ANÀLISI DE L'ADN MITOCONDRIAL EN MOSTRES DE CERVELL POSTMORTEM DE PACIENTS AMB DIAGNÓSTIC D'ESQUIZOFRÈNIA I DE PERSONES SANES

Magdalena Kostova Lefterova

TREBALL DE FI DE GRAU BIOTECNOLOGIA

Tutora acadèmica: Maria Del Carmen Portillo Guisado, doble titulació dels graus en Biotecnologia i en Bioquímica i Biologia Molecular, en Enginyeria Informàtica i en Biotecnologia, graus en Bioquímica i Biologia Molecular, Biotecnologia, Enologia, Nutrició Humana i Dietètica, departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili (URV)

Tutora del treball a l'empresa: Dra. Lourdes Martorell Bonet, Grup de Recerca Genètica i Ambient en Psiquiatria (GAP). Institut d'Investigació Sanitària Pere Virgili (IISPV-CERCA), Facultat de Medicina i Ciències de la Salut (URV) i Hospital Universitari Institut Pere Mata (HUIPM) lourdes.martorell@urv.cat

En cooperació amb: Hospital Universitari Institut Pere Mata (HUIPM)

Supervisora: Bengisu Kevser Bulduk, estudiant de doctorat del GAP. IISPV-CERCA, URV, HUIPM. bengisukevser.bulduk@estudiants.urv.cat

Data de convocatòria; Juny 2023

Jo, Magdalena Kostova Lefterova , amb NIE Y-0029648-S, soc coneixedora de la guia de prevenció del plagi a la URV Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants (aprovada el juliol 2017) (http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/) i afirmo que aquet TFG no constitueix cap de les conductes considerades com a plagi per la URV.

Tarragona, __06__ de_juny_de _2023_

(signatura)

Agraïments

En primer lloc vull agrair a la meva tutora de pràctiques Dr. Lourdes Martorell Bonet per haver-me acceptat en el seu equip d'investigació, per compartir els seus coneixements i per la seva gran ajuda per a realitzar aquest Treball de Fi de Grau. Els seus consells i recomanacions han sigut indispensables en els moments en els que no estava segura de com enfocar o continuar amb el meu projecte. A més, gràcies a ella he pogut entendre millor alguns conceptes relacionats amb el funcionament mitocondrial i l'esquizofrènia.

També vull agrair a la Bengisu Kevser Bulduk per a ajudar-me a entendre i realitzar la part computacional d'aquest treball, a més de per la seva gran paciència i comprensió amb les que va afrontar el fet de no poder entendre'ns sempre en anglès. Per altra part, m'ha encantat treballar juntes i, tant ella com Lourdes han suposat una vertadera motivació per a continuar-me formant com a científica.

Agraeixo a tot l'equip de Genètica i Ambient en Psiquiatria (GAP) per haver creat un ambient tan inspirador i agradable per a treballar. Valoro molt el gran respecte que es tenen entre tots, encara que provinguin de branques científiques i de nivells acadèmics diferents, a la vegada que també permeten l'entrada de l'humor i l'espontaneïtat que afavoreixen l'existència una dinàmica molt més amena i agradable.

Per altra part, vull agrair a la meva tutora de TFG per haver-me ajudat a realitzar correctament el meu Treball de Fi de Grau mitjançant exemples, consells de format i contingut, a més de per la seva gran paciència en front a tots els problemes que em van sorgir durant aquest període i que em van dificultar, fins a cert punt, enviar els avenços de l'estudi en les places inicialment establertes.

També vull agrair als meus professors i professores de la Facultat d'Enologia de la Universitat Rovira i Virgili per haver-me format i aportat les bases necessàries de Biotecnologia que em van permetre sol·licitar aquestes pràctiques i realitzar el meu Treball de Fi de Grau.

Per últim vull agrair als meus pares tot el suport i ajuda que em van donar durant tots els meus estudis i per haver-me animat en els moments més difícils, ja que, sense ells res de tota la meva experiència acadèmica i pràctica hagués estat possible.

ÍNDEX

Dades del Centre

Aquest treball final de grau es basa en l'estada de pràctiques extracurriculars al grup de investigació de Genètica i Ambient en Psiquiatria (GAP), de l'Institut de Investigació Sanitària Pere Virgili (IISPV). El grup està integrat per psiquiatres, psicòlegs, fisioterapeutes, infermers i biòlegs de l'Hospital Universitari Institut Pere Mata (HUIPM), molts d'ells doctors i professors associats de la Facultat de Medicina i Ciències de la Salut de la Universitat Rovira i Virgili (URV). La responsable del GAP és la Dra. Elisabet Vilella, directora de l'Àrea de Recerca de l'HUIPM i subdirectora de l'IISPV.

L'IISPV (https://www.iispv.cat) va ser fundat a l'any 2005 a partir del conveni de col·laboració científica interinstitucional entre l'Institut de la Salut Camp de Tarragona (Hospital Universitari Joan XXIII de Tarragona, l'àrea d'Atenció Primària Camp de Tarragona), el Grup SAGESSA (Hospital Universitari Sant Joan de Reus i l'àrea d'Atenció Primària SAGESSA), el Grup Pere Mata (Hospital Psiquiàtric Universitari Institut Pere Mata, Sanatori Villablanca, Fundació Pere Mata i Fundació Villablanca), l'Institut Català de la Salut Terres de l'Ebre (Hospital de Tortosa Verge de la Cinta) i la URV. La seva activitat consisteix en gestionar i centralitzar la investigació sanitària i biomèdica del Camp de Tarragona i les Terres de l'Ebre. Té com a objectius principals promoure, desenvolupar i difondre la investigació, la formació en les ciències de la vida i la salut, centrant-se actualment en quatre àmbits de investigació: Malalties metabòliques i nutrició; infecció, immunitat i medi ambient; oncologia; i neurociències i salut mental (https://www.iispv.cat/recerca/).

L'IISPV compta amb més de 400 investigadors i amb una important col·laboració internacional. A més, disposa de diverses plataformes de suport per a la recerca entre les quals hi ha la Plataforma de Metabòlica, de Suport Estadístic, de Cultius Cel·lulars, el Biobank, un Servei de Bioinformàtica, i la Unitat d'Estudis Clínics.

Aquest treball de fi de grau es basa en un projecte competitiu (PI18/00514) finançat per l'Instituto de Salud Carlos III "Implicació de l'ADN mitocondrial en les psicosis primerenques: relació amb el risc de malaltia, estrès, rendiment cognitiu, simptomatologia clínica i síndrome metabòlica" que tracta d'entendre millor la implicació de les anomalies presents a l'ADN mitocondrial (ADNmt) amb l'aparició de l'esquizofrènia. La investigadora principal (IP) és la Dra. Lourdes Martorell Bonet, qui va dirigir les meves pràctiques a la Facultat de Medicina i Ciències de la Salut de la URV i l'IISPV i el co-IP és el Dr. Gerard Muntané Medina.

Resum i paraules clau

L'esquizofrènia és un trastorn del neurodesenvolupament que es produeix per la interacció de factors genètics i ambientals. Tot i ser una afectació coneguda, encara és un repte entendre en detall els factors que intervenen en el seu desenvolupament.

Actualment, és coneixen alguns factors genètics que intervenen en l'aparició de les psicosis primerenques i més tard l'esquizofrènia. Entre aquests factors genètics hi ha les variants de número de còpia (CNVs, de l'anglès *Copy Number Variants*), els polimorfismes d'un únic nucleòtid (SNPs, de l'anglès *Single Nucleotide Polymorphisms*), les variants rares d'un únic nucleòtid (SNVs, de l'anglès *Single Nucleotide Variants*) i també algunes alteracions epigenètiques. Tot i això, existeixen certs indicis de la implicació de les alteracions de l'ADN mitocondrial (ADNmt) en l'esquizofrènia, encara que els coneixements relacionats amb aquest camp són força limitats.

Els mitocondris són orgànuls membranosos propis de les cèl·lules eucariotes. La seva principal funció és generar energia en forma d'ATP a partir de la fosforilació oxidativa mitjançant la cadena de transport electrònic i l'ATP sintasa. És un orgànul fonamental per al correcte funcionament de la majoria de tipus cel·lulars i òrgans però, la seva funció és especialment important per al manteniment de l'estructura i la funció cerebral, ja que, produeixen el 90% de l'energia requerida per les cèl·lules neuronals. Degut al seu vital paper en gairebé tots els òrgans del cos humà, les alteracions genètiques, morfològiques i funcionals en aquests orgànuls poden provocar malalties mitocondrials molt severes que disminueixin considerablement la qualitat de vida de les persones que les pateixin.

Les alteracions genètiques mitocondrials consisteixen en delecions, variants d'un únic nucleòtid, polimorfismes d'un nucleòtid i variacions en el número de copies però, encara no es coneix amb el suficient detall el seu paper en l'aparició de l'esquizofrènia.

És per aquest motiu que el grup de recerca estudia la implicació d'aquesta molècula en l'inici dels trastorns psicòtics. Aquest treball de fi de grau té per objectiu l'estudi de les delecions de l'ADNmt i ha analitzat 79 mostres cerebrals postmortem de 39 persones amb diagnòstic d'esquizofrènia i 40 persones sanes com a controls mitjançant la seqüenciació de nova generació i l'eina computacional *eKLIPse*.

Com objectiu secundari, s'han comparat dues tècniques de seqüenciació, *Ion Torrent* i *Ilumina*, per a determinar quina és la més adequada per a utilitzar en aquest tipus d'anàlisis.

Paraules clau: ADNmt, deleció, *eKLIPse*, esquizofrènia, *FastQC, Ilumina, Ion Torrent,* malalties mitocondrials, mitocondri, psicosi primerenca.

1. Introducció

1.1. Esquizofrènia

L'esquizofrènia és un trastorn psiquiàtric greu que afecta el desenvolupament neurològic i es caracteritza per la presència d'una combinació de símptomes positius, negatius i alteracions cognitives. Actualment es coneix que l'esquizofrènia és deguda a factors genètics que, a l'interaccionar amb diversos factors ambientals, poden originar una gran diversitat de simptomatologies i diferències en quant a l'inici, la presentació i el progrés de la malaltia (Khavari & Cairns, 2020).

Segons la Organització Mundial de la Salut (OMS), aproximadament 24 milions de persones en el món presenten esquizofrènia, amb una prevalença al voltant de l'1% de la població.

Es coneix que la incidència d'aquest trastorn és significativament superior en individus que presenten familiars afectats, en comparació a la resta de la població. Per exemple, infants amb un pare afectat presenten un 17% de risc de desenvolupar esquizofrènia en algun moment de la seva vida, mentre que en els casos en què els dos progenitors presenten el diagnòstic, el risc augmenta fins al 35% (Khavari & Cairns, 2020).

Segons l'Institut Nacional de Salut Mental dels Estats Units (NIH) i l'Organització Mundial de la Salut (OMS), l'esquizofrènia generalment es diagnostica entre els 16 i 30 anys, després d'un primer episodi psicòtic. No acostuma a presentar-se en nens petits, encara que existeixen alguns casos. Per altra part, aquest trastorn acostuma a manifestar-se abans en els homes que en les dones.

1.2. Simptomatologia de l'esquizofrènia

Entre els símptomes o manifestacions fenotípiques més representatives de l'esquizofrènia es troben els símptomes positius, que inclouen al·lucinacions (ja siguin visuals, auditives, olfactives, tàctils o relacionades amb els sabors), deliris, trastorns en la forma del pensament i del moviment. Per altra part, es poden presentar símptomes propis de la depressió com l'anhedonia, l'aïllament social i la falta d'expressivitat, que conformen els símptomes negatius. A més, també és possible observar en pacients amb aquest trastorn una sèrie d'alteracions a

nivell cognitiu com la manca d'atenció, la dificultat en l'aprenentatge i la resolució de problemes i una disminució en la memòria de treball (Khavari & Cairns, 2020).

1.3. Etiologia

El cervell és un òrgan de regulació complexa, de manera que existeixen molts factors tant genètics com ambientals, susceptibles de causar alteracions sistemàtiques, estructurals i/o funcionals que poden provocar una gran varietat de síndromes conductuals i cognitius entre els que es troba l'esquizofrènia (Khavari & Cairns, 2020).

1.3.1. Factors genètics implicats (SNPs i CNVs)

Les variants de número de còpia (CNVs) consisteixen en alteracions cromosòmiques estructurals i suposen la major font de variabilitat genètica, coneguda actualment. Impliquen pèrdues o guanys de grans fragments d'ADN que produeixen duplicacions o delecions de més de 1.000 pb. Tot i que les CNVs individuals es trobin en freqüències baixes (0,5%), poden provocar efectes greus en la salut i mostrar un comportament pleiotròpic per a una varietat considerable de trastorns neuropsiquiàtrics que comparteixen la seva etiologia. La primera CNV que es va associar a l'esquizofrènia va ser la deleció 22q11.2. Més tard, els estudis en models animals i cel·lulars de CNVs neuropsiquiàtrics i els estudis gènics van aportar més informació rellevant en quant als mecanismes causants d'aquests trastorns (Rees & Kirov, 2021).

Actualment, es coneixen 12 CNVs de risc per a l'aparició d'esquizofrènia però, s'especula que se'n trobaran més a mesura que s'analitzin més mostres. Un aspecte interessant és que la majoria d'aquests CNVs incrementen el risc de patir també altres trastorns neuropsiquiàtrics diferents de l'esquizofrènia, i afectacions no psiquiàtriques com la diabetis o la hipertensió. Els principals gens implicats en aquestes CNVs són els relacionats amb els complexes sinàptics de proteïnes associades al citoesquelet i N-Metil-D-àcid aspàrtic, la senyalització de GABA i glutamat i els canals de calci depenents del voltatge (Rees & Kirov, 2021).

En quant als SNPs, segons l'estudi de mapatge genòmic realitzat per (Trubetskoy et al., 2022) es van identificar 628 gens (435 codificants per a proteïnes) que presentaven al menys un SNP potencialment involucrat amb l'aparició de l'esquizofrènia. Es va evidenciar la presència de rs4766428, amb una probabilitat posterior major del 99% (PP>0.99) de ser el causant de la formació d'un SNP present en un locus que engloba 25 gens, es troba en la regió codificant per a *ATP2A2* i provoca la malaltia de Darier, que està associada al trastorn bipolar i l'esquizofrènia. Es va determinar que *ATP2A2* pot estar implicat en la patogènesis de l'esquizofrènia regulant

els nivells citoplasmàtics neuronals de calci. Per altra part, es va observar que la predisposició genètica per a l'esquizofrènia és idèntica entre homes i dones, tot i les evidències de la variabilitat en quant a l'edat d'inici, els símptomes i el desenvolupament del trastorn respecte el sexe (Trubetskoy et al., 2022).

1.3.2. Factors ambientals i canvis epigenètics cerebrals associats

Encara que l'esquizofrènia presenti un fort component genètic, amb una heretabilitat del 80%, existeixen certs factors d'exposició ambiental que poden intervenir en el seu desenvolupament i neuropatologia, com les infeccions immunes maternes, el trauma infantil, les complicacions durant l'embaràs i/o el part, la falta de nutrients i l'exposició al cànnabis o toxines. Gràcies a un llarg historial d'estudis en diferents models animals es postula que aquests factors de l'ambient poden intervenir en diverses modificacions epigenètiques que s'acumularan al cervell i altres teixits durant el desenvolupament de l'organisme i acabaran causant canvis en l'expressió gènica. (Khavari & Cairns, 2020). Aquesta especulació coincideix amb la teoria del neurodesenvolupament de l'esquizofrènia, que és la més acceptada en l'actualitat per a explicar l'aparició d'aquest trastorn. Aquesta teoria presenta al component hereditari i l'exposició ambiental durant el desenvolupament (especialment prenatal, infància primerenca i adolescència) com a causants de l'esquizofrènia. Tot i així, és necessari tenir en compte que els canvis epigenètics es poden manifestar com a resultat d'un estat patològic i no ser necessàriament el seu causant (Khavari & Cairns, 2020).

Gràcies als anàlisis neurofisiològics i neuropatològics s'ha observat que l'esquizofrènia sovint es relaciona amb una deficiència en quant a l'estructura cerebral i la connectivitat funcional. Aquesta pèrdua podria ser causada per una anomalia dels mecanismes epigenètics al rebre senyals ambientals a través de l'activitat neuronal o altres, provocant canvis en la modulació de l'expressió i resposta gènica (Khavari & Cairns, 2020).

Els principals mecanismes epigenètics que mostren certes evidencies d'estar relacionats amb l'aparició de l'esquizofrènia i la severitat dels seus símptomes són les metilacions, les modificacions d'histones, els microRNA (miRNA) i els RNAs llargs no codificants (lncRNA). Entre els seus efectes més freqüents es troba l'alteració de la neurotransmissió de GABA (component important per la neuropatologia de l'esquizofrènia) sovint observada en dones, la metabolització de la dopamina (important regulador dels símptomes positius de la esquizofrènia) com també diferencies en quant a l'expressió de diversos gens implicats en el desenvolupament del sistema nerviós central, les funcions sinàptiques i les connectivitats

neuronals. Generalment, aquests efectes s'observen al còrtex prefrontal dorsolateral i el còrtex del cingle anterior, alguns d'aquests intervenint en diferents processos metabòlics, la mielinització, l'embolcallament neuronal i la diferenciació d'oligodentrócits (Khavari & Cairns, 2020).

1.4. Els mitocondris

1.4.1. Funció mitocondrial general

Els mitocondris són orgànuls intracel·lulars que es troben en totes les cèl·lules nucleades. Resulten indispensables per a una gran varietat de funcions dins de la cèl·lula eucariota. Són responsables de la síntesis del 95% de l'ATP cel·lular a través de la fosforilació oxidativa produïda per la cadena de transport d'electrons i l'ATP sintasa, pel que són els principals productors d'energia de l'organisme (Schlieben & Prokisch, 2023). Produeixen el 90% de l'energia de les cèl·lules neuronals (Das et al., 2022). D'altra banda, també intervenen en l'amortiment de calci, la modulació de l'activitat sinàptica, la regulació de l'apoptosi (inicien l'apoptosi depenent de la caspasa), la regulació d'espècies reactives d'oxigen, la biogènesi de grups de ferro i sofre, el metabolisme de lípids, ferroptosi i el metabolisme d'aminoàcids (Schlieben & Prokisch, 2023).

La cadena de transport d'electrons està composada per 4 Complexes proteics localitzats a la membrana mitocondrial interna (Complexes I, II, III, IV). Aquestes Complexes proteics s'encarreguen de bombejar protons a l'espai intermembranós mentre que els electrons són transferits d'un Complex a l'altre oxidant diverses molècules per a finalment produir aigua. La diferència de potencial és aprofitada per l'ATP sintasa per fosforilar una molècula d'ADP i produir ATP. Part de les subunitats proteiques de la cadena respiratòria i de l'ATP sintasa es troben codificades en el genoma mitocondrial (ADNmt), tot i que la majoria estan codificades pel genoma nuclear (ADNn), que són traduïdes al nucli i transportades fins al mitocondri (Valiente-Pallejà et al., 2022)

1.4.2. Genètica mitocondrial

Els mitocondris presenten un ADN doble cadena circular, poliploide, sense introns i d'herència materna. S'anomena cadena pesada a la rica en guanines i cadena lleugera a la rica en citosines. L'ADNmt conté un total de 16.569 parelles de bases que formen 37 gens i codifiquen per a 13 proteïnes, 22 ARN de transferència (ARNt) i 2 ARN ribosòmics 12S i 16S (ARNr) (Valiente-Pallejà et al., 2022). Es codifiquen 1 polipèptid i 8 ARNt a la cadena lleugera i 12 polipèptids, 2 ARNr i 14 ARNt a la cadena pesada. Existeix també una regió no codificant amb el nom de

bucle de desplaçament o bucle D en la qual hi ha els orígens de replicació i transcripció de gairebé tot el mtDNA (Yan et al., 2019). La distribució dels gens codificats pel ADNmt es troba descrita a la Figura 1.

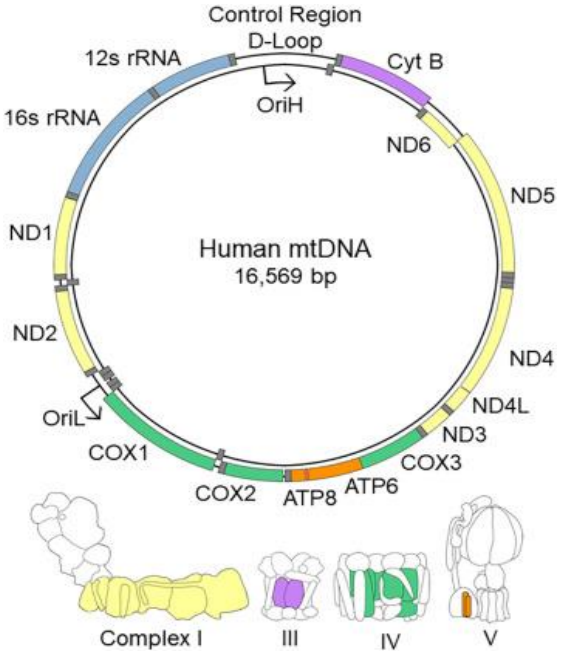


Figura 1: Esquema de l'ADNmt humà i els gens codificants per als Complexes I, III, IV i V. (Sandra et al., 2018)

L'ADNmt s'empaqueta formant complexes ADNmt-proteïna que s'anomenen nucleoides. La proteïna principal del nucleoide és el factor A de transcripció mitocondrial (TFAM), que té una important implicació en la distribució i organització del nucleoide. La distribució dels nucleoides dins del mitocondri resulta indispensable pel correcte funcionament d'aquest orgànul, pel que la presència d'alteracions en aquesta organització poden provocar diferents malalties (Yan et al., 2019).

Per altra banda, les 13 proteïnes codificades per ADNmt, constitueixen part dels Complexes I, III, IV i V de la cadena de transport d'electrons i l'ATP sintasa, la síntesis dels quals depèn dels ARN ribosòmics (ARNr) i ARN de transferència (ARNt) mitocondrials (Yan et al., 2019).

A més, cal esmentar que el genoma nuclear codifica aproximadament per uns 1500 gens que tenen algun tipus de implicació en el funcionament i la localització del mitocondri (Das et al., 2022).

Cada mitocondri conté de mitjana de 2 a 10 còpies d'ADN mitocondrial que es troben a la matriu, prop de la cadena de transport d'electrons i no estan protegits per histones, pel que son susceptibles a ser modificats per radicals lliures (que provenen sobretot de la cadena de

transport electrònic) i desenvolupar mutacions somàtiques (Yan et al., 2019). La taxa de mutacions de l'ADNmt en humans és 10 vegades superior en comparació al nuclear, probablement degut a aquests factors (Valiente-Pallejà et al., 2022).

Cada cèl·lula pot presentar de centenars a milers de còpies del ADNmt depenent de les seves necessitats energètiques. El número de còpies mutades pot variar al llarg del temps, el que implica que és possible trobar cèl·lules on totes les còpies estiguin mutades o que ninguna presenti cap mutació, fenomen que es coneix amb el nom d'homoplàsmia. Pel contrari, en el cas de que part de les molècules de l'ADNmt presentin mutacions i una altra no, s'anomena heteroplàsmia (Yan et al., 2019).

1.4.3. Malalties mitocondrials

La quantitat de malalties que s'han considerat de tipus mitocondrial ha estat creixent de forma constant a mesura que s'ha anat descobrint més sobre les anomalies genètiques i els mecanismes patològics relacionats amb el metabolisme energètic mitocondrial. Degut a la distribució dels gens d'aquest orgànul, les malalties mitocondrials poden ser causades tant per alteracions patològiques del genoma mitocondrial com del nuclear, pel que poden seguir qualsevol patró d'herència. S'han associat aproximadament uns 425 gens amb diferents patologies mitocondrials, 90 d'aquests intervenint en malalties neurodegeneratives infantils, des de que es va descobrir al primer en 1988 (Schlieben & Prokisch, 2023).

De forma general, les malalties mitocondrials consisteixen en un grup heterogeni de malalties que coincideixen en el fet d'estar causades per una disfunció mitocondrial, sobretot per una fosforilació oxidativa incorrecta. Aquestes patologies es poden presentar en forma d'un ampli espectre clínic caracteritzat per símptomes que poden tenir lloc en qualsevol edat, afectant qualsevol òrgan i teixit i manifestant alteracions a múltiples sistemes. Tot i la gran variabilitat inherent a aquestes malalties, cal destacar que predominen les afectacions causades als òrgans estretament relacionats amb el metabolisme aeròbic. També és important considerar la complexitat gènica darrere d'aquestes afeccions, fent que el seu diagnòstic molecular representi un autèntic repte. Variants en un únic gen i, fins i tot, la mateixa mutació dins d'un gen pot provocar diferents malalties mitocondrials, evidenciant d'aquesta manera fenòmens d'heterogeneïtat al·lèlica. Alguns exemples són les mutacions en gens codificants com *MT-ND5* o *MT-CO3*, responsables de la miopatia mitocondrial, encefalopatia, acidosi làctica i episodis semblants a l'ictus (MELAS), la Neuropatia Óptica Hereditària de Leber (LHON), el síndrome de l'epilèpsia mioclònica amb fibres vermelles esquinçades (MERRF) i la malaltia

de Leigh. Per altra part, un mateix fenotip patogènic pot ser causat per mutacions en diferents gens, indicant també la possibilitat d'una heterogeneïtat de locus. Alguns exemples són el síndrome de Leigh (Schlieben & Prokisch, 2023).

Altres mutacions al genoma mitocondrial es relacionen amb malalties com la diabetis, la malaltia d'Alzheimer, el Parkinson i el càncer. És important considerar que alguns estudis proposen la possibilitat d'acumulació de les mutacions al ADNmt amb el temps, de manera que presenten una important implicació en l'envelliment i la neurodegeneració que moltes vegades acompanya aquest procés. Existeixen moltes evidències que defensen aquesta idea i mostren que una dinàmica mitocondrial irregular i les mutacions causades durant la replicació del seu genoma porten a fenòmens d'envelliment. A més, s'ha observat un elevat nombre de delecions a l'ADNmt present a les cèl·lules neuronals de la substancia negra cerebral de pacients ancians amb Parkinson, que es caracteritza per la pèrdua de dopamina en aquesta regió. Per altra part, la malaltia d'Alzheimer es relaciona amb mutacions heteroplàsmiques de l'ADNmt (Yan et al., 2019). L'heteroplàsmia pot aparèixer degut a mutacions puntuals espontànies en teixits específics o per delecions múltiples provinents de variants patogèniques heretades de gens nuclears que intervenen en el manteniment de l'ADNmt (Goudenège et al., 2019).

1.4.4. Anomalies mitocondrials a l'esquizofrènia

Existeixen diversos estudis que han relacionat diferents alteracions mitocondrials amb l'aparició i l'augment del risc de patir esquizofrènia. A continuació, es mostren algunes de les anomalies més rellevants en relació a aquest trastorn:

1.4.4.1. Delecions

Les delecions dins de l'ADNmt normalment tenen lloc entre regions flanquejades per seqüències curtes repetides. Per tant, es pot distingir entre les delecions de Classe I, que estan flanquejades per repeticions directes perfectes que conformen el 60% dels casos d'esquizofrènia i les delecions de Classe II, que es troben flanquejades per repeticions imperfectes i representen el 30% dels casos (Goudenège et al., 2019).

Una deleció comuna de l'ADNmt és la deleció somàtica de 4.977 parells de bases, la qual s'ha trobat en adults però no en teixits fetals. La quantitat delecionada varia segons la localització cerebral, sent més elevada als nuclis de dopamina, entre altres. La majoria d'estudis no mostren canvis significatius en quant a aquest tipus de deleció entre els pacients amb esquizofrènia i els grups control. Tot i així, s'ha observat que conté gens codificants per a les subunitats del

citocrom oxidasa, NADH deshidrogenasa i ATP sintetasa, que són essencials pel funcionament mitocondrial i la patologia de l'esquizofrènia, pel que s'esperaria que tinguessin algun tipus d'implicació en aquest trastorn, però els resultats no demostren aquesta teoria (Roberts, 2021).

Per altra banda, alguns estudis relacionen la reducció de l'activitat, la quantitat de proteïna i/o d'RNA missatger dels Complexes I, II i IV amb l'aparició o severitat de l'esquizofrènia. L'activitat dels Complexes no canvia uniformement per a tots els nuclis, un exemple és l'observació d'una reducció de l'activitat del Complex IV i un increment de la del Complex II al putamen i al nucli accumbens en teixits post-mortem de pacients d'esquizofrènia (Roberts, 2021).

A més, l'activitat del Complex I sembla estar afectada per la ingesta de drogues, mentre que la del II s'ha associat a la severitat dels símptomes (Roberts, 2021).

No s'han trobat diferències significatives en quant a l'activitat dels diferents Complexes entre els pacients control i els que presenten esquizofrènia, pel que s'ha proposat un possible mecanisme de compensació per a restaurar l'activitat del citocrom c oxidasa tot i que hi hagi una manca en l'activitat de la subunitat II (Roberts, 2021).

Per últim, segons estudis previs realitzats pel grup d'investigació al que s'ha dut a terme aquest estudi, existeixen certs indicis d'un increment significatiu en quant al nivell d'insercions i delecions trobades en mostres d'ADNmt de pacients d'esquizofrènia (Valiente-Pallejà et al., 2022).

1.4.4.2. Polimorfismes

Segons l'estudi de (Ivanova et al.,2021) existeixen evidencies de que tant les mutacions puntuals de substitució de base com les delecions i insercions mostren correlació amb els trastorns psicòtics. El polimorfisme 1811A>G (*MT-RNR2*) és el que es mostra amb major freqüència en l'esquizofrènia (24,3%), en comparació amb els grups controls (4,3%) (p=0,07). També destaquen certs polimorfismes localitzats sobretot als gens codificants pel Complex I *MT-ND4* (11251G i 11467G), *MT-ND3* (10398G), *MT-ND1* (4216C), i *MT-ND5* (12611G i 13708A), alguns d'aquests implicats en la disfunció mitocondrial.

1.4.4.3. Número de còpies

En varis estudis com els de (Kumar et al., 2018) i (Das et al., 2022) s'han pogut realitzar estudis i reportar certes diferències en quant al número de còpies de l'ADNmt en pacients amb esquizofrènia, el trastorn bipolar i els trastorns psicòtics respecte a grups control sans. En

l'estudi de (Das et al., 2022) s'han analitzat mostres del còrtex dorsolateral prefrontal, del gir temporal superior, del còrtex visual primari i del nucli accumbens de cervells postmortem. Entre els seus resultats destaca un número de còpies de l'ADNmt significativament superior en el còrtex prefrontal dorsolateral de les mostres amb esquizofrènia i trastorn bipolar, nivells inferiors en el gir temporal superior de les mostres amb trastorn bipolar, i cap canvi significatiu al còrtex visual primari i el nucli accumbens, tant de les mostres amb esquizofrènia com les que presentaven trastorn bipolar respecte als grups control.

Per altra part, s'han reportat evidències d'una disminució en quant al número de còpies d'ADNmt en mostres de sang de teixit perifèric tant en pacients amb esquizofrènia com amb trastorn bipolar, encara que també es relaciona amb l'envelliment (Das et al., 2022).

Un possible causant de l'alteració de la quantitat de l'ADNmt és també el tractament amb antipsicòtics i altres fàrmacs. És per aquest motiu que (Das et al., 2022) van realitzar una sèrie d'anàlisis toxicològics a les mostres de teixit cerebral postmortem amb la finalitat d'agrupar-los en funció de si es troba o no la presència de drogues psicotròpiques o medicació. Es va concloure que, tot i que no es podia desestimar la possibilitat de que els resultats es vegin afectats per la prèvia exposició a la medicació, la seva presència en el moment de la mort no provoca canvis significatius al número de còpies del ADNmt al còrtex dorsolateral prefrontal, al gir temporal superior, al còrtex visual primari i al nucli acumbens. Aquests resultats suggereixen la possibilitat de que el nombre de còpies de l'ADNmt present en aquestes regions cerebrals sigui més estable davant dels canvis ambientals i els fàrmacs que el que es troba en circulació (Das et al., 2022).

Tot i els descobriments mencionats, fins ara, no es coneix amb seguretat si el canvi en el número de còpies de l'ADN mitocondrial està relacionat amb les psicosis, la comorbiditat o el tractament (Kumar et al., 2018).

En referència a l'estudi de (Valiente-Pallejà et al., 2020) es va observar una concentració significativament menor de ADNmt en mostres sanguínies dels participants amb esquizofrènia en comparació al grup control. En aquesta investigació, es va trobar una correlació entre la disminució del numero de copies d'ADNmt i l'envelliment, el consum d'alcohol i tabac, la inflamació i la severitat dels símptomes psicòtics, suggerint que podria actuar com a marcador patofisiològic per a l'esquizofrènia (Valiente-Pallejà et al., 2020).

2. Hipòtesi de treball i objectius

La hipòtesi d'aquest estudi és que, en mostres de cervell postmortem, els pacients amb diagnòstic d'esquizofrènia presenten un % d'heteroplàsmia de l'ADNmt significativament superior als individus del grup control.

Els objectius principals de la investigació són:

1) Identificar i quantificar les alteracions de mostres d'ADN de cervell postmortem de pacients amb diagnòstic d'esquizofrènia i en individus control.

2) Comparar el número d'alteracions entre els dos grups d'estudi: pacients amb diagnòstic d'esquizofrènia i individus grup control.

3) Descriure si les alteracions de l'ADNmt han estat prèviament reportades.

4) Comparar els resultats obtinguts amb dues metodologies de seqüenciació diferents, *Ion Torrent* i *Illumina.*

3. Participants, materials i mètodes

3.1. Participants

Per aquest estudi es disposa de 79 mostres de cèl·lules de còrtex dorsolateral prefrontal de cervells postmortem pertanyents a 39 pacients amb esquizofrènia i 40 pacients control sense aquest trastorn. L'ADN de còrtex dorsolateral prefrontal i les dades clíniques van ser obtingudes pel Biobanc del País Basc, Centro de Investigación Biomédica en Red en Salud Mental (CIBERSAM). Els pacients amb esquizofrènia escollits per a l'estudi van ser els que presentaven una causa de mort per suïcidi i dels que s'havia verificat que tenien trastorns psicòtics previs segons els criteris del Manual diagnòstic i estadístic dels trastorns mentals (DSM-IV o DSM-IV-TR). El grup control està conformat per mostres cerebrals de pacients amb una causa de mort no provocada per suïcidi i que no presentaven trastorns psicòtics. Es disposa de informació rellevant de cada pacient referent a la data i causa de mort, el temps transcorregut entre l'hora de la mort i el processament del cervell per obtenir la mostra d'ADN interval post-mortem (PMD, "*postmortem delay*" en anglès), medicació prèvia i consum de tòxics mesurats a partir de mostres de sang i del pH del cervell. La informació corresponent al número de participants de cada grup i la seva classificació per sexes, edats període transcorregut entre la seva mort i l'obtenció de la mostra es troba en la Taula1.

Entre els criteris d'inclusió relacionats amb els pacients amb esquizofrènia es considera que aquests presentin un trastorn psicòtic. Per tant, han de complir els criteris del DSM-5 de trastorn esquizofreniform, esquizofrènia, trastorn esquizoafectiu o trastorn psicòtic no especificat.

A més, el projecte ha d'estar aprovat per un Comitè d'Ètica, que ha d'haver assegurat també que les mostres utilitzades estiguin recollides al Biobanc del País Basc.

Per altra banda, entre els criteris d'exclusió figura el no compliment d'algun dels criteris d'inclusió mencionats anteriorment.

Taula 1: Característiques descriptives de les mostres analitzades (HC: Healthy Control, SCZ: esquizofrènia)

	Grup	Sexe	Nombre total	Mitjana	Mediana	Desviació Estàndard	Mínim	Màxim
Edat	HC	Masculí	33	47,2	46,0	15,47	23,00	84,0
		Femení	7	53,0	52,0	14,31	29,00	74,0
	SCZ	Masculí	32	47,8	49,0	15,77	23,00	83,0
		Femení	7	53,4	52,0	14,79	30,00	76,0
PMD	HC	Masculí	33	19,4	17,0	10,75	4,00	60,0
		Femení	7	20,4	19,0	5,19	16,00	31,0
	SCZ	Masculí	32	20,3	17,0	13,16	3,00	52,0
		Femení	7	15,4	15,0	3,99	10,00	22,0

3.2. Amplificació i seqüenciació

L'ADNmt de cada participant es va amplificar utilitzant dues parelles de primers que produïen dos fragments solapats. Seguidament, es va realitzar la seqüenciació de les mostres amb la tecnologia d'Illumina d'extrems parells.

3.2.1. Tècniques de seqüenciació de nova generació

Les tecnologies de seqüenciació de nova generació són mètodes de seqüenciació d'ADN d'alt rendiment que van adquirir gran importància des del 2004, quan van ser introduïdes per primera vegada per Roche amb l'equip 454 FLX Pyrosequencer (Larson et al., 2023).

La millora respecte el seu antecedent de primera generació del 1970 (Sanger), és que permeten la seqüenciació massiva d'una gran quantitat de fragments d'ADN, suposant l'equivalent d'analitzar milions d'experiments individuals de Sanger simultàniament. Aquest avanç permet una ràpida seqüenciació de genomes sencers per un cost bastant baix, el que permetria la introducció de la genòmica personalitzada i la medicina de precisió. Encara que, la gran quantitat de informació aportada per aquest tipus de tècniques necessiten la disposició de grans

i sofisticats recursos computacionals i bioinformàtics per a poder obtenir uns resultats adequats i informatius (Larson et al., 2023).

3.2.1.1. Illumina

Illumina és la companyia líder de diversos instruments de seqüenciació. La seva metodologia es basa en la seqüenciació per síntesi (SBS, "*sequencing by synthesis*" en anglès) (Pervez et al., 2022). Aquesta tècnica consisteix en carregar una llibreria de seqüències sobre una superfície sòlida recoberta d'una gran quantitat de petits oligòmers complementaris a les seqüències adaptadores dels fragments de la llibreria. Aquest mètode permet que en cada carril de la cel·la on es carreguen les mostres es permeti la realització de diferents experiments de seqüenciació. Després de carregar els fragments, aquests s'amplifiquen per PCR, originant grups de clons de seqüències de milers de còpies d'ADN. Per evitar que la polimerasa sobrepassi la seqüència motlle a l'hora d'estendre la complementaria, s'utilitzen didesoxinucleòtids marcats amb fluorescència que es detecten per un làser d'autofocus que permet reclutar les bases de la seqüència complementària. Les terminacions són reversibles, el que permet el control de la síntesi (Larson et al., 2023) i generar fragments de la mateixa longitud (en el cas de que corresponguin del mateix experiment) (Pervez et al., 2022).

La seqüenciació pot ser d'extrems únics o parells, que fa referència a les opcions de seqüenciar un o els dos extrems de l'insert. La seqüenciació d'extrems parells acostuma a ser la més utilitzada per a la majoria d'estudis, ja que ofereix una millor precisió del mapatge de lectura respecte a la d'extrems únics, un augment de la cobertura i la possibilitat de detectar reordenaments del genoma, com per exemple la fusió de gens (Larson et al., 2023).

3.2.1.2. Ion Torrent

Ion Torrent és una altra tècnica de seqüenciació de nova generació que es va introduir al 2011 (Pervez et al., 2022) i és el principal competidor d'*Illumina* (Pereira et al., 2020). També es basa en la SBS i utilitza mesures de pH per a generar seqüències nucleotídiques, que poden variar en quant a la seva longitud (Pervez et al., 2022). Es basa en els canvis de pH provocats per l'alliberament de protons (H^+) durant la polimerització de la seqüència d'ADN. Els inserts s'adhereixen a una sèrie d'esferes o "*beads"* i s'amplifiquen per PCR, resultant en esferes amb còpies d'un mateix fragment d'ADN. A continuació, les esferes es dispersen en micropous amb chips de matriu de sensors semiconductors. Cada nucleòtid de la cadena incorporat per la polimerasa a la cadena creixent provoca l'alliberament d'un protó, que alterarà el pH de la

solució. Aquests canvis es detecten pel sensor iònic incorporat al chip i es transformen en senyals de voltatge que després s'utilitzen pel reclutament de bases (Pereira et al., 2020).

El principal inconvenient d'aquesta tècnica és la seva ineficàcia per a la quantificació d'homopolímers de gran longitud. Aquest aspecte és degut a què les incorporacions múltiples d'una mateixa base a cada cadena provoca un augment important de la concentració de protons, el que genera una major senyal de incorporació i implica que s'afegirà més d'un nucleòtid a la vegada (Pereira et al., 2020).

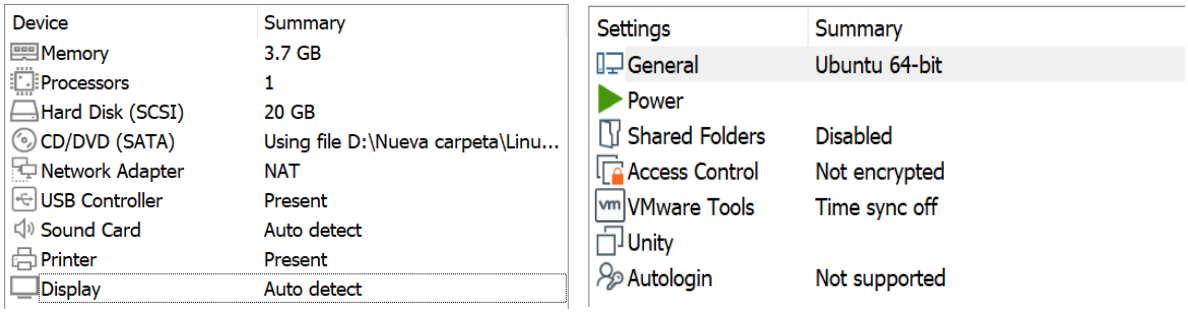
3.3. Eines Bioinformàtiques utilitzades

3.3.1. Màquina Virtual de Linux

Tots els anàlisis han estat realitzats en una màquina virtual de *Linux Lite 5.0* per assegurar el funcionament senzill i òptim dels softwares utilitzats. Per a construir la màquina virtual s'ha fet servir *VMware Workstation 17 Player* (https://www.vmware.com/products/workstation-player/workstation-player-evaluation.html), que s'ha mostrat més eficaç que l'*Oracle VM VirtualBox* i presenta menys problemes tècnics associats.

3.3.1.1. Creació de la màquina virtual

Un cop s'instal·li el programa *VMware Workstation 17 Player*, es podrà construir la màquina virtual seguint les característiques de la Imatge 1:



Imatge 1: Característiques bàsiques de la màquina virtual de Linux 5.0 Lite

A continuació, després, d'instal·lar *Linux 5.0 Lite* correctament a la màquina virtual s'han d'instal·lar la resta de paquets, actualitzacions i adicions de convidat *(gcc, make, libbz2-dev, zlib1g-dev, libncurses5-dev, libncursesw5-dev, liblzma-dev).*

Un cop preparat el sistema operatiu, és necessari instal·lar *Python (2.7.16), Biopython (1.76), HTSLIB 1.9, SAMTOOLS 1.9* i *BCFTools 1.9.*

3.3.2. *FastQC*

FastQC és un programa dissenyat pel control de qualitat creat per Simon Andrews de Babraham Bioinformatics per a analitzar dades de seqüenciació d'alt rendiment utilitzant el llenguatge de Java. El seu objectiu és realitzar de forma ràpida una sèrie d'anàlisis de qualitat sobre les dades de seqüenciació per informar dels possibles problemes que s'haurien de considerar abans de seguir amb l'estudi (Babraham Bioinformatics, 2023).

De forma general, *FastQC* té la capacitat de revisar dades en format *bam, sam* o *fastq* per a informar de quines regions de la seqüència poden presentar problemes, ja siguin originats pel seqüenciador o la pròpia llibreria d'on s'obtenen les mostres. Els resultats dels anàlisis es presenten en gràfiques i taules tant en format d'imatges com a partir d'un espai HTML on es recull de forma clara i comprensible tota la informació obtinguda (Babraham Bioinformatics, 2023).

Els resultats que retorna el software inclouen:

Estadística bàsica: Reporten dades com el nom i tipus d'arxiu, el número de seqüències processades, la longitud de les lectures, el contingut de guanina i citocina i el tipus de puntuació de qualitat (Babraham Bioinformatics, 2023).

Qualitat de seqüència per base: Mostra el rang dels valors de qualitat de totes les bases i per cada posició de la seqüència analitzada. Presenta una taula amb les puntuacions mitjanes, la qualitat mitjana i el rang interquartil. Classifica les puntuacions en grups de molt bona qualitat (verd), qualitat raonable (taronja) i baixa qualitat (vermell), de manera que les puntuacions més elevades corresponen a una major qualitat. És comú que a mesura que avanci el procés la qualitat baixi progressivament, pel que al final de la taula normalment les lectures entren a l'àrea taronja (Babraham Bioinformatics, 2023).

Puntuacions de qualitat per seqüència: Estudia si existeix algun subconjunt de les seqüències analitzades que mostri predominantment valors de qualitat baixos. Moltes vegades es provocat per errors de visualització durant la seqüenciació i només un nombre reduït de seqüències haurien de presentar aquest problema. Si pel contrari, s'observa una elevada proporció de seqüències amb una baixa qualitat general és una indicació d'algun tipus de problema sistemàtic que sovint nomes afecta a una part del gràfic (Babraham Bioinformatics, 2023).

Contingut de seqüència per base: Presenta la proporció de les posicions de cadascuna de les quatre bases de la seqüència de DNA analitzada. Si s'analitza una llibreria de seqüències

aleatòria, s'haurien d'obtenir 4 línies força paral·leles entre si, cadascuna mostrant el percentatge d'una de les bases al llarg de tota la seqüència, i en cap cas han de ser desequilibrades entre elles. Grans canvis entre les bases normalment indiquen una possible sobrerepresentació que contamina la llibreria. Pel contrari, si les alteracions són consistents en totes les bases, probablement la causa sigui un problema sistemàtic de la seqüenciació de la llibreria (Babraham Bioinformatics, 2023).

Contingut de GC per base: Mostra el percentatge de guanines i citocines per a cada posició de base dins de la seqüència resultant en un gràfic lineal. Per a llibreries aleatòries, la línia no hauria de patir grans alteracions i romandrà força recta i horitzontal al llarg de tota la gràfica. La seva interpretació és similar a la del gràfic de Contingut de seqüència per base (Babraham Bioinformatics, 2023).

Contingut de GC per seqüència: Mesura el contingut de guanines i citocines al llarg de tota la seqüència i el compara amb una distribució normal modelada del contingut de guanines i citocines. Normalment una llibreria aleatòria mostra un gràfic de distribució normal del contingut de guanines i citocines on el pic central indica el seu contingut general dins del genoma. També es mostra una altra corba que representa un model del contingut GC calculat a partir de les dades observades i que actua com a referència. Si s'observa una distribució inusual indica una llibreria contaminada o altres tipus d'errors com una interrupció de la seqüència. Una distribució normal i desplaçada indica problemes sistemàtics independents a les posicions de les bases (Babraham Bioinformatics, 2023).

Contingut de N per base: Si el seqüenciador és incapaç de reclutar les bases amb suficient confiança, aquestes es substitueixen per N en comptes d'una base a l'atzar. Per tant, en aquests anàlisis es mesura el percentatge de reclutaments de N per a cada posició dins de la seqüència. Generalment, el gràfic hauria de mostrar uns percentatges baixos de N, especialment al final de la seqüència. Percentatges superiors a 5% comencen a indicar que possiblement l'anàlisi no ha interpretat les dades amb la suficient confiança per a que sigui vàlid (Babraham Bioinformatics, 2023).

Distribució de longitud de seqüència: Molts seqüenciadors d'alt rendiment generen fragments de llargada uniforme però, altres poden contenir lectures molt variades. Encara que s'obtinguin llibreries més uniformes és possible que es tallin les seqüències, sobretot a la part final per eliminar possibles reclutaments de bases de baixa qualitat (Babraham Bioinformatics, 2023).

Seqüències duplicades: En aquest anàlisi, es presenta un gràfic que recull la llargada dels fragments estudiats. En la majoria dels casos es presenta una línia que mostra un pic només en una de les longituds de fragment, si l'arxiu conté varietat en quant als fragments, es podrà visualitzar al gràfic l'abundància relativa de cada longitud trobada. El mòdul quantifica el grau de duplicació de cada seqüència de l'arxiu analitzat i crea una taula que recull el número relatiu de seqüències dels diferents percentatges de duplicació. Per a la detecció d'una duplicació és necessari que les seqüències coincideixin completament, pel que les lectures superiors a 75bp es divideixen en fragments de 50pb. Les lectures més llargues són més propenses a contenir errors de seqüenciació que poden incrementar artificialment la diversitat observada i a infrarepresentar seqüències amb elevada duplicació (Babraham Bioinformatics, 2023).

Seqüències sobrerepresentades: En llibreries variades creades a partir de seqüenciacions d'alt rendiment, el fet de trobar una seqüència amb major representació que la resta indica que aquesta pot ser altament significativa, que sigui un contaminant o que la llibreria no sigui tant variada com s'espera. S'enumeren totes les seqüències que representen més del 0,1% del total. Per a cada sobrerepresentacio trobada, el programa el buscarà en un base de dades dels contaminants més comuns i retornarà les millors coincidències, aquestes han de ser com a mínim de 20 pb i no tenir més d'1 base diferent. El resultat és només orientatiu i no assegura que aquesta sigui la font de contaminació real. Per a la detecció d'una duplicació és necessari que les seqüències coincideixin completament, pel que les lectures superiors a 75bp es divideixen en fragments de 50pb. Les lectures més llargues són les més propenses a contenir errors de seqüenciació que poden incrementar artificialment la diversitat observada i a infrarepresentar seqüències amb elevada duplicació (Babraham Bioinformatics, 2023).

Kmers sobrerepresentats: És probable que si les seqüències analitzades són molt llargues i de baixa qualitat, els errors de seqüenciació redueixin significativament el número de duplicacions real. Aquest mòdul mesura l'enriquiment de cada 5-mer dins de la llibreria. Calcula el nivell esperat segons el contingut total de la llibreria i ho compara amb el recompte real (observat/esperat) per cada k-mer. El resultat que retorna és una llista de coincidències i un gràfic que recull els 6 millors per mostrar el patró obtingut al llarg de la seqüència. D'aquesta manera, permet visualitzar si existeix un enriquiment general o un patró de biaix en diferents localitzacions de la lectura (Babraham Bioinformatics, 2023).

3.3.2.1. Instal·lació de *FastQC*

S'ha instal·lat l'ultima versió disponible d'aquesta eina *v0.12.1(Win/Linux.zipfile*) que es troba disponible per al seu us lliure des de l'1 de març del 2023 a: (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip).

3.3.3. eKLIPse

És una eina bioinformàtica especifica i sensible per a la detecció i quantificació de reordenaments de l'ADNmt a partir de les dades obtingudes per seqüenciació de nova generació (simple i per parelles). La principal aplicació d'aquest software és estudiar les causes i conseqüències de les alteracions detectades i entendre la possible relació genotip-fenotip que pot haver en diferents característiques fenotípiques (Goudenège et al., 2019).

De forma general, l'eina parteix d'un punt d'interrupció per comparar les seqüències aportades per l'imput amb una seqüència d'ADNmt de referència de la mateixa longitud. A partir de la referència, es poden trobar les zones que no coincideixen, sobretot a causa d'una deleció o per estar localitzades en altres regions del genoma (Goudenège et al., 2019).

Aquest software es basa en el concepte de "*soft clipping*". El "*clipping"* consisteix en tallar a la seqüència analitzada els extrems dels fragments dels que no es troben coincidències amb la seqüència de referència per evitar errors d'alineament (que són més comuns als extrems). D'aquesta manera, si els extrems tallats s'eliminen es realitza un procés de "*hard clipping*" i si es mantenen per intentar realinear-los en altres regions de la seqüència de referència és "soft clipping". Aquest fet es pot donar en algun dels dos extrems o als dos simultàniament. (Goudenège et al., 2019)

eKLIPse està desenvolupat en base al llenguatge de *Python2 (versió 2.7)* i requereix la instal·lació d'eines i mòduls addicionals pel seu correcte funcionament, concretament:

- *python 2.7*
- *biopython*
- *tqdm*
- *samtools*
- *blastn i makeblastdb (>=2.3.0+)*
- *circos*

Les seqüències originades pel "*soft clipping"* s'alineen amb la seqüència de referència mitjançant el *BLASTn,* escollint només un resultat de *BLAST* per cerca. Per obtenir aquest

resultat es realitza un filtratge segons una sèrie de paràmetres editables, com per exemple, percentatge de identitat i cobertura, cost d'obrir i estendre un gap, entre altres, que depenen de la qualitat de les seqüències analitzades i la tècnica de seqüenciació. *Samtools* és necessari per a reduir el mostreig i facilitar l'anàlisi. Finalment, retorna els resultats del procés en forma de dues taules d'*Excel* i gràfics creats amb l'eina de *circos* (Goudenège et al., 2019).

El programa necessita que tots els arxius per analitzar estiguin en format *bam* o *sam*, a més de disposar de la seqüència de l'ADNmt de referència. A continuació, processa les mostres per reduir el número de lectures. Les mostres s'alineen amb l'ADNmt de referència mitjançant el mòdul de profunditat de lectura, que quantifica el número de lectures totals i de les que han patit el procés de *"soft-clipping"* per a cada nucleòtid, i el mòdul de *"soft-clipping"* que recupera els fragments retallats amb les seves coincidències de l'ADN de referència. El mòdul de profunditat de lectura calcula el percentatge de delecions de cada mostra i representa la cobertura en un gràfic conegut com *circos*. Degut a què les delecions acostumen a estar entre seqüències total o parcialment repetides, és possible que es generin errors d'alineament que provoquin errors del punt d'interrupció (*breakpoint*). Per aquest motiu, el fet d'agrupar el fragment tallat per *"soft clipping"* amb el seva coincidència localitzada *upstream*, permet al software detectar possibles repeticions i alinear les delecions dels punts d'interrupció (Goudenège et al., 2019).

Les seqüències coincidents amb la referència s'escriuen en format *FASTA* per a que puguin ser alineades amb l'eina de *BLASTn* i filtrar els resultats segons els paràmetres escollits. En aquest pas, només s'accepten els resultats de *BLASTn* bidireccionals, és a dir, que tant en la lectura directa com la inversa, la posició de *"soft-clipping"* d'una lectura ha de coincidir amb l'alineament del fragment tallat de l'altra lectura i viceversa (Goudenège et al., 2019).

Finalment, el programa genera una taula d'Excel amb la freqüència acumulativa de les delecions de cada gen per a cadascuna de les mostres, una altra taula amb la predicció de totes les delecions segons les posicions del punt d'interrupció, la carrega de les delecions, els resultats de *BLAST* i les repeticions i els gràfics de *circos* (Goudenège et al., 2019).

EL gràfic de *circos* mostra el següent aspecte:

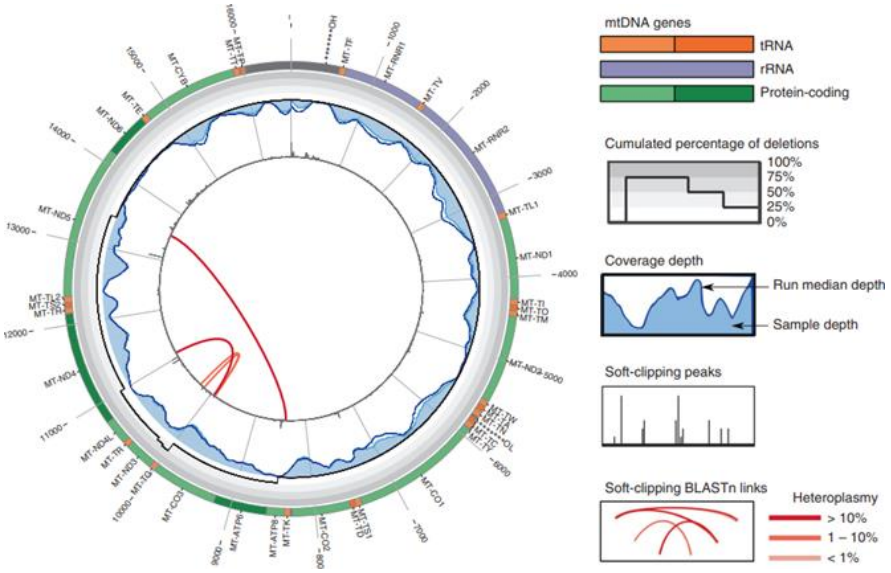


Figura 2: Representació gràfica de circos (Goudenège et al., 2019).

Tal i com es pot observar, la circumferència externa representa l'estructura de l'ADNmt indicant la posició dels gens codificants per tRNA (taronja), rRNA (violat), proteïnes (verd) i la regió no codificant del "*D-loop"* (negre). A la circumferència de sota, es visualitzen una sèrie d'anells grisos de diferents intensitats indicant, amb l'ajuda d'una línia negra, el percentatge de delecions acumulatiu al llarg de la seqüència. A la següent circumferència interna s'observa un camp blau que representa la profunditat de cobertura de cada base i una línia de color blau més fosc que mostra la cobertura mitjana de totes les mostres analitzades. A continuació, es troba una àrea blanca amb una sèrie de línies negres, l'altura de les quals és directament proporcional el número de seqüències que han patit *"soft-clipping"*. Finalment, al centre del cercle s'observen una sèrie de corbes vermelles de diferents intensitats, on cada corba indica una relació bidireccional obtinguda pel *BLASTn* i la intensitat del color mostra el percentatge d'heteroplàsmia (Goudenège et al., 2019).

3.3.3.1. Instal·lació d'*eKLIPse*

Un cop s'hagi preparat el sistema operatiu de la màquina virtual i instal·lat *FastQC*, es descarrega i instal·la *eKLIPse* (https://github.com/dooguypapua/eKLIPse), el paquet d'eines de *ncbi-BLAST v2.3.0+* i *circos*. En aquest estudi, s'ha utilitzat la versió *eKLIPse unix (v_1-0),* ja que, l'última versió disponible (v 2-1) presenta certs problemes de funcionament que no ha sigut possible solucionar (no s'ha pogut contactar amb el desenvolupador d'aquesta versió). Per assegurar que el programa funcioni correctament i de la forma més senzilla possible, és necessari que tots els programes addicionals que es requereixen per *eKLIPse* s'instal·lin dins de la carpeta d'aquest software o s'indiqui la ruta corresponent.

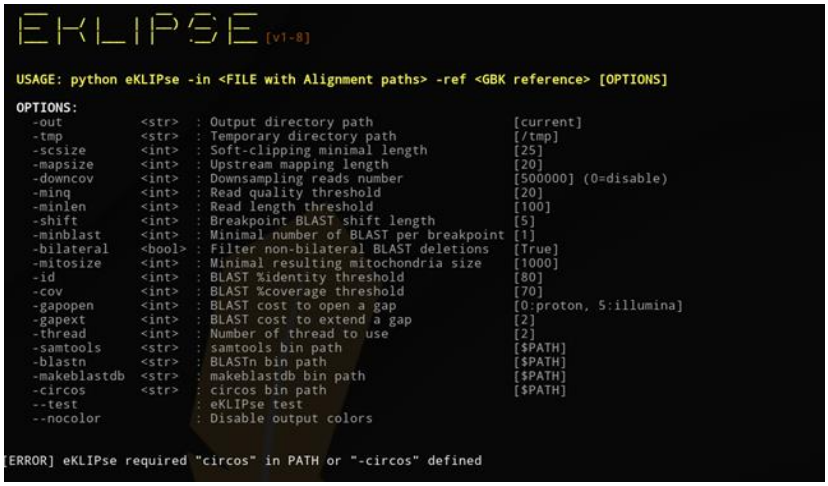
Per a la instal·lació d'*eKLIPse* s'utilitza la següent comanda:

```
unzip Qt_eKLIPse_unix_v1-0.zip
cd Qt_eKLIPse_unix_v1-0.zip
chmod a+x eKLIPse
./eKLIPse
```

Un cop instal·lat *eKLIPse* s'ha de comprovar que funciona correctament realitzant el corresponent test:

```
python2 eKLIPse.py --test
```

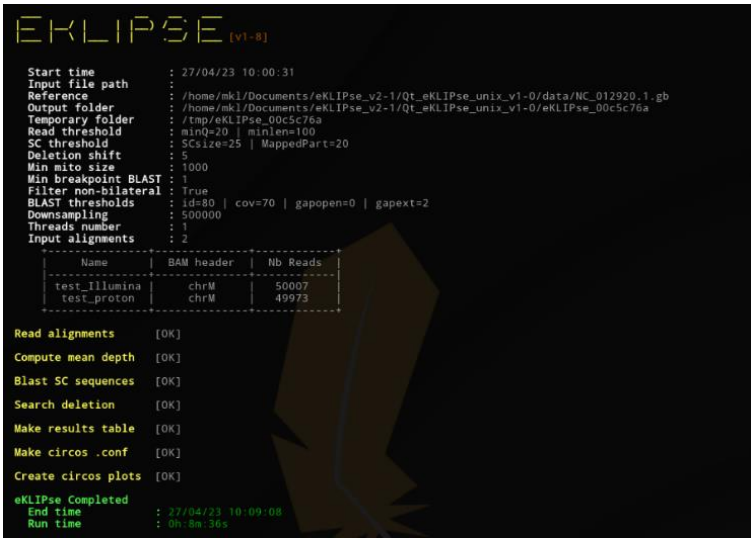
A continuació, apareixerà la següent pantalla indicada a la Imatge 2:



Imatge 2: Inici del test d'*eKLIPse*

Per a poder continuar amb el test s'ha d'especificar el *"PATH"* dels programes sol·licitats utilitzant:

```
export PATH="$PATH: ruta cap el programa o carpeta sol·licitada"
```



Imatge 3: Final del test d'*eKLIPse*

Un cop s'han definit les rutes, es completarà automàticament el test, mostrant la pantalla de la Imatge 3, i donarà com a resultat una carpeta amb dues taules d'*Excel* i un gràfic *circos*.

En aquest punt, ja es tenen preparats tots els programes i es pot seguir amb l'anàlisi dels arxius de les mostres.

3.4. Mètode

3.4.1. Mètode previ per a l'obtenció de les mostres de seqüències d'ADNmt

El protocol utilitzat es basa en els estudis previs relacionats (Torrell et al, 2017; Valiente-Pallejà et al, 2018). Primer es realitza l'amplificació del ADNmt de cada mostra i es comprova mitjançant una electroforesi en gel d'agarosa 0,7% i marcadors de DNA que aquesta s'ha produït correctament. L'amplificació s'utilitzarà per a generar dos fragments superposats que cobreixin la totalitat de l'ADNmt. EL protocol utilitza l'enzim TaKaRa LA Taq® Hot Start Version (RR042Q), idoni per a amplificacions extrallargues. La reacció tarda 5 hores en completar-se en el cas de l'amplificació amb dos fragments, pel que es recomanable que es deixi processar durant la nit. A continuació, es mostren una sèrie de taules que resumeixen els oligonucleòtids utilitzats com a *primers* per a cada amplificació, els reactius necessaris per a preparar la mescla de reacció i les condicions que ha de seguir el termociclador en cada cas.

3.4.1.1. Amplificació de l'ADN basada en dos amplicons de PCR superposats

Es preparen els oligonucleòtids amb les característiques indicades a la Taula 2, la solució de reacció segons les instruccions de la Taula 3 i es realitza una PCR seguint les indicacions de la Taula 4:

Taula 2: Primers per a amplificar l'ADNmt en dos fragments superposats

Nom del Primer	Seqüència 5' -> 3'	Posició	Tm (ºC)	Llargada del producte	Superposició
FampA	AAATCTTACCCCGCCTGTTT	2480-2499	58	8379	2480-2669
RampA	AATTAGGCTGTGGGTGGTTG	10858-10839	60		
FampB	GCCATACTAGTCTTTGCCGC	10653-10672	62	8567	10858-10653
RampB	GGCAGGTCAATTTCACTGGT	2688-2669	60		

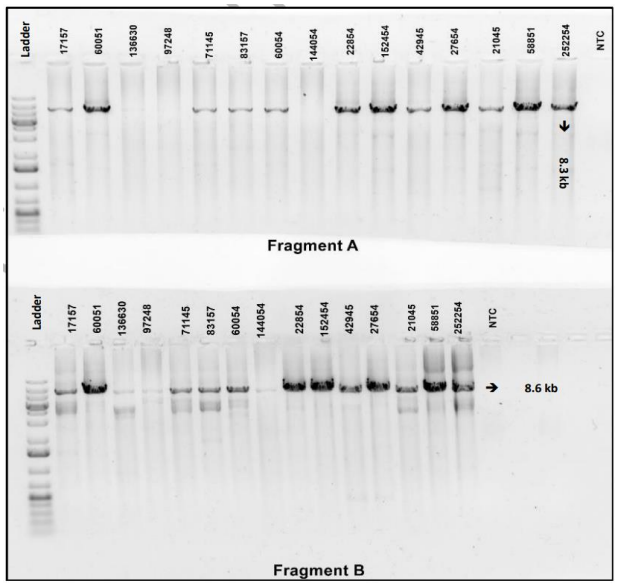
Taula 3: Solució per a amplificar l'ADNmt en dos fragments superposats

Reactiu	Volum (µl)
10× TaKaRa LA PCR buffer II	2,5
dNTP 2,5 mM	2
TaKaRa LA Taq polimerasa	0,2875
BSA	0,3125
primer directe 5 µM	0,5
primer revers 5 µM	0,5
DNA (25 ng)	2
H_2O	16,9

Taula 4: Condicions del termociclador

17 cicles	95ºC 95ºC 69ºC* 68ºC	2 min 15 sec 30 sec 5 min	*Començar per 69ºC i baixar 0,5ºC per cicle fins arribar a 60ºC
18 cicles	95ºC 60ºC 68ºC 68ºC 4ºC	15 seg 30 seg 9 min 10 min Aguantar	** Augmentar 10 segons a cada cicle

Si la PCR ha estat correcta s'haurien d'observar en el gel d'electroforesi les bandes corresponents al fragment A i B per cada mostra i cap banda en el control negatiu, com s'indica a la Imatge 4. Degut a que la reacció produeix amplicons inespecífics, és recomanable utilitzar kits de purificació de l'agarosa.



Imatge 4: Electroforesi en gel d'agarosa al 0,7%. A la imatge superior es visualitzen les bandes corresponents al fragment A, a la imatge inferior les del fragment B. NTC: *non template control* o control negatiu.

A continuació, es realitza un procés de purificació dels amplicons obtinguts que es divideix en dos passos. Primer, s'extreuen els fragments d'ADN del gel d'electroforesi i s'apliquen una sèrie de rentats i centrifugacions seguint les indicacions de GenElute™ Gel Extraction Kit and GenElute™ PCR Clean-Up Kit (Sigma-Adrich). Després, es segueix amb una purificació amb boles metàl·liques (ChargeSwitch PCR Clean-Up), adaptant el protocol a partir de ChargeSwitch™ PCR Clean-Up Kit (CS12000).

Un cop es tenen els productes de PCR purificats es fragmenten tractant-los tots com a mostres diferents. Després es preparen els adaptadors per *Illumina* o *Ion Torrent* que es lligaran als fragments (s'ha de posar el mateix adaptador als dos fragments pertanyents al mateix subjecte). Es descarten els fragments més llargs utilitzant un vòrtex que ajudarà a unir-los a les boles AMPure XP, deixant els fragments desitjats al sobrenedant. Al següent pas, s'utilitzaran un altre cop les boles AMPure XP per unir els fragments de la llargada desitjada i deixar els més curts al sobrenedant. Els fragments (lligats als adaptadors) desitjats s'eluiran de les boles i s'amplificaran per PCR utilitzant una solució de PCR preparada segons les instruccions de la Taula 5 i les condicions del termociclador mostrades a la Taula 6:

Taula 5: Solució de PCR dels fragments units a adaptadors

ADN lligat a adaptadors	1-40µl
Primers	10 µl
H_2O estèril	Variable
NEBNext High-Fidelity 2X PCR Master Mix	50 µl

Taula 6: Condicions del termociclador per la PCR dels fragments d'ADNmt units a adaptadors

Duració inicial	98ºC	30 seg
4-12 cicles	98ºC	10 seg
	58ºC	30 seg
	65ºC	30 seg
1 cicle	65ºC	5 min
Aguantar	4ª	∞

*100 ng -> 6-8 cicles; menys de 100 ng -> 8-12 cicles

Seguidament, es neteja la llibreria obtinguda per l'amplificació utilitzant les boles AMPure XP i es realitza una quantificació de la llibreria utilitzant el kit HS, seguint les instruccions del Agilent High Sensitivity D1000 ScreenTape System Quick Guide.

Després es quantifica i analitza el control de la qualitat de la llibreria per dos assajos diferents (fluorímetre Qubit 4 i Agilent High Sensitivity D1000 ScreenTape System). Tot i que Qubit 4 mesura de forma més sensible la concentració de les llibreries, la distribució dels amplicons només és detectable mitjançant el kit TapeStation HS del segon mètode. Per tant, utilitzant les dues tècniques, és possible calcular la molaritat de les mostres amb una fiabilitat suficient.

Mitjançant la concentració obtinguda pel primer anàlisi i la mida mitjana de la llibreria obtinguda pel segon, es calcula la concentració en nM dels fragments utilitzant la fórmula:

$$\frac{(\text{concentració en } \frac{\text{ng}}{\mu\text{l}})}{(\frac{660\text{g}}{\text{mol}} \times \text{mida de la llibreria mitana en pb})} \times 10^6 = \text{concentració en nM}$$

Finalment, es prepara una llibreria de mostres equimolars formada per 60 pM d'amplicons en total.

3.4.2. Anàlisi bioinformàtic de les dades de la llibreria

Aquest procés, que és en el que es centra el meu Treball de Fi de Grau comença per la realització d'un control de qualitat de la lectura de les mostres en format *fastq* utilitzant l'eina *FastQC*. Aquest pas es pot realitzar tant obrint el propi programa i introduint les mostres que es volen analitzar, com mitjançant una terminal. En aquest cas, s'ha utilitzat el menú del programa. Com a resultat s'obtindrà una carpeta d'imatges que conté tots els gràfics obtinguts i un enllaç a una pagina *HTML* amb el resum dels resultats.

D'acord als resultats aportats per *FastQC* s'han de realitzar les millores necessàries sobre la qualitat de les mostres en format *fastq* i existeixen diferents possibilitats:

Si les seqüències dels oligonucleòtids es troben sobrerepresentades o aporten diferents profunditats de lectura, aquests es poden eliminar del principi o el final de les lectures utilitzant *cutadapt v3.4*. Aquesta eina es pot utilitzar també per a millorar la qualitat de les lectures establint que totes siguin entre 50-300 bases de longitud. A més, permet eliminar un nombre específic de bases al principi o al final de la seqüència si aquestes regions mostren baixades importants de la qualitat o tallar els extrems 3' de les lectures amb terminacions de baixa qualitat. En aquest cas, s'ha realitzat una eliminació de les seqüències dels quatre oligonucleòtids, s'ha establert que la mida de les lectures es trobi a un interval de 20-310 bases i s'han eliminat 15 bases tant del principi com del final de cada lectura.

De forma addicional, es pot comprovar la profunditat de lectura (número de vegades que una base es representa en totes les lectures de la seqüenciació), utilitzant:

```
for sample in {01..80} do;
samtools depth sorted_BRAIN_${sample}.bam > sorted_BRAIN_${sample}.txt
```

Per a poder analitzar els resultats a *eKLIPse*, aquests han d'estar en format *bam*, pel que a continuació, s'utilitza *samtools* per a passar els arxius del format *fastq* a *bam*.

El format *sam* és de text pla i conté una quantitat molt elevada d'informació, pel que es necessita traduir en un format binari *bam*, ja que aquest té la capacitat de comprimir considerablement els arxius *sam*.

```
for sample in {01..80}; do
    bwa index mtDNA.fasta;
    bwa mem mtDNA.fasta
    BRAIN_${sample}.assembled.fastq > BRAIN_${sample}.assembled.sam;
    samtools view -bS BRAIN_${sample}.assembled.sam > BRAIN_${sample}.assembled.bam;
    samtools sort BRAIN_${sample}.assembled.bam -o sorted_BRAIN_${sample}.bam;
    samtools index sorted_BRAIN_${sample}.bam;
    done
```

Abans d'analitzar les mostres es necessari crear un arxiu de text on es trobi la ruta de totes les mostres i el seu número com s'indica a la Imatge 5:

Imatge 5: Arxiu de text amb totes les rutes de les mostres en format *bam*

A continuació, s'introdueix la següent comanda:

```
python2 eKLIPse.py -in ruta de l'arxiu de text amb les rutes de les mostres -ref ruta del genoma de referencia anomenat NC_012920.1.gb -scsize 25 -mapsize 20 -downcov 0 -minq 20 -minlen 100 -shift 5 -minblast 5 -bilateral True -mitosize 100 -id 80 -cov 90 -gapopen 5 -gapext 2 -thread 2
```

D'aquesta manera s'analitzen les mostres segons els paràmetres de la Taula 7:

Taula 7: Nom, definició i valors dels paràmetres d'operació d'*eKLIPse* utilitzats

Paràmetre	Definició	Valor utilitzat
-out	Ruta del directori de l'arxiu de l'output (resultat)	*current*
-tmp	Ruta de directori temporal	*/tmp*
-scsize	Longitud mínima del "*soft-clipping"*	25
-mapsize	Longitud del fragment mapejat "*upstream"*	20
-downcov	Nombre de lectures de la reducció de les mostres ("*downsampling*")	0
-minq	Llindar de qualitat de la lectura.	20
-minlen	Llindar de llargada de la lectura	100
-shift	Mida de la regió corredissa del punt d'interrupció	5
-minblast	Nombre mínim de *BLAST* per punt d'interrumpció.	5
-bilateral	Filtrar els *BLAST* unidirreccionals	*True*
-mitosize	Eliminar els ADNmt delecionats menors de	100
-id	Límit de % d'identitat de *BLAST*	80
-cov	Límit de % cobertura de *BLAST*	90
-gapopen	Cost de *BLAST* per a obrir un "*gap"* o trencament.	5
-gapext	Cost de *BLAST* per estendre el "*gap"*	2
-thread	Nombre de càrrega	2
-samtools	Ruta cap a la carpeta *bin* de *samtools*	
-blastn	Ruta cap a la carpeta *bin* de *BLASTN*	
-makeblastdb	Ruta cap a la carpeta *bin* de *makeblastdb*	
circos	Ruta cap ala carpeta *bin* de *circos*	

Aquest procés necessita aproximadament 4-5 hores per a completar-se, finalment retorna 2 taules d'*Excel* i els gràfics *circos* (un per mostra).

El procediment anterior fa referència a l'obtenció i anàlisi de les mostres seqüenciades amb les tècniques d'*Illumina* i *Ion Torrent*. Per altra banda, per a realitzar la comparació entre les dues tècniques de seqüenciació, s'ha analitzat la seva qualitat mitjançant *FastQC* i, a continuació, es van aplicar els mateixos filtres de *cutadapt* per a comprovar si millorarien. En quant a l'anàlisi amb *eKLIPse*, s'apliquen els mateixos paràmetres excepte en quant al cost d'obrir un "*gap"* (*-gapopen*), que ha de ser de 5 en el cas d'analitzar mostres d'*Illumina* i 0, si s'han seqüenciat per *Ion Torrent*.

Per als anàlisis i representacions estadístiques posteriors s'han utilitzat les eines *jamovi (Version 2.3.)*(J. Love et al., 2022) i *Microsoft Excel*.

4. Resultats

4.1. Control de la qualitat de les mostres amb *FastQC* i *cutadapt*

Inicialment, al realitzar l'anàlisi amb *FastQC* es va observar una important decaiguda de la qualitat al principi i al final de la mostra, sobretot als mòduls de Qualitat de seqüència per base i Contingut de la seqüència per base. Degut a què es presentava aquesta problemàtica en totes les mostres, era indicatiu d'un possible problema metodològic (probablement durant la seqüenciació) que havia causat una sobrerepresentació en una part de les mostres i suposava un contaminant per a la seva correcta lectura. Per aquest motiu, es va optar per tallar 15 bases del principi i del final de totes les mostres amb *cutadapt*.

Com a resultat, la qualitat va mostrar millores en els punts observats anteriorment i, com a conseqüència, algunes mostres van presentar millores també en quant a la sobrerepresentació (ja que aquesta havia baixat en alguns casos). A continuació, es mostra a la Imatge 4 un resum d'aquesta millora mitjançant els gràfics de puntuacions de la qualitat per base i de contingut de seqüència per base d'una de les mostres abans i després d'aplicar l'edició per *cutadapt*.

Imatge 6: Qualitat de les mostres abans de l'edició amb cutadapt (a dalt) i després de tallar 15 nucleòtids del principi i del final de les seqüències (a baix). Les dues imatges de l'esquerra corresponen als gràfics de Qualitat de seqüència per base i les dues de la dreta als gràfics lineals de Contingut de seqüència per base.

Aquestes mostres millorades per *cutadapt* van ser seguidament analitzades mitjançant *eKLIPse*, obtenint com a resultat una taula que indicava el percentatge acumulat de delecions en cada gen per cadascuna de les mostres, una taula amb la predicció de les possibles delecions, les seves freqüències, punts de interrupció, número de *BLAST*, profunditat de lectura i les repeticions que es troben flanquejant la deleció (si es que hi han).

Inicialment, es va provar a fer l'anàlisi amb un número de *BLAST* mínim d'1, però aquest va ser incrementat posteriorment a 5. Es va prendre aquesta decisió, ja que, en el primer cas gairebé totes les mostres presentaven delecions, pel que es va optar per introduir a l'eina unes condicions més restrictives. Aquests nous paràmetres eliminen els resultats pels que s'havia obtingut menys de 5 *BLASTs* (o coincidències), de manera que s'incrementa la fiabilitat de que la previsió sobre la regió de la deleció sigui certa.

4.2. Estudi de la distribució de les dades

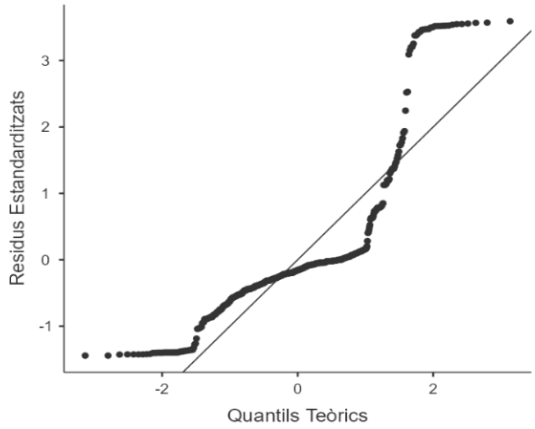
A partir dels resultats anteriors el primer que es va estudiar va ser si les dades obtingudes estaven distribuïdes normalment mitjançant la prova de Kolmogorov-Smirnov:

Taula 8: Característiques descriptives i test de Kolmogorov-Smirnov del % d'Heteroplàsmia i la Longitud de la deleció segons cada grup (HC: Healthy Control, SCZ: esquizofrènia)

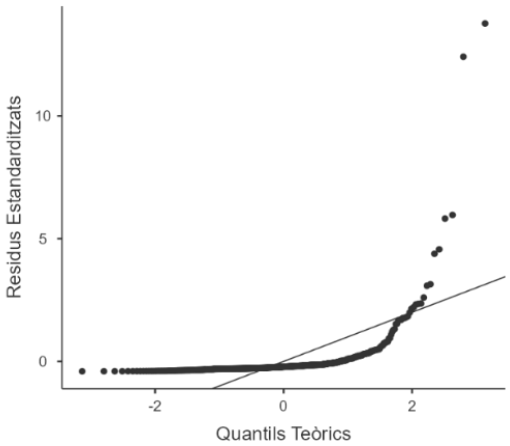
	Grup	% Heteroplàsmia	Longitud de la deleció
Mitjana	HC	0,485	4656
	SCZ	0,685	4732
Mediana	HC	0,148	4086
	SCZ	0,291	4421
Desviació estàndard	HC	1,82	3480
	SCZ	1,13	2875
Rang Interquartil	HC	0,216	1385
	SCZ	0,557	1548
Mínim	HC	0,00700	51
	SCZ	0,0460	51
Màxim	HC	22,1	16317
	SCZ	7,86	16161
Estadístic de Kolmogorov-Smirnov		0,342	0,279
Valor-p de Kolmogorov-Smirnov		< 0,001	< 0,001
25percentil	HC	0,0550	3233
	SCZ	0,144	3497
50percentil	HC	0,148	4086
	SCZ	0,291	4421
75percentil	HC	0,271	4618
	SCZ	0,702	5045

Taula 9: Recompte total de les delecions per cada grup (HC: Healthy Control, SCZ: esquizofrènia)

Grup	HC	SCZ
Nombre total de delecions	349	235



Gràfic 1: Q-Q de Longitud de la deleció



Gràfic 2: Q-Q de % d'Heteroplàsmia

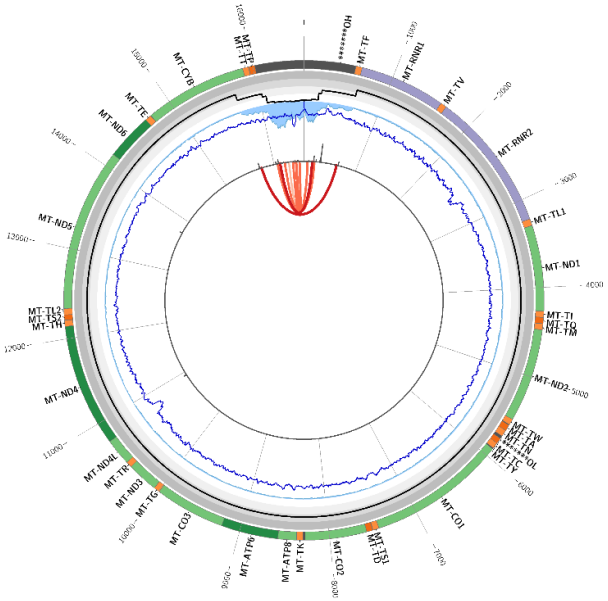
En total, *eKLIPse* mostra una previsió de 349 delecions a les mostres del grup control (o *Healthy Control* en anglès) i 235 per al grup de participants amb esquizofrènia que es troben recollides a la Taula 9. Com es pot observar pel baix nivell del valor p de Kolmogorov-Smirnov de la Taula 8 és possible rebutjar la hipòtesis nul·la (que les dades de la longitud de deleció i % d'heteroplàsmia presenten una distribució normal). S'obté un valor de p molt baix que indica que seria molt difícil obtenir el valor de l'estadístic si la distribució fos normal. Per altra part, als gràfics Q-Q (Gràfics 1 i 2), la recta indica una distribució normal teòrica en funció de les dades utilitzades i els punts il·lustren la distribució real de les mostres, que es troba bastant allunyada de la normalitat. Per tant, es conclou que les mostres utilitzades són de tipus quantitatiu, independent, continu i sense distribució normal. És per aquest motiu, que s'utilitzarà l'estudi de la mediana en comptes de la mitjana i s'optarà per utilitzar el test estadístic de U Mann Whitney.

4.3. Anàlisis i selecció de les mostres

Al consultar els resultats crus de l'anàlisi amb *eKLIPse*, es va observar que una gran part de les delecions es trobaven amb unes freqüències molt baixes, inferiors al 0,5%. El % d'heteroplàsmia mínim que pot detectar *eKLIPse* de forma fiable és 0.5, pel que es va decidir establir un llindar de 0,5 i eliminar els resultats que es trobin per sota.

A més, degut a que varies de les mostres control presentaven uns % d'heteroplàsmia molt superiors als de la resta, es va analitzar també la profunditat de lectura i la cobertura obtinguda per *samtools* de cada mostra. D'aquesta manera es pretén entendre si les diferencies dels % d'heteroplàsmia observat es deuen a la pròpia mostra o degut a errors causats a l'hora de realitzar el procediment experimental.

Analitzant la profunditat de lectura de cada mostra aïllada és possible observar que moltes d'aquestes, sobretot les que presenten elevats nivells d'heteroplàsmia acostumen a tenir una profunditat de lectura baixa i/o molt variant (encara que moltes presenten cobertures de profunditat de lectura elevades en les posicions on es reporten delecions), el que implica una pobre qualitat de lectura degut a problemes metodològics durant la seqüenciació. Un clar exemple és la següent mostra:



Imatge 7: Resultat de *circos* d'una mostra control anòmala

Aquesta mostra control presenta delecions en gairebé la totalitat del seu mtDNA, com es pot veure per la línia negra que rodeja tota la seqüència, indicant un percentatge acumulat de delecions aproximadament de >75%. Per altra banda, es pot veure que la profunditat de lectura (blau clar) de cada nucleòtid és molt baixa comparat amb la de la mitjana de les mostres (blau fosc), excepte en el bucle D, que és una zona no codificant però, presenta unes profunditats de lectura molt més elevades que la resta del ADNmt.

Per altra part, s'ha calculat la cobertura de totes les mostres i en totes s'ha obtingut un 100%, pel que es pot afirmar que tota la mostra ha estat seqüenciada i posteriorment llegida i analitzada per *eKLIPse*.

A més, s'han considerat les profunditats de lectura dels dos fragments seqüenciats de totes les mostres d'ADNmt i s'ha calculat la mediana de la profunditat de lectura de tots els nucleòtids de cada mostra, les medianes de les profunditats de lectura de les zones superposades, les relacions mediana de la profunditat total / mediana de la profunditat de la superposició, la

mediana de la profunditat de lectura de l'amplicó del fragment A, la de l'amplicó del fragment B, A/B i B/A. Les profunditats de lectura de cada fragment (A i B) són equivalents, de manera que les relacions A/B i B/A corresponen a 1. Aquest aspecte demostra que no hi ha sobrerepresentacions de cap dels fragments del ADNmt seqüenciat i que les delecions detectades es troben amb unes profunditats de lectura equivalents entre elles. Per tant, si es troben diferents freqüències de delecions, no és degut a canvis en la profunditat de lectura entre els amplicons dels dos fragments, és a dir, no es degut a problemes metodològics durant la seqüenciació o de qualitat de les mostres i no és necessari eliminar cap mostra dels càlculs.

4.4. Anàlisi estadístic dels resultats d'*eKLIPse* amb un llindar mínim de 0,5 % d'Heteroplàsmia

Després de seleccionar les mostres amb % d'heteroplàsmia superior a 0,5 es tornen a comprovar les característiques descriptives de les dades i es realitza una Prova T per a Mostres independents amb el test de U Mann Whitney per a comparar la longitud de seqüenciació i el % d'heteroplàsmia entre els dos grups.

Taula 10: Característiques descriptives i test de Kolmogorov-Smirnov del % d'Heteroplàsmia i la Longitud de la deleció segons cada grup amb llindar mínim de % d'Heteroplàsmia de 0,5 (HC: Healthy Control, SCZ: esquizofrènia)

	Grup	% Heteroplàsmia	Longitud de la deleció
Mediana	HC	1,05	3755
	SCZ	1,04	4086
Desviació estàndard	HC	4,28	6232
	SCZ	1,56	2560
Rang Interquartil	HC	1,85	12662
	SCZ	0,998	1624
Mínim	HC	0,509	101
	SCZ	0,502	100
Màxim	HC	22,1	16317
	SCZ	7,86	16095
25percentil	HC	0,754	1855
	SCZ	0,701	3196
50percentil	HC	1,05	3755
	SCZ	1.04	4086
75percentil	HC	2,60	14517
	SCZ	1,70	4820

Taula 11: Recompte total de les delecions amb llindar mínim de 0,5 % d'Heteroplàsmia per cada grup (HC: Healthy Control, SCZ: esquizofrènia)

Grup	**HC**	**SCZ**
Nombre total de delecions	80	50

Com s'observa en la Taula 11, el grup control (HC), presenta un major recompte total de delecions respecte el grup amb esquizofrènia (SCZ). Per altra banda, a la Taula 10, aquest grup (HC) mostra majors percentatges d'heteroplàsmia en quant a la mediana, mínim, màxim, rang intercuartil i els percentils 25%, 50% i 75% però, també mostra una major desviació estàndard. En canvi, el grup de participants amb esquizofrènia mostren majors longituds de deleció segons els valors calculats a la mediana i els percentils 25%, 50% i presenta una menor desviació estàndard que el grup control. Tot i així, les mostres control mostren una major longitud de deleció mínima, màxima, al Rang Interquartil i al percentil 75%.

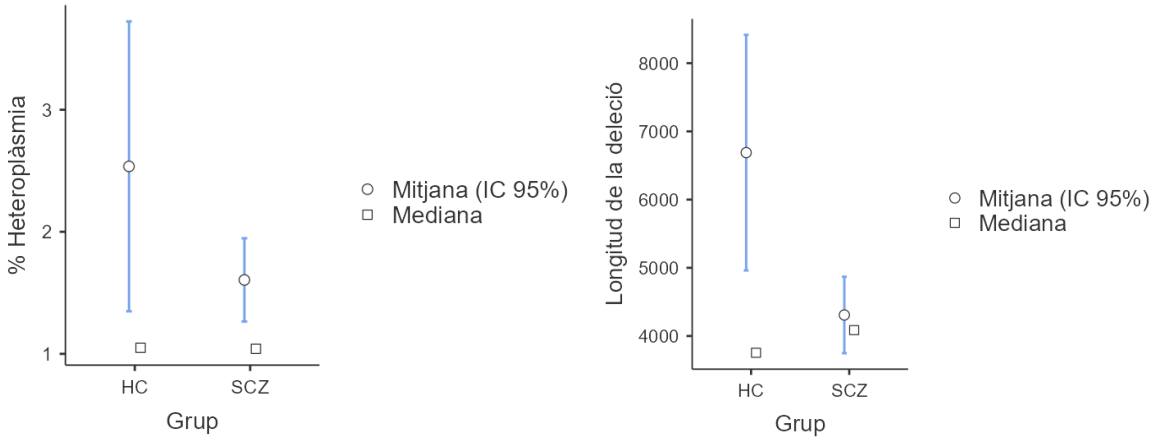
Per a comprovar si aquestes diferències són significatives es realitza una prova U de Mann Whitney, suposant que els dos grups presenten una mitjana de % d'heteroplàsmia i longitud de deleció diferent:

Taula 12: Prova U de Mann-Whitney per a la hipòtesi de que les mitjanes del dos grups (grup control i participants amb esquizofrènia) són diferents entre elles

		Estadístic	**p**
% Heteroplàsmia	U de Mann-Whitney	1828	0,412
Longitud de deleció	U de Mann-Whitney	1933	0,750

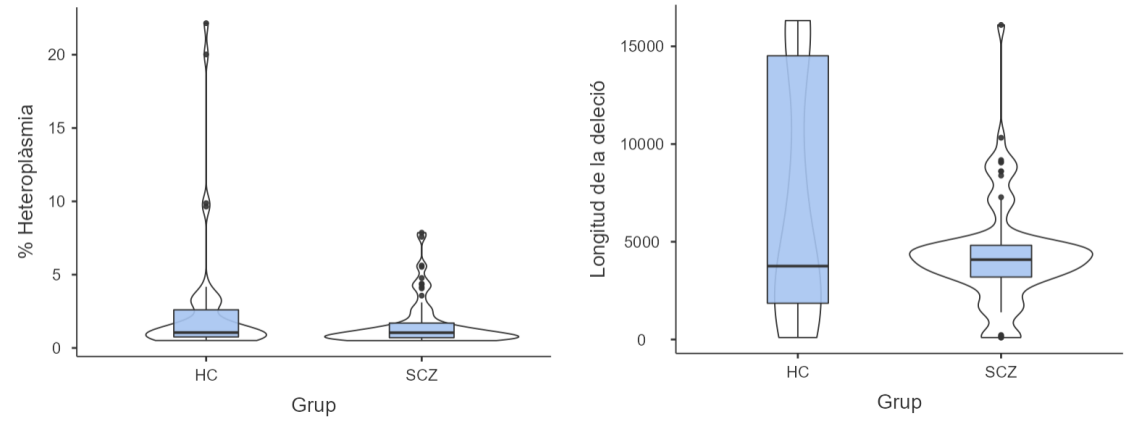
Anotació. $H_a\ \mu_{HC} \neq \mu_{SCZ}$

Si es considera un interval de confiança del 0,05, es pot observar a la Taula 12 que el valor de p, al ser superior a 0,05, indica que es pot rebutjar la hipòtesis, pel que no es presenten diferències significatives entre els dos grups (ni per la longitud de seqüència ni pel % d'Heteroplàsmia). El fet de que les diferències entre les mitjanes i medianes dels dos grups (control i participants amb esquizofrènia) en quant al % d'Heteroplàsmia i Longitud de deleció siguin poc significatives es pot demostrar també visualment pels Gràfics 4 i 5.



Gràfics 3 i 4: Gràfics descriptius de la mitjana, amb interval de confiança del 95%, i de la mediana del % d'Heteroplàsmia (Gràfic 3) i de la Longitud de deleció (Gràfic 4), del grup control (HC) i el grup amb esquizofrènia (SCZ)

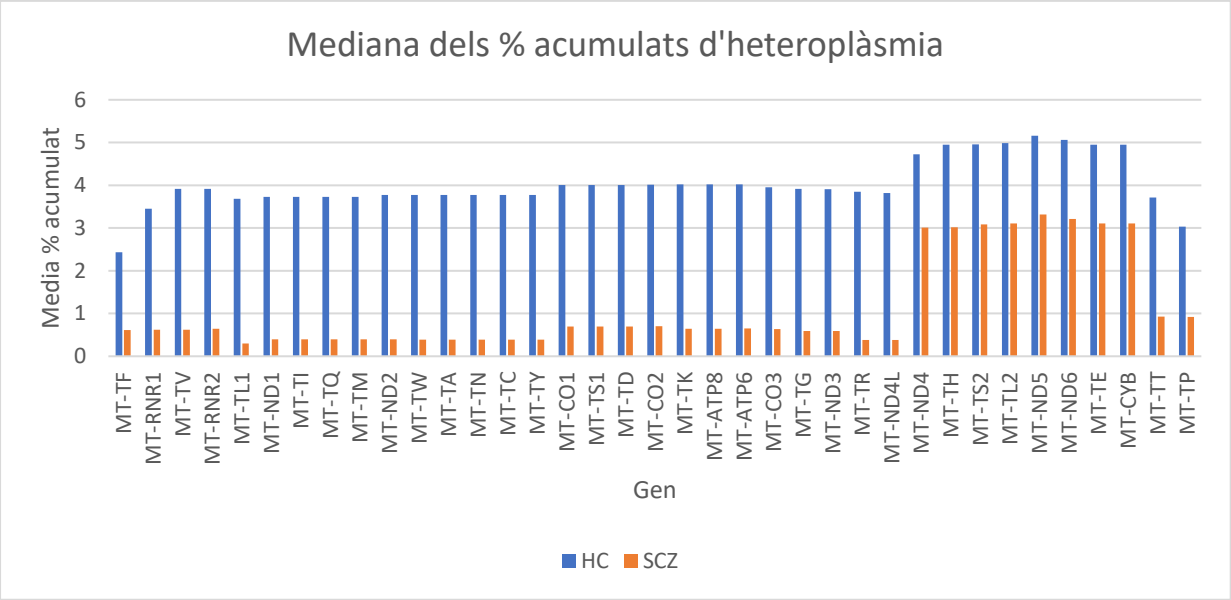
Per altra banda, els diagrames de caixes (Gràfics 6 i 7) tampoc mostren diferències significatives entre les medianes dels dos grups en quant al % d'Heteroplàsmia i la longitud de deleció. Tot i així, s'observa de forma evident les variacions entre els quartils i els màxims de longitud de deleció.



Gràfics 6 i 7: Diagrames de caixa per al % d'Heteroplàsmia (Gràfic 6) i la Longitud de la deleció (Gràfic 7) del grup control (HC) i el grup amb esquizofrènia (SCZ).

A més, a partir del resultat d'*eKLIPse* sobre els percentatges acumulats de les mutacions de cada gen per cada mostra analitzada, es va poder determinar quins són els gens més afectats per les delecions, tant en les mostres control com les que presenten esquizofrènia. Segons l'histograma de la Mediana dels % acumulats d'heteroplàsmia (Gràfic 8) aquests gens corresponen a 3 proteïnes del Complex I (MT-ND4, MT-ND5 i MT-ND6), 2 gens de RNA de transferència mitocondrials (MT-TH per a histidina, MT-TE per a glutamat, MT-TS2 per a serina 2 i MT-TL2 per a leucina 2) i MT-CYB que codifica per al citocrom B mitocondrial del Complex III. Les medianes dels percentatges acumulats de cada gen per a cada mostra són

superiors en el grup control respecte al grup amb esquizofrènia però, els dos grups coincideixen en uns majors nivells de delecions en els gens anteriorment esmentats.



Gràfic 8: Histograma de les medianes dels % acumulats d'heteroplàsmia per a cada gen i grup, HC: grup control, SCZ: grup amb esquizofrènia

Mutacions en qualsevol d'aquests gens pot provocar una gran varietat de malalties mitocondrials. A més, és interessant recordar que, com s'ha indicat prèviament en la introducció, existeixen indicis de que delecions en el Complex I, poden estar relacionats amb l'esquizofrènia o la severitat dels seus símptomes (Roberts, 2021).

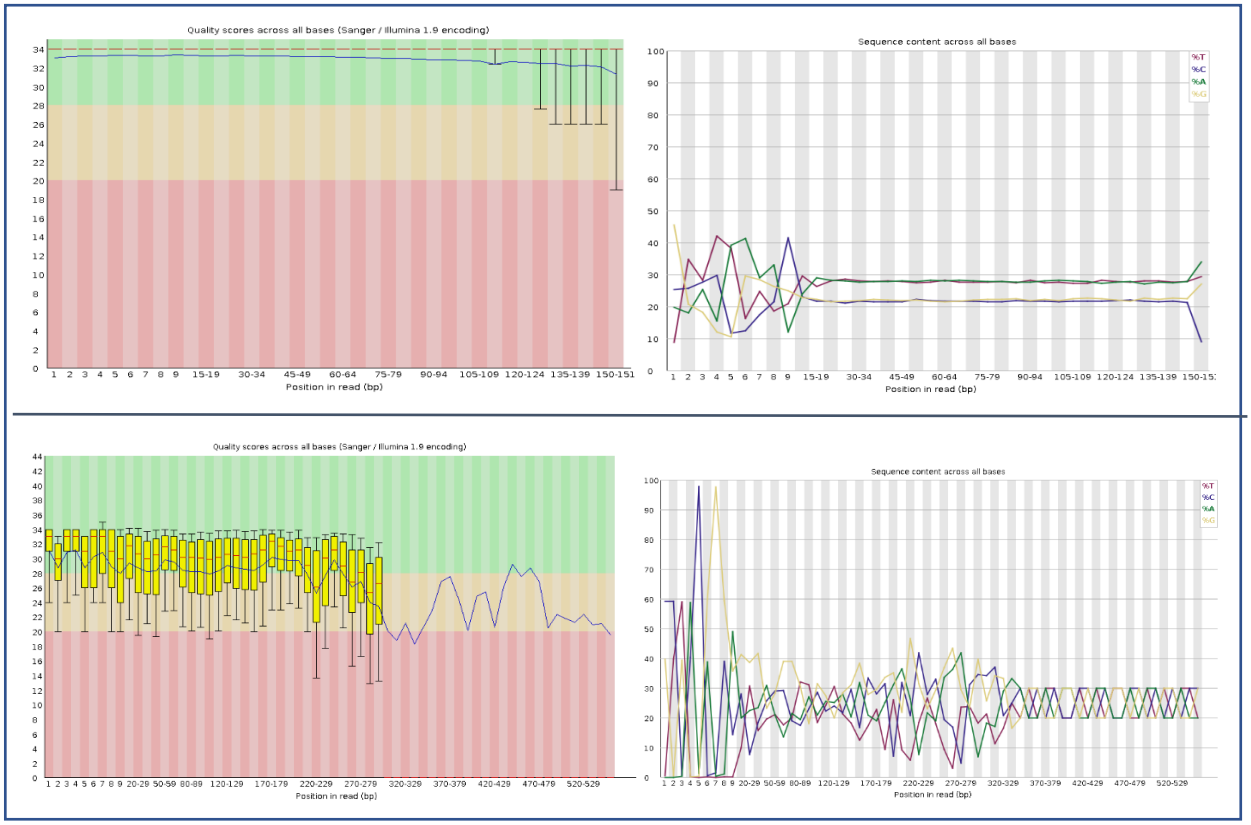
Addicionalment, es va comprovar la influència de l'Interval Post-mortem sobre el recompte total de delecions, el % d'Heteroplàsmia i la Longitud de deleció. Segons la Taula suplementaria 2 de l'Annex es va observar que, generalment, s'obtenia un major número de delecions si l'extracció de la mostra es produïa en les primeres 20 hores després de la mort del pacient, obtenint 40 delecions al grup control i 62 al grup amb esquizofrènia. Tot i així, el % d'Heteroplàmia i la Longitud de la deleció tendeix a ser major al grup control, com s'indica a la Taula Suplementaria 1 de l'Annex. D'aquesta manera, es demostra la possible degeneració de la mostra amb el temps i la importància d'extreure-la en un període de temps curt després de la mort del pacient.

Per altra banda, es van repetir els anàlisis de les característiques descriptives de les mostres però, aquest cop tenint en compte el sexe i l'edat dels participants, obtenint els resultats de la Taula suplementària 3 de l'Annex. No es van observar patrons en quant a l'edat però, les mostres de participants masculins van mostrar majors puntuacions per a la quantitat total de

delecions i, en alguns casos, també presentaven majors % d'Heteroplàsmia i Longitud de deleció (Taula suplementària 4). Aquest fet es va reafirmar mitjançant una prova U Mann Whitney (Taula suplementària 5) que va demostrar que les mitjanes del % d'Heteroplàsmia i Longitud de deleció eren significativament majors en el grup dels mascles.

4.5. Comparació entre *Ion Torrent* i *Illumina*

Un altre dels objectius del estudi és comparar les tècniques de seqüenciació d'*Illumina* i *Ion Torrent* per esbrinar quina ofereix uns resultats més fiables. Per tant, la manera de trobar aquesta informació és realitzant un anàlisi de la qualitat de les dades mitjançant l'eina *FastQC*. A continuació, es mostren gràficament els resultats obtinguts a l'analitzar la mateixa mostra seqüenciada per les dues tècniques esmentades, en el primer cas s'utilitza *Illumina*, en el segon *Ion Torrent*.



Imatge 8: Qualitat de les mostres seqüenciades per *Illumina* (a dalt) i *Ion Torrent* (a baix). Les dues imatges de l'esquerra corresponen als gràfics de Qualitat de seqüència per base i les dues de la dreta als gràfics lineals de Contingut de seqüència per base.

Com es pot observar per les puntuacions de qualitat entre bases, les dues mostres presenten una clara decaiguda en quant a la seva qualitat al final de la lectura. Aquest fenomen és força comú i la forma més senzilla de solucionar-ho es tallant el final de la seqüència amb *cutadapt*. Tot i així, s'observa que la majoria de la seqüència analitzada per *Illumina* es troba en la part superior

de l'àrea verda del gràfic, el que indica una bona qualitat i lectura al llarg de la seqüència. En el cas analitzat per *Ion Torrent*, es mostra que encara que el principi de la seqüència es representa a l'àrea verda, es troba molt a prop de la groga, a més de que gairebé la meitat de la seqüència es troba en aquesta segona franja. Aquest fet és indicatiu d'unes pitjors puntuacions en quant a la qualitat de la mostra seqüenciada.

Per altra part, en els gràfics de contingut de seqüencies per base es mostra un clar desequilibri entre el percentatge de bases al principi i al final de la seqüència analitzada per *Illumina*, fet que es podria solucionar tallant l'inici i el final de la seqüència amb *cutadapt*. A la mostra seqüenciada per *Ion Torrent* en canvi, aquest desequilibri és present amb major intensitat, tant al principi com al mig de la seqüència, el que impossibilitaria que es pugui solucionar amb *cutadapt*, ja que es perdria massa informació si es talla pel mig de la seqüència.

Globalment, i en la resta de mòduls del *FastQC*, s'han observat diferències en la qualitat de totes les mostres seqüenciades per *Illumina* i *Ion Torrent*.

En quant a la cobertura, tot i que moltes de les mostres obtenen puntuacions majors del 99%, hi ha una part important que no arriba al 10% o és queda en 0%. Per tant, es pot afirmar que la qualitat de les mostres obtingudes per *Ion Torrent* és més baixa que amb *Illumina*. Aquest fet implica també que la profunditat de lectura no sigui la mateixa entre els dos fragments seqüenciats, ja que, en molts dels casos, part de la seqüència no es llegeix.

A l'analitzar aquestes mostres amb *eKLIPse* no es va obtenir cap deleció en cap de les mostres, que és bastant esperable considerant els problemes de qualitat mencionats anteriorment.

Finalment, es va establir que per aquest estudi, *Illumina* és la tècnica de seqüenciació que ofereix una major qualitat de lectura i a partir de la qual s'obtindran les mostres pel seu posterior anàlisi amb *eKLIPse*.

5. Discussió

Segons els resultats de l'anàlisi realitzat per *eKLIPse* i la Prova T de U Mann Whitney, no és possible confirmar la presència de diferencies significatives en quant al % d'heteroplàsmia i la longitud de les delecions reportades entre el grup control i el grup de mostres provinents dels participants amb esquizofrènia. Encara que es mostri una major mediana de longitud de deleció en el grup amb esquizofrènia, aquesta diferència no és significativa. A més, tot i que les

longituds de les delecions siguin superiors en alguns casos de mostres de participants amb esquizofrènia, la freqüència en la que aquestes delecions es produeixen en el seu ADNmt és inferior que en el grup control.

Al estudiar més detalladament la qualitat de les mostres analitzades, s'ha observat que la gran majoria presenten una profunditat de lectura baixa i que pot veure's bruscament alterada (amb cobertures superiors o inferiors) en alguns punts de la seqüència. Aquest factor podria ser indicatiu d'una baixa qualitat en la lectura de les mostres, és per aquest motiu que es va analitzar la cobertura de la seqüenciació de les mostres per a comprovar aquesta teoria. Al presentar totes les mostres una cobertura de 100% i demostrar que els dos fragments (A i B) tenen la mateixa mediana de profunditat de lectura, es va rebutjar la idea de que la qualitat de lectura sigui insuficient, ja que es demostrava que no havia quedat cap part sense llegir ni cap desequilibri en la profunditat de lectura entre els dos fragments..

Un altre factor a considerar, és que el baix número de mostres analitzat impedeix construir afirmacions relacionades amb els resultats obtinguts amb la suficient seguretat com per a que es puguin extrapolar en casos externs a l'estudi. Aquesta és una dels una de les problemàtiques principals que es presenten en aquest projecte i un dels grans motius pels que encara es tenen uns coneixements tant limitats sobre l'esquizofrènia.

Aquest és el primer estudi que utilitza l'eina d'*eKLIPse* per a comprovar si els reordenaments de l'ADNmt poden tenir implicació en l'aparició de l'esquizofrènia. Per altra banda, tot i que l'ADN genòmic està força estudiat en relació als trastorns mentals (Trubetskoy et al., 2022), no és el mateix cas per a l'ADNmt. Existeixen estudis han trobat certs indicis del paper de les mutacions a l'ADNmt en alteracions com l'Autisme (Caporali et al., 2022) o el trastorn Bipolar (Angrand et al., 2021) però, existeixen pocs projectes destinats a l'estudi íntegre dels seus efectes en relació a l'esquizofrènia.

És important mencionar que alguns dels resultats obtinguts, com l'elevat % acumulat de delecions en alguns gens codificants per a subunitats del Complex I, coincideix amb els resultats obtinguts per altres estudis relacionats amb aquest àmbit (Roberts, 2021).

Respecte al programa d'anàlisi utilitzat, tot i que ofereix un estudi detallat i uns resultats molt gràfics i intuïtius per a poder entendre els reordenaments genètics presents a les mostres d'ADNmt estudiades, cal tenir en compte els possibles errors que pot cometre. És per aquest motiu que s'han desenvolupat eines alternatives a l'*eKLIPse* com *MitoSAlt*.

MitoSAlt és una eina computacional que permet la identificació, quantificació i visualització precisa de duplicacions i delecions presents a mostres provinents de la seqüenciació genètica. Permet estudiar tant mostres d'ADNmt obtingudes per tècniques de seqüenciació d'extrems simples com parells i presenta com a resultat un mapa de predicció sobre la possible presència de delecions i duplicacions, a més d'una taula amb informació detallada sobre els nivells d'heteroplàsmia i els punts d'interrupció (Basu et al., 2020).

Segons estudis previs en models de ratolins (Basu et al., 2020), *MitoSAlt* presenta la capacitat de detectar amb una elevada sensibilitat nivells baixos d'heteroplàsmia (0,5), fins i tot amb uns nivells moderats de cobertura de profunditat de lectura (Basu et al., 2020).

A més, aquest software és capaç d'estimar nivells relatius de mtDNA, que suposen una informació molt útil per a estudiar el número de còpies de mtDNA. Tot i que, per a assegurar que aquest resultats són fiables, és necessari considerar la presència d'alteracions estructurals grans, ja que poden afectar el nivell de mtDNA mesurat sense que es presentin canvis en el número de còpies (Basu et al., 2020).

Paral·lelament a aquets estudi, altres membres de l'equip de investigació van fer servir *MitoSAlt* per a analitzar els mateixos grups de mostres, obtenint diferències molt més significatives entre els dos grups. En aquest projecte, s'havia escollit utilitzar *eKLIPse* des d'un principi, ja que, es desconeixia el seu nivell d'efectivitat. Per tant, encara que el software d'*eKLIPse* no hagi demostrat la hipòtesi de que les alteracions de l'ADNmt estiguin implicades en les psicosis primerenques, és important no descartar definitivament aquesta idea i utilitzar altres eines i mètodes per a arribar a unes evidències més fiables. Per altra part, ha sigut possible mostrar mitjançant aquest estudi, alguns indicis de que *eKLIPse* pot ser no tan eficaç per a la detecció de reordenaments dins de l'ADNmt.

És important recordar que, com s'ha mencionat prèviament, l'esquizofrènia és una malaltia d'elevada complexitat en la que hi ha una important base genètica per a estudiar però, que aquesta també pot ser alterada pels diversos factors ambientals als que s'exposa l'organisme. Es per això que cal tenir en compte que, encara que una persona mostri certs caràcters genètics que indiquin una predisposició a patir aquesta afectació, l'ambient i el seu estil de vida poden resultar decisius per a que finalment es puguin donar o no els primers símptomes de psicosi.

6. Conclusions

L'esquizofrènia és un trastorn del neurodesenvolupament del que encara es coneix poc. Segons els resultats obtinguts, no es pot confirmar que el grup de mostres amb esquizofrènia presenta un major nivell de delecions, % d'heteroplàsmia i de longitud de deleció respecte el grup control. Per tant, no ha sigut possible confirmar amb seguretat la hipòtesi que els reordenaments mitocondrials intervinguin en l'aparició de les psicosis primerenques al comparar el dos grups.

Entre els gens amb major % acumulat de delecions es troben alguns codificants per a proteïnes dels Complexes I i III i alguns tRNAs mitocondrials. Tots són possibles causants de malalties mitocondrials però, els del Complex I podrien reafirmar la informació trobada prèviament a la bibliografia.

En quant a l'eina d'anàlisi utilitzada, tot i que el seu principal objectiu és predir i quantificar els possibles reordenaments mitocondrials presents en mostres d'ADNmt, aquest estudi demostra que pel seu insuficient grau de fiabilitat no pot ser utilitzat com a software de referència i ha d'estar acompanyat sempre per altres programes com MitoSAlt.

També és important recordar que l'aparició de l'esquizofrènia depèn d'una gran varietat de factors, aquest fet podria suposar que l'ADNmt tingui una menor implicació de l'esperada o que aquesta es compensi per altres tipus d'alteracions en l'organisme no estudiades encara. Aquest fenomen podria ser un important motiu per a estudiar amb una major profunditat els possibles mecanismes de compensació que es podrien presentar tant a nivell genètic com metabòlic per evitar l'inici d'aquest trastorn. També podria ser un punt de partida interessant per a l'estudi de noves teràpies per a prevenir o almenys, disminuir els símptomes més incapacitants de l'esquizofrènia. Tot i això, encara falta un llarg camí i molts projectes d'investigació per a arribar fins a aquest punt.

Part d'aquesta variabilitat genètica de l'ADNmt pot ser deguda també per factors externs o per les característiques pròpies del organisme, com es va mostrar en els estudis addicionals del número de delecions, % d'Heteroplàsmia i Longitud de deleció respecte, l'edat, el sexe i l'Interval Post-mortem.

En referència, a la tècnica de seqüenciació més adequada per aquest estudi, tot i els problemes exposats, *Illumina* és la que ofereix uns resultats de major qualitat i fiabilitat respecte *Ion Torrent*.

7. Bibliografia

Angrand, L., Boukouaci, W., Lajnef, M., Richard, J. R., Andreazza, A., Wu, C. L., Bouassida, J., Rafik, I., Foiselle, M., Mezouad, E., Naamoune, S., Chami, L., Mihoub, O., Salah, S., Benchaaben, A., Le Corvoisier, P., Barau, C., Costes, B., Yolken, R., … Tamouza, R. (2021). Low peripheral mitochondrial DNA copy number during manic episodes of bipolar disorders is associated with disease severity and inflammation. *Brain, Behavior, and Immunity*, *98*, 349–356. https://doi.org/10.1016/j.bbi.2021.09.003

Babraham Bioinformatics. (2023). *FastQC_Manual* (0.12.0). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Basu, S., Xie, X., Uhler, J. P., Hedberg-Oldfors, C., Milenkovic, D., Baris, O. R., Kimoloi, S., Matic, S., Stewart, J. B., Larsson, N. G., Wiesner, R. J., Oldfors, A., Gustafsson, C. M., Falkenberg, M., & Larsson, E. (2020). Accurate mapping of mitochondrial DNA deletions and duplications using deep sequencing. *PLoS Genetics*, *16*(12). https://doi.org/10.1371/journal.pgen.1009242

Caporali, L., Fiorini, C., Palombo, F., Romagnoli, M., Baccari, F., Zenesini, C., Visconti, P., Posar, A., Scaduto, M. C., Ormanbekova, D., Battaglia, A., Tancredi, R., Cameli, C., Viggiano, M., Olivieri, A., Torroni, A., Maestrini, E., Rochat, M. J., Bacchelli, E., … Maresca, A. (2022). Dissecting the multifaceted contribution of the mitochondrial genome to autism spectrum disorder. *Frontiers in Genetics*, *13*. https://doi.org/10.3389/fgene.2022.953762

Das, S. C., Hjelm, B. E., Rollins, B. L., Sequeira, A., Morgan, L., Omidsalar, A. A., Schatzberg, A. F., Barchas, J. D., Lee, F. S., Myers, R. M., Watson, S. J., Akil, H., Bunney, W. E., & Vawter, M. P. (2022). Mitochondria DNA copy number, mitochondria DNA total somatic deletions, Complex I activity, synapse number, and synaptic mitochondria number are altered in schizophrenia and bipolar disorder. *Translational Psychiatry*, *12*(1). https://doi.org/10.1038/s41398-022-02127-1

Goudenège, D., Bris, C., Hoffmann, V., Desquiret-Dumas, V., Jardel, C., Rucheton, B., Bannwarth, S., Paquis-Flucklinger, V., Lebre, A. S., Colin, E., Amati-Bonneau, P., Bonneau, D., Reynier, P., Lenaers, G., & Procaccio, V. (2019). eKLIPse: a sensitive tool for the detection and quantification of mitochondrial DNA deletions from next-generation sequencing data. *GENETICS in MEDICINE*, *21*(6), 1407–1416. https://doi.org/10.1038/s41436

Ivanova, E. M., Kandilarova, S. M., Lukanov, T. I., Naumova, E. J., Akabalieva, K. V., & Milanova, V. K. (2021). NGS-based mtDNA Profiling Could Reveal Genetic Alterations in Schizophrenia. Current topics in medicinal chemistry, 21(11), 938–948. https://doi.org/10.2174/1568026621666210521155500

Khavari, B., & Cairns, M. J. (2020). Epigenomic Dysregulation in Schizophrenia: In Search of Disease Etiology and Biomarkers. In *Cells* (Vol. 9, Issue 8). NLM (Medline). https://doi.org/10.3390/cells9081837

Kumar, P., Efstathopoulos, P., Millischer, V., Olsson, E., Wei, Y. B., Brüstle, O., Schalling, M., Villaescusa, J. C., Ösby, U., & Lavebratt, C. (2018). Mitochondrial DNA copy number is associated with psychosis severity and anti-psychotic treatment. Scientific reports, 8(1), 12743. https://doi.org/10.1038/s41598-018-31122-0

Larson, N. B., Oberg, A. L., Adjei, A. A., & Wang, L. (2023). A Clinician's Guide to Bioinformatics for Next-Generation Sequencing. In *Journal of Thoracic Oncology* (Vol. 18, Issue 2, pp. 143–157). Elsevier Inc. https://doi.org/10.1016/j.jtho.2022.11.006

Love, J., Dropmann, D., Selker, R., Gallucci, M., Jentschke, S., Balci, S., Seol, H., & Agosti, M. (2022). *The jamovi project* (2.3). https://www.jamovi.org/about.html

Pereira, R., Oliveira, J., & Sousa, M. (2020). Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. In *Journal of Clinical Medicine* (Vol. 9, Issue 1). MDPI. https://doi.org/10.3390/jcm9010132

Pervez, M. T., Hasnain, M. J. U., Abbas, S. H., Moustafa, M. F., Aslam, N., & Shah, S. S. M. (2022). A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. In *BioMed Research International* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/3457806

Rees, E., & Kirov, G. (2021). Copy number variation and neuropsychiatric illness. In *Current Opinion in Genetics and Development* (Vol. 68, pp. 57–63). Elsevier Ltd. https://doi.org/10.1016/j.gde.2021.02.014

Roberts, R. C. (2021). Mitochondrial dysfunction in schizophrenia: With a focus on postmortem studies. *Mitochondrion*, *56*, 91–101. https://doi.org/10.1016/j.mito.2020.11.009

Sandra, E. C., Erin Munkácsy, & Pickering, A. M. (2018). Cause or Casualty: The Role of Mitochondrial DNA in Aging and Age-Associated Disease. *Biochim Biophys Acta Mol Basis Dis*, 285–297. https://doi.org/10.1016/j.bbadis

Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., Bryois, J., Chen, C. Y., Dennison, C. A., Hall, L. S., Lam, M., Watanabe, K., Frei, O., Ge, T., Harwood, J. C., Koopmans, F., Magnusson, S., Richards, A. L., Sidorenko, J., … van Os, J. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502–508. https://doi.org/10.1038/s41586-022-04434-5

Valiente-Pallej, A., Tortajada, J., Bulduk, B. K., Vilella, E., Oria Garrabou, G., Muntan, G., & Martorell, L. (2022). Comprehensive summary of mitochondrial DNA alterations in the postmortem human brain: A systematic review. *EBioMedicine*, *76*, 103815. https://doi.org/10.1016/j

Valiente-Pallejà, A., Torrell, H., Alonso, Y., Vilella, E., Muntané, G., & Martorell, L. (2020). Increased blood lactate levels during exercise and mitochondrial DNA alterations converge on mitochondrial dysfunction in schizophrenia. *Schizophrenia Research*, *220*, 61–68. https://doi.org/10.1016/j.schres.2020.03.070

Yan, C., Duanmu, X., Zeng, L., Liu, B., & Song, Z. (2019). Mitochondrial DNA: Distribution, mutations, and elimination. In *Cells* (Vol. 8, Issue 4). MDPI. https://doi.org/10.3390/cells8040379

8. Autoavaluació

Tot i que els resultats finals del meu estudi no van resultar tant concloents i prometedors com esperava, considero que el meu període de pràctiques en l'equip d'investigació en Genètica i Ambient en Psiquiatria de la Facultat de Medicina i Ciències de la Salut de la URV, com també el treball que vaig poder realitzar van resultar molt útils per a continuar formant-me acadèmicament i adquirir una major experiència en l'àmbit de la investigació.

Durant les pràctiques vaig poder ampliar les meves bases de genètica i bioinformàtica, ja que, vaig poder experimentar amb la part més pràctica d'aquests coneixements, que és com obtenir uns resultats finals a partir de mostres obtingudes al laboratori. En aquest sentit, he aprés més en relació al llenguatge de programació de *Python* i *Biopython*, com també he aprés a crear una màquina virtual, instal·lar, utilitzar i programar en el sistema operatiu de *Linux*.

També he aprés altres mètodes estadístics que no havia vist en les meves classes teòriques, com l'estudi de distribució normal de dades de Kolmogorov-Smirnov i les proves de hipòtesis de U Mann Whitney.

Per altra part, vaig fer servir i vaig entendre les bases de funcionament d'eines d'anàlisi de seqüències genètiques com *FastQC, cutadapt* i *eKLIPse*, com també vaig aprendre a utilitzar altres eines de càlcul estadístic diferents a l'*Excel*, com *jamovi*.

També vaig obtenir coneixements més detallats en quant a la genètica i el funcionament mitocondrial i les seves implicacions en molts trastorns humans, com per exemple els mentals. A més, vaig aprendre més sobre l'esquizofrènia, el funcionament del cervell i els trastorns del neurodesenvolupament gràcies als seminaris impartits cada setmana durant les pràctiques.

És molt important destacar també el fet de que vaig poder experimentar la vertadera dinàmica de com és treballar en equip dins d'un departament de recerca i com comunicar-me millor a l'hora de realitzar un projecte científic. A més, vaig poder perfeccionar la meva comunicació tant oral com escrita en anglès, gràcies al fet de que una de les supervisores del meu treball és estrangera, pel que només ens podíem comunicar en anglès, i també perquè vaig haver de contactar amb alguns dels programadors de les eines bioinformàtiques d'anàlisis utilitzades.

9. Annexos

Taula suplementària 1: Caràcteristiques descriptives de les dades del grup control: HC i amb esquizofrènia: SCZ segons l'Interval Post-mortem (PMD), NaN: no hi ha información disponible

									Percentils		
	Grup	**PMD (h)**	**Mitjana**	**Mediana**	**Desviació Estàndard**	**Rang Interquartil**	**Mínim**	**Màxim**	**25th**	**50th**	**75th**
% Heteroplàsmia	HC	1-20	2,853	1,119	4,7261	2,3792	0,509	22,150	0,758	1,119	3,137
		20-40	1,271	0,801	0,9838	0,9020	0,514	3,613	0,752	0,801	1,654
		40-60	1,243	1,243	NaN	0,0000	1,243	1,243	1243	1.243	1,243
	SCZ	1-20	1,507	1,010	1,4891	0,7678	0,502	7,861	0,701	1,010	1,469
		20-40	2,333	1,429	1,8435	3,0237	0,553	5,626	0,913	1,429	3,936
		40-60	0,585	0,561	0,0794	0,0640	0,518	0,699	0,541	0,561	0,605
Longitud de la deleció	HC	1-20	7347,375	5424,000	6475,3856	13681,5000	101	16317	2039,500	5424,000	15721,000
		20-40	4056,556	3235	4759,5987	2232,0000	176	16083	1757,000	3235,000	3989,000
		40-60	4027,000	4027	NaN	0,0000	4027	4027	4027,000	4027,000	4027,000
	SCZ	1-20	4519,871	4172,000	1986,4494	1442,5000	109	10323	3420,500	4172,000	4863,000
		20-40	3991,143	4120,000	4389,2912	4379,7500	100	16095	211,250	4120,000	4591,000
		40-60	2141,250	1726,500	1088,3046	994,2500	1391	3721	1436,750	1726,500	2431,000

Taula suplemetaria 2: Número de delecions per grup (control: HC o amb esquizofrènia: SCZ) segons l'Interval Post-mortem

GRUP	**PMD**	**Nombre total de delecions**
HC	1-20	40
	20-40	9
	40-60	1
SCZ	1-20	62
	20-40	14
	40-60	4

Taula suplementària 3: Característiques descriptives de les mostres segons el grup (control HC o amb esquizofrènia SCZ), el sexe i l'edat, NaN: No hi ha informació disponible

										Percentils		
	Grup	**Sexe**	**Edat**	**Mitjana**	**Mediana**	**Desviació Estàndard**	**Rang Interquartil**	**Mínim**	**Màxim**	**25th**	**50th**	**75th**
% Heteroplàsmia	HC	Masculí	20-40	1,956	1,721	1,184	2,333	0,526	3,27	0,818	1,721	3,151
			40-60	1,138	0,851	0,840	0,444	0,509	4,17	0,750	0,851	1,194
			>60	5,344	1,458	7,789	3,867	0,514	22,15	0,952	1,458	4,819
		Femení	20-40	NaN	NaN	NaN	.	NaN	NaN	NaN	NaN	NaN
			40-60	1,652	0,771	1,701	1,520	0,573	3,61	0,672	0,771	2,192
			>60	2,410	0,936	3,330	1,436	0,516	9,64	0,726	0,936	2,162
	SCZ	Masculí	20-40	0,787	0,714	0,285	0,460	0,502	1,26	0,541	0,714	1,001
			40-60	1,605	1,160	1,441	1,017	0,514	7,58	0,699	1,160	1,716

Taula suplementària 3: Característiques descriptives de les mostres segons el grup (control HC o amb esquizofrènia SCZ), el sexe i l'edat, NaN: No hi ha informació disponible

										Percentils		
	Grup	Sexe	Edat	Mitjana	Mediana	Desviació Estàndard	Rang Interquartil	Mínim	Màxim	25th	50th	75th
Longitud de la deleció	HC	Femení	>60	0,895	0,788	0,358	0,119	0,589	1,60	0,740	0,788	0,859
			20-40	1,294	1,143	0,639	0,751	0,739	2,15	0,843	1,143	1,594
			40-60	1,677	0,902	1,837	0,658	0,557	7,86	0,702	0,902	1,360
			>60	4,258	4,842	1,819	2,005	1,723	5,63	3,548	4,842	5,553
		Masculí	20-40	7363,455	4027	7066,197	13356,000	101	16317	1169,000	4027,000	14525,000
			40-60	6293,765	3293	6544,225	13560,000	120	16149	2132,000	3293,000	15692,000
			>60	6925,333	3876,000	6627,860	12472,250	120	16080	2779,250	3876,000	15251,500
	SCZ	Femení	20-40	NaN	NaN	NaN	.	NaN	NaN	NaN	NaN	NaN
			40-60	5498,333	236	9166,639	7953,500	176	16083	206,000	236,000	8159,500
			>60	6691,429	7280	3025,564	675,500	109	9440	7165,000	7280,000	7840,500
		Masculí	20-40	5410,500	4826,500	5685,805	8220,250	109	16095	221,750	4826,500	8442,000
			40-60	3948,703	3775	1624,010	1211,000	1391	10323	3376,000	3775,000	4587,000
			>60	1928,667	1710,500	2044,579	3223,250	100	4591	113,500	1710,500	3336,750
		Femení	20-40	4063,750	4186,500	741,843	735,250	3072	4810	3757,500	4186,500	4492,750
			40-60	4964,476	4658	2143,895	1234,000	1433	9101	3889,000	4658,000	5123,000
			>60	5801,500	4733,500	2256,836	1369,000	4559	9180	4583,000	4733,500	5952,000

Taula suplementària 4: Número de delecions per grup (control: HC o amb esquizofrènia: SCZ) segons el sexe i l'edat.

GRUP	Sexe	Edat	Nombre total de delecions
HC	Masculí	20-40	11
		40-60	17
		>60	12
	Femení	20-40	0
		40-60	3
		>60	7
SCZ	Masculí	20-40	8
		40-60	37
		>60	6
	Femení	20-40	4
		40-60	21
		>60	4

Taula suplementària 5: Prova T de U Mann Whitney per a Mostres Independents (grup control i grup amb esquizofrènia) segons la hipòtesi de que la Mitjana del % d'heteroplàsmia i longitud de delecions és major en homes que en dones.

		Estadístic	p
% Heteroplàsmia	U de Mann-Whitney	1649	0,739
Longitud de la deleció	U de Mann-Whitney	1320	0,990

Anotació. H_a $\mu_{Masculí} > \mu_{Femení}$