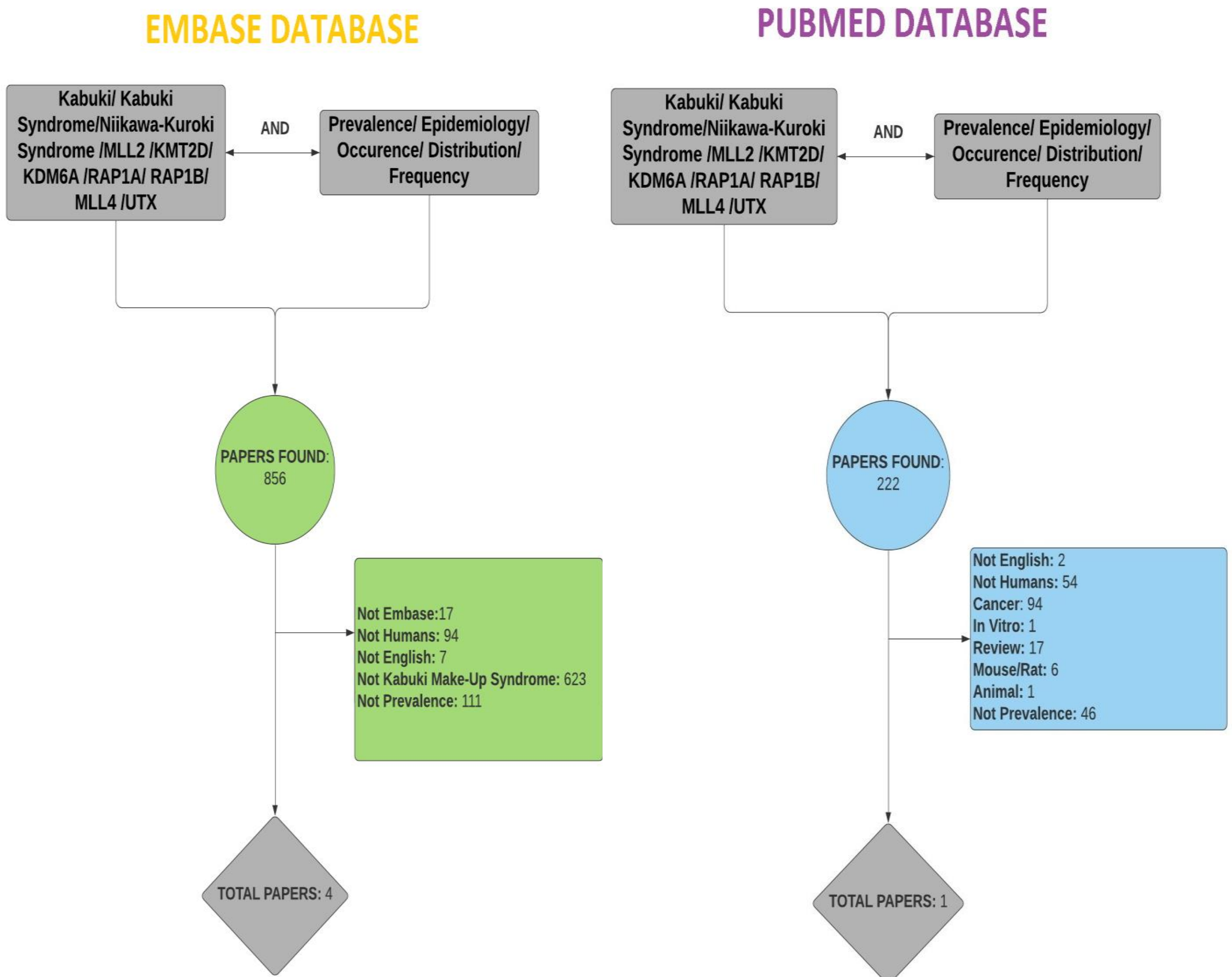


Figure 11a. (left) epidemiological results obtained from the database EMBASE. **Figure 11b. (right)** epidemiological results obtained from the database PUBMED.

From EMBASE a total of 4 papers were obtained after filtering, from PUBMED a total of 1 paper (**table 5**).

Study ID	Authors	Title	Year	Study Period	N live births	N cases
1	Niikawa et al.	Kabuki Make-Up (Niikawa-Kuroki) Syndrome: A study of 62 Patients	1988	1971-1986	1600000	50
2	Galán-Gómez et al.	Kabuki make-up (Niikawa-Kuroki) syndrome in five Spanish children	1995	1998-1990	250000*	5
3	White et al.	Growth, Behaviour, and Clinical Findings in 27 Patients with Kabuki (Niikawa-Kuroki) Syndrome	2004	1985-1999	946770	11
4	Langlois et al.	Descriptive Epidemiology of Birth Defects Thought to Arise by New Mutation	2015	1999-2009	4207898	5
5	Coi et al.	Prevalence Estimates of Rare Congenital Anomalies by Integrating Two Population-Based Registries in Tuscany, Italy	2017	2006-2013	256256	5

Table 5. **List of papers analysed to perform the meta-analysis.** * inhabitants. Study ID1 (Niikawa et al., 1988), ID2 (Galan-Gomez et al., 1995), ID3 (White et al., 2004) ID4 (Langlois & Scheuerle, 2015), ID5 (Coi et al., 2017).

The papers with study ID2 (Galan-Gomez et al., 1995) was excluded from the analysis as there are no data on the number of births in that region in reference period. Lastly, the four selected studies were subjected to meta-analysis**.** From the meta-analysis, starting from the epidemiological studies of

Kabuki syndrome, the prevalence of this syndrome is equal to 2.25 [0,4; 4,1] or 1 to 45000 (**figure 12**).

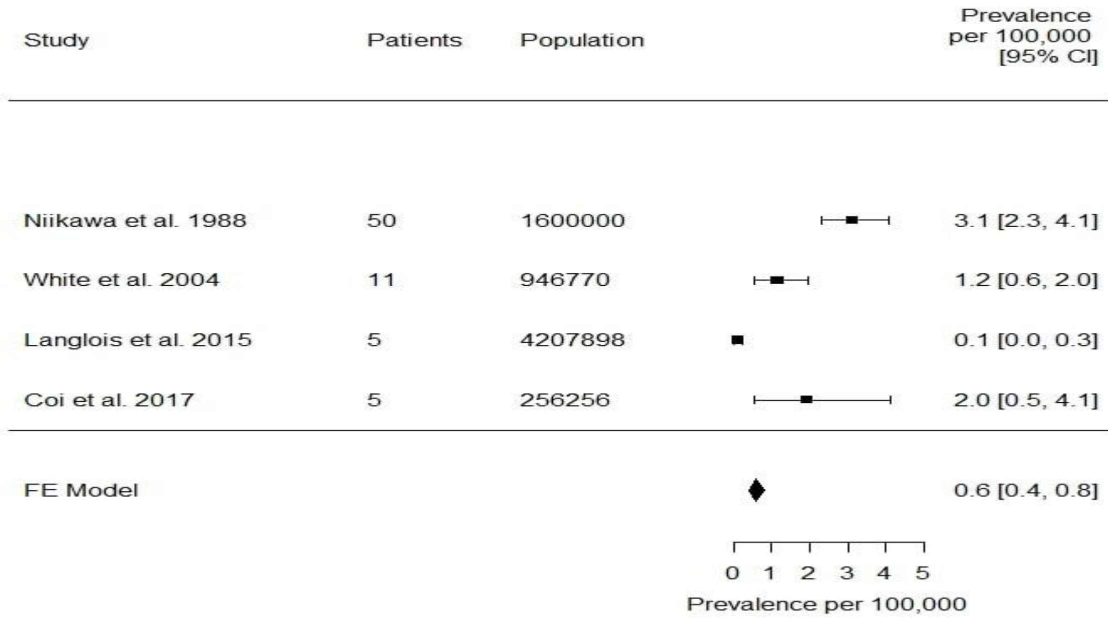
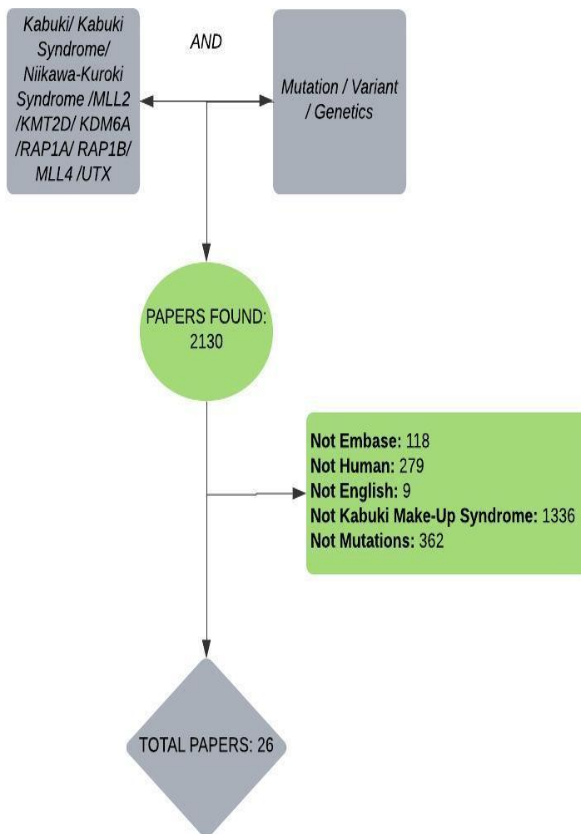


Figure 12. Forest plot where the results of the meta-analysis are shown starting from the epidemiological studies of the kabuki syndrome.

3.3 Kabuki Syndrome: Prevalence estimates from allele frequency

3.3a Known Disease Variants

All the *KMT2D* and *KDM6A* variants were collected using as reference database EMBASE and PUBMED (**figure 13**). The search ended on March 2020.



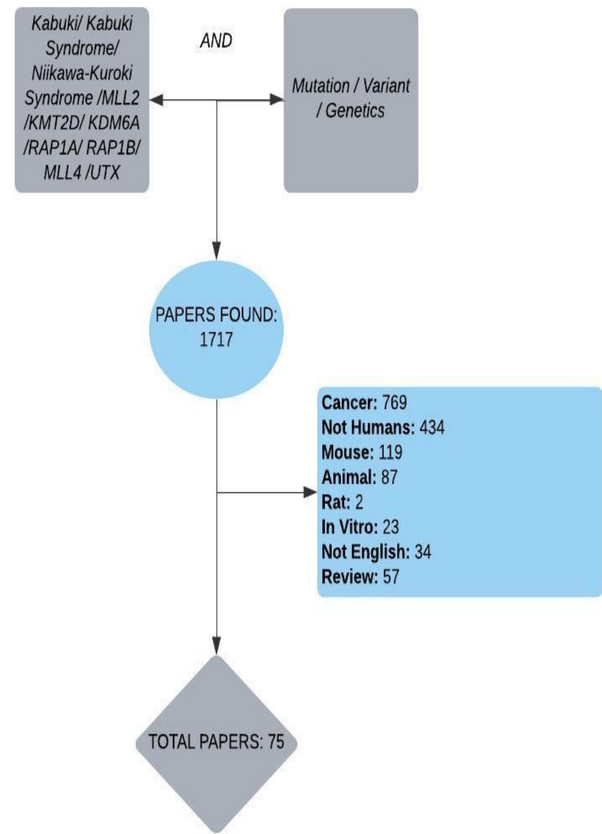


Figure 13a (left). Genetic Papers obtained from EMBASE. **Figure 13b. (right)** Genetic Papers obtained from PUBMED.

From EMBASE 26 papers, that concern mutations in *KMT2D* or *KDM6A*, were obtained instead from the PUBMED 75 papers were found. A total of 816 variants were collected for *KMT2D*, which 589 were pathogenic and 43 likely pathogenic (**table 6a**). In particular, for 12 variants (that are classified like: in-frame deletion/insertion, frameshift, splice-site,

protein_altering_variant, indel) Mutation Taster and VEP (Variant Effect Predictor) have been used to predict pathogenicity, resulting in 12 predicted deleterious variants. Instead, for *KDM6A*, 49 variants were found, which 41 were pathogenic and 1 likely pathogenic (**table 6b**). For 1 variant Mutation Taster and VEP were used and the result is that is predicted deleterious. Of all the pathogenetic and likely pathogenetic variants of literature, only 27 were present on GnomAD (25 of *KMT2D* and 2 of *KDM6A*).

TOTAL VARIANTS	**816**
Pathogenic	589 (12 predicted deleterious)
Likely Pathogenic	43
VUS	82
Likely Benign	60
Benign	30

Table 6a. Variants of *KMT2D* found in various papers analyzed, divided according their pathogenicity.

TOTAL VARIANTS	***49***
Pathogenic	41 (1 predicted deleterious)
Likely Pathogenic	1
VUS	6

Table 6b. Variants of *KDM6A* found in various papers analysed, divided according their pathogenicity.

Then, all the *KMT2D* and *KDM6A* variants were searched in Clinvar. For *KMT2D* were found 1076 variants but only 474 were used for further analysis (**table 7**), based on not enough evidence concerning their pathogenicity. Only 101 variants were found in GnomAD, with only 2 likely pathogenetic and included for further analysis. For the *KDM6A* we found 258 variants (**Table 8**), only 34 were selected, but no one of these were present in GnomAD.

Not found in GnomAD	**373**
Pathogenic	113
Likely Pathogenic	100
VUS	82
Likely Benign	62
Benign	16
Variants in GnomAd	101
Likely Pathogenic	2
VUS	99

Table 7. Variants of *KMT2D* found in Clinvar.

Not found in GnomAD	34
Pathogenic	13
Likely Pathogenic	10
VUS	9
Likely Benign	1
Benign	1

Table 8. Variants of *KDM6A* found in Clinvar but not present in GnomAD.

By combining the pathogenetic and likely pathogenetic variants from the literature and Clinvar (present on GnomAD) we have a total of 29, in particular 27 variants of *KMT2D* gene and 2 of *KDM6A* gene. After that, using the tool http://cardiodb.org/allelefrequencyapp/, Filtering AF was calculated (for each variant) and also the two parameters Maximum Credible Population AF and Maximum Tolerated Reference Ac. To calculated the lastly two parameters, the following setting was used in the tool:

- **Inheritance**: monoallelic.
- **Prevalence**: 1:45000
- **Allelic Heterogeneity:** 0,02
- **Genetic Heterogeneity**: 1
- **Penetrance:** 1
- **Confidence:** 95%.
- **Reference Population Size:** 235990.

The *Inheritance* was set on "monoallelic" because the Kabuki Syndrome is an Autosomal Dominant disease. The *Prevalence* inserted in the tool was 1:45000, because it corresponds to the average value of the interval [0.4, 4.1] obtained during the estimate of the prevalence starting from the Epidemiological Papers, that means 2.25 on 100000 or 1:45000.

The *Allelic Heterogeneity* was set equal to 0.02 because, the most frequent variant of the *KMT2D* gene, that is NM_003482.3:c.6595delT [p.(Tyr2199Ilefs * 65)], was found 18 times in 1046 patients (961 of them had variants in *KMT2D* and while 59 in *KDM6A*), so 18/1046 is equal to 0.0172, which can be approximated to 0.02. The *Genetic Heterogeneity* was set to 1 by convention (in autosomal dominant diseases). The *Penetrance* in Kabuki Syndrome is 100% so we set the value of "1". Finally, the *Reference Population Size* was set to "235990" by averaging the "Allele Number" parameter among all the variants considered (both *KMT2D* and *KDM6A*, taking into consideration the Literature and Clinvar). By setting all the parameters, a "**Maximum Credible Population AF**" equal to "2,25e-07" and a "**Maximum Tolerated Reference AC**" of "1" have been obtained. Excluding the variants that did not pass the ExomeQ or GenomeQ Quality control (for exons variants where there is an excess of heterozygotes are excluded, while for genomes variants that do not have a high quality of the sequenced genome are excluded) (https://gnomad.broadinstitute.org/), and then eliminating the variants where the "Filtering AF> Maximum Credible Population AF", 4 variants were obtained. These constitute the ***Known pathogenic variants***, which have a *Global Mutant Allele Frequency* equal to 0,0000179. Using the Hardy-Weinberg equation for Dominant conditions the

result obtained is that the Known disease variants give a global prevalence of 3,6 or 1/27933 (**table 9**).

	TOT	AFR	LAT	ASH	EAS	EFI	ENF	SAS	OTH
Global MAF	0,0000179	0,0001769	2,896E-05	NULL	NULL	NULL	8,8E-06	NULL	NULL
Probability	3,579936E-05	0,0003537	5,7918323E-05	NULL	NULL	NULL	1,7599845E.05	NULL	NULL
Prevalence	**3,6**	**35,4**	**5,8**	NULL	NULL	NULL	**1,8**	NULL	NULL
1 disease for	**27933**	**2827**	**17266**	NULL	NULL	NULL	**56819**	NULL	NULL

Table 9. Results of the prevalence estimate of Kabuki Syndrome from the Global MAF of the variants classified as *Known disease variants*. MAF= Mutant Allele Frequency, NULL= no data available, AFR= African, LAT=Latino, ASH= Ashkenazi Jewish, EAS= East Asian, EFI= European Finnish, ENF=European Not Finnish, SAS=South Asian, OTH=Other.

3.3b Predicted Disease Variants

All the variants of *KMT2D* and *KDM6A* genes, present in Clinvar, Literature (PUBMED, EMBASE), Ensembl and GnomAD were collected. For the *KMT2D* gene 11290 variants were found, instead for *KDM6A* the variants are 4806 (**table 10**). Of these, only 6878 were present in GnomAD, for *KMT2D*, and 1400 for *KDM6A*. After selecting the variants that are frameshift, stop codon, splice site acceptor, splice site donor (because were possibly pathogenetic), and excluding these that had already been used for the Known Disease Variants prevalence calculation, 79 and 18 variants were obtained for *KMT2D* and *KDM6A* respectively. For each of them, the Filtering AF was calculated, and a selection was made, eliminating those variants where the Filtering AF was greater than the Maximum Credible Population AF (2,25^-

07). Also, the variants where the Allele Count is bigger than Maximum Tolerated Reference AC (equal to 1) where excluded for further analysis. These 26 variants obtained with the filters where then subjected to the analysis with LoFtool (https://omictools.com/loftool-tool), with 5 variants belonging to *KDM6A* (all with score of 0,187) reported as likely pathogenic.

KMT2D		***KDM6A***	
TOTAL	31544	TOTAL	33030
GnomAD	6878	GnomAD	1400
ClinVar	1076	ClinVar	258
Ensembl	23700	Ensembl	31376

Table 10. All the variants of *KMT2D* and *KDM6A* found in the various database (Clinvar, GnomAD, Ensembl).

Therefore, the **Predicted pathogenic variants**, have a *Global Mutant Allele Frequency* equal to 0,000028. Using the Hardy-Weinberg equation for Dominant conditions the result obtained is that the Predicted disease variants give a global prevalence of 5,6 or 1/17852 (**table 11**). Adding the results obtained from the prevalence estimates starting from the Known Disease Variants and Predicted Disease Variants, it results that the estimated global genetic prevalence of Kabuki Syndrome is equal to 9,2 or 1/10891 (**table 12**).

	TOT	AFR	LAT	ASH	EAS	EFI	ENF	SAS	OTH
Global MAF	0,000028	NULL	NULL	NULL	0,0000072	NULL	0,000050	NULL	NULL
Probability	0,000056	NULL	NULL	NULL	0,000144	NULL	0,00010	NULL	NULL
Prevalence	**5,6**	**NULL**	**NULL**	**NULL**	**14,4**	**NULL**	**10,0**	**NULL**	**NULL**
1 disease for	**17852**	**NULL**	**NULL**	**NULL**	**6924**	**NULL**	**9964**	**NULL**	**NULL**

Table 11. Results of the prevalence estimate of Kabuki Syndrome from the Global MAF of the variants classified as *Predicted disease variants*. MAF= Mutant Allele Frequency, NULL= no data available, AFR= African, LAT=Latino, ASH= Ashkenazi Jewish, EAS= East Asian, EFI= European Finnish, ENF=European Not Finnish, SAS=South Asian, OTH=Other.

	TOT	AFR	LAT	ASH	EAS	EFI	ENF	SAS	OTH
Global MAF	0,000046	0,000177	0,000029	NULL	0,000072	NULL	0,000059	NULL	NULL
Probability	0,000092	0,000354	0,000058	NULL	0,000144	NULL	0,000118	NULL	NULL
Prevalence	**9,2**	**35,4**	**5,8**	**NULL**	**14,4**	**NULL**	**11,8**	**NULL**	**NULL**
1 disease for	**10891**	**2827**	**17266**	**NULL**	**6924**	**NULL**	**8478**	**NULL**	**NULL**

Table 12. Results of the Final prevalence estimate of Kabuki Syndrome. MAF= Mutant Allele Frequency, NULL= no data available, AFR= African, LAT=Latino, ASH= Ashkenazi Jewish, EAS= East Asian, EFI= European Finnish, ENF=European Not Finnish, SAS=South Asian, OTH=Other.

Chapter four

DISCUSSION

In my thesis path, a sub-group of patients with Chromathinopathies were analysed. In particular, all the mutations found were within exons and led to non-synonymous amino acid changes. The patients were clinical diagnosticated as affected by Chromatinopathy, the NGS analysis confirmed the clinical suspicious by identifying causing gene mutations for each case. This result is consistent with the usual differential diagnosis between them, because of the phenotype overlapping among Chromatinopathies.

The analysis of 172 samples with a clinical suspicious of Kabuki syndrome was carried out since exons 40-41 of the *KMT2D* gene were excluded from an earlier mutational screening of a larger KS cohort due to technical problems due to the difficulty of amplification of the fragments between the two exons. Therefore, the re-analysis allows to exclude causative variants from this large cohorts of samples, a cohort that will be processed by NGS with the Chromatinopathies panel in the coming months. In my study, I have deal principally with the prevalence estimates of Kabuki syndrome globally and in the major ethnic groups, getting a global higher frequency than the currents prevalence of 1:32000 (Niikawa et al., 1988). Of the others papers that I have

collected in the epidemiological approach, only one concern the prevalence of the KS syndrome, which give a prevalence of 1:86000 (White et al., 2004), instead for the others there were data only on the cases of KS, in a given period of time, and the number of born on the place taken into consideration. Obviously, the divergences found in the epidemiological papers may be due to differences in population, diagnosis and methodology, like can be seen between Texas state (1:841000) (Langlois & Scheuerle, 2015) and Japan (1:32000) (Niikawa et al., 1988). Indeed, with the genetic approach, using the GnomAD data set the global prevalence of KS at birth is between “3.6“and “9.2“per 100000. From my study can be highlighted that African and East Asian patients have bigger risk of have KS at birth, instead Latino ethnicity have lower risk. Moreover, it results that the final genetic prevalence estimates (9.2) are higher than epidemiological estimates (2,25) and this can be explained by the methodological differences, underdiagnosis due to phenotypical overlapping with other Chromatinopaties, false positives present in the database of allele frequencies used for the genetic prevalence estimates. In fact, the approach used has advantages and disadvantages: the lack of in vitro data supporting the consequence-predicted variants, the difficulty of obtaining precise estimates of the parameters like Allelic Heterogeneity, Genetic Heterogeneity and Maximum allelic contribution and finally the

selection of true pathogenic variants in the Predicted Disease variant's method.

The field of Chromatinopathies is in continuous updating, proving that a major focus in these syndromes may be necessary due to the mainly role of genes implicated in these genetic diseases. In my thesis, using a series of bioinformatics tools, performing meta-analysis and searching in specific databases, we have updated the prevalence of Kabuki Syndrome, both globally and in almost all ethnic groups. I think that the data presented will of support and usefulness for improving diagnosis and counselling on Kabuki syndrome. Finally, a continuous survey of genetic data on Kabuki is needed to constantly update the data presented here.

GLOSSARY

Allele Count (AC): is the counts of each alternate allele for each site across all samples.

Allele Frequency (AF): is the relative frequency of an allele at a particular locus in a population, expressed as a fraction or percentage.

Allelic heterogeneity: is the maximum proportion of variation within a gene that is attributable to a single allele.

Allele Number (AN): is the number of alleles successfully sequenced at the site.

Confidence: select in the range 0.9 – 0.999, this value represents the probability of observing a sample Allele count ≤ the reported maximum Allele Count. Increasing the confidence level increases the maximum Allele Count that would be considered compatible with disease-causation. Defaults to 0.95.

Filter allele frequency: is the highest disease-specific maximum credible population Allele Frequency for which the observed Allele Count is not compatible with pathogenicity.

Genetic heterogeneity: is the maximum proportion of disease attributable to variation in a single gene.

Inheritance: is the passing on of traits from parents to their offspring; either through asexual reproduction or sexual reproduction, the offspring cells or organisms acquire the genetic information of their parents.

Maximum credible population AF: correspond to an actual observed Allele Count in the reference sample.

Maximum tolerated reference AC: is the Allele Count at the upper bound of the one-tailed 95% confidence interval (95%CI Allele Count) of a Poisson distribution, for the specified maximum credible Allele Frequency, given the sample size (observed allele number, AN).

Penetrance: is the proportion of individuals carrying a particular variant of a gene that also express an associated trait .

Prevalence: is the proportion of a particular population found to be affected by a medical condition (typically a disease or a risk factor such as smoking or seat-belt use) at a specific time.

Reference population size: is the size of a subset of a target population that serves as a standard against which research findings are evaluated.

BIBLIOGRAPHY

Arney, K. L., & Fisher, A. G. (2004). Epigenetic aspects of differentiation. *J Cell Sci, 117*(Pt 19), 4355-4363. doi:10.1242/jcs.01390

Aygun, D., & Bjornsson, H. T. (2020). Clinical epigenetics: a primer for the practitioner. *Dev Med Child Neurol, 62*(2), 192-200. doi:10.1111/dmcn.14398

Aziz, A., Liu, Q. C., & Dilworth, F. J. (2010). Regulating a master regulator: establishing tissue-specific gene expression in skeletal muscle. *Epigenetics, 5*(8), 691-695. doi:10.4161/epi.5.8.13045

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res, 21*(3), 381-395. doi:10.1038/cr.2011.22

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., . . . Schreiber, S. L. (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. *Proc Natl Acad Sci U S A, 99*(13), 8695-8700. doi:10.1073/pnas.082249499

Bestor, T., Laudano, A., Mattaliano, R., & Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J Mol Biol, 203*(4), 971-983. doi:10.1016/0022-2836(88)90122-2

Bestor, T. H. (2000). The DNA methyltransferases of mammals. *Hum Mol Genet, 9*(16), 2395-2402. doi:10.1093/hmg/9.16.2395

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30*(15), 2114-2120. doi:10.1093/bioinformatics/btu170

Borun, T. W., Pearson, D., & Paik, W. K. (1972). Studies of histone methylation during the HeLa S-3 cell cycle. *J Biol Chem, 247*(13), 4288-4298. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5035694

Byvoet, P., Shepherd, G. R., Hardin, J. M., & Noland, B. J. (1972). The distribution and turnover of labeled methyl groups in histone fractions of cultured mammalian cells. *Arch Biochem Biophys, 148*(2), 558-567. doi:10.1016/0003-9861(72)90174-9

Cedar, H., & Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet, 10*(5), 295-304. doi:10.1038/nrg2540

Chi, P., Allis, C. D., & Wang, G. G. (2010). Covalent histone modifications--miswritten, misinterpreted and mis-erased in human cancers. *Nat Rev Cancer, 10*(7), 457-469. doi:10.1038/nrc2876

Cloos, P. A., Christensen, J., Agger, K., Maiolica, A., Rappsilber, J., Antal, T., . . . Helin, K. (2006). The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. *Nature, 442*(7100), 307-311. doi:10.1038/nature04837

Cocciadiferro, D., Augello, B., De Nittis, P., Zhang, J., Mandriani, B., Malerba, N., . . . Merla, G. (2018). Dissecting *KMT2D* missense mutations in Kabuki syndrome patients. *Hum Mol Genet, 27*(21), 3651-3668. doi:10.1093/hmg/ddy241

Coi, A., Santoro, M., Pierini, A., Marrucci, S., Pieroni, F., & Bianchi, F. (2017). Prevalence Estimates of Rare Congenital Anomalies by Integrating Two Population-Based Registries in Tuscany, Italy. *Public Health Genomics, 20*(4), 229-234. doi:10.1159/000481358

Fahrner, J. A., & Bjornsson, H. T. (2019). Mendelian disorders of the epigenetic machinery: postnatal malleability and therapeutic prospects. *Hum Mol Genet, 28*(R2), R254-R264. doi:10.1093/hmg/ddz174

Feng, Q., Wang, H., Ng, H. H., Erdjument-Bromage, H., Tempst, P., Struhl, K., & Zhang, Y. (2002). Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Curr Biol, 12*(12), 1052-1058. doi:10.1016/s0960-9822(02)00901-6

Galan-Gomez, E., Cardesa-Garcia, J. J., Campo-Sampedro, F. M., Salamanca-Maesso, C., Martinez-Frias, M. L., & Frias, J. L. (1995). Kabuki make-up (Niikawa-Kuroki) syndrome in five Spanish children. *Am J Med Genet, 59*(3), 276-282. doi:10.1002/ajmg.1320590303

Gao, J., Brackley, S., & Mann, J. P. (2019). The global prevalence of Wilson disease from next-generation sequencing data. *Genet Med, 21*(5), 1155-1163. doi:10.1038/s41436-018-0309-9

Gershey, E. L., Haslett, G. W., Vidali, G., & Allfrey, V. G. (1969). Chemical studies of histone methylation. Evidence for the occurrence of 3-methylhistidine in avian erythrocyte histone fractions. *J Biol Chem, 244*(18), 4871-4877. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5824561

Goll, M. G., Kirpekar, F., Maggert, K. A., Yoder, J. A., Hsieh, C. L., Zhang, X., . . . Bestor, T. H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. *Science, 311*(5759), 395-398. doi:10.1126/science.1120976

Gopalakrishnan, S., Van Emburgh, B. O., & Robertson, K. D. (2008). DNA methylation in development and human disease. *Mutat Res, 647*(1-2), 30-38. doi:10.1016/j.mrfmmm.2008.08.006

Hannibal, M. C., Buckingham, K. J., Ng, S. B., Ming, J. E., Beck, A. E., McMillin, M. J., . . . Bamshad, M. J. (2011). Spectrum of MLL2 (ALR) mutations in 110 cases of Kabuki syndrome. *Am J Med Genet A, 155A*(7), 1511-1516. doi:10.1002/ajmg.a.34074

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet, 39*(3), 311-318. doi:10.1038/ng1966

Hempel, K., Lange, H. W., & Birkofer, L. (1968). [Epsilon-N-trimethyllysine, a new amino acid in histones]. *Naturwissenschaften, 55*(1), 37. doi:10.1007/BF00593411

Ho, H. H., & Eaves, L. C. (1997). Kabuki make-up (Niikawa-Kuroki) syndrome: cognitive abilities and autistic features. *Dev Med Child Neurol, 39*(7), 487-490. doi:10.1111/j.1469-8749.1997.tb07470.x

Hong, S., Cho, Y. W., Yu, L. R., Yu, H., Veenstra, T. D., & Ge, K. (2007). Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. *Proc Natl Acad Sci U S A, 104*(47), 18439-18444. doi:10.1073/pnas.0707292104

Jenuwein, T., Laible, G., Dorn, R., & Reuter, G. (1998). SET domain proteins modulate chromatin domains in eu- and heterochromatin. *Cell Mol Life Sci, 54*(1), 80-93. doi:10.1007/s000180050127

Jin, B., Li, Y., & Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer, 2*(6), 607-617. doi:10.1177/1947601910393957

Jin, B., Tao, Q., Peng, J., Soo, H. M., Wu, W., Ying, J., . . . Robertson, K. D. (2008). DNA methyltransferase 3B (DNMT3B) mutations in ICF syndrome lead to altered epigenetic modifications and aberrant expression of genes regulating development, neurogenesis and immune function. *Hum Mol Genet, 17*(5), 690-709. doi:10.1093/hmg/ddm341

Jin, B., Yao, B., Li, J. L., Fields, C. R., Delmas, A. L., Liu, C., & Robertson, K. D. (2009). DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer. *Cancer Res, 69*(18), 7412-7421. doi:10.1158/0008-5472.CAN-09-0116

Kareta, M. S., Botello, Z. M., Ennis, J. J., Chou, C., & Chedin, F. (2006). Reconstitution and mechanism of the stimulation of de novo methylation by human DNMT3L. *J Biol Chem, 281*(36), 25893-25902. doi:10.1074/jbc.M603140200

Kuroki, Y., Suzuki, Y., Chyo, H., Hata, A., & Matsui, I. (1981). A new malformation syndrome of long palpebral fissures, large ears, depressed nasal tip, and skeletal anomalies associated with postnatal dwarfism and mental retardation. *J Pediatr, 99*(4), 570-573. doi:10.1016/s0022-3476(81)80256-9

Lan, F., Bayliss, P. E., Rinn, J. L., Whetstine, J. R., Wang, J. K., Chen, S., . . . Shi, Y. (2007). A histone H3 lysine 27 demethylase regulates animal posterior development. *Nature, 449*(7163), 689-694. doi:10.1038/nature06192

Langlois, P. H., & Scheuerle, A. E. (2015). Descriptive epidemiology of birth defects thought to arise by new mutation. *Birth Defects Res A Clin Mol Teratol, 103*(11), 913-927. doi:10.1002/bdra.23412

Lee, M. G., Villa, R., Trojer, P., Norman, J., Yan, K. P., Reinberg, D., . . . Shiekhattar, R. (2007). Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. *Science, 318*(5849), 447-450. doi:10.1126/science.1149042

Li, Y., Bogershausen, N., Alanay, Y., Simsek Kiper, P. O., Plume, N., Keupp, K., . . . Wollnik, B. (2011). A mutation screen in patients with Kabuki syndrome. *Hum Genet, 130*(6), 715-724. doi:10.1007/s00439-011-1004-y

Malik, S., & Bhaumik, S. R. (2010). Mixed lineage leukemia: histone H3 lysine 4 methyltransferases from yeast to human. *FEBS J, 277*(8), 1805-1821. doi:10.1111/j.1742-4658.2010.07607.x

Micale, L., Augello, B., Fusco, C., Selicorni, A., Loviglio, M. N., Silengo, M. C., . . . Merla, G. (2011). Mutation spectrum of MLL2 in a cohort of Kabuki syndrome patients. *Orphanet J Rare Dis, 6*, 38. doi:10.1186/1750-1172-6-38

Micale, L. Merla, G. (2012). Molecular Genetics of Kabuki Syndrome. In *eLS*. doi: 10.1002/9780470015902.a0023848

Murray, K. (1964). The Occurrence of Epsilon-N-Methyl Lysine in Histones. *Biochemistry, 3*, 10-15. doi:10.1021/bi00889a003

Niikawa, N., Kuroki, Y., Kajii, T., Matsuura, N., Ishikiriyama, S., Tonoki, H., . . . et al. (1988). Kabuki make-up (Niikawa-Kuroki) syndrome: a study of 62 patients. *Am J Med Genet, 31*(3), 565-589. doi:10.1002/ajmg.1320310312

Niikawa, N., Matsuura, N., Fukushima, Y., Ohsawa, T., & Kajii, T. (1981). Kabuki make-up syndrome: a syndrome of mental retardation, unusual facies, large and protruding ears, and postnatal growth deficiency. *J Pediatr, 99*(4), 565-569. doi:10.1016/s0022-3476(81)80255-7

Okano, M., Xie, S., & Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat Genet, 19*(3), 219-220. doi:10.1038/890

Oksanen, V. E., Arvio, M. A., Peippo, M. M., Valanne, L. K., & Sainio, K. O. (2004). Temporo-occipital spikes: a typical EEG finding in Kabuki syndrome. *Pediatr Neurol, 30*(1), 67-70. doi:10.1016/s0887-8994(03)00419-3

Paik, W. K., & Kim, S. (1967). E-N-dimethyllysine in histones. *Biochemical and Biophysical Research Communications, 27*(4), 479-483. doi:https://doi.org/10.1016/S0006-291X(67)80010-X

Paulussen, A. D., Stegmann, A. P., Blok, M. J., Tserpelis, D., Posma-Velter, C., Detisch, Y., . . . Schrander-Stumpel, C. T. (2011). MLL2 mutation spectrum in 45 patients with Kabuki syndrome. *Hum Mutat, 32*(2), E2018-2025. doi:10.1002/humu.21416

Probst, A. V., Dunleavy, E., & Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol, 10*(3), 192-206. doi:10.1038/nrm2640

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B. D., Sun, Z. W., Schmid, M., . . . Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature, 406*(6796), 593-599. doi:10.1038/35020506

Robertson, K. D. (2005). DNA methylation and human disease. *Nat Rev Genet, 6*(8), 597-610. doi:10.1038/nrg1655

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C., . . . Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature, 419*(6905), 407-411. doi:10.1038/nature01080

Schuettengruber, B., Chourrout, D., Vervoort, M., Leblanc, B., & Cavalli, G. (2007). Genome regulation by polycomb and trithorax proteins. *Cell, 128*(4), 735-745. doi:10.1016/j.cell.2007.02.009

Shi, Y., Lan, F., Matson, C., Mulligan, P., Whetstine, J. R., Cole, P. A., . . . Shi, Y. (2004). Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell, 119*(7), 941-953. doi:10.1016/j.cell.2004.12.012

Tsukada, Y., Fang, J., Erdjument-Bromage, H., Warren, M. E., Borchers, C. H., Tempst, P., & Zhang, Y. (2006). Histone demethylation by a family of JmjC domain-containing proteins. *Nature, 439*(7078), 811-816. doi:10.1038/nature04433

Whetstine, J. R., Nottke, A., Lan, F., Huarte, M., Smolikov, S., Chen, Z., . . . Shi, Y. (2006). Reversal of histone lysine trimethylation by the JMJD2 family of histone demethylases. *Cell, 125*(3), 467-481. doi:10.1016/j.cell.2006.03.028

White, S. M., Thompson, E. M., Kidd, A., Savarirayan, R., Turner, A., Amor, D., . . . Bankier, A. (2004). Growth, behavior, and clinical findings in 27 patients with Kabuki (Niikawa-Kuroki) syndrome. *Am J Med Genet A, 127A*(2), 118-127. doi:10.1002/ajmg.a.20674

Yoder, J. A., & Bestor, T. H. (1998). A candidate mammalian DNA methyltransferase related to pmt1p of fission yeast. *Hum Mol Genet, 7*(2), 279-284. doi:10.1093/hmg/7.2.279