Recherche de gènes candidats responsables du Syndrome d'Aicardi : Complémentarité des approches expérimentales et bioinformatiques

Saliha Yilmaz

Nancy-Université

FACULTE DES SCIENCES & TECHNIQUES

U.F.R Sciences & Techniques biologiques
Ecole Doctorale Biologie, Santé et Environnement (BioSE)

Thèse

présentée pour l'obtention du titre de

Docteur de l'Université Henri Poincaré, Nancy-I

Spécialité Génomique

par **Saliha YILMAZ**

Recherche de gènes candidats responsables du Syndrome d'Aicardi : Complémentarité des approches expérimentales et bioinformatiques

Soutenue publiquement le 7 Novembre2007

Directeur de thèse : Professeur Philippe JONVEAUX

Membres du jury :

Rapporteurs :	Professeur Yves MOREAU Professeur Joris Robert VERMEESCH
Examinateurs :	Docteur Marie-Dominique DEVIGNES Docteur John Louis McGREGOR Docteur Roberto INCITTI Professeur Bruno LEHEUP Professeur Philippe JONVEAUX
Invité d'honneur :	Professeur Jean AICARDI

Laboratoire de Génétique Médicale, CHU Nancy-Brabois, Rue du Morvan, 54511 Vandoeuvre-lès-Nancy

A Annick Perroux, ma tata…

Sans toi ce travail n'existerait pas

Quel bonheur de t'avoir connue

A la jolie Anne-Lorène…

… à toutes les filles Aicardi de part le monde et aux familles qui ont participé à ce projet…

Remerciements

Je remercie tout particulièrement mon Directeur de thèse, le Professeur Philippe Jonveaux. Merci de m'avoir accueillie dans votre laboratoire, de m'avoir accordé votre confiance et de m'avoir permis de confirmer mon goût (devenue passion) pour la génétique humaine.

Mes remerciements vont également à Christophe Philippe. Tu as été présent à chaque fois que j'ai eu besoin de toi, merci pour ta générosité et ton altruisme.

Je remercie également Marie-José Grégoire. Merci de m'avoir fait profiter de votre esprit cartésien. Clarté et simplicité …

Toute ma profonde gratitude pour Monsieur Jean Aicardi. Merci Monsieur de me faire l'honneur de votre présence dans mon jury.

Yves Moreau et Joris Robert Vermeesch m'ont fait l'honneur d'être les rapporteurs de cette thèse, et je les en remercie, de même que pour leur participation au Jury.

Bien sur, je remercierai les deux représentantes de l'équipe orpailleur Marie-Dominique Devignes et Malika Smaïl-Tabbone. Merci à vous, très sincèrement je n'ai cessé d'apprendre et j'ai encore soif…

Je tiens à remercier Roberto Incitti pour les conseils stimulants et les suggestions enrichissantes que j'ai reçus de sa part.

Je remercie également Bruno Leheup d'avoir accepté de participer au Jury de soutenance. Votre savoir m'a toujours impressionnée.

*Je tiens à dire un grand Merci à John Louis McGregor. Merci de m'avoir si naturellement et si gentiment accueillie dans votre équipe à Londres. Je suis impatiente à l'idée de commencer un tel projet avec vous et Sophie. Une nouvelle aventure dans une nouvelle vie et une ville où, quoi qu'on en dise, il ne pleut pas** * autant qu'à Nancy.

J'en profite pour remercier Flo, Marion, Sophie, Chris, Kira et Thibault pour tous les bons moments que nous avons partagés ensemble durant mes séjours à Londres.

Un grand merci à Karène pour sa gentillesse et sa disponibilité. Merci à Hervé et Cédric. Et bien sur à Stéphane, un vrai cadeau en tant que premier stagiaire…

Et puis, il y a tout ce monde autour de moi, dont beaucoup de personnes d'exception. Des personnes que je connaissais ou que j'ai rencontrées et qui ont fait que ces années ont été aussi riches et agréables.

Mes amis de toujours, Eléonore et Manu, Valérie et Gaby. Je suis à la traîne mais j'ai presque fini…

Marie Jo, merci pour ton amitié et ta sincérité.

Ilham. Je ne désespère pas de te voir un jour chez moi…idem, tu vas me dire…

Micheline. Toi heureusement que tu es là…et pour beaucoup de monde j'en suis sur.

Aline et son petit bout de chou…qui va bientôt nous rejoindre.

Merci à Khaled, Cindy, Üstün, Nizar, Adrien, Lazslo…avec qui j'ai partagé cette dernière année de thèse si agréable.

Je souhaite remercier toutes les personnes, qui d'une façon ou une autre m'ont épaulée : soit au sein du laboratoire de génétique que celui du Loria … grâce à leur gentillesse, leur disponibilité et leur compétence, ils m'ont enrichie pour mener à bien ce fabuleux projet.

Merci à mes parents de m'avoir toujours poussé à réaliser ce que je souhaitais faire. D'avoir été là, toujours à mes cotés.

Merci à mes frères Okan et Ozcan. Promis, à l'avenir je serai moins stressée et plus disponible.

Merci à ma sœur Sevgi. Tu tombes toujours à pic … merci

Et merci à mon porte bonheur. Quelle patience tu as…

Table des matières

Table des matières

Matériel et Méthodes

Résultats

Discussion

Conclusion

Listes des figures et tableaux

Liste des figures

Liste des tableaux

Abréviations

Abréviations

ACGR : Approach for Candidate Gene Retrieval
ACP : Analyse en Composantes Principales
ACRC : acidic repeat containing
ADN : acide désoxyribonucléique
ADNc : ADN complémentaire
AI : Amélogenèse imparfaite
AICARDI DS-GO : Aicardi disease-specific GO
AIF1 : allograft inflammatory factor 1
AMELX : amelogenin
AMPA : a-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
AMPA-R : alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor
ANXA1 : annexin A1
ANOVA : ANalyse Of Variance
AR : Androgen receptor
ARHGAP6 : Rho GTPase activating protein 6
ARHGEF6 : Rac/Cdc42 guanine nucleotide exchange factor 6
ARMCX3 : armadillo repeat containing, X-linked 3
ARN : acide ribonucléique
ARNc : ARN complémentaire
ARNm : ARN messager
ARX : aristaless
BAC : Bateriale Artificial Chromosome
Bandes GTG : bandes G, Trypsine, Giemsa
BD : base de données
Bdnf : brain-derived neurotrophic factor
BET : bromure d'étidium
BP : Biological Process (Processus Biologique)
CBX5 *:* chromobox homolog 5 (HP1 alpha homolog, Drosophila
CC : Cellular Component (composant cellualaire)
CCDC68 : coiled-coil domain containing 68
CDKL5 : cyclin-dependent kinase-like 5
CENPA : centromere protein A
CENPJ : centromere protein J
CGH : Comparaive Genomic Hybridization
CHARGE : Coloboma, Heart anomalies, choanal Atresia, reatardation, Genital and Ear anomalies
CHD7 : chromodomain helicase DNA binding protein 7
CIA : communication interauriculaire
CLEC2B : C-type lectin domain family 2, member B
CNPs : Copy Number Polymorphisms
CNV : copy number variation

COG : Clusters of Orthologous Groups
CREB : cAMP responsive element binding protein
CRSP : cofactor required for SP1 activation
CRSP2 : cofactor required for Sp1 transcriptional activation, subunit 2
Ct : threshold cycle
CUGN (Communauté Urbain du Grand Nancy)
CXCL13 : (C-X-C motif) ligand 13
ddNTP : didésoxynucléotides
DLG3 : (discs, large homolog 3)
DLX5 : distal
DMD : dystrophin (muscular dystrophy, Duchenne and Becker types
DMSO : dimétylsulfoxyde
DO : densité optique
DS : déviation standard
Dup : duplication
EASE : Expression Analysis Systematic Explorer
EMBL : European Molecular Biology Laboratory
EST : expressed sequence tag
EVOC : Expressed Sequence Annotation for Humans
Fbbi : image ontology
FBbt : fly_anatomy
FBdv : fly_development
FGFR3 : fibroblast growth factor receptor 3
FLNA : filamin A
Flybase : Database of Drosophila Genes & Genomes
FMR1 : fragile X mental retardation protein1
FMRP : fragile X mental retardation 1
GABA : gamma-aminobutyric acid
GAGE7 : G antigen 7
GEO : Gene Expression Omnibus ou array express
GO : Gene Ontology
GOA : The Gene Ontology Annotation
H2AFB3: H2A histone family, member B3
HCCS : holocytochrome c synthase
HD : Huntington Disease
HDAC1 : Histones désacétylases 1
HDAC2 : histone deacetylase 2
HEK : Human Embryonic Kidney
HIST1H1E : histone 1, H1e
HIST1H2AH : histone 1, H2ah
HIST1H2BK : histone 1, H2bk
HIST1H3H : histone 1, H3h
HIST1H4I : histone 1, H4i
HIST1H4L : histone 1, H4l
HIST2H2AA : histone 2, H2aa

HNRPDL : heterogeneous nuclear ribonucleoprotein D-like
HP1 : Heterochromatin protein-1
HUMARA : Human Androgen Receptor A
IFI27 : interferon, alpha-inducible protein 27
IGKC : immunoglobulin kappa constant
IL27 : interleukin 27
IRM : Imagerie par Résonance Magnétique
ISVs : Intermediate Size Variants [104]
kb : kilobase
KEGG : Kyoto Encyclopedia of Genes and Genomes
LCR : Low Copy Repeat
LCRs : Large-scale Copy Repeat [99]
ligand 13 : B-cell chemoattractant
LINES : Long Interspersed
LORIA : Laboratoire Lorrain d'Informatique est ses Applications
MGD : Mouse genome Database
MAGED1 : melanoma antigen family D, 1
MAGE-ML : Microarray Gene Expression Markup Language
MAGUK : membrane associated guanylate kinase
MAOA *:* monoamine oxidase A
MAPK : mitogen-activated protein kinase
MECP2 : methyl CpG binding protein 2
MCPH1 : microcephaly, primary autosomal recessive 1
MF : Molecular Fonction (fonction moléculaire)
MGED : Microarray Gene Expression Data Society
MGI : Mouse Genome Informatics
mGluR : metabotropic glutamate receptor
mGluR5-R : mGluR5 receptor complex
MIAME : Minimum Information About a Microarray Experiment
MID1 : midline 1
min : minute
ml : millilitre
MLS : Microphtalmia with Linear Skin defects
mM : millimolaire
MP : Mammalian Phenotype
MSL3L1 : male-specific lethal 3-like 1
NCBI : National Center for Biotechnology Information
NF1 : Neurofibromin 1
NIH : National Institutes of Health
NLGN4 : neuroligin 4
NMDA : ion-channel–forming Nmethyl-D-aspartic acid
NRC/MASC : NMDA receptor complex/MAGUK-associated signaling complex
NUF2 : NDC80 kinetochore complex component, homolog (S. cerevisiae)
OBO : Open Biomedical Ontologies
OMIM : Online Mendelian Inheritance in Man

OPHN1 : oligophrenin 1
p : bras court d'un chromosome
PAK3 : p21 activated kinase 3
pb : paire de bases
PBS : phosphate buffered saline
PCM1 : pericentriolar material 1
PCR : polymerase chaine reaction
PIR-PSD : Protein InformationResource-Internationnal Protein Sequence Database
PMP22 : periferal myeline protein 22
PORCN : porcupine homolog
PSD : post synaptic density
PSP : proteome post-synaptique
PSP : Protéome post-synaptique :
PLXNA3 : plexin A3
PLXNB3 : plexin B3
q : bras long d'un chromosome
QI : Quotient intellectuel
Q-PCR : quantitative-PCR
RM : Retard mental
RMLX : Retard mental lié au chromosome X
RMX : Retard mental non syndromique ou non spécifique
RMXS : Retard mental syndromique
RPL21 : ribosomal protein L21
RPL26 : ribosomal protein L26
RPL31 : ribosomal protein L31
RPL34 : ribosomal protein L34
RT : Reverse transcriptase
RT-PCR : reverse transcriptase polymerase chain reaction
sec : seconde
SGD : Saccharomyces Genome Databse
SINE : Short INTerspersed
SNP : single nucleotide polymorphisms
SO : sequence ontology
SP1 : Sp1 transcription factor
SQL : Structured query language
SSC : solution saline citrate
SUV39H1 : suppressor of variegation 3-9 homolog 1
SUV39H3: suppressor of variegation 3-9 homolog 1
SVF : Sérum de veau fœtal
SYN1 : Synapsin 1
t(X ;A) : translocation entre un chromosome X et un autosome
Taq : *thermophilus aquaticus*
TCF4 : transcription factor 4
TCTE1L : dynein, light chain, Tctex-type 3
TIGR : The Institute for Genomic Research

TRA1 : tumor rejection antigen (gp96) 1
TRI : Thrombosis Research Institute
Tris-HCL : tris(hydroxyméthyl)-aminométhane chlorhydrate
UBD : ubiquitin D
UBE3A : ubiquitin protein ligase E3A
VLDRLR : very low density lipoprotein receptor
WBP5 : WW domain binding protein 5
WISC-R : Intelligence Scale for Children-Revised
XIC : X Inactivation Center
XIST : X Inactivation Specific Transcript
ZNF92 : zinc finger protein 92
°C : degré celsius
µg : microgramme
µl : microlitre
% : pourcent

Préambule

Préambule

Le syndrome d'Aicardi est une maladie génétique rare (estimée à 1 cas sur 500 000). Le diagnostic est porté classiquement devant la survenue précoce de spasmes infantiles, une agénésie du corps calleux et des lacunes chorio-rétiniennes avec un handicap cognitif souvent sévère. Toutefois, les données récentes font apparaître un spectre clinique plus large. Cette affection s'observe essentiellement chez les filles et chez les rares garçons avec la formule chromosomique 47,XXY. A une exception près, aucun cas familial n'a été décrit. Le mode d'hérédité évoqué est dominant lié au chromosome X.

Le laboratoire de génétique (EA 4002-IFR111) s'est intéressé au syndrome d'Aicardi sous l'initiative de deux personnes. Conseillés par Laurence Hirsch, directrice de la communication de la CUGN (Communauté Urbain du Grand Nancy), Annick Perroux fondatrice de l'association AAL Syndrome d'Aicardi et son président Olivier Vauchelet ont rencontré le Professeur Jonveaux pour faire connaître leur volonté d'être impliqués dans la recherche. Le premier contact établi, le projet rentrant bien dans le domaine de compétence du laboratoire, les travaux de recherche ont été initiés par le recrutement de familles. J'ai débuté mon travail de thèse sur le syndrome d'Aicardi en hiver 2004, avec, il est vrai, 3 patientes AIC, expliquant notre stratégie initiale de recherche. Grâce à l'Association AAL-Syndrome d'Aicardi et aux réseaux de neuropédiatres et généticiens, nous disposons en août 2007 de 24 familles Aicardi originaires de toute l'Europe.

Ce travail de thèse a donc été orienté vers la recherche de gènes candidats impliqués dans la survenue du syndrome d'Aicardi.

Une première partie de ce travail de thèse introduit les retards mentaux et particulièrement les retards mentaux liés au chromosome X au sein desquels se positionne le syndrome d'Aicardi. Une partie de cette introduction est plus particulièrement consacrée aux connaissances actuelles sur le syndrome d'Aicardi et les études entreprises pour la recherche de gènes candidats. Compte tenu du caractère sporadique de cette affection, grevant toute possibilité d'étude par liaison génétique, des nouvelles approches ont donc été envisagées.

Les puces à ADN représentent probablement la révolution technologique la plus importante de ces dernières années. Elles donnent accès à une vision globale de la cellule tant au niveau

de l'ADN (puces génomiques) qu'au niveau de l'ARN (puces transcriptomiques). Les puces génomiques ont abouti à la découverte de gènes impliqués dans des maladies monogéniques. L'étude du transcriptome présente également un intérêt certain dans la compréhension des mécanismes sous-jacents la maladie. L'étude à un temps « t » de l'état transcriptomique de la cellule apporte des informations précieuses concernant les dérégulations dues à une maladie. L'analyse et l'intégration des résultats issus de ces expériences sur puces passent par l'utilisation des connaissances accumulées dans les bases de données biologiques publiques. Les apports de la bioinformatique à la recherche de gènes candidats représentent la dernière partie de cette introduction.

Nous avons utilisé ces approches globales et intégratives dans le cadre du Syndrome d'Aicardi. Une première étude transcriptomique (AS1) a été menée sur trois patientes (AIC) avec les lignées lymphoblastiques initialement disponibles au laboratoire. La stratégie utilisée des puces 22K a conduit à la sélection initiale de 16 gènes et le séquençage de 6 d'entre eux à la recherche de mutations délétères chez les patientes AIC. Une étude (AS2) à partir cette fois d'ARN de sang a pu être conduite dans un second temps sur 10 familles Aicardi et à l'aide de puces 41K (« whole human genome »). L'analyse de ces puces a tenu compte des nouveaux outils disponibles à ce moment de l'étude. Le caryotype moléculaire, de 18 filles AIC a par la suite été étudié à l'aide d'une puce génomique spécifique du chromosome X issue du Flanders Interuniversity Institute of Human Genetics. L'interprétation des puces transcriptomique nécessitant l'intégration des données disponibles dans les bases de données biologiques, une étude *in silico* a ainsi été initiée avec le LORIA (Laboratoire lorrain d'Informatique et ses Applications). Un prototype ACGR (Approach for Candidate Gene Retrieval) a été mis en place et utilisé pour la recherche de gènes candidats potentiels pour le syndrome d'Aicardi grâce à l'intégration des connaissances des bases de données biologiques. Le déroulement des expériences a été en majorité tributaire de la disponibilité des échantillons biologiques et de l'accessibilité du plateau technique de puce transcriptomique. Les expériences de puces transcriptomiques ont été réalisées sous la direction de John McGregor au TRI (Thrombosis Research Institute) à Londres. Une partie des analyses (prétraitement, sélection des gènes candidats) ont été réalisées sur place et les étapes postérieures (validation des résultats, interprétations des données) au laboratoire de génétique de Nancy. La partie in silico a été réalisée au sein des

locaux du LORIA à Vandoeuvre-les-Nancy et sous la direction de Marie-Dominique Devignes et Malika Smaïl-Tabbone.

Les résultats et gènes candidats potentiels impliqués dans le syndrome d'Aicardi seront discutés au regard des données de la littérature.

Ces travaux ont été rendus possibles par le soutien financier de la Communauté Urbaine du Grand Nancy, de la Région Lorraine et de l'association AAL-Syndrome d'Aicardi.

Introduction

Cette session est une entrée en matière qui me permettra de repositionner mon travail de thèse dans le contexte des connaissances actuelles. Je commencerai par une présentation des retards mentaux, puis des retards mentaux liés à l'X dont fait partie le syndrome d'Aicardi qui fera l'objet du troisième chapitre de cette introduction. Je consacrerai les deux derniers chapitres à l'exposition des stratégies existantes et utilisées pour l'étude de cette maladie.

1. Retard mental

1.1. Définition

Selon les critères du DSM-IV [1], le retard mental (RM) est défini comme un « **fonctionnement intellectuel significativement** inférieur à la moyenne, associé à des limitations dans au moins deux domaines du **fonctionnement adaptatif** : communication, soins personnels, compétences domestiques, habiletés sociales, utilisation des ressources communautaires, autonomie, santé et sécurité, aptitudes scolaires fonctionnelles, loisirs et travail. Le retard mental se manifeste **avant l'âge de 18 ans. »**

Le fonctionnement intellectuel est évalué par le quotient intellectuel (QI).

Ainsi, en se basant sur l'échelle d'intelligence de Wechsler (WISC-R) [2], les RM sont classés en 4 catégories :

- léger : QI de 50-69
- modéré : QI de 35-49
- sévère : QI de 20-34
- profond : Inférieur à 20

Cette première classification est souvent simplifiée et les RM sont répartis en deux grandes catégories : les RM légers avec un QI compris entre 50 et 70 et les RM dits sévères avec un QI inférieur à 50.

1.2. Prévalence des RM

La prévalence des RM est difficile à évaluer. Son estimation a été influencée par l'évolution de leurs définitions (la standardisation des critères de diagnostic), ainsi que par l'amélioration des soins médicaux.

Il est souvent fait mention des chiffres de 2 à 3 % pour la prévalence du RM léger (QI entre 50 et 70) et de 0,3-0,5% pour les RM sévères (QI<50) [3-5]. La fréquence des RM légers varie en fonction des études et du groupe d'âge étudié, contrairement aux taux des RM sévères qui restent très homogènes. Vingt à vingt cinq pour cent des RM sévères sont d'origine génétique [6]. On rapporte systématiquement que les garçons sont plus souvent atteints que les filles. La fréquence des RM chez les hommes est de 25 à 40% supérieure à celle des femmes.

1.3. Etiologie des RM

Les causes des RM sont extrêmement hétérogènes ; il peut s'agir de retards acquis ou innés. Les RM peuvent également résulter d'une combinaison de ces deux composantes.

Une cause est retrouvée dans au moins 50% des cas de retard mental sévère. Les causes périnatales et les événements postnatals comme les traumatismes expliquent environ 11% et 3 à 12 % des retards mentaux sévères respectivement. Dans environ 30 à 40% des retards mentaux sévères, des événements prénatals et (le plus souvent des anomalies génétiques) sont en cause [6]. La trisomie 21 est retrouvée dans 30% des retards mentaux sévères.

L'étiologie des retards mentaux légers est beaucoup moins documentée. Quarante cinq à soixante trois pour cent des cas n'ont pas de causes définies. Les principales causes connues des retards légers sont les souffrances fœtales et les syndromes polymalformatifs [7]. Seules quelques cas de retards légers ont été attribués à des facteurs postnatals comme des infections [6]. Les causes génétiques représenteraient 20 à 25% des cas de RM légers [8]. Les causes génétiques des RM peuvent être divisées en trois catégories :

- Les anomalies chromosomiques.

Les anomalies chromosomiques touchent un chromosome, entièrement ou partiellement. Elles affectent donc un grand nombre de gènes, qui sont soit trop exprimés (trisomie), soit insuffisamment exprimés (monosomie). Parmi les aneuploïdies, la trisomie 21 (1 enfant sur 700 naissances) est la plus fréquente des causes génétiques de RM. Les autres anomalies de nombre sont exceptionnellement viables (trisomies 13 et 18), parfois en mosaïque (trisomie 8). Les anomalies de structure telles que les translocations ont une contribution très faible dans les causes de RM mais leur récurrence est élevée puisque une fois sur deux le remaniement chez les parents est retrouvé équilibré. Les microremaniements chromosomiques comptent pour environ 10% des RM. Citons la maladie du cri du chat due à une délétion partielle du bras court du chromosome 5. Enfin, 5 à 7% des cas de RM syndromiques seraient le résultat de réarrangements sub-télomériques [8, 9].

Anomalies Chromosomiques		%
de nombre	Trisomie 21	76
	Mosaïques de chromosomes surnuméraires	01
	Dysgonosomies	04
de structure	Délétions et microdélétions	08
	Duplications	02
	Translocations déséquilibrées	03
	Translocations et inversions équilibrées	03
	Autres	01
	Télomères	Pas étudié

Tableau 1 : Anomalies chromosomiques responsables de déficience mentale dans une population de sujets institutionnalisés (Caroline du Sud, 1996)

Le tableau représente la répartition en pourcentage des différents types d'anomalies.

- Pathologies liées à l'empreinte parentale

L'empreinte parentale est un mécanisme par lequel certains loci sont réprimés pour un seul des deux allèles, en fonction de son origine parentale. Les gènes soumis à empreinte parentale sont donc différents des autres gènes, puisqu'un seul des deux allèles est exprimé dans les cellules somatiques. A ce jour, environ 70 gènes (humains ou murins) ont été identifiés comme étant soumis au phénomène d'empreinte parentale (http ://www.geneimprint.com).

Le mécanisme de l'empreinte est acquis dans la lignée germinale et fait intervenir la méthylation différentielle de l'ADN. La survenue d'un RM est alors associée à la perte d'expression du seul allèle normalement actif. Les syndromes de Prader-Willi et d'Angelman font partie des pathologies liées à ce mécanisme de l'empreinte parentale de la région 15q11-13. Un sujet sain n'exprime que l'exemplaire provenant du chromosome 15 paternel pour la région la plus centromérique (territoire du Prader Willi), l'autre région étant inactivée. De même, un individu sain n'exprime que l'exemplaire maternel du « territoire Angelman ». Ainsi, le syndrome de Prader-Willi se produit s'il n'existe aucun exemplaire paternel des gènes correspondant à ce territoire le plus centromérique. Ceci peut résulter soit d'une délétion de l'exemplaire paternel, soit d'une disomie uniparentale maternelle, à savoir l'existence exclusive de deux exemplaires d'origine maternelle. De la même façon, le syndrome d'Angelman se produit s'il n'existe aucun exemplaire maternel de la région la plus télomérique. Le plus souvent ceci provient d'une délétion de l'exemplaire maternel, d'une disomie uniparentale paternelle et plus rarement d'une mutation ponctuelle du gène *UBE3A* (ubiquitin protein ligase E3A).

- Les causes monogéniques de retards mentaux

Dans cette dernière catégorie, on peut classer les gènes selon leurs fonctions distinctes et les regrouper dans différentes classes fonctionnelles :

- les gènes impliqués dans la neurogenèse (*CENPJ* : centromere protein J, *MCPH1* : microcephaly, primary autosomal recessive 1…),
- les gènes impliqués dans la migration neuronale (*VLDRLR* : very low density lipoprotein receptor, *FLNA* : filamin A…)
- les gènes impliqués dans les fonctions neuronales et synaptiques (*FMR1* : fragile X mental retardation 1, *OPHN1* : oligophrenin 1…)
- les gènes impliqués dans la régulation de la transcription et le remodelage de la chromatine (*NF1* : Neurofibromin 1, *CDKL5* : cyclin-dependent kinase-like 5, *MECP2* : methyl CpG binding protein 2…).

Historiquement, la communauté scientifique a concentré ses efforts sur le chromosome X car celui-ci est riche en gènes exprimés dans le cerveau et d'autre part les familles avec un retard mental lié au chromosome X (RMLX) sont aisément repérables de par le mode de transmission particulier des gènes portés par l'X. De nombreux gènes responsables de RMLX ont été clonés, ces succès ont une double origine. D'une part, des méthodes innovantes ont apporté de nouveaux outils performants à la communauté scientifique. On peut notamment donner l'exemple des puces génomiques qui ont connu une expansion impressionnante, facilitant ainsi la découverte de gènes de retards mentaux. D'autre part, un effort important au niveau européen a permis d'appréhender les causes génétiques de RMLX syndromiques et non syndromiques. La création en 1995, du consortium européen sur les retards mentaux regroupant 6 centres principaux présentés sur le site EuroMRX (http ://www.euromrx.com/) a permis la mise en commun des moyens humains, technologiques et des échantillons biologiques du consortium.

2. Retards mentaux liés au chromosome X

2.1. Prévalence des RMLX

En se basant sur des données épidémiologiques, 10 à 16% des retards mentaux sévères auraient une cause liée à l'X [10]. Une étude récente [11] a permis d'estimer la fréquence des mutations dans les gènes de l'X connus pour être impliqués dans les RMLX au sein de l'ensemble des familles du consortium de l'EuroMRX. Après exclusion du syndrome de l'X fragile, 90 gènes ont été analysés chez 250 familles RMLX (avec des femmes vectrices obligatoires) conduisant à un diagnostic moléculaire positif pour 42% de ces formes familiales.

Les retards mentaux sont plus fréquents chez les hommes que chez les femmes, ce fait est maintenant corroboré par de nombreuses études épidémiologiques [5, 11-13]. L'explication avancée postule que les hommes hémizygotes ne pourraient pas compenser la présence d'un allèle délétère récessif sur le chromosome X. Utilisant les extrapolations basées sur les déséquilibres des fréquences des retards mentaux entre les garçons et les filles, Herst et

Miller [14] ont estimé la fréquence des RMLX à environ 1,83 pour 1 000 garçons. Chez les garçons, 25% des retards mentaux sévères, seraient liés au chromosome X. Fishburn et al. Rapportent une fréquence de 5/10 000 garçons pour les retards mentaux sévères [13]. Les données épidémiologiques concernant les retards mentaux légers sont moins documentées. Un ratio de 1,9 entre les garçons et les filles est rapporté pour les RM légers [5]. Ceci signifierait que plus de 50% des retards mentaux légers seraient liés à l'X [10]. Cependant, selon des études récentes, ce biais entre les deux sexes serait moins important. Ainsi, une analyse dans les familles qui ne contiennent que des hommes affectés a révélé que le biais du sex ratio ne pouvait pas être expliqué seulement par les gènes impliqués dans les retards mentaux liés à l'X [9, 15, 16].

2.2. Nosologie

Bien qu'il existe une grande variabilité génétique et clinique, une distinction est faite entre les RMLX syndromiques et non syndromiques [17].

Les patients avec un retard mental non syndromique ou non spécifique (RMX) présentent seulement des capacités intellectuelles diminuées alors que les patients avec retard mental syndromique (RMXS) associent au retard mental des anomalies physiques neurologiques ou métaboliques. Les RMX représentent les 2/3 des RMLX.

Dans une revue de 2004, Chiurazzi et al. [18]proposent de subdiviser les RMXS en 4 classes :

- les syndromes malformatifs (le syndrome de l'X fragile, le syndrome de Coffin-Lowry, le syndrome d'Opitz…)
- les maladies neuromusculaires (la dystrophie musculaire de Duchenne, l'ataxie cérébelleuse de type 2, le déficit en créatine lié au locus *SLC6A8*…)
- les maladies métaboliques (l'adrenoleucodystrophie liée au locus *ABCD1*, la déficience en ornithine transcarbamylase…)
- les maladies dominantes telles que le syndrome d'Aicardi, le syndrome de Goltz, le syndrome de Rett, l'incontinentia pigmenti de type II… [18].

Il faut cependant noter que cette distinction entre RM non spécifiques et syndromiques peut se modifier avec le temps, l'accumulation des études, l'expertise clinique… Dans sa première

description en 1943 le syndrome de l'X fragile était considéré comme non spécifique alors que maintenant il s'agit du RMXS le plus fréquent et le mieux connu.

2.3. Etiologie

Mise à part les RMLX récessifs, le décalage du sex ratio entre les garçons et les filles pourrait être expliqué par une hyper-vulnérabilité périnatale des garçons [10]. Une autre hypothèse impliquerait l'existence de facteurs génétiques liés à l'X de prédisposition au retard mental [16]. Ainsi, certaines mutations dans ces gènes entraîneraient une susceptibilité au retard mental. Avec ce modèle, la proportion des garçons avec retard mental passerait de 3 % à plus de 16 %. Si 5 à 10 % des hommes portaient ces facteurs, cela serait suffisant pour expliquer l'excès des RM chez les garçons avec un QI moyen de la population masculine peu différent de celui des filles. De plus, ces facteurs auraient une incidence beaucoup plus faible sur les retards mentaux sévères, ce qui est en concordance avec une plus faible proportion d'excès de garçons affectés pour les retards mentaux sévères [10].

Par exemple, des facteurs de prédisposition influençant le comportement ont été déjà décrits.

Les mutations dans le gène *MAOA* (monoamine oxidase A) ont déjà été associées avec un comportement impulsif et violent [19]. Une étude sur les enfants maltraités a été menée afin de comprendre pourquoi parmi les enfants maltraités, certains développent un comportement antisocial et d'autres pas. Les enfants maltraités avec un génotype qui confère un niveau d'expression élevé de *MAOA* semble moins susceptible de développer un comportement antisocial [20].

Plusieurs sites mis à jour régulièrement font état des connaissances sur les gènes responsables de retard mental lié à l'X, citons par exemple : http ://xlmr.interfree.it/home.htm et http ://www.ggc.org/xlmr.htm.

Par exemple, les figures 1, 2 et 3, présentent une mise à jour des connaissances datant de février 2007.

A cette date, parmi les retards mentaux syndromiques, 51 gènes clonés et 35 régions candidates étaient répertoriés (figures 1&2). Pour les retards mentaux non syndromiques, 28 gènes ont été impliqués et 54 familles RMLX cartographiées (figure 3).

Il reste néanmoins plus d'une centaine de gènes RMLX à découvrir [11] et ceci est bien le reflet de l'extrême complexité du système cognitif humain.

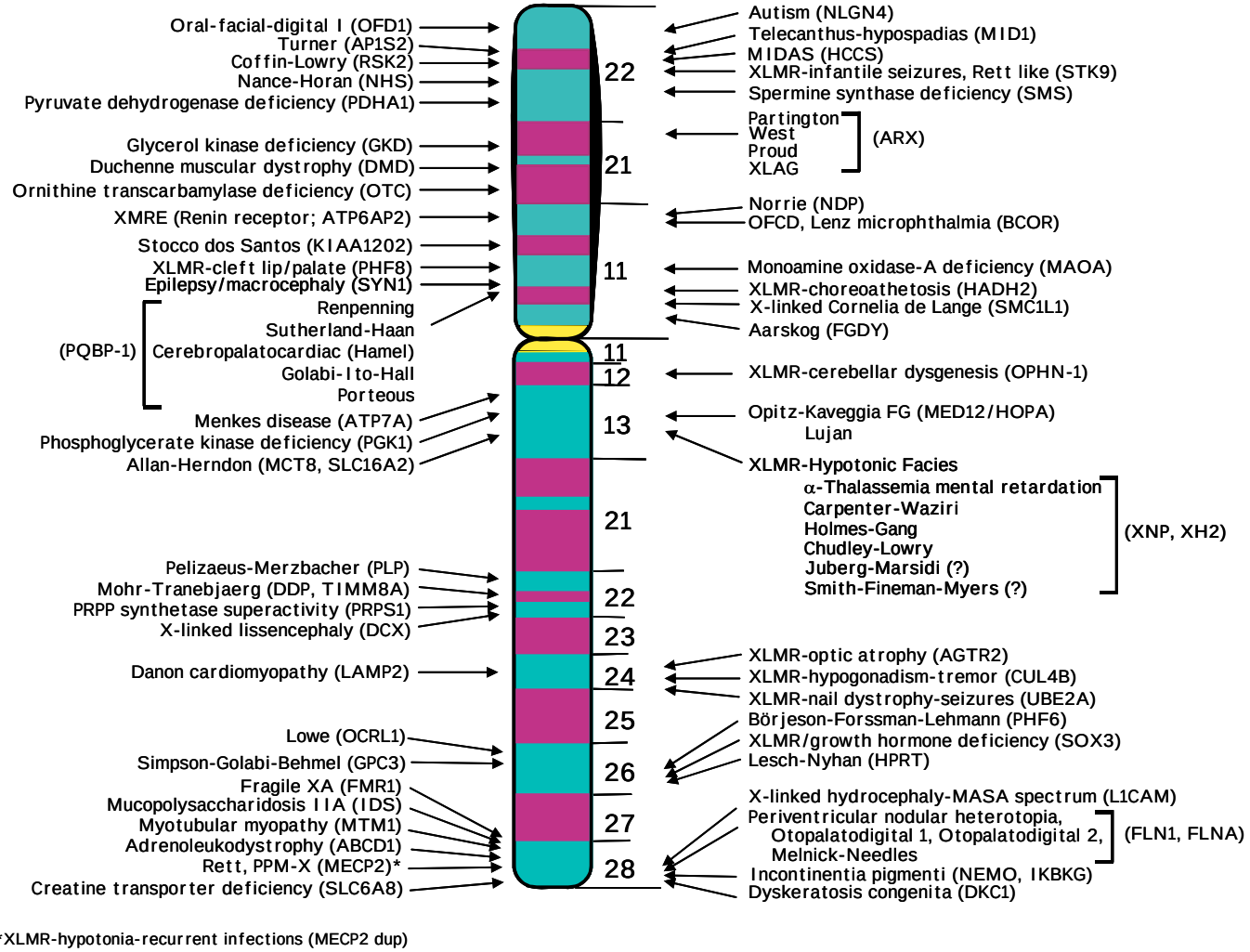


Figure 1 : Retards mentaux syndromiques et gènes clonés.

Localisation des 54 gènes de retard mental lié à l'X syndromiques (RMXS) connus (Greenwood Genetic Center, updated Feb. 2007).

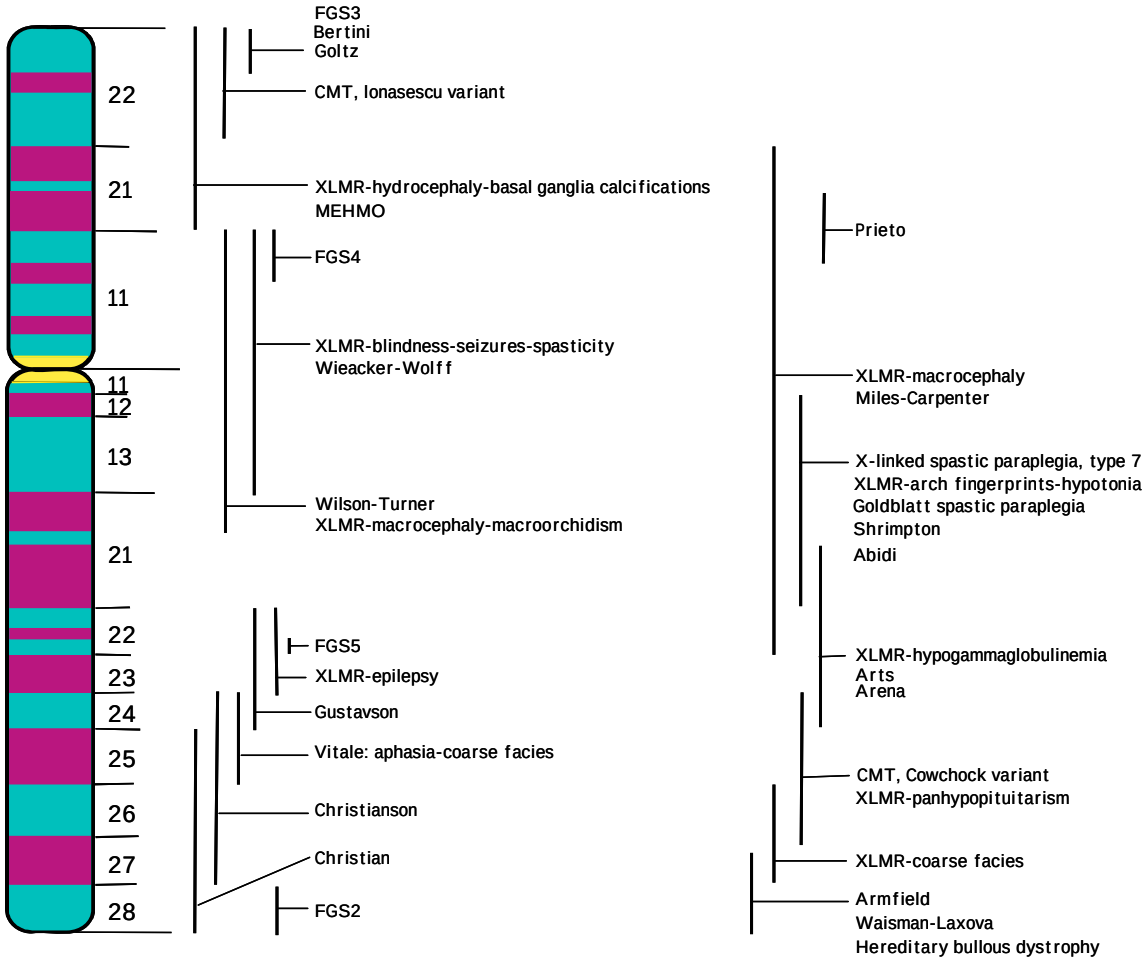


Figure 2 : Retards mentaux syndromiques et analyses de liaison.

Les traits verticaux délimitent la région candidate pour 35 retards mentaux liés à l'X syndromiques (RMXS) (lod score > 2), dont les gènes ne sont pas encore connus (Greenwood Genetic Center, updated Feb. 2007).

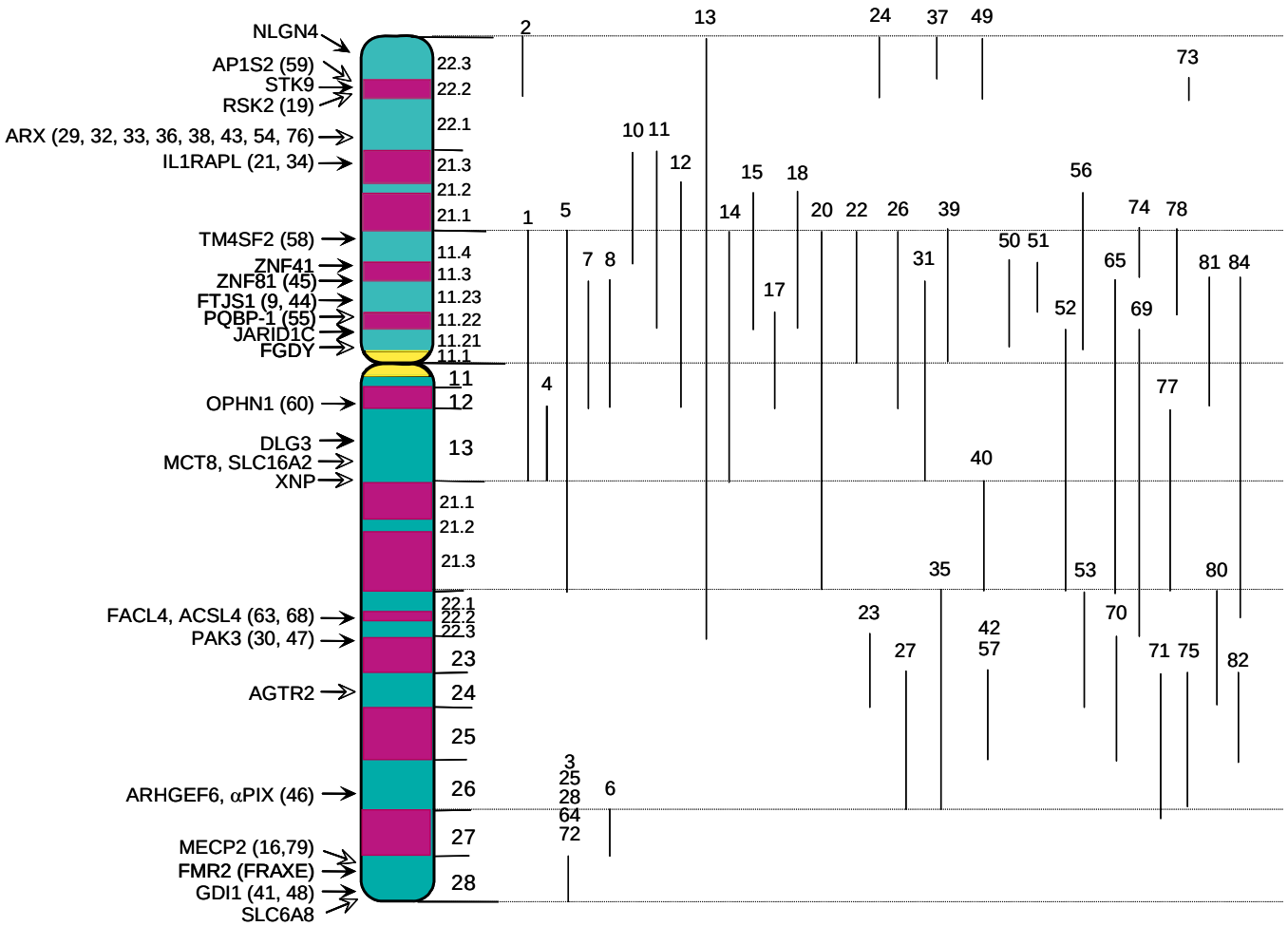


Figure 3 : Retards mentaux non syndromiques.

A droite de l'idiogramme du chromosome X est présenté la localisation génétique pour les 52 retards mentaux liés à l'X non syndromiques (RMX) (lod score > 2), dont les gènes ne sont pas encore découverts. A gauche de l'idiogramme du chromosome X, localisation des 15 RMX qui ont été cartographiés. Neuf gènes qui causent aussi bien des RMX que de RMLX syndromiques (RMXS) sont indiqués par des flèches vides (Greenwood Genetic Center, updated Feb. 2007).

2.4. Physiopathologie des RMLX

Les gènes impliqués dans les RMLX codent des protéines qui, à première vue, semblent impliquées dans des processus cellulaires très variés. Les fonctions de ces protéines peuvent être très générales (régulation de la transcription et remodelage de la chromatine), ou bien plus spécifiques (régulation du cytosquelette d'actine et donc de la morphogenèse neuronale).

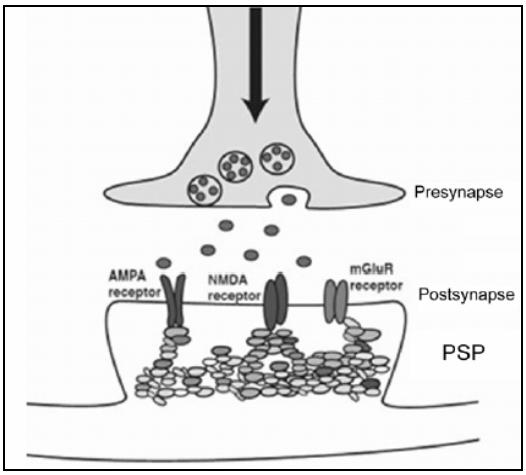
Depuis peu émergent clairement des voies biologiques et des structures qui semblent essentielles au bon fonctionnement cognitif. Parmi elles, la structure et la fonction de la synapse fait l'unanimité [9, 10, 15, 21, 22]. D'autre part, la dérégulation de la transcription et le remodelage de la chromatine semblent fréquemment en cause dans les RMLX. Mieux encore, ces « trois voies » seraient interconnectées, formant ainsi le complexe cognitif essentiel au bon fonctionnement intellectuel.

La majorité des produits des gènes impliqués dans les RM, si l'on met à part les gènes intervenant dans la transcription et le remodelage de la chromatine, est effectivement localisée dans l'espace pré/ou post synaptique. Citons par exemple les protéines FMRP (X mental retardation 1), OPHN1 (oligophrenin 1), NLGN4 (neuroligin 4), DLG3 (discs, large homolog 3), RabGDI, Neurotrypsin, PAK3 (p21 activated kinase 3).

2.4.1. RMLX et synapse

Le glutamate est un des principaux neurotransmetteurs du système nerveux central. Il est relargué dans la fente synaptique (suite au potentiel d'action provenant de la menbrane présynaptique) pour être reconnu par des récepteurs canaux ioniques et des récepteurs couplés aux protéines G (ou métabotropiques). Ainsi, la transduction du signal dans la membrane post synaptique est déclenchée. Le NMDA (ion-channel-forming N-methyl-D-aspartic acid) et les récepteurs métabotropiques du glutamate (mGluR) sont les deux principaux récepteurs du glutamate (Figure 4).

A)



B)

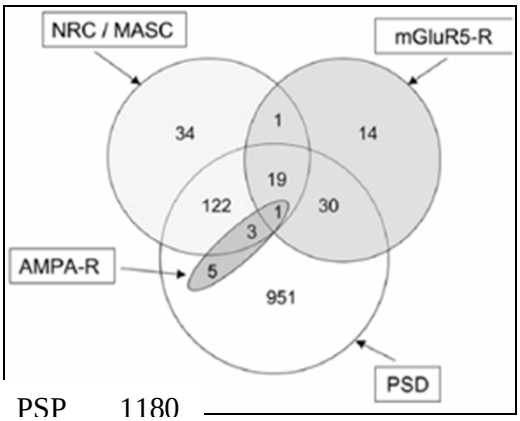


Figure 4 : Les principaux récepteurs du glutamate de la membrane post synaptique.

Le récepteur AMPA (a-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) contrôle la dépolarisation membranaire. Par opposition, les récepteurs NMDA (ion-channel–forming Nmethyl-D-aspartic acid) et mGluR (G protein–coupled metabotropic) ne jouent pas de rôle significatif dans la dépolarisation membranaire mais initient la voie de signalisation postsynaptique. Ces deux derniers sont aussi liés physiquement par des protéines d'échafaudage. B) Ces complexes font partie du protéome post-synaptique (PSP) qui contiendrait environ 1180 protéines regroupées en 4 sous-complexes : NRC/MASC (NMDA receptor complex/MAGUK-associated signaling complex), mGluR5-R (mGluR5 receptor complex), PSD (post synaptic density), AMPA-R (alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor) [22].

De l'ensemble des complexes du PSP, le complexe NRC/MASC est le mieux étudié.

Le chromosome X porte 39 des 1124 gènes codant des protéines du PSD et 7 des 186 gènes correspondant aux protéines du NRC/MASC. Les gènes du protéome post synaptique ont été fréquemment impliqués dans des RMLX. En effet, des mutations dans 6 des 7 gènes du NRC/MASC (85%) et 19 des 39 gènes du PSD (49%) ont été rapportées dans des familles avec RMLX. Les autres gènes sur le chromosome X liés au complexe PSP sont donc des candidats potentiels pour les RMLX (tableau 2). La figure 5 montre l'implication des différents gènes de ces complexes dans les RMLX dans l'espèce humaine et la souris.

Gene	Locus	KIAA or MGCFull-Length Clone	Cellular Fonction
PGRMC1	Xq24	MGC8891	Transmenbrane receptor
ATP2B3	Xq28	NA	Transmenbrane receptor
CASK	Xp11.4	MGC1500920	Scaffolder
CNKSR2	Xp22.12	KIAA0902	Scaffolder
SH3KBP1	Xp22.12	MGC9446	Scaffolder
SEPT6	Xq24	KIAA0128	GTP-binding protein potentially involved in cytokinesis
PSMD10	Xq22.3	MGC9114	Involved in the ATP-dependant degradation of ubiquitinated proteins
DDX3X	Xp11.4	MGC20129	Regulation of transcription, splicing and translation
OGT	Xq13.1	MGC22921, MGC39117	Protein amino acid O-linked glycosyation
RP2	Xp11.3	KIAA0215	Involved in beta-tubulin folding
IDH3G	Xq28	MGC5393, MGC2102	Involved in critic acid cycle in mitochondrion
PDCD8	Xq25	MGC111425	Mitochondrial apoptosis-inducing factor
PDHA1	Xp22.12	MGC8609	Glycolysis, gluconeogenesis, acetyCOA metabolism
HNRPH2	Xq22.1	…	Heterogenous nuclear ribonucleoprotein complex (pre-mRNA maturation)
SMARCA1	Xq25	MGC151056	Chromatin remodeling, regulation of transcription
RPS4X	Xq13.1	MGC8636, MGC87857	Translation
SYP	Xp11.23	MGC0359	Presynaptic vesicle
SLC25A5	Xq24	MGC65136	Ion transporter
MGC4825	Xp22.11	MGC4825	Unknown
IQSEC2	Xp11.22	KIAA0522	Unknown

Tableau 2 : Gènes du protéome post-synaptique (PSP) cartographiés sur le chromosome X qui n'ont pas encore été impliqués dans les Retards mentaux [22].

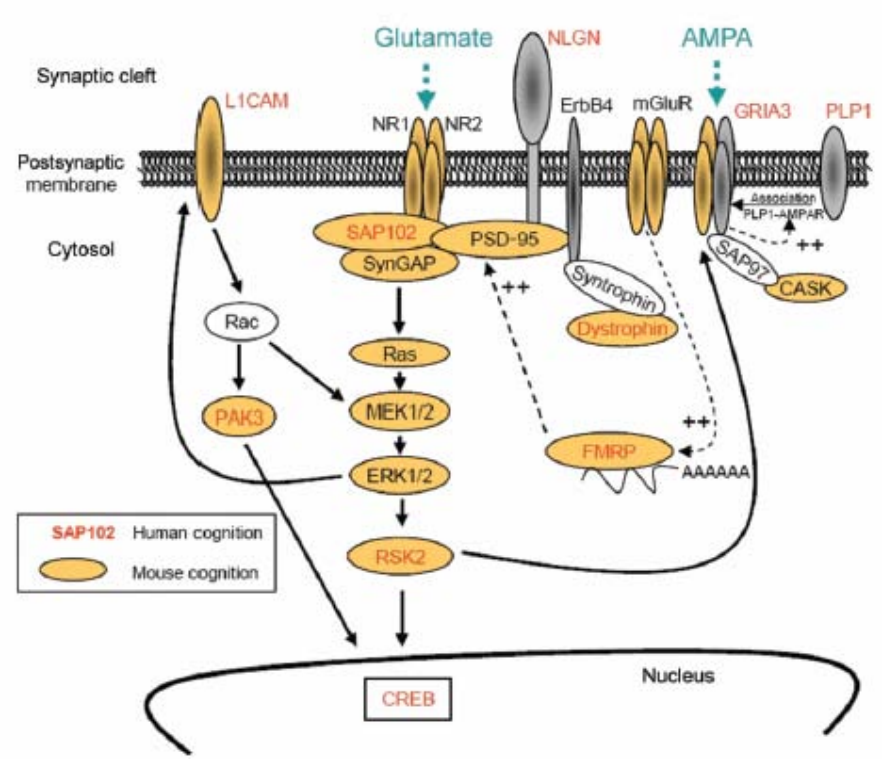


Figure 5 : Retards mentaux et voies de signalisation impliquant des protéines du PSP

Les complexes NMDA et MAGUK coordonnent la réponse post synaptique à l'activation des récepteurs NR1 et NR2 (sous unités de NMDA). Les sous unités NR1 et NR2 sont liées aux protéines SAP102 and PSD-95 du complexe MAGUK (membrane associated guanylate kinase). Ces deux dernières se lient directement à SynGAP qui régule la voie Ras-ERK-RSK. Cette voie de signalisation régule la transcription (tel CREB : cAMP responsive element binding protein 1), l'adhésion cellulaire et les récepteurs AMPA. FMRP régule la traduction de PSD-95 via l'activation de mGluR. La couleur rouge représente les gènes responsables de RMLX dans l'espèce humaine et la couleur jaune ceux pour la souris.

2.4.2. RMLX, régulation de la transcription et remodelage de la chromatine

Certains gènes régulant l'expression d'autres gènes peuvent être responsables lorsqu'ils sont mutés de RMLX. Leur nombre va croissant et pour certains d'entre eux le lien avec la plasticité synaptique est déjà établi. L'exemple de *MECP2* (methyl CpG binding protein 2) est caractéristique. Les mutations du gène *MECP2* sont responsables du syndrome de Rett, maladie qui affecte quasi exclusivement les filles. La protéine MECP2 se lie aux dinucléotides

CpG methylés, permettant ainsi le recrutement de corépresseurs (tel Sin3A) et d'histones désacétylases (HDAC1 et HDAC2). La présence de ce complexe protéique conduit à la condensation de la chromatine et entraîne la répression de l'expression de gènes cibles intervenant dans la maturation post mitotique des neurones. Les modèles de souris invalidées pour Mecp2 ont permis de mettre en évidence un certain nombre de gènes dont l'expression est affectée lorsque la protéine Mecp2 est absente tels DLX5/DLX6 et Bdnf (brain-derived neurotrophic factor).

2.4.3. RMLX et possibilités thérapeutiques

Malgré la complexité de la physiopathologie des RMLX, de nouvelles études montrent qu'une approche thérapeutique pharmacologique serait possible, pour certaines formes de RMLX. Si l'on considère le modèle du complexe neuronal présenté par Laumonnier et al., des molécules pourraient par exemple activer des voies de signalisation compensatrices. Par exemple, la plasticité synaptique peut être restaurée par administration d'antagoniste des récepteurs au glutamate ou par le lithium [23]. Par ailleurs, chez la drosophile adulte mutée pour Fmr1 (fragile X mental retardation 1) et traitée par les antagonistes des mGluR, la mémoire à court terme et le comportement normal semblent se rétablir.

3. Le Syndrome d'Aicardi

3.1. Historique

Le syndrome d'Aicardi (OMIM 304050) a été décrit pour la première fois en 1965 par le docteur Jean Aicardi suite à l'observation de filles présentant des spasmes en flexion [24]. En 1969, une seconde étude portant sur un plus grand nombre de cas lui a permis de proposer une triade de signes cliniques pour le diagnostic du syndrome d'Aicardi. Ainsi, ce syndrome était défini par une agénésie du corps calleux, des spasmes infantiles, et des lacunes

choriorétiniennes [25]. A cette triade sont souvent associées d'autres anomalies cérébrales, oculaires et vertébrales ainsi qu'une dysmorphie cranio-faciale [25-28].
Le retard mental accompagnant ce syndrome est le plus souvent sévère. C'est un syndrome rare dont la prévalence est estimée à 1 cas sur 500 000, avec un nombre revue récemment à 500 cas en Europe (http ://www.orpha.net/orphacom/cahiers/docs/FR/Prevalence_des_maladies_rares.pdf).

3.2. Spectre clinique

L'accumulation des cas de syndrome d'Aicardi (AIC) publiés dans la littérature et l'apport des nouvelles techniques d'imagerie ont permis d'appréhender l'étendue du spectre clinique. Une minorité des patientes ne présente pas la triade classique.

3.2.1. Les anomalies neurologiques

Premières manifestations de la maladie, les causes convulsives apparaissent souvent tôt, dans les 3 premiers mois et dominent la scène. L'aspect le plus habituel est celui de spasmes infantiles souvent de présentation asymétrique et même unilatérale. Un retard mental, des signes pyramidaux voir une hémiplégie se dessinent plus tard. Parmi les 77 patientes Aicardi étudiées par Rosser et al, 91 % des patientes montrent un niveau intellectuel équivalent à celui d'un nourrisson d'un an, seulement une patiente atteint un niveau intellectuel comparable à celui d'un enfant de trois ans [29]. Le tracé EEG le plus fréquent est caractérisé par des bouffées d'ondes lentes, étroites et de haute amplitude séparées d'ondes intervalles de bases amplitude [30].
Le retard mental est le plus souvent sévère mais il peut aussi être léger [29, 31, 32]. Un seul cas de syndrome d'Aicardi typique a été décrit avec des fonctions cognitives normales [33].
Chez les patientes Aicardi l'agénésie du corps calleux n'est pas toujours complète [26, 34]. Il ne s'agit donc pas d'un critère constant et suffisant pour le diagnostic de la maladie et il n'est probablement jamais isolé chez les patientes AIC [30](figure 6).

A) B)

Figure 6 : Agénésie du corps calleux dans le syndrome d'Aicardi

La figure représente l'image IRM (Imagerie par Résonance Magnétique) en coupe sagittale d'un témoin normal. B) est l'image IRM coupe sagittale chez une patiente avec une agénésie partielle du corps calleux. Les flèches rouges indiquent l'emplacement du corps calleux [35].
Les nouvelles techniques d'imagerie telle l'IRM, ont révélé que l'agénésie du corps calleux n'est qu'une des malformations du système nerveux central chez les patientes présentant un syndrome d'Aicardi. On peut ainsi citer, la présence d'hétérotopies de la substance grise, de polymicrogyries, de ventriculomegalies et de kystes. Les hétérotopies périventriculaires seraient probablement constantes chez les patientes AIC [30].

3.2.2. Les anomalies oculaires

On s'accorde pour dire que les lacunes choriorétiniennes sont pathognomoniques et ceci malgré les deux cas de syndrome d'Aicardi ne présentant pas ces lacunes [30].

En fondoscopie, les lacunes apparaissent comme des tâches arrondies rosâtres ou blanchâtres avec une forme bien délimitée et des bords hyperpigmentés [36] (figure 7). La taille des lacunes ne varie pas avec l'âge. Les colobomes du disque optique et les microphtalmies sont les autres anomalies oculaires les plus fréquemment rencontrées.

A) B)

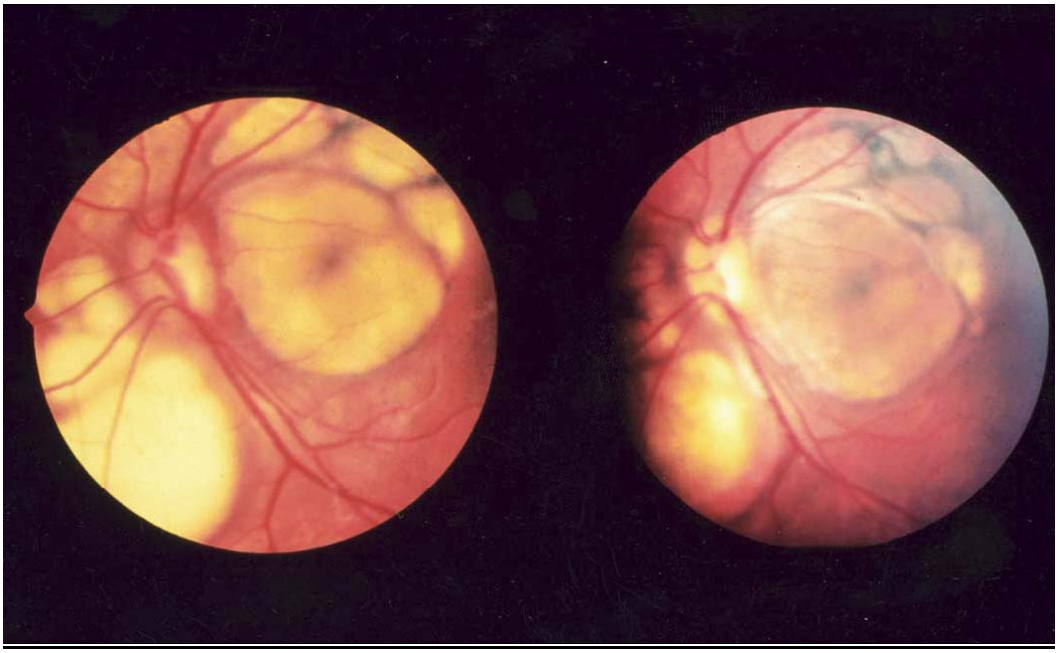


Figure 7 : Les lacunes choriorétiniennes dans le syndrome d'Aicardi

Fonds d'œil. A) Rétine de la patiente à 1 an. B) la même patiente à 5 ans. La forme et la taille des lacunes restent inchangées mais elles semblent plus prononcées (dépôt de pigments sur les contours des lacunes) [30].

3.2.3. Les anomalies extra neurologiques

Parmi les anomalies non neurologiques figurent les anomalies des côtes et vertèbres (côtes manquantes, hémivertèbres, vertèbres soudées) présentes dans la moitié des cas conduisant à une scoliose parfois marquée [27]. Il existerait également des anomalies faciales caractéristiques chez les patientes Aicardi [28]. Sur la base d'une analyse clinique de 40 enfants Aicardi, 65,5% des patientes présentent un prémaxillaire proéminent avec une pointe du nez retroussée, des sourcils clairsemés dans leur partie externe. Parfois est noté aussi des lésions dermatologiques (22,5%) et des malformations des mains (7,5%).

Une fente labiopalatine a également été rapportée chez trois patientes Aicardi [37, 38]. Le syndrome d'Aicardi est aussi associé à une incidence accrue de tumeurs cérébrales, et plus spécifiquement de papillomes du plexus choroïde [39].

3.2.4. Pronostic et histoire naturelle

Le syndrome d'Aicardi débute par la survenue, avant l'âge de trois mois, de spasmes infantiles.

Par la suite, l'évolution clinique est variable d'une patiente à l'autre. Par exemple, certaines filles acquièrent la marche et prononcent quelques mots alors que d'autres n'acquièrent ni la marche, ni aucune sorte de langage.

Une étude conduite par Menezes et al. Tente de faire les liens entre 28 caractéristiques neurologiques et le pronostic clinique[27]. Les auteurs concluent qu'aucun de ces signes neurologiques ne permet de prédire l'évolution de la maladie.

Bien qu'il soit très difficile de prédire le pronostic des patientes Aicardi, une évolution favorable reste exceptionnelle. A ce jour, seuls 7 cas ont été rapportés dans la littérature avec une atteinte a minima, cependant aucun de ces 7 cas ne présentent la triade caractéristique de ce syndrome [31, 33, 35, 40-43].

Les anomalies oculaires avaient été présentées comme un facteur de mauvais pronostic pour le AIC [44] mais le cas d'une évolution favorable chez une patiente avec des lacunes choriorétiniennes importantes semble invalider cette observation [33].

Il n'est pas possible d'établir un lien strict entre la gravité des anomalies cérébrales (agénésie partielle versus complète du corps calleux) et l'atteinte des fonctions cognitives chez les patientes AIC [27]. Cependant, parmi les 7 cas avec une évolution clinique favorable, 6 ont un retard mental léger et dans 6 cas sur 7 une agénésie partielle du corps calleux. De la même façon, l'épilepsie ne semble pas être un facteur pronostic car un bon contrôle pharmacologique des crises n'est pas associé à un meilleur développement psychomoteur. La plupart des filles AIC ne marchent pas (79%) et l'acquisition de la parole reste rare (96%). Il semblerait que les acquisitions au niveau du développement psychomoteur ne soient pas perdues. L'étude de Menezes [27] montre une espérance de vie limitée avec une mortalité à 6 ans de 25% et à 15 ans de 60%.

3.2.5. Proposition de critères pour le diagnostic clinique du syndrome d'Aicardi

Sur la base de nouvelles observations cliniques Jean Aicardi a proposé en 2005 des critères enrichis de diagnostic du syndrome. Ainsi, il définit des critères majeurs qui incluent la triade et des critères mineurs qui permettent de conforter le diagnostic clinique de la maladie (tableau 3).

Le diagnostic du syndrome AIC reste toujours basé sur la triade : agénésie du corps calleux, spasmes en flexion, et lacunes choriorétiniennes. Cependant, les patientes sans agénésie du corps calleux mais présentant d'autres critères majeurs peuvent être considérées comme porteuses du syndrome d'Aicardi. Ainsi, la présence d'au moins deux critères de la triade associés à deux autres critères (majeurs ou mineurs) suggère fortement le diagnostic clinique du syndrome AIC [28].

Major features
Infantile spasms[a]
Chorioretinal ' lacunae' [b]
Coloboma of the optic disc (and nerve) often unilateral
Agenesis of the corpus callosum (total or partial)
Cortical malformations (mostly microgyria) [b]
Periventricular (and subcortical) heterotopia [b]
Intracranial cysts (probably ependymal) interhemispheric or around third ventricle
Papillomas of choroid plexuses
Supporting features
Vertebral and costal abnormalities
Microphthlalmia and/or other eye abnormalities
'Split brain' EEG (dissociated suppression-burst tracing)
Gross hemispheric asymmetry

[a] May be replaced by other types of seizures (usually focal).
[b] Present (or probably present) in all cases.

Figure 8 : Les critères révisés diagnostiques du syndrome d'Aicardi [30]

3.3. Données génétiques

3.3.1. Mode de transmission de la maladie

Le syndrome d'Aicardi s'observe chez les filles sans distinction d'origine ethnique ou géographique. Le seul cas exceptionnel de deux sœurs atteintes de AIC [45] plaide pour une origine génétique de ce syndrome. Pour expliquer le cas rapporté de jumelles monozygotes discordantes pour le syndrome d'Aicardi, on peut évoquer une mutation somatique précoce au cours du développement embryonnaire chez la jumelle atteinte [45].
Cinq paires de jumelles dizygotes discordantes pour le SA ont été décrites, ces cas ne semblent pas en faveur de facteurs périnatals à l'origine du syndrome d'Aicardi [46].

3.3.1.1. Transmission dominante liée au chromosome X d'origine maternelle

Ce syndrome rare ne touche que les filles ou de façon exceptionnelle les garçons présentant par ailleurs un syndrome de Klinefelter (47, XXY) [47].
Le syndrome d'Aicardi est sporadique, excepté le cas décrit par Molina évoquant un mosaïcisme germinal chez l'un des deux parents [45]. L'hypothèse d'une transmission dominante liée au chromosome X d'origine maternelle expliquerait l'atteinte exclusive des filles et l'effet létal chez les garçons hémizygotes pour le chromosome X. Toutefois, l'absence de biais dans le sex-ratio en faveur des filles dans les fratries des cas sporadiques de Syndrome d'Aicardi va à l'encontre de ce modèle [48].

3.3.1.2. Transmission dominante liée au chromosome X d'origine paternelle

L'absence de biais dans le sex-ratio peut s'expliquer par des néomutations d'origine paternelle. Ce modèle expliquerait l'atteinte exclusive des filles car le père transmet son chromosome X uniquement à ses filles. Cependant, la fréquence des néomutations

ponctuelles semble augmenter avec l'âge chez les hommes et aucune observation n'a relevé un âge élevé des pères de filles AIC lors de la procréation [48].

3.3.2. Inactivation de l'X et syndrome d'Aicardi

3.3.2.1. Notions d'inactivation du chromosome X

L'inactivation du chromosome X est un processus cellulaire normal mis en place tôt au cours de l'embryogenèse (14 jours après la fécondation) chez la femme et dont la fonction est d'assurer un dosage génique équivalent entre les deux sexes pour les gènes portés par l'X [49]. Une fois établie, l'inactivation est stable et héritée de façon clonale au cours des divisions cellulaires des cellules somatiques. Le chromosome X inactif présente toutes les caractéristiques de l'hétérochromatine. Il se réplique de façon tardive au cours du cycle cellulaire, il est hyperméthylé, sa chromatine est condensée, ses histones H4 hypoacétylées, il peut être visualisé à l'interphase sous la forme du corpuscule de Barr accolé à la face interne de la membrane nucléaire [50, 51].

Le processus peut être décomposé en plusieurs étapes : l'initiation de l'inactivation (le comptage des chromosomes X, le choix de l'X à inactiver, l'inactivation), l'expansion de l'inactivation et le maintien de l'inactivation.

Une partie du chromosome X en q13 est indispensable pour la mise en place du processus d'inactivation, cette région a été dénommé XIC (X Inactivation Center) [52]. Dans cette région, le transcrit XIST (X Inactivation Specific Transcript) est exprimé seulement à partir du chromosome X inactif [53]. La preuve directe de l'implication de ce transcrit dans l'initiation et la propagation de l'inactivation de l'X a été apportée avec des souris invalidées pour Xist [54]. XIST code un ARN nucléaire non traduit.

Au moment de la formation du zygote femelle, les deux chromosomes X sont actifs et le niveau d'expression de XIST est faible [55]. Au cours de la différenciation des cellules, le taux de l'ARN Xist augmente car il est stabilisé au niveau de l'X inactif. L'ARN Xist se propage en cis tout le long du chromosome X et le recouvre, entraînant ainsi son inactivation. Chez la souris, les transcrits Tsix et Xite régulent le comptage des X et le nombre d'X à inactiver. Comptage et choix des chromosomes à inactiver impliqueraient l'appariement des

chromosomes X homologues [56]. Ainsi, chez la souris dans les cellules somatiques en division, les deux centres d'inactivation des chromosomes X s'apparient de manière transitoire juste avant l'inactivation de l'X [57].

Quinze pour cent des gènes échappent à l'inactivation [58] et sont exprimés à la fois à partir du chromosome X actif et inactif. Ces gènes sont notamment localisés au niveau des régions pseudo-autosomiques (PAR1 en Xp22.3 et PAR2 en Xq28) et possèdent donc des régions homologues sur le chromosome Y. D'autres gènes de l'X échappent aussi à l'inactivation de l'X et certains d'entre eux ont un équivalent fonctionnel sur l'Y.

Chez les femmes phénotypiquement normales, sans antécédent familial de pathologie de retard mental lié à l'X, l'inactivation du chromosome X est un processus aléatoire touchant les chromosomes X paternel et maternel dans une proportion équivalente (50 % - 50 %) de cellules somatiques. La distribution de l'inactivation de l'X dans la population féminine normale suit une répartition Gaussienne dans laquelle la grande majorité des femmes (80 % à 90 %) a un profil d'inactivation aléatoire (50 % - 50 % à 80 % - 20 %).

Dans la population générale de par les phénomènes stochastiques, les extrémités de la courbe Gaussienne correspondent aux femmes (10 à 20 % de la population générale) présentant par pur hasard une inactivation fortement biaisée (90 % - 10 % à 100 % - 0 %) en faveur de l'un ou l'autre des chromosomes X.

Certains mécanismes (anomalies chromosomiques, mutation au sein du gène XIST, facteurs génétiques en cis ou trans) peuvent induire une inactivation biaisée du chromosome X.

Dans les translocations (X ; autosome) équilibrées, on observe le plus souvent une inactivation biaisée de l'X normal.

Chez une patiente atteinte par une maladie dominante liée à l'X et létale chez les garçons, il existe une contre sélection des cellules portant l'X muté actif. La majorité des maladies dominantes liées à l'X, comme l'incontinentia pigmenti, le syndrome orol-facio-digital de type I et le syndrome Goltz, est associée à une inactivation biaisée du chromosome X [59-61]. Des exceptions existent comme par exemple le syndrome de Rett où la majorité des patientes ne présente pas de biais d'inactivation.

3.3.2.2. Etudes d'inactivation pour le Syndrome d'Aicardi

Les maladies dominantes liées à l'X présentent dans la plupart des cas une inactivation biaisée. Pour le Syndrome d'Aicardi, on pourrait penser que l'inactivation biaisée du chromosome X explique la grande hétérogénéité clinique chez les filles. Ainsi, les filles les moins atteintes pourraient être celles qui présentent en majorité un chromosome anormal inactif dans leurs cellules. Ce raisonnement est en partie valable pour le syndrome de Rett ou un profil d'inactivation variable chez certaines patientes serait une des explications possibles pour une grande hétérogénéité clinique.

Dans une première étude sur 7 patientes AIC une inactivation biaisée a été retrouvée chez les 3 filles les plus sévèrement atteintes (retard mental le plus sévère et une épilepsie pharmaco-résistante). Inversement, 2 des filles avec un retard mental moins prononcé avaient une inactivation aléatoire [62]. Cependant, une seconde étude ne retrouve pas ce lien entre biais d'inactivation et phénotype clinique. Ainsi, parmi les 10 patientes AIC (triade classique), étudiées un profil d'inactivation aléatoire est observé [48]. Trois des mères asymptomatiques des patientes Aicardi présentent une inactivation biaisée. Cette observation est en concordance avec une distribution binomiale de l'inactivation de l'X dans la population générale, 20% des femmes saines présenterait une inactivation biaisée. Les trois cas de patientes AIC avec biais d'inactivation décrits précédemment par Neidich, vu le nombre restreint de patientes étudiées, pourraient faire partie de ce biais stochastique [48]. Si l'on admet que l'inactivation du chromosome X chez les filles AIC est aléatoire, alors en découle un certain nombre de possibilités et d'hypothèses.

- L'inactivation de l'X pourrait ne pas intervenir dans la pathologie,
- il est possible qu'une très faible proportion de cellules avec l'X anormal soit suffisant à conduire au phénotype du syndrome d'Aicardi et qu'un biais complet en faveur de l'X anormal conduise à une létalité (comme chez les hommes hémyzigotes),
- Il est aussi possible que le gène du syndrome d'Aicardi ne soit pas exprimé ou que son expression ne soit pas essentielle dans les cellules hématopoïétiques. Ainsi, une inactivation biaisée liée à la maladie ne serait pas observable dans le tissu sanguin.

- Enfin, un défaut d'inactivation conduisant à une disomie fonctionnelle de la région du gène AIC pourrait expliquer l'atteinte des femmes sans biais d'inactivation et l'absence de cas de Syndrome d'Aicardi chez les hommes.

3.4. Recherche de gènes candidats pour le syndrome d'Aicardi

Le clonage positionnel vise à identifier le gène responsable de la maladie génétique sans connaissance a priori du produit et de la fonction du gène [63]. Il s'agit d'un processus à plusieurs étapes qui débute par le positionnement du gène dans une région chromosomique donnée. Le but est ensuite d'affiner la localisation primaire en construisant la carte physique précise de la région. Dans la grande majorité des cas, une carte intégrée (génétique, physique et transcriptionnel) de la région candidate est disponible dans les bases de données biologiques publiques. Si ce n'est pas le cas, une analyse moléculaire de la région est nécessaire. Il devient alors possible d'entreprendre une recherche de séquence exprimée dans l'intervalle critique afin d'identifier un ou plusieurs gènes candidats. L'implication définitive d'un gène candidat dans la maladie se fait par la mise en évidence de mutations délétères chez les sujets malades.

La première étape du clonage positionnel consiste donc en la localisation chromosomique du gène, par des études de liaison génétique dans les familles où ségrége la maladie et à l'aide de marqueurs polymorphes. Cette localisation peut être facilitée par la découverte, chez un individu présentant la maladie génétique, d'une anomalie chromosomique (translocation, délétion, inversion).

3.4.1. Carte génétique

Les gènes morbides peuvent être localisés par des études de liaison génétique dans les familles où ségrége la maladie. L'efficacité et la puissance de cette approche ont été significativement améliorées par le développement des cartes géniques de très haute

résolution, fondée sur le polymorphisme de locus microsatellites hautement informatifs et couvrant l'ensemble du génome [64]. Toutefois, la densité trop faible de ces microsatellites est un facteur limitant pour les cartes génétiques. Une approche plus récente utilise les SNP (Single Nucleotide Polymorphism) plus nombreux. Des informations sur le nombre, la cartographie et le pourcentage d'hétérozygotie des SNP sont disponibles dans les sites comme le NCBI (National Center for Biotechnology Information)(http ://www.ncbi.nlm.nih.gov/SNP/). Ainsi, il est possible de suivre la co-ségrégration de SNP informatifs avec la transmission de la maladie à travers les générations. Les analyses de liaison génétique permettent de déterminer pour les maladies monogéniques un intervalle dans lequel il existe une forte probabilité que soit contenu le gène morbide.

Dans le cas particulier du syndrome d'Aicardi cette approche n'est pas possible puisqu'aucun cas de transmission familiale n'a été répertorié à ce jour.

3.4.2. Anomalies chromosomiques associées au syndrome d'Aicardi

Trois patientes AIC ont été décrites avec une anomalie de la région p22 du chromosome X.
La première est une patiente présentant le syndrome d'Aicardi avec une translocation chromosomique équilibrée de novo, t(X ;3) et un point de cassure en Xp22.3 pouvant interrompre le locus Aicardi responsable du phénotype clinique observé [65].
Parmi les signes cliniques du syndrome AIC, on retrouve chez cette patiente les lacunes choriorétiniennes, une microphtalmie unilatérale, une agénésie du corps calleux, des anomalies du squelette (scoliose et côtes supplémentaires), un profil EEG anormal, un retard mental mais les signes épileptiques et spasmes en flexions n'ont pas été rapportés. Par ailleurs, la patiente présente également un symblépharon partiel, une opacité cornéenne de l'œil droit, une lagophtalmie de l'au-delà gauche, un ptosis bilatéral, une hydrocéphalie, une anomalie de Dandy Walker et une asymétrie prononcée de la tête. Le caryotype s'écrit : 46,XX,t(X ;3)(p22.3 ;q12).

Les études cytogénétiques, aussi bien sur les lymphocytes que sur les fibroblastes montrent une inactivation systématique du chromosome X transloqué.

Plusieurs hypothèses sont émises pour expliquer le phénotype clinique chez cette patiente.

- les symptômes atypiques seraient dus à la monosomie fonctionnelle 3q.
- La perte de fonction du gène AIC serait plus désavantageuse au niveau cellulaire que la monosomie fonctionnelle 3q, d'où une inactivation préférentielle de l'X transloqué.
- Le gène du syndrome d'Aicardi serait soumis à l'inactivation de l'X.

Deux problèmes majeurs ressortent de cette étude. La première concerne les études d'inactivation. Dans la majorité des cas de translocations équilibrées, le chromosome X normal est celui qui est le plus souvent inactivé. En effet, la monosomie fonctionnelle qu'entrainerait l'inactivation du chromosome transloqué, semble présenter un désavantage important alors que l'inactivation de l'X normal permet de conserver le dosage génique. Dans ce cas d'étude précis, le chromosome anormal est systématiquement inactivé. La patiente présente une monosomie fonctionnelle 3q partielle. Si comme le présentent les auteurs, l'absence du gène AIC est plus désavantageuse que la monosomie, une inactivation biaisée devrait être systématiquement observée pour le syndrome AIC, ce qui n'est de toute évidence pas le cas. D'autre part, le phénotype qui serait lié au point de cassure n'est pas évident. En effet, il est possible d'impliquer la monosomie fonctionnelle 3q au phénotype observé. De plus, concernant ce cas précis, le professeur Aicardi douterait de la validité du diagnostic de syndrome d'Aicardi (d'après OMIM 304050).

Deux autres patientes Aicardi avec des délétions dans la région Xp22 ont été rapportées dans une même étude [66]. La cohorte présentée était composée de six patientes avec un caryotype 46, XX et d'une patiente avec une délétion cytogénétique del(X)(p22pter). Les détails du diagnostic de syndrome d'Aicardi n'étaient pas rapportés. L'analyse moléculaire de la région Xp22 révèlent l'absence des loci DXS278 et DXS85 chez la patiente avec la délétion del(X)(p22pter). Pour une des 6 patientes avec un caryotype normal, l'analyse moléculaire de la région Xp22 révèle une perte du marqueur DXS278. En 1990, la même équipe recherche des délétions moléculaires de cette région chez 7 filles Aicardi porteuses d'un caryotype normal mais ne retrouve pas d'anomalies pour les 8 marqueurs polymorphes étudiés (incluant DXS278)[62].

3.4.3. Un même gène pour deux syndromes ?

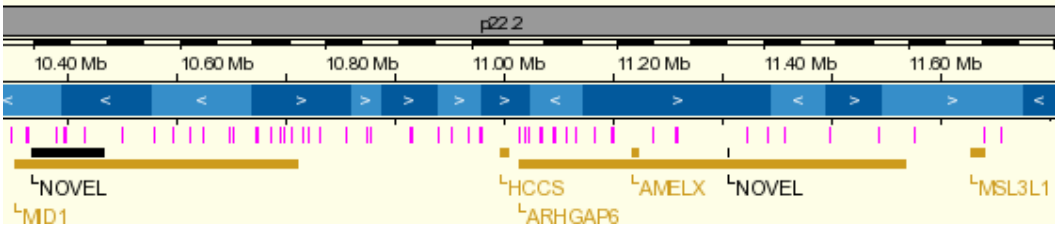
Microphtalmie, colobomes, agénésie du corps calleux et épilepsie sont aussi décrits dans d'autres RMLX dominants comme le syndrome de Goltz et le syndrome MLS (Microphtalmia with Linear Skin defects).

Certains auteurs ont émis l'hypothèse selon laquelle le syndrome d'Aicardi serait allélique du syndrome MLS. Ce dernier est associé à des délétions de la région Xp22.2 (selon les cartes disponibles sur ensembl et genome browser) [67-69].

Le syndrome MLS (OMIM 309801) aussi connu sous le nom de microphtalmie syndromique 7 et syndrome MIDAS est caractérisé par une hypoplasie de la peau du cou et du visage et des anomalies au niveau des yeux (microphtalmie, une sclérocornée et une opacité cornéenne). Parmi les symptômes associés on trouve, une agénésie du corps calleux, une ventriculomégalie, des anomalies cardiaques, un retard du développement et une petite taille [70-73]. Le syndrome est dominant et lié à l'X avec une létalité chez les garçons hémizygotes [74]. Il est associé à une monosomie de la région en Xp22. Ainsi, la région monosomique minimale de 610 kb en Xp22.2 conduisant à ce syndrome contient trois gènes : *MID1 (*midline 1*), HCCS (*holocytochrome c synthase) et *ARHGAP6 (*Rho GTPase activating protein 6) ([73] ; revue de la littérature). Le gène *HCCS* a été impliqué dans la maladie avec la description d'une délétion de 8,6 kb incluant une partie du gène chez une mère et deux filles affectées. En outre, une mutation non sens (R197X) et une mutation faux sens (R217C) ont été rapportées de novo chez deux filles non apparentées avec des caryotypes normaux.

A ce jour, aucune microdélétion n'a pu être mise en évidence dans cette région chez les filles Aicardi étudiées [62, 75]. En outre, les études de séquençage systématique des gènes de cette région (*ARHGAP6* ; *HCCS* ; *MID1* ; *MSL3L1* : male-specific lethal 3-like 1*)* montrent que le syndrome d'Aicardi n'est pas allélique au syndrome MLS [74, 76] (figure 9).

La figure 9 présente une carte intégrée de la région Xp22.2



http ://www.ensembl.org/

Figure 9 : Carte intégrée de la région Xp22.2 contenant les gènes séquencés pour le syndrome d'Aicardi.

Les gènes *ARHGAP6*, *HCCS*, *MID1, MSL3L* ont été séquencés chez des patientes AIC. Aucune mutation délétère n'a été identifiée. Le nouveau gène (NOVEL) le plus télomérique ENSG00000205582 possède 2 exons et code une protéine de 54 résidus. La cartographie de ce gène a été déterminée par alignement de la séquence de son ADNc. Le second gène (NOVEL) ENSG00000205558 de 53 résidus a été cartographié de façon identique [77]. *AMELX* (amelogenin) est responsable de l'amélogenèse imparfaite (AI) liée au chromosome X. L'AI constitue un groupe d'anomalies du développement affectant la structure et l'apparence clinique de l'émail de toutes ou de quasiment toutes les dents, de façon plus ou moins identique. Compte tenu des données fonctionnelles, ce dernier gène ne constitue pas a priori un candidat intéressant pour le Syndrome AIC. Mise à part les deux nouveaux gènes identifiés sur Ensembl, pour lesquels aucune information fonctionnelle n'est disponible, la région peut donc être *a priori* être exclue.

Le syndrome de Goltz (ou Focal dermal hypoplasia, syndrome de Goltz-Gorlin) (OMIM : 305600) se transmet selon le mode dominant lié au chromosome X, avec létalité in utero pour les garçons. Il est caractérisé par une atteinte cutanée et des anomalies très variées pouvant affecter les yeux, les dents, le squelette, le système nerveux central et les systèmes urinaire, gastro-intestinal et cardiovasculaire. Les signes cliniques comportent des zones d'atrophies cutanées et des papillomes, en particulier des lèvres et des régions génitale et/ou anale. Scoliose, hypoplasie claviculaire et costale, et déformations thoraciques font partie des anomalies osseuses rencontrées. Les anomalies dentaires sont la règle et peuvent associer malpositions dentaires, dents surnuméraires et anomalies de l'émail. L'atteinte oculaire à type de colobome de l'iris, microphtalmie, et/ou strabisme est classique. Un retard psychomoteur peut être observé.

Deux femmes avec une délétion de la région terminale dans le bras court du chromosome X ont été décrites comme combinant un phénotype de syndrome d'Aicardi et de Goltz [68].

La région Xp22.31 serait donc impliquée dans un syndrome des gènes contigus. L'hypothèse alternative serait que les gènes des deux syndromes seraient très proches. Cependant,

Gorlin affirme que les conditions décrites par Naritomi *et al.* [68] correspondent clairement à un syndrome MLS (OMIM 305600 : (Gorlin, R. J.,Personal Communication. Minneapolis, Minn., 10/19/1998)). Les cas présentés n'étaient probablement pas un syndrome de Goltz, puisque le gène PORCN (porcupine homolog) localisé en Xp11.23, a été tout récemment impliqué dans le syndrome de Goltz [78, 79].

Certains auteurs [76] ont avancé l'hypothèse que le gène Aicardi pouvait se situer dans une région autre que Xp22. Sans exclure totalement celle-ci, il ne paraît néanmoins plus raisonnable de concentrer les recherches seulement sur cette région de l'X.

3.4.3.1. Approche fonctionnelle

Cette approche est basée sur la fonction des gènes et non plus sur les anomalies structurales de l'ADN. Jusqu'à présent, cette méthode a aboutit au séquençage du gène *FLNA (*filamin A*).* Une étude post mortem des cerveaux de deux petites filles Aicardi, avait révélé l'accumulation de filamine dans les astrocytes. La filamin A est impliquée dans les hétérotopies familiales nodulaires bilatérale périventriculaires transmises selon un mode dominant liée à l'X avec une létalité chez le garçon. L'observation de ce dépôt de filamine A dans les astrocytes des patientes AIC a conduit les auteurs à émettre l'hypothèse que le gène de la filamine A serait aussi impliqué dans le syndrome d'Aicardi [76]. Le séquençage des 48 exons, incluant les jonctions exon-intron n'a révélée aucune mutation délétère chez les 10 filles AIC testées [76].

4. De nouvelles approches expérimentales pour l'identification de gènes candidats

De nouveaux outils sont actuellement disponibles autorisant une étude globale du génome ou du transcriptome de ce génome. Ces outils font appel à la construction de microréseaux

(microarrays) de molécules cibles d'ADN qui seront reconnues par hybridation comparative par de l'ADN génomique (puces dites CGH pour Comparaive Genomic Hybridization) ou par des ARN cellulaires (puces dites transcriptomiques).

Ainsi dans ce chapitre, j'aborderai ces deux types de puces et leur application dans la compréhension des maladies génétiques. Par souci de clarté, je prends le parti d'utiliser le terme de puce génomique pour les puces génomiques qui permettent une étude du génome à la recherche de variations quantitatives et le terme de puces transcriptomiques pour les puces dédiées à l'étude des transcrits.

4.1. Puces génomiques

4.1.1. Principe

La technique d'hybridation génomique comparative sur puces à ADN, est basée sur le même principe que l'hybridation génomique comparative sur chromosomes métaphasiques. Dans cette approche, les chromosomes sont remplacés par des segments génomiques dont la localisation physique est connue [80]. Dans une expérience typique de puce génomique, un ADN test et un ADN référence sont marqués par des fluorochromes distincts, les séquences répétées sont bloquées par de l'ADN Cot1 humain. L'hybridation (sur une sonde) de deux ADN marqués en quantité équimolaire sera statistiquement égale. Les hybridations aspécifiques sont éliminées par une série de lavage puis les intensités de fluorescence sont lues à l'aide d'un scanner fonctionnant sur le principe d'un microscope confocal couplé à un laser. Un rapport des intensités test/référence est calculé pour chacune des sondes présentes sur la puce. En théorie, les signaux sont représentatifs du nombre de copies initiales respectifs dans les deux ADN cohybridés, en pratique ce n'est pas exactement le cas. En effet, plusieurs facteurs expérimentaux rendent la normalisation des résultats nécessaire. Par exemple, l'incorporation des deux fluorochromes ne se fait pas avec la même efficacité, et une différence d'intensité peut aussi être observée selon les régions de la lame du fait des compositions en base, de la proportion des séquences répétées… [81]. La normalisation s'effectue sur des sondes contrôles, présentes en nombres de copie égaux et consiste à leur

attribuer un rapport test/référence égale à 1 (ou à 0 sur une échelle logarithmique). Une autre méthode consiste à normaliser par rapport à la moyenne des intensités de la lame, en considérant que la majorité des ratios sont proche de 1 (clones non déviants dans l'ensemble).

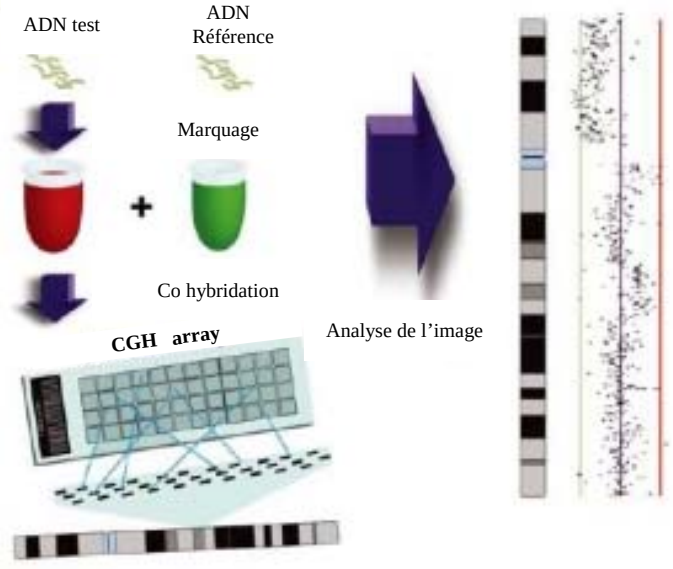


Figure 10 : Représentation d'une expérience classique de CGH array.

Après extraction, les deux populations d'ADN sont marquées (généralement la cyanine 3 et de la cyanine 5 sont utilisées). Les deux populations de cibles marquées sont hybridées sur les sondes de la puce. Le ratio d'intensité entre la fluorescence émise par le fluorochrome rouge (cyanine 5) représentant l'ADN test et celle émise par le vert (cyanine 3) représentant l'ADN référence est calculé. Pour chaque sonde, ce ratio permet de détecter les variations du nombre de copies de la cible chez un patient (test) par rapport à un témoin référence. Ce ratio représente les variations présentes dans l'ADN test par rapport à la l'ADN de référence.

4.1.2. Considérations techniques et analytiques

4.1.2.1. Les différents types de puces génomiques

La résolution de la technique dépend de la taille des sondes déposées sur la lame, du nombre et de l'espacement régulier des sondes entre elles, en terme de couverture du génome [82].

L'approche initiale des puces génomiques fut l'utilisation d'ADN de grande taille par l'intermédiaire des clones BAC (Bacterial Artificial Chromosome)[80, 83]. L'utilisation de clones génomiques de haut poids moléculaire a été décrite pour l'analyse des réarrangements dans les tumeurs [84, 85] ou la détection de remaniements constitutionnels [86]. Les clones peuvent être spécifiques d'une région [87], d'un chromosome [88] ou représenter le génome complet [89]. Des puces composées de sondes moins complexes constituées d'ADNc [90], de produits de PCR [91] ou d'oligonucléotides [92] sont aussi disponibles. Les clones d'ADNc d'une longueur comprise de 0,5 à 2 Kb présentent l'inconvénient d'être non homogène dans leur taille. Les signaux les plus intenses seront observés surtout pour les ADNc les plus longs. D'autre part, du fait de leur composition exclusivement exonique la détection des anomalies introniques n'est pas possible. Un avantage d'une telle puce réside dans le fait que les délétions et les amplifications peuvent être corrélées au changement d'expression en utilisant la même puce mais avec des cibles de type ARN [93]. Les puces à oligonucléotides sont elles plus récentes. Elles détectent des polymorphismes mononucléotidiques ou SNP. Leur emploi nécessite un traitement préalable de la cible afin de réduire sa complexité génomique et diminuer la probabilité d'hybridation croisée aux multiples sondes oligonucléotidiques de la puce et donc de signaux non spécifiques [94].

Les puces génomiques composées de clones de haut poids moléculaire donnent des signaux d'intensité plus élevée que les puces avec des séquences plus courtes. Les différences d'intensités sont plus nettes et donc les mesures plus précises. Par contre, les puces composées de séquences d'ADN plus courtes ont potentiellement une résolution supérieure à celle des puces à BAC. En effet, la taille des BAC étant comprise entre 100 et 200 Kb, les variations seront visibles si plus de 50% de la séquence est impliquée, ce qui correspond à en moyenne à 80 kb. Inversement, pour les séquences courtes, l'intensité du signal issue de l'hybridation est proche de celui du bruit de fond. Les améliorations visent donc à utiliser des séquences courtes avec une intensité de fluorescence élevée. Aujourd'hui, des sociétés comme Agilent proposent des puces de 244 000 oligonucléotides couvrant l'ensemble du génome. Les oligonucléotides sont synthétisés *in situ* par la technologie jet d'encre. Le principe est d'envoyer des agents chimiques à l'aide d'éjecteurs piézoélectriques qui vont déprotéger les nucléotides synthétisées [95, 96]. Une autre technologie utilisée par NimbleGEn est la photolithographie. Le procédé consiste à illuminer sélectivement un

substrat de verre comportant des nucléotides avec leur protecteur photolabile. Cette exposition à travers un masque enlève les groupes protecteurs, ce qui donne accès à des régions déprotégées où le couplage chimique avec de nouveaux nucléotides va pouvoir s'effectuer.

4.1.2.2. Eléments répétées

La variation des ratios est affectée par plusieurs facteurs induisant des biais dans les résultats expérimentaux. Les éléments répétés qui représentent plus de la moitié du génome ont une grande importance. Parmi eux, les ADN satellites au niveau des centromères, composés de séquences répétées en tandem et les séquences provenant d'éléments transposables tel les LINES (Long Interspersed) , SINE (Short INTerspersed), séquence Alu …
Le blocage par l'ADN cot1 n'est pas efficace à 100% et donc il en résulte un biais résiduel. D'autre part, le contenu en séquences répétées peut varier entre les sondes de la puce. Ainsi chaque puce est optimisée par rapport d'une part à la quantité d'ADN Cot1 et d'autre par le choix des ADN déposées, en évitant les séquences avec un grand nombre de séquences répétés. On détermine ainsi des valeurs seuils pour lesquelles le ratio est « déviant ». Classiquement pour des puces à BAC, la « normalité » est comprise entre 0,80 et 1,20 en ratio d'intensité.

4.1.2.3. Variation du nombre de copie

D'autres variations génomiques pouvant affecter les résultats obtenues avec les puces génomiques sont les variations structurales de l'ADN comme les délétions/duplication/inversion (tableau 4).
Les duplications segmentaires ou LCR (Low Copy Repeat), sont des blocs d'ADN de 1 à 400 kb présents à de nombreux emplacements sur le génome mais surtout concentrée au niveau des télomères et centromères. Cinq pour cent du génome est composé de ces séquences dupliquées à haut degré de similarité (> 90 %) [97]. Les études moléculaires ont montrés que la présence de grandes séquences LCR fortement homologues flanquantes prédisposaient

les réarrangements chromosomiques par recombinaison homologue non allèlique (NAHR). De nombreuses études ont rapporté une association entre la localisation de ces structures et les réarrangements chromosomiques [98-101].

Variation	Rearrangement type	Size range[a]
Single base-pair changes	Single nucleotïde polymorphisms, point mutations	1 bp
Small insertions/deletions	Binary insertion/deletion events of short sequences (majority < 10 bp in size)	1-50 bp
Short tandem repeats	Microsatellites and other simple repeats	1-500 bp
Fine-scale structural variation	Deletions, duplications, tandem repeats, inversions	50 bp to 5 kb
Retroelement insertions	SINEs, LINEs, LTRs, ERVs[b]	300 bp to 10 kb
Intermediate-scale structural variation	Deletions, duplications, tandem repeats, inversions	5 kb to 50 kb
Large-scale structural variation	Deletions, duplications, large tandem repeats, inversions	50 kb to 5 Mb
Chromosomal variation	Euchromatic variants, large cytogenetically visible deletions, duplications, translocations, inversions, and ancuploidy	~5 Mb to entire chromosomes

[a] Size ranges quoted are indicative only of the scale of each type of rearrangement, and are not definitive.
[b] SINE, short interspersed element; LINE, long interspersed element; LTR, long terminal repeat; ERV, endogenous repeat virus.

Tableau 3 : Le spectre de variation structurale du génome humain

La présence de variation du nombre de copie (CNV pour copy number variation) de certaines régions d'ADN génomique chez des individus sains a été décrite de façon indépendante par deux études. Ainsi en 2004, Sebat *et al.,* grâce à des puces à oligonucléotides à haute densité, identifient 76 régions génomiques sujettes à des variations quantitatives [102]. La même année, cette fois avec des puces à BAC d'environ 1 Mb recouvrant l'ensemble du génome Iafrate *et al.* [100] identifient 255 loci avec variation du nombre de copie. Leurs observations ont rapidement été confirmées et complétées par plusieurs équipes [82, 103]. Les bases de données publiques telles que le Genome Browser (http ://genome.ucsc.edu/) incluent ces données sur les CNV et les bases de données Database of Genomic Variants (http ://projects.tcag.ca/variation/) ainsi que Human Structural Variation Database (http://humanparalogy.gs.washington.edu/structuralvariation/) sont dédiées à ces variants structuraux.

Le terme CNV est réservé aux variants génomiques de taille supérieure à 1 kb [104] et remplace les termes précédemment utilisés pour les désigner tel que LCRs (Large-scale Copy Repeat) [100], CNPs (Copy Number Polymorphisms) [102] et ISVs (Intermediate Size Variants) [105]. Les CNV représentent de l'ordre de 12% du génome et touchent plusieurs

milliers de gènes. Ces variations structurales influenceraient l'expression des gènes de diverses façons. Les insertions/délétions peuvent ajouter ou enlever des copies de gènes, influençant ainsi le phénotype si le gène en question est sensible à l'effet de dosage. A noter que dans certains cas, la corrélation entre l'expression et la variation du nombre de copies n'est pas obligatoire. C'est le cas pour les gènes de l'alpha défensine (*DEFA1/DEFA3*) où il n'y a pas de relation entre le niveau d'ARNm et le nombre de copies du gène [106]. Une majorité de gènes serait tolérante envers des variations en nombre de copie [107] . Les insertions/délétions dans les régions codantes peuvent modifier l'épissage, fusionner des gènes... Ceux hors des régions codantes peuvent également influencer l'expression en touchant les séquences régulatrices [108].

Il est intéressant de noter que ces variants ont été associés à des familles de gènes qui présentent une forte évolution inter-espèces [109] ainsi qu'à des gènes impliqués dans l'immunité, suggérant que ces CNV joueraient un rôle important dans l'adaptation à l'environnement [110]. De plus, les CNV pourraient influencer la pénétrance des maladies mendéliennes dominantes en rapport avec des mutations dans des gènes cartographiés dans des CNV. Différents modèles théoriques ont été proposés par l'équipe de Beckmann pour expliquer l'influence des CNV sur la pénétrance et la sévérité variable des phénotypes [111]. Ces données confèrent au génome une image étonnement dynamique non suspectée.

De nombreuses études rapportent une association entre la localisation des CNV et les duplications segmentaires qui sont clairement des points chauds pour les remaniements chromosomiques. Pour des CNV de plus de 10 kb, le nombre de duplications segmentaires présent serait 4 à 10 fois supérieure [100, 101].

Les duplications segmentaires et les CNV, présents chez tous les individus affectent également les résultats des puces génomiques [112, 113]. Si un locus contenant ce genre de sondes varie en nombre de copies, le ratio correspondant peut être sous estimé du fait de l'absence de ces structures dans l'autre allèle [114, 115]. Inversement tous les loci contenant ce genre de séquence peuvent donner un ratio déviant de la normalité [115-117]. Ces séquences ne peuvent pas être éliminées par l'ADN Cot1 comme pour les séquences Alu et LINES. Ainsi, la composition des séquences déposées sur les lames de CGH est importante, puisque comme le rapportent Locke *et al.,* [116] les variations telles que les duplications segmentaires peuvent induire des ratios déviants (1/2 ou 3/2), les résultats sont significativement compromis lorsque les duplications sont de grande taille (50 à 100 % du

BAC) et hautement identiques (> 98 %). Il en va de même pour les CNV recouvrant souvent plusieurs clones BAC, présent chez les individus témoins et présent dans une grande proportion du génome (environ 12 %)[118]. Il s'agit dans ce cas de déterminer comme pour le cas des SNP, par accumulation des connaissances, quels CNV sont associés aux maladies et lesquels ne le sont pas.

4.1.2.4. Analyse des données issues de puces génomiques

Les premières approches de normalisation des ratios d'intensité étaient basées sur un ajustement simple de la médiane des ratios d'intensités à la valeur 1 (ou bien, le log de la médiane des ratios à 0). D'autres approches de normalisation plus complexe sont aussi utilisées. Quoiqu'il en soit, appliquer une normalisation donnée sans comprendre les facteurs de variabilité possible peut introduire des erreurs systématiques. Outre la normalisation, il n'est pas rare de réaliser des expériences où les fluorochromes utilisées sont inversés (dye swap ou hybridation inverse). Ainsi, les variants retenus sont ceux qui apparaissent dans les deux expériences.

4.1.3. Les apports des puces génomiques

Les anomalies chromosomiques sont une cause fréquente de maladies génétiques. Les techniques classiques de cytogénétique comme le caryotype sont limitées en résolution ou comme pour la technique de FISH servent à la détection d'une région spécifique du génome.
Les techniques de CGH permettent au contraire la recherche de variants quantitatifs sur l'ensemble du génome et avec une résolution pouvant aller jusqu'à quelques dizaines de bases.
Les puces génomiques ont permis d'identifier des microremaniements chromosomiques jusqu'alors non détectables. Ceci a conduit à la découverte de nouveaux gènes impliqués dans les maladies génétiques [119, 120] ou bien à définir la région critique pour une maladie donnée [121]. Par exemple , dans le diagnostic des retards mentaux modérés à sévères (avec ou sans dysmorphie), l'introduction de la technique de puce génomique, a amélioré la

détection des remaniements d'environ 8% [122]. Ainsi de nombreuses études à la recherche d'anomalies submicroscopiques chez des patients avec anomalies du développement ont été rapportées [123-127]. Le gène du syndrome CHARGE (Coloboma, Heart anomalies, choanal Atresia, retardation, Genital and Ear anomalies), (OMIM 214800) a ainsi été cloné avec succès grâce à la technologie des puces génomiques [120]. Dix huit patients ont été criblés avec des puces génomiques de 1Mb de résolution contenant des clones BAC. Une microdélétion de novo de 4,8 Mb a ainsi été identifiée en 8p12. Le séquençage des 9 gènes de la région candidate a révélé des mutations délétères dans le gène CHD7 (chromodomain helicase DNA binding protein 7) dans la majorité des patients. Pour le syndrome de Pitt-Hopkins, durant le criblage par puces génomiques, une délétion de 1,2 Mb en 18q12 a été détectée chez un des patients étudiés. Cette région contient trois gènes connus *RAB27B* (famille des oncogènes RAS), *CCDC68* (coiled-coil domain containing 68), et *TCF4* (transcription factor 4) qui se révèle être impliqué dans le syndrome. De même, dans la quête des gènes des retards mentaux non syndromiques liés au chromosome X, le gène *ZNF674* a été découvert grâce à la mise en évidence d'une délétion de 1 Mb en Xp11 à l'aide d'une puce constituée de BACs spécifiques de l'X [128].

4.1.4. Les limites de la technique

En premier lieu, les remaniements de structure équilibrés ne seront pas reconnus par cette approche purement quantitative. Ainsi, la mise en évidence d'une anomalie déséquilibrée et de survenue de novo chez un enfant nécessitera une analyse cytogénétique parentale pour chercher un éventuel remaniement de structure équilibré. De même, bien que plus rares, les remaniements de structure équilibrés à l'origine d'une pathologie de point de cassure, par interruption de la structure d'un gène ou de sa régulation par effet de position, ne seront pas diagnostiqués par la puce. Finalement, un certain nombre de cas restent difficiles d'interprétation compte tenu des variations observées dans la population générale. Ceci nécessite donc une meilleure compréhension de ce que l'on pourra appelé les variants génomiques.

C'est pourquoi des plates-formes spécifiques ont été récemment mises en place, par le Wellcome Trust Sanger Institute (www.sanger.ac.uk/PostGenomics/decipher/), l'European Cytogenetics Association (www.ecaruca.net/) dont le but est de rassembler les données de cytogénétique moléculaire corrélées aux informations cliniques qui leur sont associées. Ces plates-formes fourniront les bases indispensables à la compréhension du rôle de ces variations génomiques de grande taille et de leur rôle dans la pathologie constitutionnelle. Enfin, une difficulté inhérente à la technologie existe dans certains cas de figures. L'analyse des données passe par une quantification relative au nombre de copie d'une référence. Il n'y a pas de distinction de répartition du nombre de copie sur les deux chromosomes homologues.

En résumé les puces génomiques constituent manifestement un outil efficace pour la recherche de variants quantitatifs délétères impliqués dans les maladies monogéniques. Cette technologie, représente à l'heure actuelle la méthode de clonage positionnel la plus efficace dans le cas des retards mentaux sporadiques où les techniques classiques ne sont pas utilisables. Il est donc pertinent de se tourner vers cette technique pour la recherche de gène candidats du syndrome d'Aicardi.

Outre cette approche constitutionnelle, les puces génomiques permettent également l'étude globale du transcriptome. Puisque le transcriptome n'est pas une structure figée mais au contraire dynamique, la possibilité donnée d'une vision du transcriptome dans sa globalité a eu un impact très fort sur la compréhension des fonctions et voies dérégulées impliquées dans les mécanismes pathologiques. Ces études ont été fructueuses dans l'amélioration du diagnostic et du pronostic clinique pour certaines tumeurs solides [129-131] et hémopathies [132, 133] et ont apporté des résultats essentiels à la compréhension de ces pathologies.

4.2. Etude du transcriptome

La naissance de la technique, résulte de la nécessité, d'une étude globale du transcriptome. Jusqu'alors, l'étude des transcrits passait par une approche analytique d'un gène. Et c'est par l'accumulation de ces expériences unitaires qu'une compréhension *a posteriori* de la cellule était envisagée. Les nouvelles approches globales quant à elles fournissent une image à un temps t du niveau d'expression de plusieurs milliers de gènes dans un type cellulaire et un contexte physiologique et/ou pathologique donné. Cette vision globale instantanée est très importante en particulier pour le transcriptome pour lequel au temps t+1 l'image obtenue peut être très différente. Dans la présentation, concernant l'étude du transcriptome, je me focaliserai principalement sur les différentes étapes d'une expérience des puces à ADN deux couleurs et leurs applications.

4.2.1. Principe des puces transcriptomiques

Dans l'étude globale des transcrits, on distinguera les microarrays à une couleur (type Affymetrix) où l'on mesure une valeur d'expression absolue et les puces deux couleurs (par exemple Agilent) qui sont basées sur l'expression relative d'un échantillon test par rapport à un échantillon référence.

Dans tous les cas, l'expérience commence par une extraction des ARN suivie de la synthèse des ADNc, leur marquage et hybridation sur les puces à ADN, lavages et séchage des lames et lecture par un scanner. Le signal généré pour chaque fluorochrome est converti en fausses couleurs (généralement rouge pour la cyanine 5 et vert pour la cyanine 3). Les deux images sont superposées. Les étapes présentées dans la Figure 9 sont ainsi valables. Seules les cibles changent de nature.

4.2.2. Le plan expérimental

La conception de l'expérience dépend en premier lieu de la question biologique posée [119, 134, 135]. Cette réflexion est en particulier importante pour les puces à deux couleurs pour lesquelles on obtient une expression relative. Les résultats seront fonction du choix des échantillons à hybrider simultanément sur la puce. Le plan expérimental doit être pensé afin de rendre l'analyse et l'interprétation des données aussi simples et efficaces que possible, compte tenu du problème posé ainsi que des contraintes expérimentales et matérielles.
Les points suivants doivent être pris en considération pour optimiser la conception initiale du projet :

4.2.2.1. Eviter le plus possible les confusions d'effets.

Par exemple, il ne faut pas confondre l'effet expérimentateur et l'effet biologique (recherché). Si 20 lames sont hybridées par 2 personnes (10 lames chacune), les différences observées dans ces deux groupes de lames seront aussi forcément influencées par l'expérimentateur. Il est donc préférable qu'une même et seule personne intervienne lors des manipulations. Une des solutions proposées pour distinguer les variabilités techniques et les variations biologiques est de maximiser les réplicats biologiques et optimiser les réplicats techniques [136, 137]. Réaliser des réplicats biologiques consiste à analyser le plus grand nombre d'échantillons possibles. Les réplicats biologiques peuvent être intra et inter-individus. Les réplicats intra individu correspondent à différents prélèvements du même tissu chez un même patient tandis que les réplicats inter-individus sont réalisés avec différents prélèvements du même tissu chez différents individus. Les réplicats techniques, quant à eux, peuvent correspondre aux dépôts multiples d'un même gène sur une même lame ou à l'hybridation de plusieurs lames avec les mêmes échantillons et une inversion des marquages (hybridation inverse). Les réplicats techniques permettent de valider la qualité des différentes étapes de conception d'une puce à ADN. Dans une étude de puce transcriptomiques, où l'on teste une hypothèse statistique, les réplicats biologiques sont essentiels alors que les réplicats techniques ne le sont presque jamais [138, 139].

4.2.2.2. Eviter les biais inconscients des expérimentateurs.

Une randomisation de l'hybridation peut souvent se révéler utile pour éviter de regrouper les échantillons de façon inconsciente. Si vingt lames sont à hybrider en 4 séries de 5 lames, il est préférable de randomiser des séries.

4.2.2.3. Référence commune

Lorsque l'on souhaite comparer plus de deux échantillons, il est préférable de choisir une référence commune permettant de normaliser les résultats de chaque hybridation. Cette référence peut être constituée de l'un des échantillons à comparer, un échantillon de référence distinct tel qu'un mélange arbitraire d'ARN, ou encore un mélange des ARN de tous les échantillons à comparer.

4.2.2.4. Choix imposés par les limitations matérielles et techniques.

Il convient de tenir compte du type et de la quantité de matériel dont on dispose (nombre de puces, facilité d'obtention des échantillons d'ARN et quantité disponible…) et de les confronter à ce qui sera requis pour réaliser les expériences de manière optimale.
D'autre part, pour une même question plusieurs réponses sont possibles. Il faut savoir qu'il n'y a pas de plan optimal mais seulement des plans plus judicieux (ou moins mauvais) que d'autres. Voici trois principaux types de plan expérimental [140]. Le dessin avec référence commune, le dessin en boucle et un plan d'expérience en « bloc » (Figure 11).

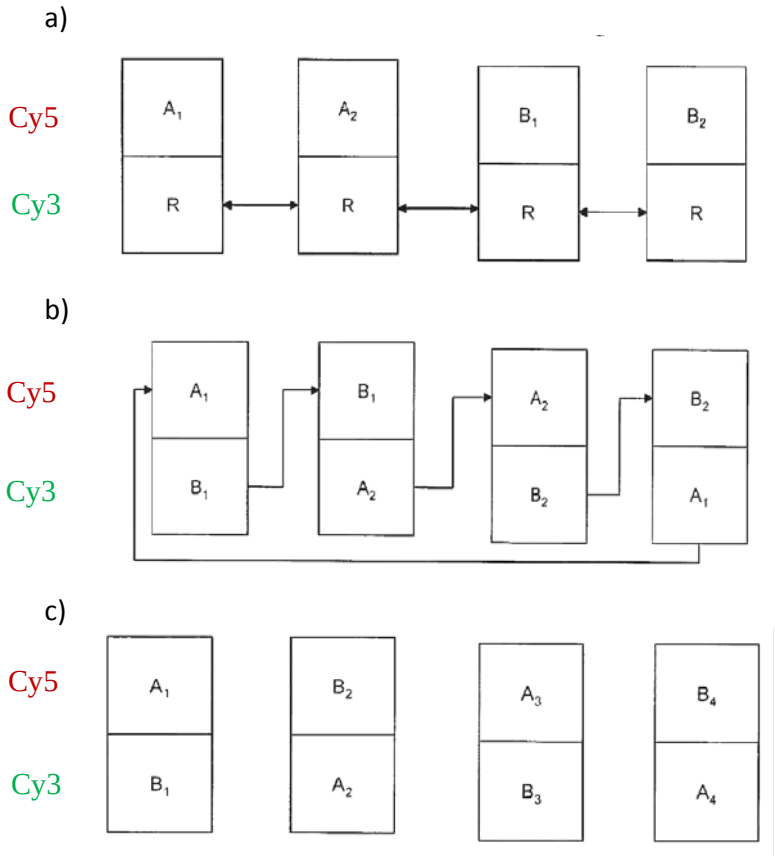


Figure 11 : Exemples de plans expérimentaux pour des expériences de comparaison de classe

Les rectangles représentent les puces à ADN. A1 est l'échantillon de classe A, B1 est l'échantillon de la classe B etc… R est l'échantillon de référence pour les comparaisons indirectes. Les flèches connectent les mêmes échantillons d'une expérience à l'autre. La cyanine5 (Cy5) et la cyanine3 (Cy3) sont les fluorochromes utilisées pour le marquage des échantillons. A) Dessin avec référence commune (Reference design), tous les échantillons sont comparés à la même référence. B) Dessin en boucle (Loop design). C) Dessin en bloc (Block design) pour une comparaison directe des échantillons.

Le dessin avec référence commune est le plus couramment utilisé. Il rend les comparaisons inter-lames possible et les analyses des résultats plus simples que les plans expérimentaux plus complexes.

D'autre part, on s'affranchit des expériences de dye swap (inversion des fluorochomes), puisque tous les échantillons test sont marqués avec le même fluorochrome. Les biais éventuels liés au marquage sont alors les mêmes pour l'ensemble des puces et n'affectent pas les comparaisons entre les échantillons. L'inconvénient de ce plan expérimental est que

la comparaison des classes n'est pas directe et passe par la référence. On introduit donc forcément des variabilités inter-lames.

Un plan d'expérience en « bloc » (block design) permet également de gérer l'influence de certains facteurs de variation sur une expérience. Par exemple, les signaux d'une lame subissent les mêmes traitements (hybridation, lavage...) et les mesures au sein de cette lame sont plus homogènes entre elles qu'entre les autres lames de l'expérience. La comparaison entre deux échantillons hybridés sur une même lame est donc plus directe. Dans le cas d'une comparaison intra-lame, seule une normalisation des données à l'intérieur de la puce est nécessaire tandis qu'une comparaison entre lame nécessite également un ajustement des mesures entre les lames. Comme pour le plan d'expérience en boucle, ce plan expérimental a l'avantage d'une comparaison directe cependant il est plus difficile d'effectuer des analyses statistiques des groupes puisque la comparaison inter-lame ne sera pas efficace.

4.2.3. Analyse des données d'expression

Dans la grande majorité des expériences le but est de trouver les gènes dont l'expression varie entre deux groupes (par exemple un groupe témoin et un groupe traité par un médicament).

Initialement la sélection des gènes signatures (gènes sous ou/et sur régulés) a été basée sur l'utilisation d'un seuil de log ratio. Pour chaque gène, c'est le rapport entre ses niveaux d'expression mesurés dans des conditions différentes qui est informatif : on parle de facteur de régulation ou variation de ratio (« fold change »). Avec cette méthode de variation de ratio, on se place à un seuil (souvent 2 en valeur absolue) et on sélectionne tous les gènes dont la variation de ratio est supérieur en valeur absolue à cette valeur seuil. Cependant, ce n'est pas un test statistique et n'intègre pas la variance [141]. Par exemple, la variabilité des différences d'expression sur l'ensemble de la puce n'est pas prise en compte. Ainsi, un gène (par exemple un facteur de transcription) dont la variation n'est que de 1,5 fois peut avoir des conséquences drastiques. D'autre part, une différence d'expression de 2 peut résulter d'un ratio d'intensité de 10 / 5 ou 10 000 / 5 000. Cette méthode est aujourd'hui considérée comme non valide pour la sélection de gènes signatures [138]. Une autre approche a été

proposée dans le logiciel Luminator® (Rosetta Biosoftware, Inc.). Les auteurs utilisent la construction d'un modèle d'erreur dérivé des données répétées, obtenues par l'ensemble des expériences menées avec une technologie de puce à ADN donnée. Rosetta Luminator® par ce modèle estime l'erreur sur la mesure de chaque gène. Ainsi, la valeur p associée au ratio de chaque gène (probabilité que la différence d'expression observée soit due au hasard) intègre la correction associée à la plateforme utilisée. Cette méthode comme la précédente ne prend pas en compte les réplicats biologiques dans le calcul des probabilités d'expression différentielle. Or, en utilisant les répétitions, il est possible de déterminer si un gène est ou n'est pas différentiellement exprimé, en utilisant les tests d'hypothèses (test statistique).

Les tests statistiques attribuent une valeur p associée à chaque gène cette fois-ci en combinant les résultats de plusieurs expériences (plusieurs lames). Plus robuste que les simples seuils présentés plus haut, il faut néanmoins savoir que dans la grande majorité des cas, le nombre de réplicats reste faible pour les tests statistiques. Ceci provient essentiellement de la forte dissymétrie des données, en effet le nombre de gènes analysé est toujours bien supérieur au nombre d'expériences (nombre de lames). On pourra notamment citer le test t de Student, l'analyse de la variance ANOVA et l'Analyse en Composantes Principales (ACP) ou l'approche Bayesienne.

- Le test t repose sur la comparaison des moyennes de chaque groupe. Le test de Student est la méthode la plus couramment utilisée pour évaluer si la différence observée entre 2 échantillons est significative. Cette approche requiert une normalité de la distribution et un nombre important d'échantillons.
- Une des nombreuses variantes du test t est l'approche bayésienne. Baldi et Long par exemple, présentent un test t modifié où la variance est estimée grâce au théorème de Bayes. Leur approche est exécutée par un logiciel nommé Cyber-T disponible sous la forme d'un service Web (http://cybert.ics.uci.edu/)[142].
- L'analyse de la variance (ANOVA) permet de tester si un facteur (variable qualitative) a un effet significatif sur une variable réponse quantitative [143, 144]. Dans une expérience de puces transcriptomiques, lorsque de multiples facteurs (âge, sexe..) sont à analyser, le test t et ses variantes ne suffisent généralement pas à l'interprétation. Une analyse type ANOVA permet d'évaluer si les moyennes de un

ou plusieurs groupes d'échantillons sont significativement différentes et si un ou plusieurs facteurs affectent les mesures.

- l'Analyse en Composantes Principales permet de fournir une vision globale des relations entre les gènes et les échantillons. Il récalcule une suite d'axes orthogonaux qui reflètent les variations dans les données.

A ces méthodes d'analyses se rajoutent les tests statistiques qui eux répondent à la question suivante : comment classer les gènes relativement à leurs mesures d'expression dans plusieurs expériences ? Dans ce cas interviennent les méthodes de classification (ou clustering). Les méthodes de classification sont utiles lorsque le but est de découvrir des groupes de gènes dans les données d'expression.

4.2.4. Comparaison inter-plates-formes

Il est fréquent d'observer une apparente incohérence des données lorsque l'on compare plusieurs plates-formes entre elles [145]. Les gènes différentiellements exprimés dans l'une et l'autre des plates-formes ont une intersection relativement faible. Toutefois, si la comparaison est faite sur des données pré-filtrées, les comparaisons révèlent une bonne corrélation des résultats [146]. Ainsi, pour qu'une comparaison inter-plateforme soit valable, il faut d'une part que les échantillons biologiques testés proviennent de la même extraction et aussi que les puces utilisées soient comparables dans leur contenu. Il faut s'assurer que les différents oligonucléotides sur les puces d'une plateforme à l'autre représentent les mêmes ARNm ce qui n'est pas des plus aisés puisque les sociétés ne donnent pas forcément la séquence des oligonucléotides utilisés. D'autre part, les gènes dont l'expression est très faible sont plus difficilement détectables par certaines plates-formes (puces oligonucléotides 25 bases). En effet, l'intensité du signal est par exemple deux fois supérieure pour des oligonucléotides 30 bases que des oligonucléotides 25 bases [147]. Ainsi, si on filtre d'une part sur les mêmes séquences et que l'on omet les gènes de faible expression, les comparaisons inter-plates-formes sont beaucoup plus fiables qu'il n'y paraît. Au problème des multiples variations des expériences des puces à ADN s'ajoutent encore une étape

critique qui est l'analyse des données. Utiliser les outils adaptés non seulement à la question biologique posée mais aussi à la plateforme utilisée n'est pas des plus simples. Il est donc nécessaire de maîtriser l'ensemble de ces points pour prétendre faire une comparaison inter-plateforme valable.

4.2.5. Signification des données d'expression

Les analyses statistiques permettent de dégager les gènes différentiellement exprimés mais ne donnent pas la signification biologique des résultats. Une intégration des données d'analyse du transcriptome avec les connaissances préétablies est nécessaire pour interpréter des résultats de puces à ADN et construire de nouvelles connaissances. Là encore la quantité de données est très importante et une analyse manuelle ne peut être une solution. Les méthodes utilisées vont de l'annotation fonctionnelle des gènes dérégulés à la construction de voies biologiques en passant par l'étude des promoteurs. Divers outils bio informatiques ont donc été développés pour effectuer l'intégration et l'interprétation des résultats des données transcriptomiques. Nous aborderons ces outils dans la partie bio informatique.

4.2.6. Applications possibles des puces à ADN dans l'étude du transcriptome

La plupart des maladies chez l'homme comme le cancer, les maladies cardiovasculaires, les maladies neurologiques ont des étiologies complexes. Elles font intervenir de nombreux gènes en interaction aussi bien qu'une composante environnementale. Les techniques d'études du transcriptome sont une approche pour élucider les mécanismes physiopathologiques de ces maladies, leur classification, la recherche de marqueurs pronostiques ou le développement de médicaments [129, 131]. Les premières études comprennent ainsi, les comparaisons de tissus malades et des tissus sains [148], les facteurs pronostiques des cancers [130, 149, 150] et leurs classifications [151, 152].

La majorité des revues s'accordent à dire qu'il y a une disparité entre la quantité et la qualité des études des données d'expression [153, 154]. Les publications ne fournissent pas les données nécessaires à cette estimation de qualité des données. Par exemple, une étude sur les publications concernant les données d'expression entre 2003 et 2005 a révélé que sur 293 publications, seulement une étude discutait de l'efficacité du test statistique et du facteur taille des échantillons [155].

Une homogénéisation des données est nécessaire et est d'ailleurs en cours avec des structures comme MIAME (Minimum Information About a Microarray Experiment) ou GEO (Gene Expression Omnibus) ou array express. En effet, ce sont des bases de données stockant les données d'expression avec la description de la méthodologie de l'expérience et ceci dans une structure et un vocabulaire unique.

Si l'analyse du transcriptome par la technologie des puces à ADN offre un aperçu des « corrélations » entre les gènes et les phénomènes biologiques (*« guilty by association »*), elle ne permet pas à elle seule de révéler la causalité des mécanismes de régulation [156]. Aussi, l'intégration des données complémentaires telles que les annotations des gènes issues de différentes sources contrôlées comme les ontologies, les résumés d'articles scientifiques ou les banques de données protéiques [157], est devenue indispensable pour interpréter les données issues des expériences de transcriptomique. La diffusion de ces informations par le web permet une large diffusion des données mais celle-ci s'est faite de façon indépendante, en séparant les données par entité biologique (ADN, ARN, Protéine), par niveau d'organisation différente (cellule, tissus, organisme…) et par technologie différente (analyse du transcriptome, du génome…). L'effort consiste à intégrer des données hétérogènes afin d'en extraire de nouvelles connaissances, qui mènent à la découverte. Depuis quelques années, des nouvelles stratégies de génomique intégrative ont vu le jour pour tirer profit de cette masse d'information [158]. Ainsi, l'intégration d'études de liaison et de données d'expression a aboutit à la découverte du gène impliqué dans la déficience en cytochrome c

oxydase chez l'homme [159]. L'intégration de données de QTL et données d'expression a conduit, chez la souris, à l'identification du gène Abcc6 comme le principal gène impliqué dans la calcification cardiaque [160]. C'est dans cette optique d'exploitation des bases de données publiques (7.2) dans le but de créer des systèmes de prédiction de gène candidats (7.3) que s'inscrira cette dernière partie de l'introduction.

5. L'approche bioinformatique pour la recherche de gènes candidats

5.1. Définitions de la bioinformatique

Le NIH (National Institutes of Health) propose la définition suivante : ***"****Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data*" "Recherche, développement, ou application des outils et des approches informatiques pour optimiser l'utilisation des données biologiques, médicales, comportementales ou sanitaires, y compris les données à acquérir, stocker, organiser, archiver, analyser, ou visualiser » (http ://www.bisti.nih.gov/CompuBioDef.pdf). Le NCBI intègre la notion de pluridisciplinarité et définit la discipline de la façon suivante : « *Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."*

« La bioinformatique est un champ de la science dans laquelle la biologie, l'informatique, et technologie de l'information fusionnent dans une même discipline. Il y a trois sous-

disciplines importantes dans la bioinformatique : le développement de nouveaux algorithmes et statistiques avec lesquels il est possible d'évaluer des rapports parmi les éléments de grands jeux de données ; l'analyse et l'interprétation de divers types de données dont les séquences nucléotidiques et d'acide aminés, les domaines protéiques et des structures protéiques ; et le développement et la réalisation d'outils qui permettent un accès et une gestion efficaces des différents types d'information ».

Une définition personnalisée correspondant à notre cas d'étude peut être : « La bioinformatique est un domaine de recherche qui propose et développe des modèles, des méthodes et des outils afin de stocker et d'exploiter les données biologiques produites notamment par les expérimentations à grande échelle ».

Les définitions se multiplient selon les utilisateurs et les domaines d'applications. Un informaticien donnera une définition différente d'un biologiste puisque leur point de vue et leur lieu d'intervention dans la discipline sont différents. Les nombreuses définitions du terme reflètent la nature ubiquitaire de cette discipline plutôt qu'un problème d'uniformité. Une définition n'est pas forcément meilleure mais est plus adaptée à l'utilisateur.

Une chose est certaine, pour être effective, la bioinformatique nécessite une collaboration étroite entre les disciplines impliquées (biologie, mathématique, informatique…). Les résultats des expériences à haut débit telles que les expériences de transcriptomiques sont des sources de données utilisées pour des analyses *in silico*. L'approche expérimentale (wet lab) ne peut se passer de l'approche *in silico* qui lui permet de manipuler les quantités de données générées. Réciproquement une hypothèse prédite par une approche *in silico* (un logiciel) a besoin d'une validation expérimentale. Cette complémentarité nécessite des connaissances propres à la bio informatique. Un biologiste seul (comme un informaticien seul) ne saura tirer profit de cette discipline. Ainsi, il est illusoire de prétendre interpréter et prédire les faits biologiques sans faire appel aux techniques informatiques d'exploitation des données et des connaissances du domaine. Inversement, les outils informatiques ont des limites d'utilisation et les ignorer conduit à donner un sens à ce qui n'en a pas réellement. Conformément à notre définition personnalisée de la bioinformatique. Nous traiterons ici successivement du stockage (5.2) puis de l'exploitation (5.3) des données biologiques.

5.2. Stockage des données biologiques

5.2.1. Nécessité des bases de données biologiques

L'accumulation des données biologiques est un phénomène notoire des 10 dernières années.

D'abord la génomique et le séquençage des génomes, puis la transcriptomique, la protéomique, plus récemment l'interactomique, la métabolomique viennent à tour de rôle submerger le biologiste de données nouvelles et abondantes. Ces données sont produites par des dispositifs expérimentaux de plus en plus complexes et autonomes. Au final ceci conduit à des données en masse qu'il faut : collecter, stocker (bases de données primaires tel GenBank), intégrer (bases de données secondaires tel GENE), distribuer et finalement explorer. L'exemple du séquençage des génomes est bien représentatif des bases de données primaires tel que GenBank. L'annotation puis l'intégration de ces données vise à donner un sens à l'information stockée. C'est le passage du génome séquencé au génome annoté et aux bases de données secondaires tel que GENE (NCBI). Pour une recherche de gène candidat, l'ensemble des données primaires que l'on peut associer à un gène et exploiter en bio informatique est schématisé sur la figure 12 (page suivante)

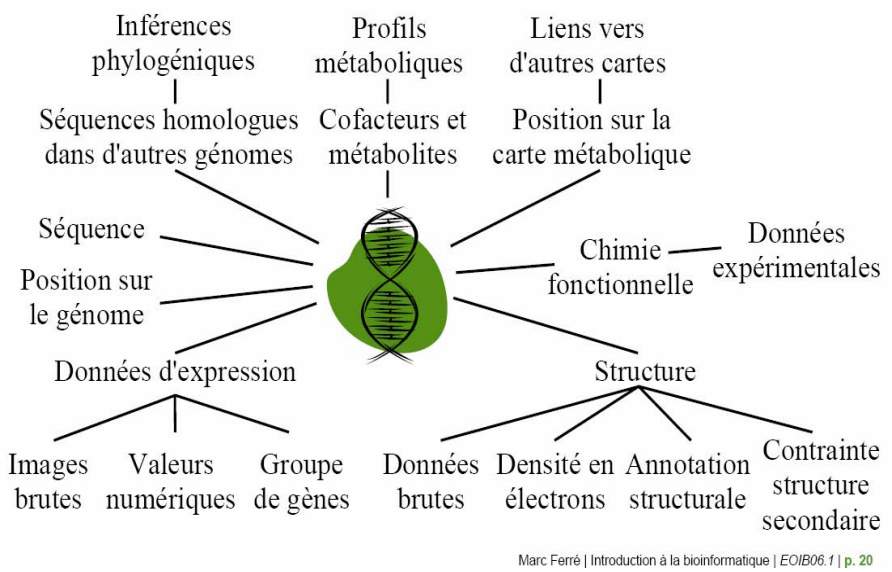


Figure 12 : Types de données pouvant êtres associées aux gènes

Pour chaque type de donnée, il existe aujourd'hui une ou plusieurs bases de données. Par exemple au NCBI, les données de séquence sont stockées dans une banque généraliste de séquence comme GenBank, les données d'expression issues des expériences de transcriptomique dans des bases de données d'expression telles que GEO (Gene Expression Omnibus), les données sur les gènes orthologues dans HomoloGene.

Les premières bases de données ont été crées dans les années 80. Ainsi, en 1986 sont créées simultanément les banques nucléiques, GenBank [161] et EMBL (European Molecular Biology Laboratory) [162]. La même année la première banque protéique, PIR-PSD (Protein InformationResource-Internationnal Protein Sequence Database) [163] est également mise en place. Il s'agissait en fait de banques de données constituées d'une succession d'enregistrements sous forme de fichiers textes lettrés (chaque item est repéré par un code en début de ligne). Par la suite j'utiliserai l'expression base de donnée (BD) pour désigner toutes les sources de données biologiques quelque soit leur structuration (banque de données, bases de données relationnelles…).

Le numéro spécial du journal Nucleic Acids Research sur les BD pour la biologie moléculaire référence, au début de l'année 2007, 968 BD publiques, soit 110 de plus que l'année précédente [164]. Cette collection se répartit actuellement en 14 catégories allant des BD de séquences nucléotidiques à celles des données immunologiques en passant par les banques

de données d'expression et de voies métaboliques. Ces BD sont plus ou moins généralistes, dédiées à un ou plusieurs organismes. Les BD ainsi référencées doivent être publiques et directement accessibles par les utilisateurs via le Web. Ainsi, celles du domaine privé ou nécessitant l'installation de logiciels en local ne sont pas recensées. Dans ce qui suit je présenterai essentiellement les bases de données sélectionnées dans notre stratégie de recherche de gène candidat pour le syndrome d'Aicardi (ACGR) « Approche for Candidate Gene Retrieval ».

Lors de la création du logiciel ACGR, notre choix s'est orienté d'une part vers des bases de données régulièrement mises à jour et d'autre part, nous avons cherché à minimiser le nombre de BD interrogées. Le site du NCBI est celui qui regroupe le mieux les données que nous recherchons. Ainsi, la BD GENE sera interrogé pour récupérer les données sur les gènes, leurs annotations issues de Gene Ontology (GO), et leur intéractants, la BD HomoloGene nous servira pour les données sur les gènes orthologues (HomoloGene), et la BD OMIM pour l'implication des gènes dans les maladies et la description de ces maladies. Les BD MGD (Mouse genome Database) et FlyBase ont été sélectionnées car ce sont à l'heure actuelle les bases de données de référence pour la souris et la drosophile. Elles font partie avec SGD (Saccharomyces Genome Databse) des trois banques de données d'organismes modèles qui ont initié le projet GO en 1998.

5.2.2. Présentation de quelques bases de données

5.2.2.1. Bases de données factuelles en biologie

Dans une base de données, les informations peuvent être codées, soit dans un langage naturel, par exemple l'anglais, soit codé dans un langage symbolique ou numérique. Dans le premier cas, nous parlerons de données textuelles, dans le second cas de données factuelles. Nous parlerons de bases de données textuelles pour les bases de données qui renferment essentiellement du texte. Il s'agit principalement de bases de données bibliographiques telles que Medline ou OMIM. Les BD factuelles sont les BD qui contiennent peu de texte. Il s'agit par exemple de bases de données comme Entrez GENE.

GENE, MGI et Flybase sont trois exemples de BD intégrées sur les gènes et HomoloGene un portail donnant les relations d'homologies entre les espèces.

- Entrez GENE

Entrez Gene (www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=gene) est une base de données du NCBI répertoriant les informations spécifiques aux gènes de différentes espèces [165]. C'est une BD intégrée dans le sens où elle inclut des informations provenant d'autres BD telles que la base de données primaire Genbank ou des BD spécifiques d'organismes modèles. Les données sont centralisées et nettoyées pour éliminer les redondances présentes notamment dans GenBank. Les entrées de GENE sont assignées d'un numéro identifiant GENE_ID unique. Le contenu de la BD est accessible par l'outil de recherche Entrez du NCBI. La fiche détail d'un gène inclut, lorsqu'elles sont disponibles, les données de nomenclature (symbole du gène, Alias, nom complet), sa localisation physique, la bibliographie, les gènes en interactions, les données sur le génotype (polymorphismes du gène), les voies biologiques dans lesquelles intervient le gène (KEGG), les homologies avec les autres espèces comme la souris et la drosophile. Suit une indexation du gène par des termes GO (voir plus loin), les numéros d'accession protéiques et nucléiques, les séquences références, les marqueurs, les phénotypes associés. Cette page contient également une collection de liens vers d'autres BD tel OMIM (Online Mendelian Inheritance in Man) et Homologene au NCBI, KEGG (Kyoto Encyclopedia of Genes and Genomes), … La figure suivante donne un aperçu des données disponibles sur une page de rapport complet.

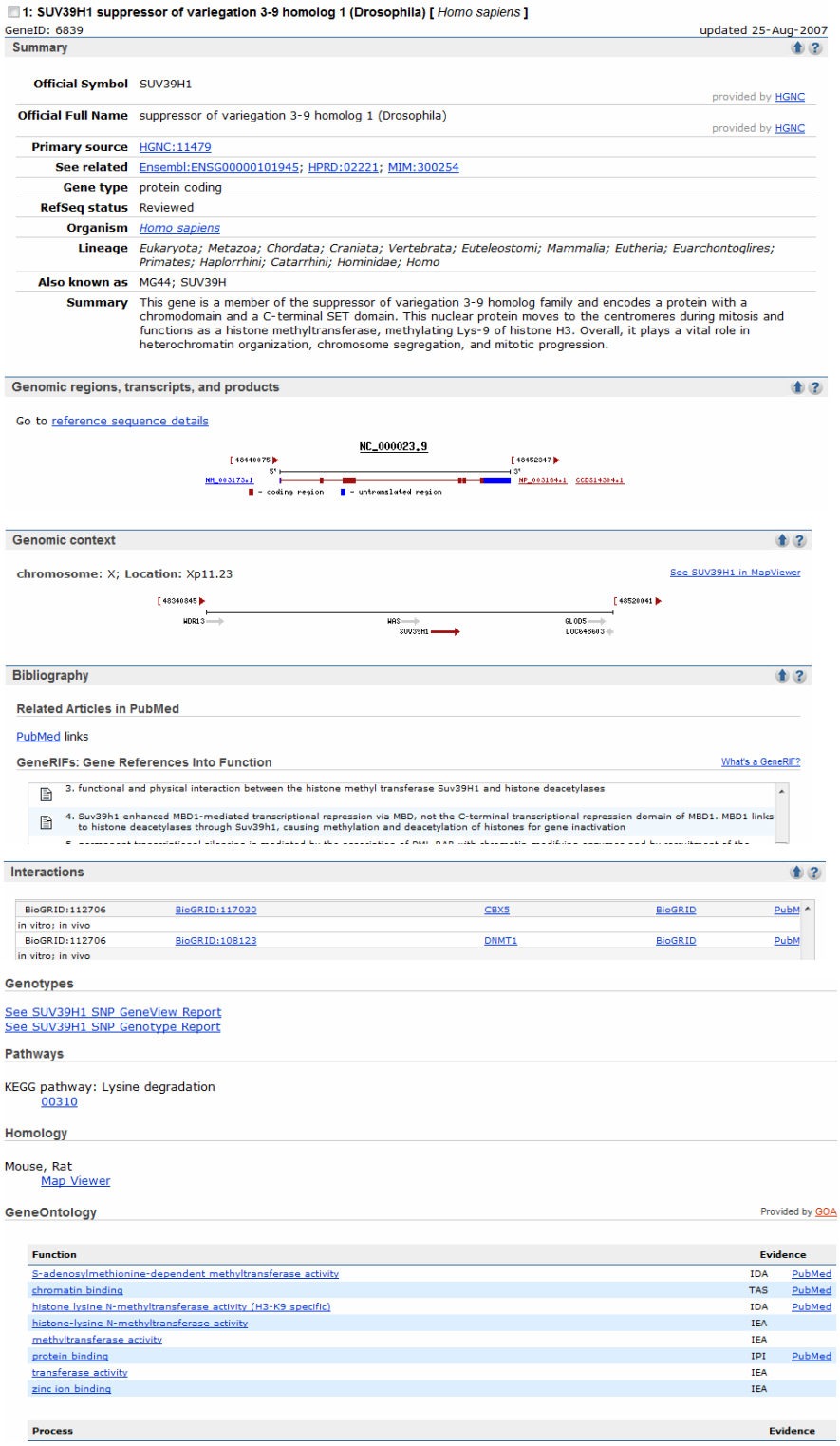


Figure 13 : Aperçu de la page Entrez GENE concernant le gène *SUV39H1*

www.ncbi.nlm.nih.gov/sites/entrez ?Db=gene&Cmd=ShowDetailView&TermToSearch=6839&ordinalpos=1&itool=EntrezSystem2.Pentrez.Gene.Gene_ResultsPanel.Gene_RVDocSum

(Suite de la figure 13 sur la page suivante)

cell cycle	IEA
cell differentiation	IEA
chromatin assembly or disassembly	IEA
chromatin modification	IEA
regulation of transcription, DNA-dependent	IEA
transcription	IEA

Component	Evidence	
chromatin	IEA	
chromosome	IEA	
chromosome, telomeric region	IEA	
condensed nuclear chromosome	TAS	PubMed
nucleus	IEA	
nucleus	TAS	PubMed

General protein information

Names
suppressor of variegation 3-9 homolog 1
H3-K9-HMTase 1
Su(var)3-9 homolog 1
histone H3-K9 methyltransferase 1
histone-lysine N-methyltransferase, H3 lysine-9 specific 1

NP_003164.1
EC 2.1.1.43

NCBI Reference Sequences (RefSeq)

Reference assembly

Genomic

1. **NC_000023.9 Reference assembly**
   Range: 48440075..48452347
   Download: GenBank FASTA
2. **NT_079573.3**
   Range: 11406898..11419170
   Download: GenBank FASTA

Alternate assembly (based on Celera assembly)

Genomic

1. **AC_000066.1 Alternate assembly (based on Celera assembly)**
   Range: 53108876..53096604, complement
   Download: GenBank FASTA
2. **NW_927703.1**
   Range: 619120..606848, complement
   Download: GenBank FASTA

Related Sequences

Nucleotide		Protein
Genomic	A58331.1	CAA03483.1
Genomic	AF196970.2 (44540..56812)	None
Genomic	CH471224.1	EAW50756.1

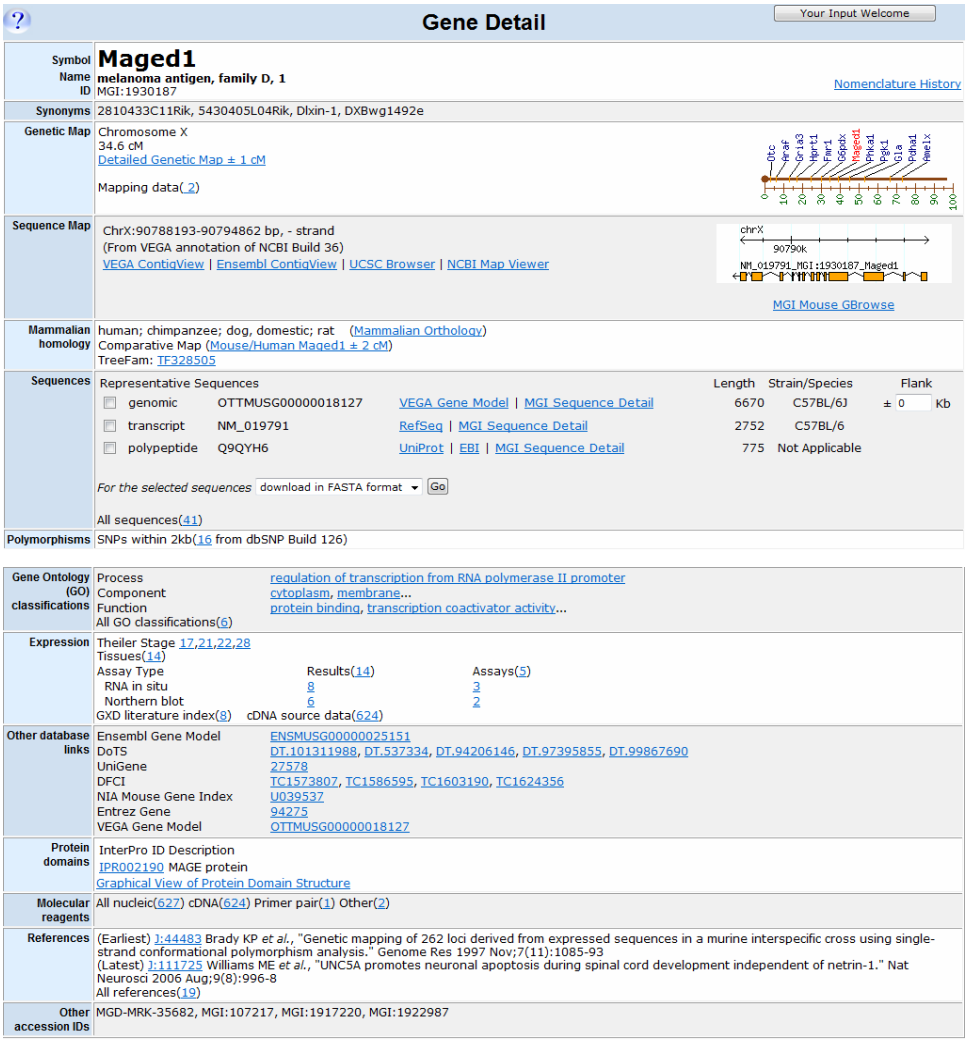
Additional Links

- MIM 300254
- HPRD 02221
- UniGene Hs.522639

- Mouse Genome infomatics

La BD spécialisée MGD accessible par l'interface MGI (www.informatics.jax.org) est dédiée à la souris, un organisme modèle clé dans l'interprétation et la compréhension du génome humain [166]. Un numéro identifiant MGI est attribué à chaque entrée. Les ressources de la BD comprennent une annotation issue de la littérature et de données expérimentales. Grâce aux informations fournies par les nombreux mutants produits chez la souris, l'information principale de la BD est la relation entre le génotype (une séquence) et le phénotype. La base de données contient notamment une ontologie (Mammalian Phenotype : MP) basée sur les

relations génotypes-phénotype. Ainsi, la catégorie « abnormal brain morphology » assignée de l'identifiant MP :0002152 inclut 1529 génotypes et 3700 annotations associées. La BD MGD contient également des données de cartographie physique, les homologies de séquence avec les mammifères, les polymorphismes, les annotations GO, des données d'expression, les domaines protéiques et des liens vers d'autres bases de données tel Entrez Gene et Medline.



Gene Detail

Your Input Welcome

Symbol	**Maged1**
Name	**melanoma antigen, family D, 1**
ID	MGI:1930187
	Nomenclature History
Synonyms	2810433C11Rik, 5430405L04Rik, Dlxin-1, DXBwg1492e
Genetic Map	Chromosome X 34.6 cM Detailed Genetic Map ± 1 cM Mapping data(2)
Sequence Map	ChrX:90788193-90794862 bp, - strand (From VEGA annotation of NCBI Build 36) VEGA ContigView | Ensembl ContigView | UCSC Browser | NCBI Map Viewer MGI Mouse GBrowse
Mammalian homology	human; chimpanzee; dog, domestic; rat (Mammalian Orthology) Comparative Map (Mouse/Human Maged1 ± 2 cM) TreeFam: TF328505
Sequences	Representative Sequences genomic OTTMUSG00000018127 VEGA Gene Model | MGI Sequence Detail 6670 C57BL/6J ± 0 Kb transcript NM_019791 RefSeq | MGI Sequence Detail 2752 C57BL/6 polypeptide Q9QYH6 UniProt | EBI | MGI Sequence Detail 775 Not Applicable *For the selected sequences* download in FASTA format Go All sequences(41)
Polymorphisms	SNPs within 2kb(16 from dbSNP Build 126)

Gene Ontology (GO) classifications	Process: regulation of transcription from RNA polymerase II promoter Component: cytoplasm, membrane... Function: protein binding, transcription coactivator activity... All GO classifications(6)
Expression	Theiler Stage 17,21,22,28 Tissues(14) Assay Type: Results(14): RNA in situ 8, Northern blot 6; Assays(5): 3, 2 GXD literature index(8) cDNA source data(624)
Other database links	Ensembl Gene Model: ENSMUSG00000025151 DoTS: DT.101311988, DT.537334, DT.94206146, DT.97395855, DT.99867690 UniGene: 27578 DFCI: TC1573807, TC1586595, TC1603190, TC1624356 NIA Mouse Gene Index: U039537 Entrez Gene: 94275 VEGA Gene Model: OTTMUSG00000018127
Protein domains	InterPro ID Description IPR002190 MAGE protein Graphical View of Protein Domain Structure
Molecular reagents	All nucleic(627) cDNA(624) Primer pair(1) Other(2)
References	(Earliest) J:44483 Brady KP *et al.*, "Genetic mapping of 262 loci derived from expressed sequences in a murine interspecific cross using single-strand conformational polymorphism analysis." Genome Res 1997 Nov;7(11):1085-93 (Latest) J:111725 Williams ME *et al.*, "UNC5A promotes neuronal apoptosis during spinal cord development independent of netrin-1." Nat Neurosci 2006 Aug;9(8):996-8 All references(19)
Other accession IDs	MGD-MRK-35682, MGI:107217, MGI:1917220, MGI:1922987

Figure 14 : Aperçu d'un rapport détaillé MGD concernant le gène Maged1

http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerDetail&key=65381

- FlyBase

FlyBase (www.flybase.org/) est une base de données dédiée à la drosophile et la famille des insectes Drosophilidae. Le consortium Flybase annote, nettoie, intègre et maintient les données de cette BD secondaire spécifique à la drosophile. Les données accessibles sont

comparables à ceux de la BD MGD. Un numéro identifiant FlyBase ID permet la traçabilité de chacune des entrées. Les informations disponibles pour une entrée incluent les données d'annotation, la cartographie physique, les fonctions, les profils d'expression, les données phénotypiques sur les mutants, la bibliographie. Des ontologies spécifiques de la BD sont également disponibles. Elles annotent par exemple les séquences des gènes (SO : sequence ontology), l'anatomie (FBbt : fly_anatomy), le developpement (FBdv : fly_development), des schémas anatomiques (Fbbi : image ontology). FlyBase contient des données d'interactions avec les autres gènes cette information n'est par exemple pas disponible dans la BD MGD pour les gènes souris. La drosophile et le portail Flybase sont une référence pour l'étude et la compréhension des processus biologiques humain à travers cet organisme modèle très étudié.

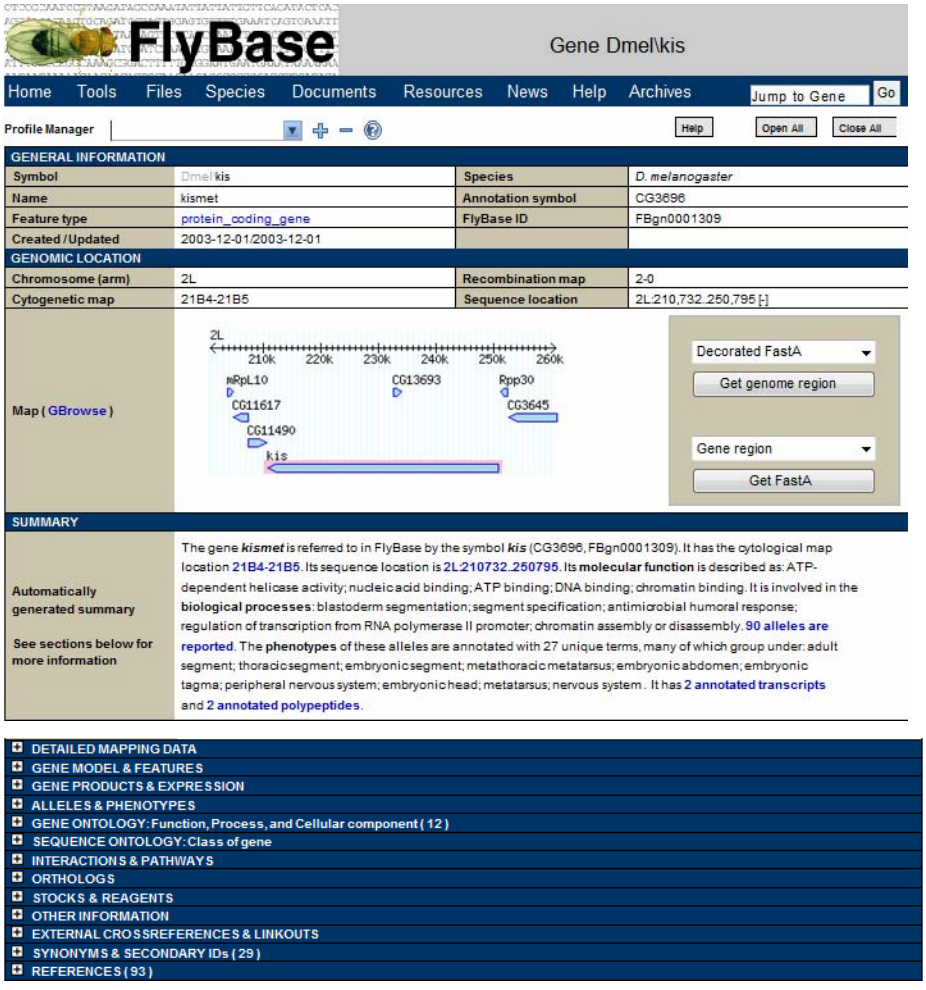


Figure 15 : Aperçu d'un rapport détaillé Flybase concernant le gène kismet

www.flybase.org/reports/FBgn0001309.html

- HomoloGene

HomoloGene (www.ncbi.nlm.nih.gov/sites/entrez/query.fcgi ?db=homologene) est un système pour la détection automatique d'homologues parmi les 18 espèces eucaryotes complètement séquencées, représentées notamment par *Homo sapiens, Pan troglodytes, Mus musculus, Rattus norvegicus, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Eremothecium gossypii, Neurospora crassa, Magnaporthe grisea, Arabidopsis thaliana et Oryza sativa* [167]. La procédure suit les similarités de séquences d'ADN lorsque les deux espèces sont proches d'un point de vue taxonomique et utilise la comparaison des séquences protéiques lorsque les espèces sont éloignées dans la taxonomie. Le rapport généré inclut notamment les homologies et les informations phénotypiques issues de OMIM, Mouse Genome Informatics [166], Zebrafish Information Network [168], Saccharomyces Genome Database [169], Clusters of Orthologous Groups (COG) [170] et FlyBase [171].

5.2.2.2. Les bases de données textuelles en biologie

Par opposition aux BD factuelles, les BD textuelles contiennent beaucoup de champs de text non structuré. OMIM et Medline représentent les plus connues d'entre elles.

- OMIM

OMIM est une BD de gènes et maladies génétiques, créé en 1966 par Dr. Victor A. McKusick de l'Université Johns Hopkins, Baltimore, Maryland (USA)[172]. Elle est accessible par l'interface du NCBI depuis 1987 [172]. Cette BD dérive de la littérature scientifique et les données contiennent des informations sur le phénotype de maladies et les gènes. La fiche détaillée est écrite sous la forme d'un texte descriptif avec le nom du gène, le mode d'hérédité, la localisation physique, les polymorphismes et une bibliographie détaillée. Des fiches résumées « clinical synopsis » et « gene map » sont aussi disponibles. Le premier est une description sous forme de mots clefs des phénotypes associés et le second présente la

localisation cytogénétique du gène parmi d'autres gènes de la région. Chaque entrée OMIM possède un numéro identifiant avec un préfixe servant à décrire la nature de l'entrée.

(*) Indique un gène de séquence connue

(#) Indique une entrée décrivant un phénotype pour lequel au moins un gène responsable est connu.

(+) Indique une entrée contenant une description d'un gène de séquence et phénotype connu

(%) Indique que l'entrée décrit un phénotype mendélien dont les bases moléculaires ne sont pas connues.

Une entrée sans préfixe indique une description phénotypique dont les bases mendéliennes n'ont pas été prouvées ou que cette entrée n'est pas clairement séparée d'une autre.

(^) Indique que l'entrée a été retirée de la BD.

Le tableau suivant représente les nombres d'entrées par préfixe [172].

	No. of Entries by Category				
Entry Classification	Autosomal	X Linked	Y Linked	Mitochondrial	Total
*** Gene with knouwn sequence**	10,644	495	48	37	11,224
+ Gene with knouwn sequence and phenotype	356	32	0	0	388
# Phenotype description, molecular basis knouwn	1,851	169	2	26	2,048
% Mendelian phenotype or locus, molecular basis unknown	1,411	134	4	0	1,550
Other, mainly phenotypes with suspected Mendelian basis	2,014	144	2	0	2,160
Total	16,276	974	56	63	17,370

Figure 16 : Nombre d'entrée dans OMIM en fonction du préfixe de l'identifiant

- Medline et Pubmed

Pubmed est le moteur de recherche sur la BD Medline. Cette BD est la référence en matière de littérature scientifique biomédicale. Elle inclut plus de 16.5 million de citations pour plus de 19 000 journaux de biologie [167]. Les enregistrements sont reliés les uns aux autres (« Related Articles ») sur la base d'une mesure de similarités d'indexation par les termes MeSH. Les termes MeSH indexent les entrées de Medline et donc peuvent être utilisés pour effectuer les recherches (voir plus loin). Comme pour les autres BD de NCBI, les requêtes peuvent être filtrées par des critères définis par l'utilisateur. Le système Pubmed fait partie de l'interface d'interrogation du NCBI (voir plus loin) qui permet de spécifier des requêtes très élaborées.

5.2.3. Interrogation des BD biologiques

Les sources dont nous disposons pour la recherche de gène candidats sont multiples, leurs contenus riches et variés, mais elles restent très hétérogènes. En effet, la mise en place d'une base de données n'est pas dictée par une règle stricte ou un code de conduite universel mais suit plutôt des choix personnels. Ceci conduit à une hétérogénéité du contenu, des formats, et de l'accessibilité rendant difficile la comparaison et l'utilisation des différentes sources de données par le bioinformaticien. Lincoln Stein a ainsi proposé un code de conduite à adopter lors de la création d'une base de données pour aboutir à « une nation bioinformatique ».

En pratique ce code de conduite n'est pas suivi. Malgré tout, l'interrogation des BD à été facilitée d'une part grâce à un effort pour uniformiser l'indexation par des vocabulaires contrôlés ou des ontologies, d'autre part grâce au développement d'interfaces d'interrogation uniformisées tel le portail Entrez au NCBI.

5.2.3.1. Vocabulaire d'indexation pour l'uniformisation des données

La mise en place de vocabulaires d'indexation (ou Ontologies) est donc rapidement devenue indispensable pour faciliter l'accès à la masse de données stockées dans les BD biologiques. Les ontologies ont permis d'unifier les annotations pour favoriser le partage et l'échange de données. On peut définir la notion d'ontologie comme un modèle conceptuel qui représente un ensemble de concepts dans un domaine et les rapports entre ces concepts référence. Elle est employée pour raisonner au sujet des objets dans ce domaine. Les concepts sont organisés dans un graphe dont les relations peuvent être :

- des relations sémantiques (synonymie) ;
- des relations de composition et d'héritage (au sens objet)

L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné. En particulier cela peut être un vocabulaire structuré et contrôlé. Afin de formaliser les différents domaines de la biologie, de nombreuses ontologies ont vu le jour.

Certaines sont génériques et s'appliquent à différents organismes, alors que d'autres sont plus spécifiques d'un domaine ou d'une espèce.

Le site OBO (Open Biomedical Ontologies, http ://obofoundry.org/) répertorie en un site Web, les ontologies publiques et les vocabulaires contrôlés couvrant les domaines de la biologie. On retrouve ainsi SO (Sequence Ontology) qui a pour but la description des séquences, EVOC (Expressed Sequence Annotation for Humans) qui vise à définir les séquences exprimées dans des conditions d'expériences données (plate forme de puces à ADN, mode de préparation des tissus, traitements …). Le groupe de travail MGED s'attache à standardiser l'annotation des expériences de puces à ADN [173]. Il coordonne également ses développements avec MAGE-ML qui vise à établir un format standard pour l'échange des données issues des expériences de puces à ADN [173], Citons encore le MESH pour l'indexation d'articles biomédicaux et GO pour l'annotation des gènes des organismes.

- MeSH

MeSH est l'acronyme de « Medical Subject Headings ». Le MeSH, produit par la NLM, est un vocabulaire contrôlé et structuré de termes utilisé pour l'analyse documentaire dans le domaine biomédical. Le vocabulaire MeSH est un outil d'indexation d'articles pour MEDLINE. L'indexation de la documentation biomédicale est ainsi effectuée de manière homogène et cohérente. Les descripteurs MeSH sont organisés de façon hiérarchique dans le MeSH Tree Structures (descripteurs en structures arborescentes). L'ontologie est organisée en 16 catégories. La catégorie des termes anatomiques est désignée par la lettre A, celle des organismes par la lettre B, celle des maladies par la lettre C etc… Chaque catégorie est divisée en sub-catégories. A l'intérieur de chaque catégorie les termes sont présentés de façon hiérarchique du plus général vers le plus spécifique par rapport aux articles indexés.

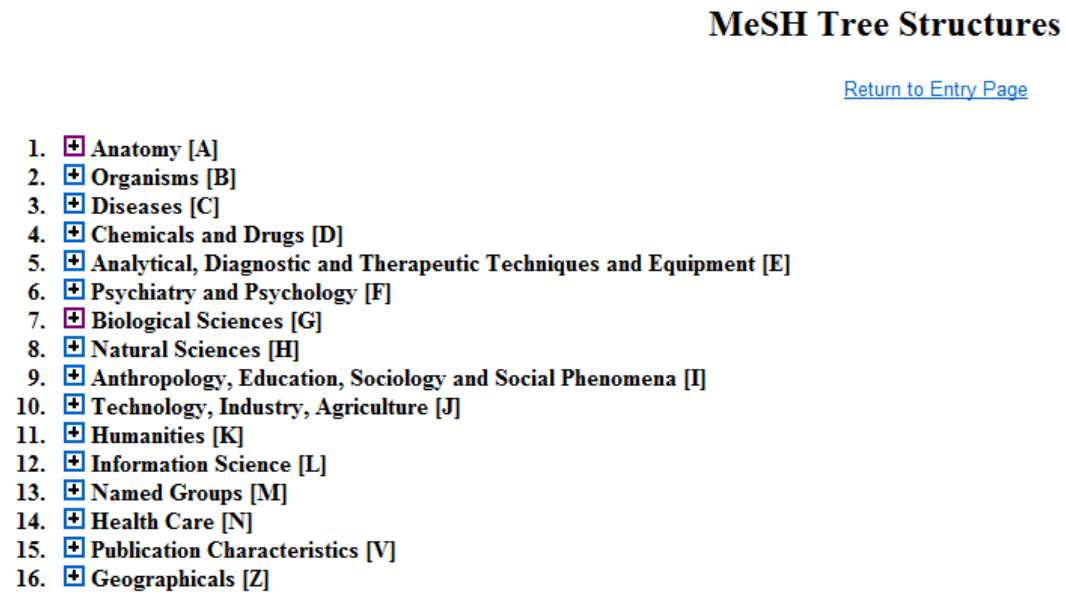


Figure 17 : Structure de l'arborescence du thesaurus MeSH

www.nlm.nih.gov/cgi/mesh/2007/MB_cgi

La recherche d'articles scientifiques en utilisant les termes MeSH devient ainsi extrêmement efficace. La BD MeSH contient notamment trois tutoriaux en format quicktime décrivant l'utilisation de l'interface pour la recherche d'articles dans Medline (http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=mesh). Typiquement, le mot-clé « microarray » dans Pubmed indique 21 179 articles scientifiques et les limites que l'on peut imposer à la requête sont : une recherche par auteur, par journal, par disponibilité de l'article (libre d'accès, résumé, …), par date de publication, par espèce (animal ou humain), par sous catégories (cancers, bioéthique…), par type de journal (lettre, revue, …), par âge (pour les études portant sur l'Homme). Si aucun de ces restrictions n'est pertinent pour l'utilisateur, il se retrouve devant l'obligation d'utiliser une combinaison de mots clefs. Or cette recherche risque d'entrainer une perte d'information. Le même mot-clé, Microarray dans l'interface MeSH donne trois termes MeSH : Microarray Analysis, Protein Array Analysis, Oligonucleotide Array Sequence Analysis. Dans chaque sous-catégorie des termes fils peuvent également être sélectionnés et une recherche avec cette sélection permet de préciser la requête de façon plus pertinente qu'un système de mot-clé librement choisis par l'utilisateur..

- Gene Ontology

Gene Ontology (GO,www.geneontology.org/) est un vocabulaire contrôlé développé par le GeneOntology Consortium, groupe de travail international basé à l'EBI, pour aider à l'annotation des génomes [174, 175]. Son objectif est d'établir un vocabulaire structuré, contrôlé et dynamique pour décrire le rôle des gènes et des produits de gènes de l'ensemble des eucaryotes.

Le vocabulaire GO se compose de trois ontologies qui définissent les processus biologiques, les fonctions moléculaires et la localisation cellulaire des produits de gènes. Le processus biologique est celui auquel un gène ou produit de gène participe (tel que la croissance cellulaire ou la transduction du signal). Un processus biologique est le résultat d'une ou plusieurs fonctions moléculaires associées dans un ordre donné. La fonction moléculaire décrit l'activité biochimique ou l'action du produit d'un gène (enzyme, transporteur, ligand). La localisation cellulaire présente l'endroit de la cellule où se trouve la forme active du produit d'un gène. GO permet d'exprimer les phénomènes de régulation et offre la possibilité de représenter des données incomplètes. Enfin, GO est un vocabulaire dynamique dont le nombre de termes évolue rapidement.

Chaque terme GO peut être un « enfant » de un ou plusieurs « parents ». En effet, la majorité des gènes est pléiotrope, c'est-à-dire qu'un gène peut avoir plusieurs produits et les produits d'un gène possèdent une ou plusieurs fonctions biochimiques. Le terme « enfant » est toujours plus spécifique que le ou les termes « parents ». La relation entre un enfant et son « parent» peut être du type « est unau-delà » (*is_a*), lorsque le terme enfant est une spécialisation du terme parent (Figure X). Elle peut aussi être de la forme « fait partie de » (*part_of*), si le terme enfant est un élément du parent. Si un terme à plusieurs « parents », il peut avoir différentes relations avec chacun de ses « parents ».

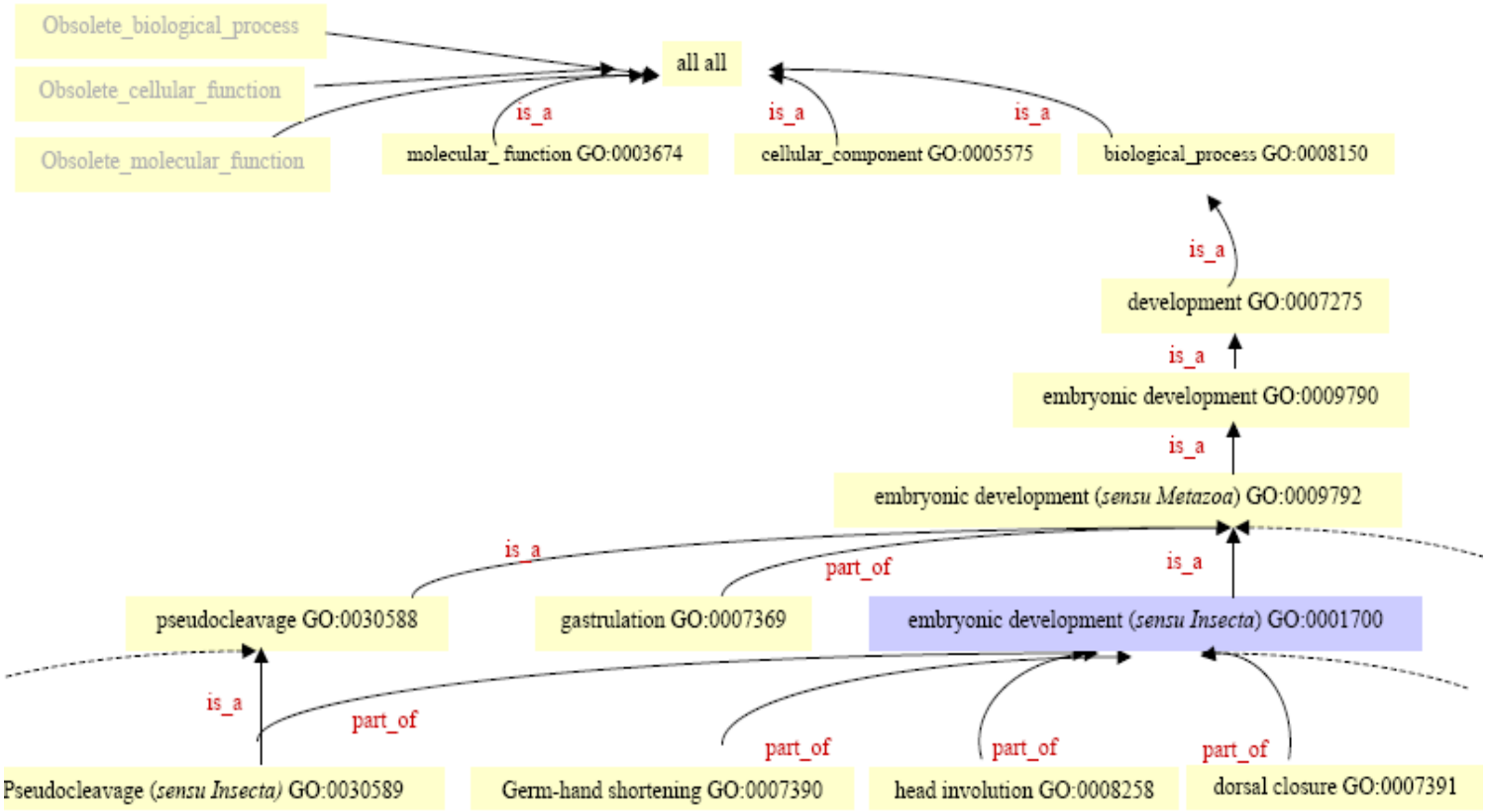


Figure 18 : Extrait du graphe de Gene Ontology.

Le graphe a pour origine les trois ontologies « molecular function », « cellular component », « biological process ». Les mentions « is_a » et « part_of » indiquent le type de relation entre les termes parents et enfants. Le terme « embryonic development (sensu Insecta) GO : 0001700» présente une limite de GO, à savoir la nécessité de définir des catégories spécifiques à certaines espèces. Les termes obsolètes sont matérialisés par les catégories « obsolete_molecular_function », «obsolete_cellular_component », « obsolete_biological_process».

La base de donnée GO n'est pas constituée des produits des gènes mais uniquement des termes et concepts (association de termes) qui les caractérisent. Dans le but d'annoter les génomes, des correspondances (mapping) entre les termes GO et les gènes ou produits de gènes sont proposées par les différentes bases de données associées au projet telles que Flybase, SGD , MGD ou GOA [176]. GOA (The Gene Ontology Annotation) est notamment un projet dirigé par EBI et le groupe SWISSPROT. Son but est de fournir, pour l'ensemble des organismes, une correspondance entre les termes GO et les produits de gènes (protéomes) référencés dans UniProt (http ://www.expasy.uniprot.org/). Les annotateurs indiquent par un symbole de 3 lettres la nature de l'origine de l'annotation telle que TAS pour « traceable author statement (tableau 4). L'avantage d'un vocabulaire contrôlé est de pouvoir être interprété par une machine. Par conséquent, l'annotation des gènes peut se faire automatiquement grâce à des outils de bio-informatique. Afin de visualiser les annotations, le consortium GO propose AmiGO, une application Web qui permet d'explorer les liaisons entre GO et les BD associées au projet. Les requêtes peuvent être faites à partir des termes GO ou des produits de gènes. Dans le même esprit, certaines grandes BD, comme EBI, proposent des outils de recherche des termes GO sur leur propre BD.

Symbole	Evidence de l'annotation
IMP	Inferred from mutant phenotype
IGI	Inferred from genetic interaction
IPI	Inferred from physical interaction
ISS	Inferred from sequence similarity
IDA	Inferred from direct assay
IEP	Inferred from expression pattem
IEA	Inferred from electronic
TAS	Traceable author statement
NAS	Non-traceable author statement
ND	No biological data available
IC	Inferred by curator

Tableau 4. Origines des annotations GeneOntology

www.geneontology.org

Une utilisation de GO est l'interprétation des données des puces transcriptomiques et la mise en évidence de catégories fonctionnelles significativement représentées dans des profils d'expression. Dans ce but, les termes GO associés aux gènes estimés « différentiels »

peuvent être comparés à l'ensemble des catégories fonctionnelles du génome étudié ou aux gènes présents sur la puce. Il existe de plus en plus d'outils pour interpréter les données de puces à ADN *via* le vocabulaire GO. Ces outils se distinguent par les données d'entrée et leur formatage, les organismes supportés, les méta-données utilisées (données d'expression, localisation chromosomique…), l'emploi ou non de statistiques et le type d'application [177, 178] . Les outils les plus connus sont certainement GOMiner [179, 180], Onto-Express [181] et Fatigo [182].GO offre la possibilité de caractériser et comparer la composition des puces à ADN en termes de catégories fonctionnelles. Draghici *et al.* (2003), grâce à leur outil OntoCompare, ont ainsi montré que différentes puces commerciales dédiées à une même question biologique ne recouvrent pas nécessairement les mêmes catégories GO. Aussi, selon la question biologique posée, le contenu d'une puce peut être plus pertinent qu'un autre.

5.2.3.2. Interface d'accès unifiée

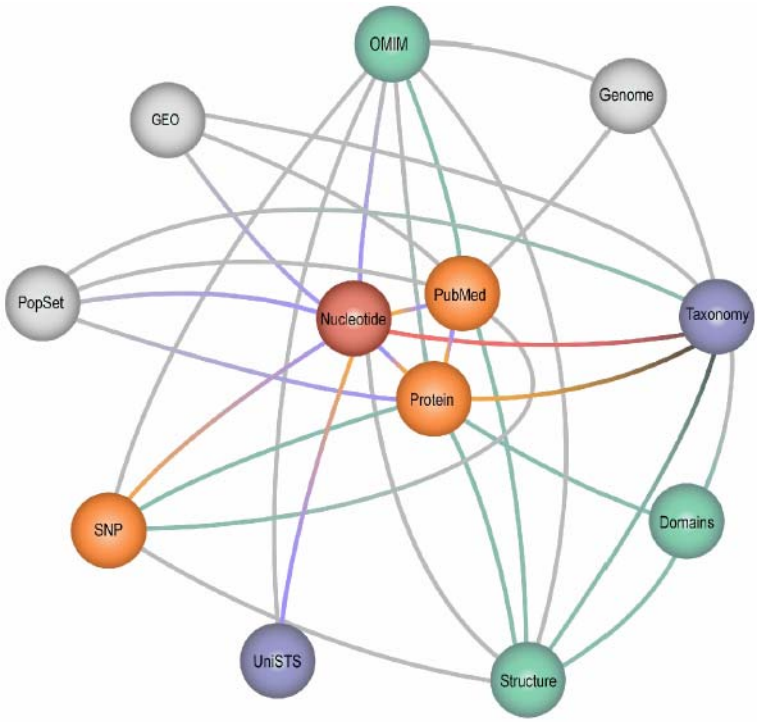
Si les centres de ressources sont très hétérogènes les uns par rapport aux autres, un effort est fait à l'intérieur de chacun d'eux pour uniformiser leurs contenus et formats. Des portails facilitent grandement l'accessibilité et la navigation à travers les bases de données. Le portail Entrez est un des plus utilisés dans le domaine de la genomique. Il intègre des données de nombreuses sources, formats et BD dans un modèle unique. La première version du portail a été distribuée en 1991 sur CD-ROM. A cette date, il permettait de naviguer/consulter les informations sur les séquences nucléotidiques issues de GenBank, les séquences protéiques issues de PDB (Protein Data Bank ; www.rcsb.org/pdb) en les associant aux articles publiés (Medline). Depuis les BD se sont multipliées et le portail permet aujourd'hui la consultation d'une vingtaine de BD (figure 8).

Figure 19 : Portail Entrez du NCBI

http ://www.ncbi.nlm.nih.gov/sites/gquery

L'organisation centrale du portail est basée sur la syntaxe des numéros identifiants de chaque entrée dans les BD du site : Gene_ID pour Entrez Gene,un numéro PMID pour Pubmed.... Un gène sera identifié par un Gene_ID unique et cet identifiant est utilisée lors de la recherche par le portail. En effet, les données telles que la cartographie du gène, sa séquence, ses annotations, sont sujettes aux changements alors que le numéro identifiant est plus stable dans le temps.

A)



B)

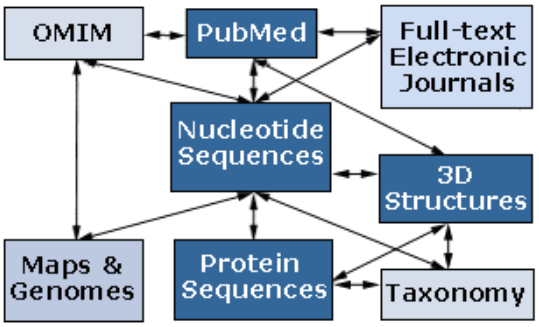


Figure 20 : Schéma des interconnexions sous jacentes à l'outil Entrez

A) Interconnection entre les BD du NCBI
B) Relation entre les types de données contenues dans les BD du NCBI

Les liens entre les bases de données du NCBI sont également disponibles dans les fiches détails des différentes BD. Par exemple, dans le rapport Gene du gène SUV39H1 il existe des liens (« Links ») vers toutes les BD du NCBI où une information sur le gène est disponible (Pubmed, GEO, GenBank…). Cette approche conduit également à créer de nouveaux liens entre les données des différentes BD. Ainsi, une entrée dans GenBank cite un article de

Pubmed mais dans cet article aucun lien n'est fait vers la séquence du gène jusqu'à ce que Entrez fasse le lien. Il est donc possible de lier des articles à un gène découvert a posteriori.

Entrez est également un outil de recherche élaborée. En plus des opérateurs logiques « AND », « OR », « NOT » des options existent, pour limiter une recherche (onglet « Limits »), avoir un aperçu du nombre de résultats (onlget « Preview »), retrouver l'historique des requêtes (onglet « History »), récupérer les résultats d'une recherche (onglet « Clipboard ») ou visualiser les détails des requêtes (onglet « Details »). Dans la BD Entrez Gene il est par exemple possible d'exclure les données concernant la mitochondrie (« exclude » et « mitochondria »), limiter la recherche aux données validées (« Limit by RefSeq Status » et « Validated ») et à l'espèce humaine (« Limit by Taxonomy » et « Homo sapiens »). Ainsi la page résultat débute par l'information : « Limits : No Mitochondria, Validated, Homo sapiens ».

5.3. Exploitation des bases de données biologiques

5.3.1. La problématique des gènes candidats

La découverte des gènes impliqués dans une maladie en génétique humaine passe classiquement par une approche positionnelle, au-delà par des études de liaison ou par la recherche d'anomalies de structure du génome chez les patients atteints, notamment avec les techniques de puces génomiques. Les approches fonctionnelles de gènes candidats sont également employées lorsqu'il est possible, et ceci est extrêmement rare, de faire une hypothèse sur la fonction *a priori* du gène candidat potentiel. Cette approche utilise souvent la recherche de domaines fonctionnels ou des tissus où s'exprime le gène. C'est de cette manière que le gène du retinitis pigmentosa a pu être identifié [183]. Les vastes ensembles de données accumulées au cours des dernières années contiennent vraisemblablement quantité de gènes candidats recherchés par les scientifiques, cependant ils sont dissimulés. La distribution des données sur le web facilite grandement l'accessibilité de l'information et ouvre la porte à de nouvelles méthodes de recherche de gènes candidats. Toute la difficulté du « gene discovery » à travers les stratégies de génomique intégrative

réside en l'exploitation efficace de ces informations disponibles. De nombreux outils bioinformatiques permettant une exploitation des BD ont vu le jour. Les données utilisées par ces outils proviennent de diverses BD cependant il est possible de regrouper les stratégies en 3 groupes reposant sur trois hypothèses différentes. Un premier groupe de logiciels utilise l'hypothèse selon laquelle les gènes impliqués dans les maladies possèderaient des spécificités propres par rapport aux gènes non impliqués dans les maladies. Un second groupe utilise une approche comparative qui consiste à rapprocher la maladie étudiée d'une autre maladie dont le gène responsable est connu. Dans cette hypothèse, les auteurs considèrent que si deux maladies sont proches d'un point de vue phénotypique alors les gènes responsables sont proches d'un point de vue fonctionnel. Ce groupe inclut les logiciels utilisant cette hypothèse pour prioriser les listes de gènes : la priorisation consistant à affecter un score d'intérêt aux gènes pour ensuite pouvoir les ranger selon cet ordre d'importance. Le dernier groupe de logiciels dans lequel s'inscrit ACGR est un groupe qui teste par des requêtes complexes diverses hypothèses concernant les gènes candidats qu'ils recherchent. C'est une recherche multicritère ou multi-hypothèses.

5.3.2. Approche globale : les gènes de maladies

Des études ont identifiés des tendances intéressantes caractérisant les gènes impliqués dans des maladies. Ces travaux comparent par des méthodes statistiques ou de fouille de données les caractéristiques des gènes connus pour être impliqués dans des maladies avec celles de gènes qui ne le sont pas. Une première étude basée sur l'analyse de 923 gènes de maladie a par exemple révélé que les gènes codant pour les facteurs de transcription étaient sur-représentés dans les maladies commençant *in utero* reflétant ainsi le rôle majeur des facteurs de transcription dans le développement. Ainsi, les gènes codant les facteurs de transcription comptent pour 30 % des gènes associées à des malformations [184]. Dans la même optique, une seconde étude postule qu'il est possible de corréler un certain groupe de gènes avec un type spécifique de maladie. Sur les 1647 gènes de maladie étudiés, ces auteurs montrent par exemple que les gènes codant les molécules structurales sont

fortement corrélés avec les maladies de la peau [185]. Les gènes de maladies coderaient des protéines de grande taille, et la majorité des gènes impliqués dans les maladies possèderaient des résidus conservés entre les espèces [185, 186]. Des logiciels de prédiction de gènes candidats ont utilisé cette approche des gènes de maladie [185-191]. Le logiciel PROSPECTR représente un des premiers exemples d'application. Il affecte un score à une liste de gènes en fonction des critères définissant les gènes de maladie. La structure (longueur du gène, du cDNA, de la protéine, du 3'UTR, du nombre d'exon), le contenu (ilôt CpG) et la conservation à travers l'évolution des gènes sont ainsi étudiés (identité de la protéine humaine et murine). Une limite à cette approche est que les gènes connus pour ne pas être impliqués dans les maladies peuvent potentiellement l'être. Puisqu'ils ont une fonction dans la cellule et que des mutations ponctuelles peuvent apparaître dans n'importe quel endroit du génome, délimiter des gènes de maladie et des gènes non impliqués dans les maladies pourrait tout simplement être un reflet de la limite de nos connaissances ou des techniques d'analyses. Citons l'exemple des SNP (« Single Nucleotide Polymorphism ») qui ont longtemps été considérés comme des variants neutres et qui par la suite se sont révélés importants dans les susceptibilités à certaines maladies. Il en va de même pour les CNV (« Copy Number Variant ») dont on ne suspectait pas l'existance il y a quelques années de cela.

5.3.3. Les approches comparatives

Il est assez intuitif de dire que deux maladies proches auraient *a priori* des gènes de fonctions similaire. Cette hypothèse suppose que les gènes de fonctions similaires sont impliqués dans des voies biologiques similaires. Les mutations qui affectent leur fonction entrainent donc des dérégulations dans les voies biologiques similaires provoquant les phénotypes proches. Un certain nombre d'observations serait en faveur de cette hypothèse. En ce qui concerne les pathologies neuronales, les gènes *OPHN1* (oligophrenin 1), *PAK3* (p21 (CDKN1A)-activated kinase 3), *ARHGEF6* (Rac/Cdc42 guanine nucleotide exchange factor 6) et *TM4SF2* (tetraspanin 7) impliqués dans des RMLX (retards mentaux liées à l'X) sont tous impliqués dans la régulation du cytosquelette d'actine [192]. Une revue sur les complexes neuronaux et les retards mentaux rapporte les gènes du protéome post-synaptique (PSP :

« post-synaptic proteome ») comme fréquemment impliqués dans des RMLX et les auteurs proposent ainsi que d'autres gènes sur le chromosome X liés au complexe PSP seraient donc des candidats potentiels pour les RMLX. Cependant, des observations ne vont pas dans le sens de cette hypothèse. Des mutations différentes dans un même gène peuvent conduire à des maladies très différentes. Certains gènes, comme le gène ARX (« aristaless-related homeobox »), montrent pour une même mutation, des phénotypes allant du retard mental isolé à des cas très sévères de syndrome de WEST.

Cette approche comparative est tout de même très répandue. Elle a été utilisée par de nombreux logiciels de prédiction de gènes candidats [193-205]. Ces logiciel fonctionnent pour la plupart sur une méthodologie visant à donner un « score d'importance/d'intérêt » aux gènes candidats par rapport à un groupe ou un gène référence connu. C'est ainsi qu'a été introduit dans le domaine, le terme anglais de « prioritization » que l'on traduira en français par priorisation. Une alternative à la priorisation est le regroupement (« clustering ») de gènes candidats par rapport à un set de gènes références. La similarité entre la liste candidate et le set référence peut être calculée par diverses approches statistiques et tenir compte des différentes propriétés des gènes comme la similarité de la séquence, les domaines protéiques, le tissus d'expression, les annotations GO. Les logiciels sont plus ou moins sophistiqués et font appel à des ressources différentes. Citons le plus complet d'entre eux le système Endeavour [201], qui utilise :

- des données de la littérature (synthèses d'articles Medline dans les fiches Entrez GENE),
- les annotations GO,
- les données d'expression issue des puces transcriptomiques (GNF),
- les données d'expression des ESTs (« expressed sequence tag ») (Ensembl),
- les domaines protéiques (InterPro),
- les données d'interactions protéines-protéine (BIND),
- les données d'interaction issues des voies biologiques (KEGG),
- les éléments Cis – régulateurs des gènes (TOUCAN),
- les motifs transcriptionnels (TRANSFAC),
- les similarités de séquences (BLAST).

Que ce soit pour la priorisation ou le regroupement, la plus grande difficulté de ces méthodes réside dans la définition du set de gènes références qui est déterminant pour la priorisation ou le regroupement de la liste candidate. Il arrive fréquemment que nous n'ayons aucun *a priori* sur la séquence et la fonction du gène candidat recherché, c'est le cas pour le syndrome d'Aicardi.

5.3.4. Approche par requêtes multicritères

Un troisième groupe de logiciels utilisent des requêtes (multi hypothèses) complexes pour définir un gène candidat. Ces méthodes ne sont pas basées sur les deux hypothèses précédentes mais tentent d'intégrer les données issues des BD pour donner un sens biologiques aux gènes sélectionnés. Elles permettent de tester un certain nombre d'hypothèses pouvant définir le gène candidat recherché. GeneSeeker [206], GeneSorter [207] sont deux systems de ce groupe de logiciels. Ils utilisent diverses possibilités de requêtes définissant le gène candidat recherché. GeneSeeker est basé sur l'intégration des données de cartographie, d'expression (puces transcriptomiques). Ainsi peuvent être effectuées des requêtes telles que : quels sont les gènes de la région Xp22, exprimés dans le cerveau ?. GeneSorter, intègre en plus les annotations GO et les similarités entre les protéines. Un système présenté par Tiffin *et al.* (2005) utilise une ontologie eVOC annotant les EST (Ensembl annotation) pour rechercher les gènes exprimés dans le tissu affecté par la maladie [208]. Cette troisième catégorie d'approches semble être la plus adaptée pour la recherche de gènes candidats pour les maladies complexes à étudier où aucun a priori sur la fonction du gène ni rapprochement avec un groupe de gènes n'est possible. Les solutions proposées sont néanmoins restées décevantes dans le cas du syndrome d'Aicardi. Une approche multicritère reposant sur de nouvelles définitions de gènes candidats a donc été envisagée.

0bjectifs

Objectifs de la thèse

Ce travail de thèse a été orienté vers la recherche de gènes candidats impliqués dans la survenue du syndrome d'Aicardi. Les approches mises en œuvre ont pour but de trouver le gène impliqué dans ce syndrome. Cependant, la compréhension de la physiopathologie passe également par l'étude des réseaux biologiques impliqués dans une maladie. Ainsi, l'étude des fonctions biologiques affectées a aussi été entreprise.

Notre stratégie a utilisé une approche à l'aide de puces génomique spécifiques de l'X (provenant du Flanders Interuniversity Institute of Human Genetics) pour un cibler l'ADN de 18 patientes AIC à la recherche d'un déséquilibre quantitatif. Prolongation logique des études du caryotype, l'analyse par CGH à l'aide de puces à ADN permet un criblage plus résolutif des anomalies quantitatives. Première étape du clonage positionnel, elle permet de se focaliser sur une région chromosomique précise. Bien évidemment, sa réussite dépend fortement du nombre de cas étudiés.

Par ailleurs, nous avons élaboré une approche transcriptomique visant l'exploration des voies biologiques dérégulées. Deux études différentes ont été menées, une sur l'ARN extrait de lignées lymphoblastiques de trois patientes et à l'aide de puce à oligonucléotides recouvrant environ 18 000 gènes. Une seconde étude à partir d'ARN extrait de sang de 10 patientes AIC et à l'aide d'une puce criblant l'ensemble du transcriptome. L'analyse des données a été effectuée avec Luminator (Rosetta), un serveur mettant à disposition des outils d'annotation et des outils statistiques (ANOVA : ANalyse Of Variance) pour la gestion des données de transcritomique. EASE (Expression Analysis Systematic Explorer) a été mis à profit pour les études de groupement fonctionnel.

Une approche *in silico* a également été utilisée. Elle aide à l'intégration des données transcriptomiques et propose également une approche de recherche de gènes candidats par l'exploitation des bases de données biologiques publiques. Les bases de données dédiées à la souris et la drosophile, organismes modèles pour l'étude des maladies humaines ont ainsi été exploitées pour enrichir les données récoltées dans les bases de données dédiées aux gènes humains. Ainsi, un logiciel baptisé ACGR a été conçu. Il est organisé autour d'une base

de données développée pour stocker des informations sur les gènes susceptibles d'être candidats. Un module de classement des gènes selon des critères de préférence de l'utilisateur est intégré au logiciel afin de présenter à l'utilisateur une liste de gènes à tester en priorité.

Matériel et Méthodes

1. Matériel biologique

Le matériel génétique étudié a été extrait à partir de prélèvements sanguins ou à partir de lignées de cellules lymphoblastiques (étude AS1). Pour chaque patient prélevé, un consentement libre et éclairé est signé pour permettre l'utilisation de l'ADN et/ou de l'ARN à des fins de recherche. Pour les filles Aicardi, le consentement est signé par les parents.

2. Culture cellulaire

Ces lignées continues sont établies à partir de sang veineux prélevé sur héparine-lithium.
La transformation lymphoblastique est réalisée sous l'action du virus Epstein-Barr [209]. Les cellules sont conservées dans l'azote liquide à –196°C dans un milieu de congélation contenant 20 % de sérum de veau fœtal (SVF), 10 % de DMSO (dimétylsulfoxyde). Les lymphoblastes sont cultivés dans du milieu RPMI 1640 avec 20 % SVF en présence d'antibiotiques- d'antifongiques (GIBCO™).

3. Techniques de cytogénétiques

3.1. Préparation des chromosomes métaphasiques

Les lignées lymphoblastoïdes sont décongelées et remises en culture jusqu'à obtention d'amas de cellules en suspension dont la qualité est appréciée à l'aide d'un microscope inversé.

3.1.1. Colchicine et blocage en métaphase

La technique cytogénétique est réalisée à partir de 7 ml de culture de lymphoblastes. Les cellules sont bloquées en métaphase par l'action de 50 µl de colchicine (20 mg/ml) pendant 50 minutes. Parallèlement à l'action de l'agent anti-mitotique et afin d'améliorer la résolution des étalements, 200 µl de bromure d'étidium (BET) (0,7 mg/ml) sont ajoutés.

3.1.2. Choc hypotonique

Après centrifugation (1200 t/min, 5 min) de la suspension de cellules, le surnageant est éliminé et le culot remis soigneusement en suspension et homogénéisé dans 1 ml de solution hypotonique (KCL à 0.075 M). Le volume total est complété à 12 ml de solution hypotonique et le tube placé à 37°C pendant 17 min. Ce choc hypotonique permet le gonflement des noyaux, la lyse de la membrane nucléaire et la dispersion optimale des chromosomes.

3.1.3. Fixation

Une préfixation avec 1 ml de fixateur (3 volumes d'éthanol absolu pour 1 volume d'acide acétique) préparé extemporanément bloque l'action du choc hypotonique. On récupère le culot cellulaire par centrifugation (1200 t/min, 5 min). Suivent 3 lavages avec le fixateur (12 ml de fixateur, 1200 t/min, 5 min). Ils permettent l'élimination d'un maximum de cytoplasme. Les tubes sont conservés au moins 20 min à 4 °C jusqu'à l'étalement.

3.1.4. Etalement

Après centrifugation (1200 t/min, 5 min) et élimination du surnageant, le culot est remis en suspension dans un volume de fixateur variable en fonction de la densité cellulaire souhaitée à l'étalement. Une à deux gouttes de suspension cellulaire sont déposées sur des lames Superfrost™ dégraissées. L'étalement est réalisé dans une enceinte close régulée au niveau de la température (20 °C) et du degré d'hygrométrie (40 %).

A cette étape, les lames sont soit traitées pour la réalisation d'un caryotype, soit pour les techniques de cytogénétique moléculaire. Dans ce cas, les lames sont vieillies pendant 24 à

48 h à l'abri de la lumière, puis déshydratées dans trois bains d'éthanol de concentration croissante (70 %, 80 %, 100 %), 1 min dans chaque bain. Les lames peuvent alors être conservées à -20 °C pendant 6 mois.

3.2. Réalisation du caryotype en bandes GTG (bandes G, Trypsine, Giemsa)

Il est possible d'identifier chaque paire chromosomique à l'aide du marquage chromosomique ou « *banding* ». Le marquage résulte d'une dénaturation physico-chimique ou enzymatique plus ou moins poussée de la structure du chromosome. Il met en évidence une alternance de bandes claires et sombres délimitant ainsi des régions et des sous-régions caractéristiques de chaque chromosome, rendant plus précise l'identification de segments chromosomiques.

Les bandes G sont le reflet de l'organisation physique des chromosomes mitotiques, dont les protéines constitutives, plus ou moins sensibles à l'action de la trypsine, révèlent une alternance de bandes claires et sombres après coloration.

Les lames étalées sont vieillies à 65°C pendant une nuit. Elles sont ensuite incubées dans la solution de trypsine à 37°C pendant 10 sec, ce temps étant ajusté en fonction du degré de digestion obtenu sur une lame d'essai. Les lames sont rincées quelques secondes dans deux bains de PBS 1X (phosphate buffered saline, NaCl 170 mM, KCl 3 mM, Na_2HPO_4 10 mM, KH_2PO_4 1,8 mM, pH 7,3). Elles sont ensuite colorées dans une solution contenant 9 ml de Giemsa (Biolyon™), 9 ml d'acide citrique (3 M) et 9 ml de méthanol (1 M). Le pH est ajusté à 6,7 avec du Na_2HPO_4 (2,84 g/l). Le colorant contient des thiazines qui se lient aux groupements phosphates de l'ADN. Après un rinçage rapide à l'eau, les lames sont séchées et analysées. L'établissement du caryotype est réalisé à l'aide du logiciel Quips® Lab manager (Vysis™).

4. Techniques de biologie moléculaire

4.1. Extraction d'ADN génomique

Elle est réalisée à l'aide du kit Nucléon BACC3 (Amersham™) selon les instructions du fournisseur. Après une précipitation par l'éthanol, l'ADN est resuspendu dans du tampon TE (Tris-HCl 10 mM pH8, EDTA 1 mM pH 8).

4.2. Extraction des ARN totaux

4.2.1. A partir de lignées lymphoblastiques

L'extraction des l'ARN totaux des lymphoblastes est effectuée le lendemain d'un changement de milieu pour être en phase de croissance exponentielle.

Les cellules sont comptées pour être en nombre limité (2 à 3.10^6 cellules) et ne pas saturer la colonne Qiagen™ (Rneasy kit). Les cellules sont ensuite lavées dans 10 ml de « phosphate buffer saline » (PBS) afin d'éliminer les protéines et le milieu de culture inhibiteur de l'extraction.

Le principe du kit Qiagen Rneasy repose sur les propriétés d'un gel de silice. Les cellules sont lysées et homogénéisées en présence d'un tampon hautement dénaturant qui permet d'éliminer les protéines, l'ADN, les protéases et tout ce qui pourrait interagir lors de la synthèse de l'ADN complémentaire (ADNc) par reverse transcriptase (RT) suivi d'une réaction de polymérisation en chaîne (PCR). L'ARN est ensuite élué dans de l'eau milliQ sans Rnase et conservés à – 80°C.

4.2.2. A partir de sang

Les prélèvements de sang ont été effectués dans des tubes PAXGene (PreAnalytiX) selon le protocole du fournisseur [210, 211]. Les ARN ont été extraits avec le kit d'extraction PAXGene blood RNA isolation kit (PreAnalytiX™) du même fournisseur, toujours en accord

avec le protocole fournit. Les ARN ont été élués dans 30 µl d'eau Rnase-free, puis contrôlés à l'aide du Bioanalyzer Agilent (Agilent Technologies Au-delà) et conservés à – 80°C.

4.3. Mesure de concentration de l'ADN et de l'ARN

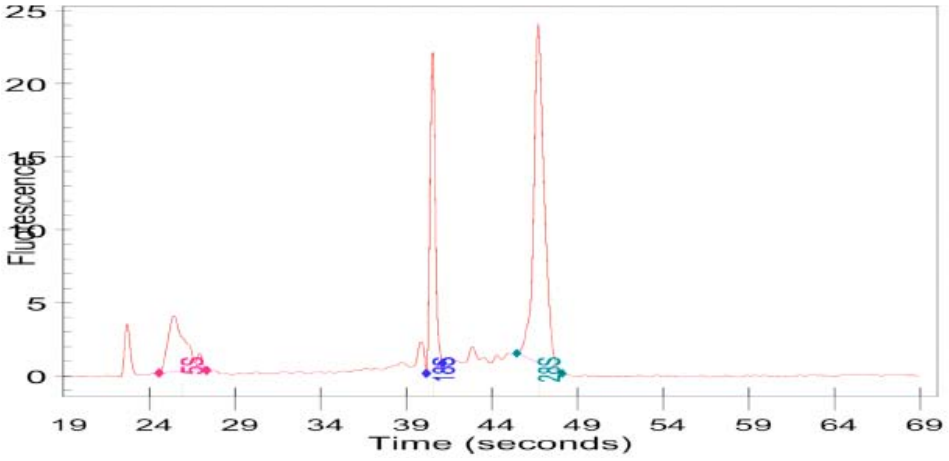
Le Nanodrop®, permet de déterminer les concentrations en acide nucléiques (ARN et ADN) après extraction en utilisant 2 µL d'échantillon. Une fois l'échantillon mis en place sur la colonne de mesure, il est maintenu grâce à la tension de surface. Une colonne de mesure composée de deux fibres optiques permet de mesurer la densité optique entre 220 et 350 nm en moins de dix secondes, la courbe Absorbance = f(λnm) peut alors être visualisée sur l'ordinateur.

4.4. Contrôle de la qualité des ARN

Le 2100 bioanalyzer (Agilent) a servit pour le contrôle de qualité des ARN. Le kit RNA 6000 Nano Assay (Agilent) à été utilisé. Cette technique utilise la technologie « lab-on-a-chip » (laboratoire sur puce). La puce en question est composée de plusieurs puits, dont 12 destinés aux échantillons, et 1 au marqueur de poids moléculaire (PM). Au niveau de ces puits, il existe des microréseaux. On dépose d'abord un gel filtré mélangé à un flurochrome (intercalant de l'ARN spécifique et fluorescent). Celui-ci est réparti équitablement au niveau de chaque puits, et remplit les microréseaux. Un tampon est alors ajouté dans chaque puits, afin de placer ensuite les échantillons (et le marqueur de poids moléculaire (ladder) dans un milieu stable, à un pH donné. La puce est vortexée afin d'homogénéiser tampon et échantillon, puis placée dans l'appareil. Les puits sont analysés un par un, en commençant par le marqueur de PM, qui sert en fait d'étalon, aux niveaux qualitatif et quantitatif. Une électrode est en contact avec chaque puits. Au moment de l'analyse, cette électrode est soumise à un haut voltage, qui va permettre une séparation très rapide des ARN de l'échantillon en fonction de leur taille (passage à travers le gel, au sein des microréseaux plus ou moins rapide). Au moment de leur passage dans le gel, les molécules d'ARN vont se lier spécifiquement au flurochrome. En fin de migration, les molécules sont détectées par fluorescence. Les molécules sont excitées par un rayonnement UV, et seuls les ARN (liés au

dye) vont réémettre un rayonnement fluorescent. L'appareil mesure, en temps réel, l'intensité de la fluorescence émise en fonction du temps de rétention de l'échantillon au sein du microréseau. Plus les ARN sont courts, plus ils migrent rapidement et donc plus leur temps de rétention est faible. Plus l'intensité de fluorescence est importante, plus la quantité d'ARN est importante. Ces données qualitatives (temps de rétention) et quantitatives (intensité de fluorescence) sont validées par rapport au marqueur de poids moléculaire qui est connu très précisément par le logiciel. Le logiciel d'analyse identifie les ARN 18S et 28S L'intégration des pics 18S et 28S conditionne le rapport 28S/18S qui est l'un des critères de contrôle qualité des ARN. Les figure suivantes représentent les profils « normal » (a) et « dégradé » (b) d'ARN.

a)



b)

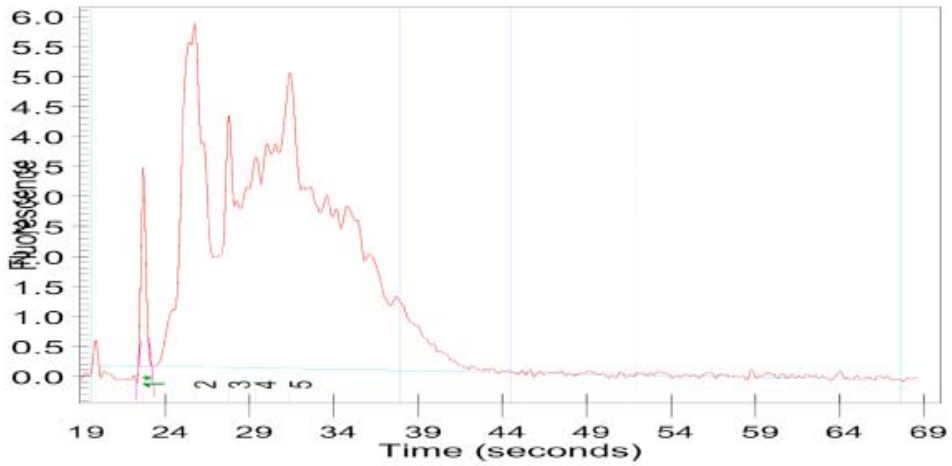


Figure 21 : profil des ARNc marqués par le 2100 bioanalyzer

4.5. PCR (polymerase chain reaction)

La composition du mélange réactionnel est la suivante, dans un volume final de 25 µl :

- 100 ng d'ADN génomique total,
- un mélange contenant les amorces sens et anti-sens (10 µM chacune), les 4 dNTP (250 µM),
- $MgCl_2$ (de 0.75 à 4 mM),
- 2 U d'enzyme Taq (*thermophilus aquaticus*) polymérase,
- tampon de réaction 1X (spécifique de l'enzyme utilisée).

Le choix des amorces utilisées et les conditions précises de chaque PCR seront détaillés ultérieurement. La qualité et la quantité du produit de PCR sont vérifiées par électrophorèse sur gel d'agarose à 0,8 % dans du tampon TBE 1X (10X : Tris borate 0.9 mM, EDTA 20 mM, pH 8,2) en présence de 0.5 mg/ml de bromure d'étidium (BET : agent intercalant de l'ADN double brin). La taille des produits de PCR est estimée par comparaison avec un marqueur de taille approprié (Roche Diagnostics™).

Les conditions initiales de PCR précédemment présentées sont mises au point pour chaque couple d'amorce, en faisant varier tout d'abord la température d'hybridation des amorces puis, si nécessaire, la concentration en $MgCl_2$. Les conditions d'amplification utilisées sont les suivantes : une dénaturation initiale de l'ADN génomique total pendant 5 min à 95 °C, suivie de 30 cycles comportant chacun, une étape de dénaturation à 94 °C, une étape d'hybridation des amorces à une température spécifique, et une étape d'élongation à 72 °C. La durée de chacune de ces étapes a été fixée à 30 sec. La PCR se termine par une extension finale à 72 °C pendant 10 min pour terminer la synthèse des brins néo-synthétisés.

4.6. RT-PCR

La transcription inverse est réalisée à partir d'1 µg d'ARN totaux à l'aide du kit Reverse-Transcriptase Superscript II (Gibco-Brl) en utilisant 500 ng d'hexanucléotides selon le protocole du fabricant.

4.7. PCR quantitative en temps réel

La Q-PCR SYBR® a servit pour la vérification des résultats des puces génomiques alors que la technologie TaqMan® a été utilisée pour valider les études d'expression. Les résultats on été analysés par le logiciel SDS (Applied Biosystems) par la méthode des ΔΔCt.

4.7.1. SYBR® green et validation des résultats de puce génomique

La validation par SYBR® green a été utilisée pour la validation des résultats de puces génomiques. C'est une molécule qui a la propriété de s'intercaler entre les paires de bases d'un ADN double brin et d'être fluorescente une fois intercalée. Ces propriétés permettent de mesurer la quantité d'ADN double brin en temps réel à la fin de chaque cycle d'amplification de la Q-PCR. La PCR en temps réel à parti d'ADNc permet la quantification d'un transcrit cible dans un échantillon grâce à la mesure de Ct (« Cycle threshold »), qui sont inversement proportionnels à la quantité initiale de cible présentes dans cet échantillon. Nous avons opté pour une quantification relative en utilisant un standard interne ou endogène (gène *WWOX* sur le chromosome 16).

- Choix des amorces pour la Q-PCR par SYBR® green

Les amorces PCR ont été sélectionnées à l'aide du logiciel Primer express (Applied Biosystems). Pour chaque clone, au moins 3 couples d'amorces ont été choisies (tableau X). Ainsi les clones RP11-388L20, RP5-1178I21, RP11-66N11, RP11-54I20, RP11-441L6, RP11-142K4, RP5-1000K24, RP11-97N5 et CTD-2511C7 ont été testé. Un couple d'amorce témoin a été sélectionnée dans les régions unique du gène *WWOX.*

- Contrôle de l'efficacité et de la spécificité de la PCR

Pour pouvoir réalisé une quantification relative basée sur la méthode d'analyse des DDCt, il est nécessaire de contrôler au préalable l'efficacité et la spécificité des différentes PCR pour les transcrits cibles et les transcrits des endogènes, en réalisant une plaque d'efficacité pour

construire une courbe d'étalonnage à l'aide de cinq dilutions en série. Nous avons utilisé les conditions de PCR suivantes, dans un volume final de 25 µl : 5 µl d'ADNc totaux purs ou dilués, 0.4 mM du couple d'amorces spécifique du transcrit à amplifier, 12,5 µL de solution *Power* MasterMix SYBR Green contenant l'enzyme AmpliTaq Gold DNA Polymerase LD (Applied Biosystems™). La quantité d'ADNc apportée variait selon une dilution en série de manière à obtenir 5 quantités différentes, correspondant à 125 ng ; 12,5 ng ; 1,25 ng ; 0,25 ng et 0,125 ng d'ARN totaux rétrotranscrits. Après une activation enzymatique de 10 min à 50°C et une dénaturation de 10 min à 95°C, 40 cycles ont été réalisés comprenant chacun 15 sec à 95°C et 1 min à 60°C. Le couple d'amorce choisi pour chaque transcrit doit être spécifique, c'est-à-dire que lors de la PCR seul le transcrit d'intérêt doit être amplifié. Pour vérifier ce paramètre nous avons réalisé une étape de dissociation à la fin de la PCR (15 sec à 95°C, refroidissement rapide jusqu'à 60°C, puis ascension progressive lente de la température jusqu'à 95°C), permettant d'obtenir une courbe de fusion du/des produit(s) de PCR. Cette courbe correspond à la dérivée de l'équation y= f(x), où y est la quantité de fluorescence émise par le SYBR Green et x la température des produits de PCR analysés. L'amplification est spécifique si cette courbe ne comporte qu'un seul pic. La température pour laquelle la valeur –d(Fluorescence)/dT s'annule correspond au Tm de l'amplicon.

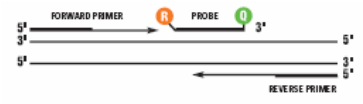
- Q-PCR par SYBR® green

Le milieu réactionnel (25 µl) est composé de 12,5 µl de Power SYBR® Green PCR Master Mix (Applied Biosystems™), de 0,4 µM de chacun des 2 amorces et de 10 à 50 ng d'ADN génomique total. Les réactions de PCR ont été réalisées à l'aide de l'appareil 7500 Fast Real-Time PCR System (Applied Biosystem™), et avec une température d'activation de 2 min à 50°C, une température initiale de dénaturation de 95°C pendant 10 min, 40 cycles d'amplification composés d'une étape de dénaturation de 15 s à 95°C, d'une étape d'hybridation et d'élongation de 1min à 72°C. La spécificité des fragments est vérifiée par un cycle de dénaturation, renaturation, dénaturation.

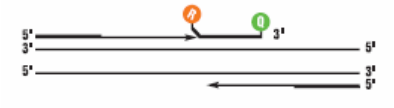
4.7.2. TaqMan® et validation des résultats de puce transcriptomiques

Il s'agit d'une sonde sur laquelle sont liés, aux extrémités 5' et 3', un agent fluorescent et un quencher qui l'empêche d'émettre sa fluorescence.

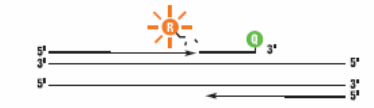
Polymérisation : un fluorochrome AU-DELA et un quencher (Q) sont attachés respectivement à l'extrémité 5' et 3' de la sonde (probe)



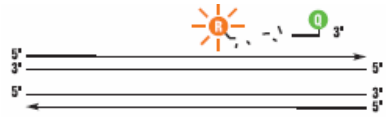
Déplacement du brin : quand la sonde est intacte, le quencher empêche la sonde d'émettre de la lumière



Clivage : durant l'extension, l'ADN polymérase par son activité exonucléase, dégrade la sonde qui se retrouv ainsi séparé de son quencher



Fin de la polymérisation : une fois la polymérisation terminée, le fluorochrome émet une fluorescence caractéristique.



La sonde (Taqman Applied Biosystems) se fixe à l'ADNc d'intérêt si celui-ci est présent dans notre échantillon. A ce moment, aucune fluorescence n'est observée. Le quencher n'est relargué par la sonde que lorsque le fragment d'ADNc d'intérêt est amplifié. La polymérase, permettant la synthèse du brin, la sonde qui rompt sa liaison à l'ADN, séparant ainsi le quencher et le fluorochrome qui émet alors de la lumière.

L'appareil utilisé est un système Applied Biosystems™ 7300 qui possède un capteur CCD détectant l'émission fluorescente à l'aide de filtres spécifiques. Un algorithme recueille alors l'information et permet le traitement des résultats.

Nous nous intéresserons dans nos expériences à la valeur de Ct (threshold cycle) qui est proportionnel à la quantité d'ADN dans l'échantillon. Le Ct correspond au nombre de cycles à partir duquel la fluorescence émise sera supérieure à un seuil de détection. Plus la quantité d'ADN initiale est importante dans l'échantillon testé et plus le Ct sera faible. Nous avons opté pour une quantification relative en utilisant un standard interne ou endogène. Une référence endogène dont l'expression est stable, est utilisée afin de rendre comparable l'expression entre plusieurs extraits différents. Le gène codant l'ARN 18S s'avère être un gène fortement exprimé dans les cellules, et ceci de façon sensiblement équivalente entre-elles. Il a servit dans l'étude AS1 à la comparaison d'expression des gènes entre les cellules.

Dans les expériences AS1, la réaction de Q-PCR est répétée trois fois. L'ADNc de départ est utilisé à une concentration comprise entre 50 et 500 ng/µL, avec les réactifs préconisés par Applied Biosystems. L'étude des 16 gènes sélectionnées pour l'étude AS1 (ASMT, CXCR3, CTAG1, FLJ21174, ITM2A, KIAA1280, MST4, NGFRAP1, NSBP1, PIR, PLXNB3, PORCN, PRDX4, RPS4X, SSR4, SYN1) a été effectués grâce aux sondes Taqman® Gene Expression Assays et du Taqman Universal PCR Master Mix (Applied Biosystems™) selon le protocole du fournisseur. Chaque étude possède une référence endogène (ARN 18S). Ainsi, un ratio d'expression RQ correspondant à l'expression relative du gène étudié chez la fille AIC par rapport à sa mère est donné pour chaque gène étudié.

4.8. Séquençage

Le principe de la réaction de séquence est basé sur la technique de Sanger (1977). Le séquençage est effectué sur une machine ABI PRISM 3100® (Applied Biosystems™).
Les produits de PCR ont été purifiés par l'automate TECAN Genesis RSP 100 avec le PCR_{96} Cleanup Plates (Millipores). L'ADN double brin est séquencé selon la méthode de Sanger en utilisant le kit BigDye® Terminator V1.1 (Applied Biosystems™). Ce protocole permet de réaliser une réaction unique, chacun des 4 didésoxynucléotides (ddNTP) étant lié de façon covalente à un fluorochrome différent. Le mélange réactionnel (5 µl) contient 2,5 µl de produits de PCR, 1 µl d'amorces à 5 pM, 0,5 µl de tampon et 1 µl du mélange « Big Dyes » (Taq polymérase, dNTP et ddNTP liés aux fluorochromes). Trente cycles de PCR ont été réalisés avec la dénaturation à 95°C pendant 10 s, l'hybridation à 50 °C pendant 5 s et

l'élongation à 60 °C pendant 4 min. Les produits de PCR sont purifiés sur une résine Sephadex™ G-50 Superfine. La séparation par électrophorèse capillaire et la détection des pics de fluorescence ont été réalisées par l'automate Abi PRISM 3100 Genetic Analyser (Applied Biosystems™). Chaque produit de PCR a été séquencé sur les deux brins.

Les variations de séquence seront recherchées chez les filles Aicardi par comparaison à une séquence normale du gène. Dans le cas d'une présence de variation dans la séquence, les bases de données publiques sont interrogées pour savoir s'il s'agit d'un variant polymorphe normal (http ://www.ncbi.nlm.nih.gov/SNP/). En sachant que les mutations dans le syndrome d'Aicardi sont sporadiques, pour être considéré délétère, le variant de séquence ne doit pas être retrouvé chez les parents. Les premières études de séquençage ont été réalisées à la suite de l'étude AS1 et donc sur trois patientes AIC. Cinq gènes (*ASMT*, *PLXNB3*, *MST4*, *SYN1*, et *NSBP1)* suivant ont été séquencés. Chaque gène est recherché sur le site Genome Browser (http ://genome.ucsc.edu/). Sur la fiche de chaque gène figure un lien direct vers le logiciel ExonPrimer (http ://ihg.gsf.de/ihg/ExonPrimer.html) qui propose les oligonucléotides à utiliser pour l'amplification par PCR. Le logiciel nommé Exon Primer a été utilisé pour le choix des amorces PCR. Il propose les couples d'oligonucléotides en tenant compte d'un certain nombre de paramètres dont notamment la température d'hybridation, le pourcentage en G-C et la taille des amorces. Le logiciel requiert la séquence de l'ADNc et de la séquence d'interêt. Ceci permet au logiciel de rechercher les amorces des segments d'ADN qui recouvrent tous les exons.

4.9. Inactivation du chromosome X

L'étude du profil de l'inactivation du chromosome X chez la femme a été réalisée au locus HUMARA (Human Androgen Receptor A) en Xq13 (Allen et al., 1992). Un microlitre d'ADN génomique (100ng/µl) est digéré par 2 µl d'enzyme *Hpa*II (10 U/µl) pendant 2 heures dans un volume réactionnel de 50 µl. Au bout de 2 heures à 37°C, 2µl d'enzyme *Hpa*II sont additionnés pour une digestion d'une durée au moins équivalente. L'enzyme *Hpa*II est sensible à la méthylation, elle ne digère l'ADN que si la séquence de son site n'est pas méthylée. Une fois la digestion terminée, l'enzyme est inactivée 5 min à 95 °C et l'ADN est purifié sur colonne (kit Wizard® DNA Clean-Up system, Amersham™). Deux microlitres d'ADN génomique digéré puis purifié et 200 ng d'ADN génomique non digéré sont amplifiés

en parallèle par PCR fluorescente à l'aide d'amorces spécifiques du locus HUMARA, dans les conditions suivantes : tampon de réaction 1X, $MgCl_2$ 1,5 mM, dNTP 200 µM, amorces 0,2 µM et 2,5 U de Taq polymérase. Après une dénaturation de 5 min à 95 °C, 35 cycles sont effectués comprenant chacun : 30 sec à 95 °C, 30 sec à 59 °C et 45 sec à 72 °C suivis d'une élongation à 72 °C pendant 7 min. Les qualités et quantités des produits d'amplification sont vérifiées sur un gel d'agarose de 2 %. Des bandes de 250-300 pb sont attendues. Selon leur intensité, les produits de PCR sont plus ou moins dilués avant d'être séparés par électrophorèse capillaire sur un séquenceur ABI 310 par les logiciels Genescan et Genotyper (Applied Biosystems™).

5. Puces à CGH

Les puces génomiques utilisées lors de ce travail ont été conçues par Marijke BAUTER et son équipe du Human Genome Laboratory, Departement of Human Genetics, Flanders Interuniversity Institute for Biotechnology (Leuven, Au-delà). Chaque puce est composée de 1875 clones du chromosome X (chaque clone est déposé deux fois) offrant une résolution théorique de 82kb [212]. Les puces possèdent 96 clones autosomiques et 153 cibles vides permettant de contrôler la qualité de l'hybridation.

Chaque lame est composée de 2 puces, et chaque puce est composée de 16 blocs eux-mêmes composés de 2 sous-ensembles, l'un étant le replicat de l'autre (fig 3). Chaque sous-ensemble comporte 144 cibles (clones du chromosome X, clones autosomiques ou cibles vides).

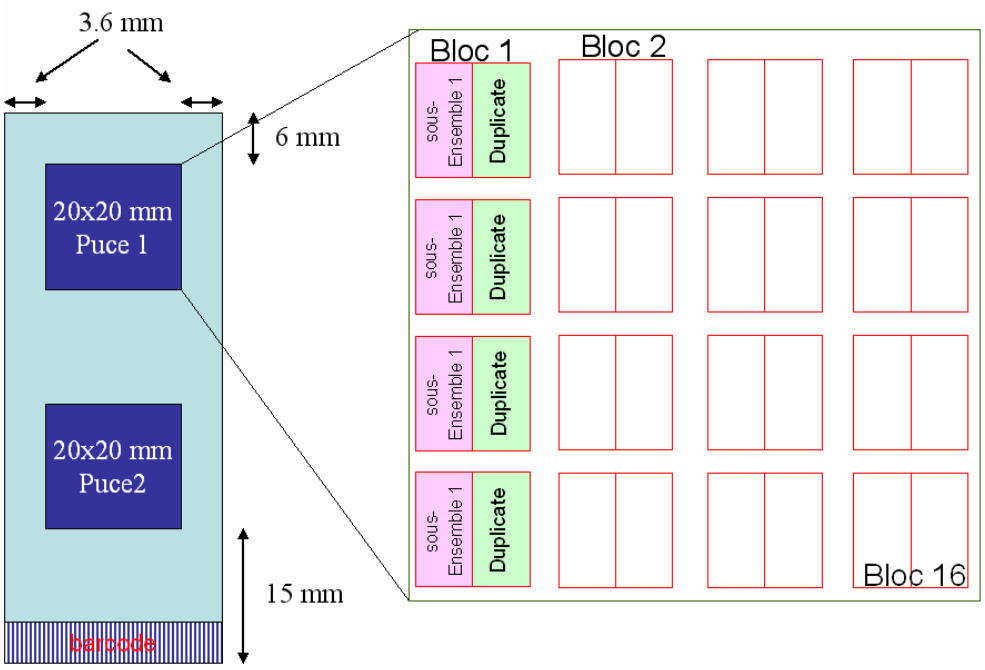


Figure 22 Description de la puce génomique spécifique de l'X utilisée

Chaque lame est composée de 2 puces, et chaque puce est composée de 16 blocs eux-mêmes composé de 2 sous-ensembles, l'un étant la réplique de l'autre

L'ADN de chaque patiente a été marqué une fois en Cy3 et cohybridé en quantité équimolaires avec de l'ADN d'un témoin de sexe féminin marqué en Cy5, puis une fois en Cy5 et cohybridé avec de l'ADN du même témoin marqué en Cy3 (Dye Swap).

Les cyanines étant sensibles à la lumière, les manipulations se sont faites dans la pénombre et les produits contenant des cyanines ont été protégés le plus possible de la lumière.

5.1. Marquage des ADN

Le marquage a été réalisé à partir de 150 ng d'ADN dilué dans 10 µl d'eau et à l'aide du kit BioPrime® Array CGH Genomic Labeling System (Invitrogen™) en suivant le protocole fourni par le fabriquant des puces. Les 150ng d'ADN ont été mélangés à 10 µl de random primers, incubés 15 min à 95°C puis mis au moins 5min sur de la glace pour dénaturer l'ADN. Après dénaturation, 1,5 µl de cyanine (Cy3 ou Cy5), 2,5 µl de dCTP et 1 µl d'Exo Klenow ont été ajoutés à l'ADN dénaturé. Le tout a été incubé une nuit à 37°C. Le marquage est arrêté par l'ajout de 2,5 µl de Buffer Stop. Les ADN ainsi marqués ont été purifiés par le kit BioPrime®

Array CGH Genomic Labeling System (Invitrogen™) suivant les instructions du fournisseur. Les ADN marqués et purifiés, ainsi que l'incorporation des cyanines ont été dosés par le spectrophotomètre ND-100 (NanoDrop®) pour évaluer le taux d'incorporation des cyanines. Pour chaque réaction, l'activité spécifique du marquage a été calculée par la formule suivante (donnée par le fournisseur des puces) :

$$\text{Activité spécifique} = \frac{1000 \times \text{la concentration de l'ADN (ng/µl)}}{324{,}5 \times \text{la concentration en cyanine (pmol/µl)}}$$

L'activité spécifique doit être inférieure à 45 pour que le marquage soit correct.

5.2. Préhybridation

Les sites d'hybridation aspécifiques ont été bloqués avec du Human Cot-1 DNA® (Invitrogen™) et/ou de l'ADN de sperme de saumon.

5.2.1. Traitement des ADN

La quantité d'ADN marqué à utiliser a été optimisée pour donner les meilleurs résultats possibles (communication personnelle, M. BAUTER, 2005). 1,05 µg d'ADN marqué en Cy5 ont été mélangés à 2,45 µg d'ADN marqué en Cy3 et à 50 µl de Human Cot-1 DNA® (Invitrogen™). Ce mélange a été précipité à l'éthanol en y ajoutant 1/10 du volume d'acétate de sodium 3M (NaOAc) et 2,5 volumes d'éthanol froid. Cette préparation a été incubée au moins 10 min à -20°C, puis centrifugée de 30 min à 4°C et 13 200 tr/min. Le culot a été récupéré, séché puis ressuspendu avec 2 µl d'ARNt et 15 µl de tampon d'hybridation à 75°C (50% de formamide SigmaUltra, SSC 2X, 10% de Sulfate de dextran, 0,1% de Tween20, 10mM de TrisHCl, pH à 7,5). Les ADN ainsi préparés ont été dénaturés 10 min à 75°C, déposés sur glace pendant 5 min, puis préhybridés pendant 1H à 37°C.

5.2.2. Solution de blocage

La solution de blocage permet de bloquer les sites d'hybridation aspécifiques du support de la puce. Elle a été préparée avec 10 µl d'ADN de sperme de saumon et 16,7 µl de Human Cot-1 DNA® (Invitrogen™), puis précipitée à l'éthanol. Les culots obtenus ont été ressuspendus dans 20 µl de tampon d'hybridation à 75°C, dénaturés 10 min à 75°C puis déposés sur glace pendant 5min. Les 20 µl de solution de blocage ont été déposés sur une seule puce, nécessitant de préparer deux solutions de blocage pour chacune des deux puces contenues sur une même lame. Une lamelle en verre de 27x60 mm a été déposée sur les lames, puis ces dernières ont été mises à incuber 1H à 37°C.

5.3. Hybridation

Apres incubation de la solution de blocage, les lamelles ont été retirées des lames. L'intégralité des 17 µl correspondant aux ADN cibles marqués par les Cy3/Cy5 préhybridé ont été déposés sur une puce. Chaque puce est ensuite recouverte par une lamelle en verre de 24x24 mm. Les lames ont été placées dans des chambres hermétiques (Corning) contenant 2x12 µl de tampon Corning dans des encoches prévues à cet effet. Les chambres ont ensuite été mises deux nuits à 37°C.

5.4. Lavage et lecture

Après hybridation, les lamelles ont été retirées des lames et ces dernières lavées dans trois bains successifs. Les lames ont été plongées dans un premier bain (PBS 1x / 0,05% Tween20) pendant 10min à température ambiante, puis elles ont été mises dans un second bain (50% de formamide / SSC 2X pendant 30min à 42°C, puis de nouveau 10 min dans un troisième bain (PBS 1X / 0,05% Tween20) à température ambiante. Les lames sont immédiatement rincées trois fois dans de l'eau mQ puis centrifugées 1 min à 1 200 tr/min pour les sécher.
Les lames ont été lues (532 nm pour la Cy3 et 635 nm pour la Cy5) avec le scanner Axon 4000B (Axon instrument) du Laboratoire Interaction Arbres/micro-organismes (UMR 1136,

centre INRA de Nancy) et les résultats interprétés à l'aide des logiciels GenePix Pro 5.0 (Axon instrument) et Microsoft Excel par le fichier « NEW AXON TEMPLATE.xls » fourni avec les puces.

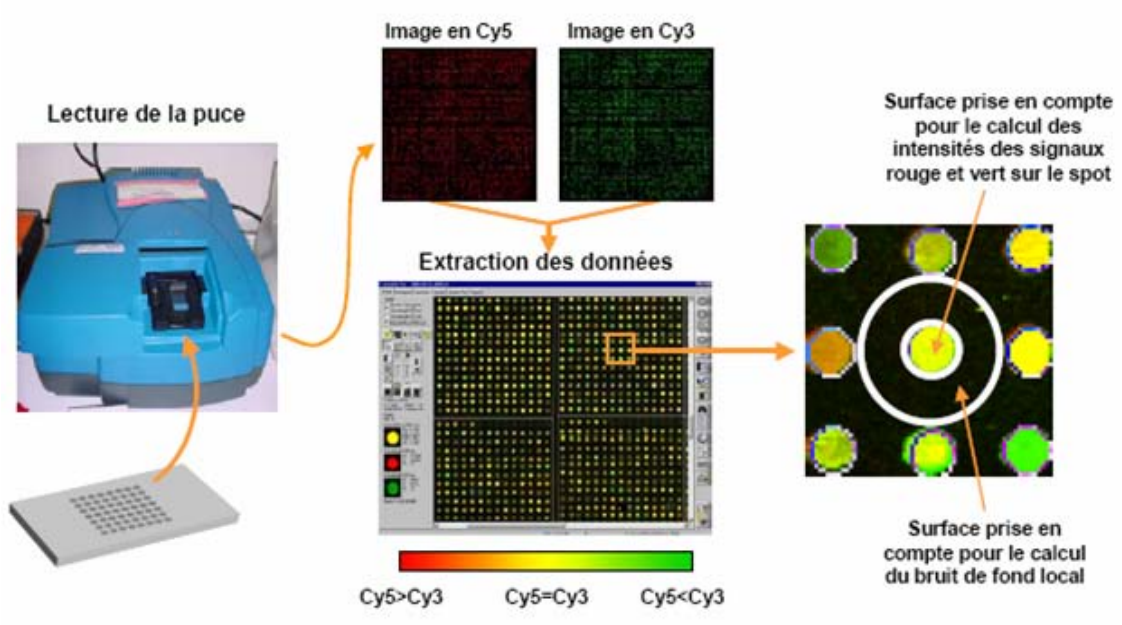


Figure 23 : Acquisition des données sur le scanner GenePix Pro

Les images obtenues pour chaque canal Cy3 et Cy5 sont enregistrées au format TIFF 16 bits en niveaux de gris. Le signal généré pour chaque fluorochrome est converti en fausses couleurs (généralement rouge pour la cyanine 5 et vert pour la cyanine 3). Les deux images sont superposées. Le rapport signal/bruit de fond devait être supérieur à 1,5-2 pour que le signal soit quantifiable.

5.5. Prétraitement des données

5.5.1. Acquisition de l'image

Les images sont analysées grâce au logiciel GenePix Pro 4 (Axon instruments, union city, CA, USA) afin d'extraire les données numériques correspondant à chaque spot. Les images sont colorées artificiellement (Cy3 en vert et celle de Cy5 en rouge) et superposées pour leur visualisation. Ainsi, un spot de couleur verte indique une quantité d'ADN plus élevée dans l'échantillon marqué avec le Cy3 par rapport à celui marqué avec le Cy5, et inversement

pour un spot de couleur rouge. Le spot apparaît jaune lorsque la quantité d'ADN cible est identique dans les deux échantillons comparés.

Le logiciel permet de définir une grille sur l'image afin d'identifier chaque spot en lui assignant des coordonnées uniques (adressage des spots) et de délimiter la surface du spot par rapport au reste de la lame (« segmentation » du signal contenant le spot par rapport au à celui définissant le bruit de fond). GenePix Pro intègre un algorithme de placement automatique des grilles, ce qui accélère considérablement l'analyse par rapport à un placement manuelle. Le logiciel génère les données numériques correspondant aux valeurs moyennes et médianes du signal émis par les pixels de chaque spot et du bruit de fond local et divers autres paramètres (l'écart-type associé aux intensités des pixels d'un spot, le rapport signal/bruit de fond, etc.). Seules les signaux avec une intensité au moins deux fois supérieur à celle du bruit de fond ont été inclues dans les analyses.

5.5.2. Extraction des données

Le logiciel d'extraction permet de repérer de façon automatique les spots non exploitables (zones de la lame couvertes de bruit de fond, spot absent…) par un système de balisage (dépôt de drapeaux ou « flag ») qui assigne un code numérique selon la qualité du spot considéré, afin de faciliter le filtrage des données non significatives.

5.5.3. Transformation des données

Les ratios repartissent les données dans un intervalle non symétrique ente les gènes sur-régulés (∋]1 ;∞[) et sous-régulés (∋]0 ;1[). Les données sont transformées avec un logarithme qui a l'avantage de produire un spectre de valeur continue et symétrique par rapport à la valeur 0. D'autre part, cette transformation rend les données plus maniables ce qui facilite les analyses statistiques (par exemple, on ne peut calculer les moyennes arithmétiques des ratios de plusieurs expériences que dans un espace logarithmique).

5.5.4. Normalisation des données

Dans une expérience de puces à ADN, les sources de variations artéfactuelles peuvent provenir de :

- la préparation des échantillons,
- l'efficacité des extractions d'ADN entre les échantillons,
- les rendements d'amplification (si une amplification est réalisée),
- l'efficacité des marquages,
- l'efficacité d'hybridation globale et la stringence des lavages,
- le au-delà de détection de la fluorescence.

Il peut également exister un biais spécifique à la plateforme, comme par exemple les caractéristiques des aiguilles de dépôts ou les propriétés des sondes, « l'effet paillasse » (tampon et lames).

L'expérimentateur ainsi que la date de l'expérience sont aussi d'importants facteurs de variabilité entre les expériences [213]. Dans l'idéal, ces sources multiples de variabilité doivent être limitées en optimisant les expérimentations. Cependant, ceci n'est que partiellement possible et donc pour rendre les résultats comparables, le reste des variabilités non biologiques est corrigé (ou limité) par le processus de normalisation [134, 139, 214].

La normalisation consiste à ajuster l'intensité globale des images acquises sur chacun des deux canaux rouge et vert, de manière à corriger des biais techniques systématiques qui tendent à déséquilibrer le signal de l'un des canaux par rapport à l'autre. Ces biais sont dus en particulier aux caractéristiques des deux fluorochromes Cy3 et Cy5 qui ne possèdent pas le même coefficient d'extinction molaire (à incorporation égale, Cy5 émet un signal plus fort que Cy3), aux différences d'incorporation des fluorochromes lors de la synthèse des cibles, et aux paramètres de lecture par le scanner (réglages de la puissance des lasers…). Il existe plusieurs méthodes pour normaliser les données. Dans nos expériences de puces génomiques la normalisation des données a été effectuée en divisant le ratio des intensités de chaque spot par la moyenne des ratios pour les clones autosomiques. Une moyenne des ratios normalisés à été effectuée entre les duplicates et les $\log_2$ ont été calculées. Les clones pour lesquels la variation des deux ratios d'intensités était supérieure à 10 % ont été exclus de l'analyse.

5.6. Analyses des résultats des puces à CGH

Les ratios obtenus pour une délétion est de ½ et pour une duplication de 3/2. Lors d'une expérience, la majorité des clones de la puce sont dans la normalité et donc les ratios d'intensités entre patient et témoin sont proches de 1. Ceci permet de calculer une déviation standard de l'ensemble des intensités reflétant ainsi la qualité d'une hybridation. L'analyse des données nécessite la sélection d'un seuil permettant de retenir les faux positifs sans éliminer les vrais positifs. Ce seuil est variable dans le protocole d'analyse des puces que nous avons utilisé. Etant donnée que la distribution des ratios d'intensité normalisé suit une courbe de Gausse, la DS peut être utilisée pour le choix du seuil. Un seuil est définit comme plus ou moins 3 ou 4 fois la DS.

En utilisant 3 DS comme seuil, 99,7% des clones seront contenues dans l'intervalle normal. Pour une puce de 2 000 clones ceci représente 6 faux positifs. En utilisant 4DS, 99, 994% des clones seront dans l'intervalle normal et ceci conduit à environ 0,12 faux positifs donc environ un faux positif tout les 8 analyses. Un second point est à prendre en compte lors de l'analyse des résultats. Les duplications ont un log2 ratio plus proche de la normalité que les délétions et sont donc plus difficile a détecter. De ce fait, 4DS doit être placé sous la limite de détection d'une duplication.

Un second seuil est ainsi définit ainsi : 4 DS ≤ $\log_2$ (3/2)-2DS ou DS ≤0,096

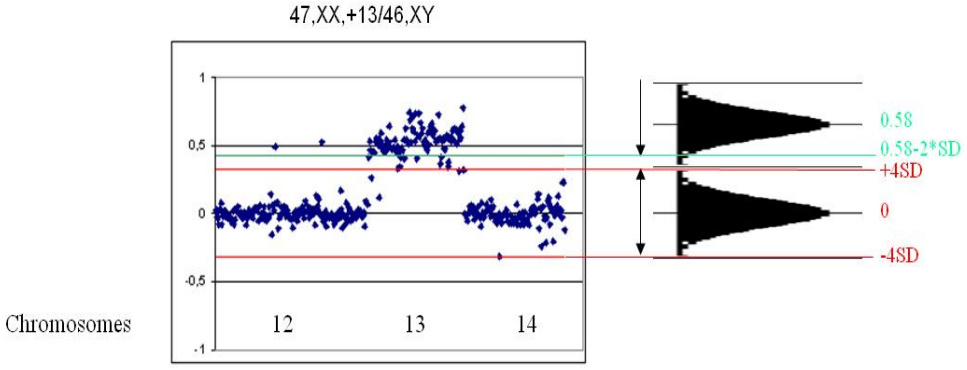


Figure 24 : Répartition de la population des clones normaux et de clones dupliqués quand SD > 0,096

La population des clones normaux et celle des clones dupliqués sont répartir selon les courbes en noir. Quand SD > 0,096, une grande partie des clones dupliqué se retrouve sous la limite de +4xSD. Dans ce cas, le $\log_2$ (1,5)-2xSD permet de faire ressortir une plus grande partie de la population des clones dupliqués.

Dans cette étude avons travaillé dans l'optique de tirer un maximum de profit de ses données de puces génomiques. Un seul clone déviant à été suffisant pour le sélectionner alors que le protocole d'analyse des puces recommande au moins deux clones chevauchants déviant dans l'intervalle entre les deux seuils. De plus, pour ne pas passer à coté d'une variation, nous avons volontairement sélectionnés des clones dans les limites des seuils fixés, augmentant ainsi volontairement le nombre de faux positifs puisque la technique de Q-PCR nous permettait par la suite de les éliminer.

Finalement, pour l'analyse des données issue de ces puces génomiques des critères de qualité et deux seuil sont définit tel que l'efficacité d'hybridation est supérieur à 97 %, l'intensité du clones sur celui du bruit de fond est supérieur à 2 et la déviation standard inférieur ou égale à 0,096 (soit ± 4 DS $\leq \log_2 (3/2) - 2$ DS).

6. Etudes transcriptomiques

Le but est de rechercher les gènes différentiellements exprimés entre les filles AIC et leur mère indemne. Ainsi, les gènes sur ou sous régulés (gènes signatures) chez les filles Aicardi par rapport à leur mère sont des candidats potentiels. Deux séries d'expériences AS1 et AS2 ont été menées. L'étude AS1 à été réalisée sur des puces Agilent A1(V2) (Agilent technologies) 22K, avec 22 000 oligonucléotides de 60 bases. L'étude AS2 a été menée avec les puces Whole Human Genome (Agilent technologies) avec 41 000 oligonucléotides. Mise à part quelques différences minimes, les protocoles expérimentaux entre ces deux études sont identiques. Ainsi, depuis le marquage jusqu'à l'extraction des données de transcriptomique (obtention des fichiers MAGE-ML) nous avons suivit les protocoles du fournisseur.

6.1. Plans expérimentaux

L'étude AS1 concerne trois patientes Aicardi (HK, BM et CC) et leur mère respective (HAK, RL, FM).

Nous avons suivi un plan d'expérience avec référence commune. La référence commune utilisée est un pool d'ARN lymphoblastiques de 10 témoins femme. L'ARN des patientes AIC et de leur mère est marqué par la Cyanine 3 et celui du pool par la cyanine 5. Six hybridations sont ainsi effectuées.

Les expériences AS2 sont réalisées sur de l'ARN extraits de sang. Un pool d'ARN extraits de sang et issue de 18 femmes saines sert de référence (marqué en cy5). Chaque patiente (fille Aicardi ou mère de fille AIC) est marqué en cyanine 3. Vingt lames whole human genome ont été utilisées. Une dégradation de l'ARN correspondant à la mère de la patiente PS n'a pas permit de l'utiliser dans les expériences. Ainsi, un pool des ARN des filles AIC a été utilisé cette hybridation. Quatre séries d'hybridation de 5 lames (regroupées au hasard) on été ainsi effectuées (tableau 6).

Serie I	Serie II	Serie III	Serie IV
LS vs **référence**	**PAL** vs **référence**	**PI** vs **référence**	**Pool filles Aicardi** vs **référence**
LA vs **référence**	**CJ** vs **référence**	**PS** vs **référence**	**CG** vs **référence**
PB vs **référence**	**DE** vs **référence**	**LC** vs **référence**	**DMT** vs **référence**
LMC vs **référence**	**TA** vs **référence**	**TE** vs **référence**	**AM** vs **référence**
AP vs **référence**	**CC** vs **référence**	**AS** vs **référence**	**KA** vs **référence**

Tableau 5 : Hybridations de l'étude AS2.

Dans chaque série, les ARN d'une patiente AIC ou sa mère (en vert) marqué par de la cyanine 3 est mise en compétition avec le pool d'ARN référence (en rouge) marqué par de la cyanine 5.

6.2. Marquage des ARN totaux

Le marquage a été réalisé à partir de 250 ng d'ARN dans 11, 5 µl d'eau RNAse free à l'aide du kit Low RNA imput Fluorescent linear amplification (Agilent™) et en suivant le protocole fournit par le fabriquant des puces. Les ARN marqués ont été purifiés avec les colonnes Rneasy (Quiagen) et l'efficacité de l'hybridation est contrôlée par le bioanalyseur 2100 (Agilent™). Pour chaque réaction, l'efficacité d'incorporation des fluorochromes à été calculé. Les valeurs doivent être comprises entre 8 à 15 pmol de Cyanine/µg ARNc).

Efficacité d'incorporation de la Cy5 (pmol Cy5/µg ARNc)= $\frac{\text{DO 633 x dilution}}{\text{250 x Concentration ARNc (ng/µl)}}$

Efficacité d'incorporation de la Cy3 (pmol Cy3/µg cRNA)= $\frac{\text{DO 633 x dilution}}{\text{150 x Concentration ARNc (ng/µl)}}$

6.3. Hybridation et lavages

Une solution d'ARNc cibles marqués est préparée dans un tube de 1,5 µl. Cette solution d'un volume total de 250 µl, contient 0, 750 µg d'ARNc marqués par de la cy3 ; 0, 750 µg d'ARNc marqués par de la Cy5 et 5 µl d'une solution contenant des ARNc cibles contrôles et fournie dans le kit d'hybridation. Dix microlitres du tampon de fragmentation (25X Fragmentation Buffer) sont ajoutés au mélange précédent et le tout est placé dans un four à 60°C pendant 30 minutes. Pour arrêter la réaction de fragmentation, 250 µl d'un tampon d'hybridation sont ajoutés au mélange précédent. Le mélange est déposée sur la puce transcriptomique qui est montée dans une chambre d'hybridation fournie par le fournisseur des puces. L'hybridation est réalisée sur 17 heures à 60°C.

Après hybridation, les lames ont été lavées dans deux bains successifs. Les lames ont été plongées dans un bain de 20X SSC/ Triton X-102 pendant 10 min à température ambiante, puis elles ont été mises dans un bain de 20X SSC/ 10 % Triton X-102 pendant 5 min. Les lames sont séchées avec de l'azote grâce à un pistolet à air comprimé.

6.4. Acquisition des données

Les lames ont été lues (532 nm pour la Cy3 et 635 nm pour la Cy5) avec le scanner Agilent (Dual DNA Microarray Scanner, G2565BA). Les étapes de pré traitement des données : acquisition d'une image, extraction des données, transformation et normalisation ont été effectuées avec le logiciel Feature extraction (v7.1) (Agilent). Les paramétrés par défaut déterminés pour les deux types de puces ont été utilisées (A1v2 et whole genome). De façon succincte, le logiciel délimite les spots à analyser des régions avec bruit de fond. Le bruit de fond local est soustrait et une normalisation des biais des fluorochromes est réalisée par la méthode de normalisation linear lowess (Locally weighted scatter plot smoothing). Cette méthode statistique permet, grâce à un système de « fenêtre glissante », de calculer une

courbe de normalisation ajustée à la forme du nuage par régression linéaire locale. Cette méthode est aujourd'hui admise pour être la plus robuste pour les puces à deux couleurs [215]. Les log ratios des signaux normalisés sont calculées. Pour les faibles valeurs d'intensité in surrogate est utilisé.

Parmi les fichiers résultats générés par le logiciel feature extraction figurent un format texte et un fichier MAGE-ML (Microarray Gene Expression Markup Language). MAGE-ML est un languages destiné aux donnée de puces à ADN (microarray designs, microarray manufacturing information, microarray experiment setup and execution information, gene expression data, and data analysis results). Des fichiers TIFF et image JPEG sont aussi générés.

6.5. Analyse des données de transcriptomique

6.5.1. Mise en évidence de gènes différentiellement exprimés

Les fichiers générés par le logiciel feature extraction sont exploités dans le logiciel Luminator (Rosetta). Ce logiciel attribue à chaque ratio calculé une valeur p. La valeur p est calculée à partir d'un modèle d'erreur spécifique des puces Agilent utilisées. Une valeur p de 0.01 signifie qu'il y a 1 chance sur 100 d'observer au moins ce niveau d'expression différentielle uniquement par le simple fait du hasard. Plus simplement, elle représente la probabilité que le gène ne soit pas exprimé de façon différentielle entre les deux échantillons co-hybridés. En fixant des seuils de variation de ratio et de valeur p, nous obtenons les gènes sur/sous régulés (signatures) pour chaque couple fille vs mère. Une fonctionnalité de luminator appelé « re-ratio » permet de calculer des ratios à partir de canaux de deux ou plusieurs ratios existants qui possèdent une référence commune. Ainsi, des gènes signatures ont été sélectionnés pour chaque couple fille versus mère étudié. Une autre façon de sélectionner des listes de gènes signatures est l'analyse par des outils statistiques, cette fois de l'ensemble des résultats (toutes les hybridations réalisées). Pour pouvoir utiliser des outils statistiques, un minimum de 4 échantillons est nécessaire. Dans l'étude AS2, nous avons

utilisé une méthode ANOVA implémenté dans le logiciel Luminator avec les paramètres par défauts (valeur p de 1,0 E-02).

6.5.2. Stratégies d'études

Les questions auxquelles nous voulons répondre sont les suivantes :

- Quel(s) sont les gènes différentiellement exprimés chez les filles AIC ?
- Pour chaque couple fille *versus* mère, quels sont les fonctions biologiques dérégulées ?

Dans la première étude le nombre limitant d'échantillons (3 couples) n'a pas permis l'utilisation de méthodes statistiques d'analyse des données. La méthode ANOVA (ANalyse Of Variance) a été utilisée pour l'étude AS2.
Dans les deux études, une méthode de sélection avec un seuil de variation de ratio de 1,5 et une valeur p de 1,0 E-02) ont été utilisés. Nous avons ainsi voulu analyser les différences entre les transcriptomes d'une fille AIC par rapport à sa mère. Cette méthode n'est pas statistique mais elle est suivie par une analyse fonctionnelle utilisant cette fois une méthode statistique avec le logiciel EASE (Expression Analysis Systematic Explorer).

6.5.2.1. Analyse fonctionnelle des résultats de transcriptomiques

Un groupe de gènes est dit « fonctionnellement enrichi » en une fonction biologique si la proportion de gènes dans le groupe connus pour être impliqués dans cette fonction biologique excède le nombre attendu par le hasard. Ainsi, pour 10 gènes signatures à étudier si 7 font partie de la catégorie fonctionnelle « biosynthèse des ribosomes », la fréquence observée (fobs) est de 70% (7/10). Sur l'ensemble du génome on sait que 215 gènes sur 6000 appartiennent à la classe fonctionnelle étudiée (fréquence référence (fref) = 215/6 000 soit 3,5%). Il existe différents outils Web qui calculent la probabilité d'observer une fréquence fobs par le hasard (compte tenu de la valeur de fref). Plus cette probabilité est faible, plus le groupe est «fonctionnellement enrichi ». Je parlerai dans ce cas de catégorie sur représenté.
Dans nos études nous avons utilisé une analyse fonctionnelle des gènes signatures par le logiciel EASE. Le logiciel à besoin de deux listes d'entrée. Une liste représentant l'ensemble

des identifiants de la puce à ADN utilisée et une liste de gènes signatures. La première liste sert de référence pour le calcul de fonctions biologiques sur représentées dans la liste signature rentrée secondairement. Les gènes signatures sont utilisés comme entrée dans le logiciel. Ils sont classés dans des catégories fonctionnelles des trois systèmes composant Gene Ontology. Un score est affecté à chaque catégorie sur représentée. Ce score est calculé par rapport à une liste référence. La liste référence que nous avons utilisée est dans le cas des études pour AS1 la liste de gènes sur la puce A1(V2) et pour l'étude AS2, la liste des gènes issue de la puce whole human genome. Ces listes références, pour les puces A1(V2) et whole human genome ont été téléchargées sur le site de ressource TIGR (The Institute for Genomic Research) (http ://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl). Le score EASE donné correspond au score donné par un test de Fisher. Pour chaque classification possible, les deux éléments sont : les gènes qui appartiennent à la catégorie et les gènes qui n'appartiennent pas à la classification. Avec le nombre de gène appartenant à chaque population, il est possible de calculer la probabilité d'appartenir à une catégorie par chance. Les analyses on été faite sur les listes signatures de chacun des couples fille versus mère.

6.5.2.2. Etude ANOVA

Dans une expérience de puces à ADN, lorsque de multiples facteurs (age, sexe..) sont à analyser, le test *t* et ses variantes ne suffisent généralement pas à l'interprétation. Un modèle plus complexe doit être construit et une analyse de variance, ou ANOVA, peut être utilisée pour mettre en évidence l'impact de chaque facteur. En effet, une étude ANOVA permet d'évaluer si les moyennes de un ou plusieurs groupes d'échantillons sont significativement différentes et si un ou plusieurs facteurs affectent les mesures. Dans l'étude AS2, nous avons ainsi étudié les facteurs de variation maladie et âge des patientes avec une méthode ANOVA du Logiciel Luminator. L'outil ANOVA implanté dans Luminator, possède la particularité d'intégrer dans le calcul les valeurs p attribués aux ratios. Ainsi la valeur p fournit des informations additionnelles pour des évaluations plus fiables des variances intra groupe. Une méthode ANOVA à deux facteurs (maladie, âge) est ainsi utilisée sur les 10 hybridations fille versus référence. Les groupes sont partitionnés de la façon suivante :

- Si les patientes sont pré-pubères elles font partie du groupe « jeune » (5 filles AIC : LA, PB, AM, LC, PAL) , si elles sont pubères elle feront partie du groupe « adulte » (14 patientes = 5 filles +9 mères)
- Pour les facteurs maladie les deux groupes sont : les 10 filles AIC (groupe maladie) et le second groupe « non maladie » est composée de 9 témoins (ce groupe contient les mères des filles AIC).

7. Etudes *in silico* et logiciel ACGR (Approach for Candidate Gene Retrieval)

Le matériel et les méthodes utilisées dans la partie bioinformatique de ma thèse ne seront pas présentés dans cette session résultat puisqu'ils ont été détaillés dans la publication.

Résultats

Dans cette section « Résultats », nous suivrons un ordre logique de présentation plutôt qu'un ordre chronologique. En effet, les restrictions que nous avons rencontrées nous ont imposés un déroulement expérimental particulier. Par exemple, les puces pour l'étude transcriptomique ont été disponibles avant les puces génomiques. Ceci explique le nombre supérieur de patientes participant à l'étude à l'aide de puces génomiques (18 patientes), alors que seules 3 patientes ont participé à l'étude d'expression (AS1) antérieure

Dix huit filles diagnostiquées syndrome d'Aicardi ont participé aux différentes études menées au cours de ce projet (paragraphe 1). Les études de cytogénétiques constitutionnelles représentent une approche systématique à aborder lors d'une étude de clonage positionnel (paragraphe 2). Le mode d'hérédité de ce syndrome nécessite également une analyse de l'inactivation du chromosome X (paragraphe 3). Les puces génomiques sont utilisées pour rechercher des microdélétions et microduplications à un niveau de résolution plus élevé que celle obtenue par l'étude du caryotype (paragraphe 4), cette approche constitue donc une suite logique aux études du caryotype. Les approches transcriptomiques, quant à elles, sont utilisées pour capturer à un temps « t », l'état du transcriptome des cellules des patientes AIC. Ces informations couplées à une intégration des informations disponibles dans les bases de données publiques peuvent conduire à dévoiler des voies biologiques impliquées dans le syndrome d'Aicardi (paragraphe 5). Une approche *in silico* aide à cette intégration et propose également une approche de recherche de gènes candidats qui lui est propre (paragraphe 6.).

1. Clinique

Il est important de rappeler que ce projet a vu le jour grâce au soutien constant et dynamique de l'association AAL- syndrome d'Aicardi (www.aicardi.info). S'agissant d'une affection dont la fréquence est rare, les patientes atteintes du syndrome d'Aicardi ont été progressivement recrutées par l'intermédiaire de cette association et aussi par le réseau de neuropédiatres et généticiens via le serveur d'information sur les maladies rares Orphanet (www.orphanet.net). Les travaux ont été initiés à partir d'un petit nombre de patientes pour s'étendre ensuite à une série de 18 familles.

Le diagnostic du syndrome d'Aicardi repose sur une expertise, auprès de neuropédiatres et généticiens, de l'ensemble des dossiers médicaux et des examens complémentaires de neuro-imagerie et d'ophtalmologie. Treize patientes répondent aux critères classiques du syndrome AIC (spasmes infantiles, agénésie du corps calleux, lacunes choriorétiniennes). Les 5 autres patientes répondent aux critères étendus du AIC [34]. Au moment de l'entrée dans l'étude, les patientes sont âgées de 6 à 25 ans (moyenne des âges 13,5 ans ; médiane des âges 10,5 ans). Le tableau 6 résume les principales données cliniques et paracliniques des 18 patientes.

Pour chaque patiente et ses parents un prélèvement sanguin a été réalisé après une information spécifique sur les explorations génétiques envisagées et recueil du consentement parental selon les termes de la loi de bioéthique en vigueur.

Patiente	Age			Spasmes infantiles	Agénésie du CC	Lacunes CR	Malformations corticales	Hétérotopies	Kystes V3/plexus choroïdes	Colobome DO/NO	Dysmorphie cranio-faciale	Anomalies costales vertébrales	Autres
	de la patiente AIC	de la mère	du père										
PB	6	29	29	+	+					+	+		CIA
LA	6	32	26	+	+	+		+					
CC	7	33	34	+	+	+					+		
CJ	20	25	26	+	+	+		+		+	+		
PS	23	21	18	+	+	+			+				
PAL	16	36	47	+	+				+				
RC	17	36	36	+	+	+							
TE	25	26	26	+	+	+							
BM	8	32	32	+	+	+			+			+	
AM	10	26	26	+	+	+		+					
DE	23	29	29	+	+	+						+	
HK	8	24	24	+	+	+						+	
KA	20	37	34	+	+			+	+	+			
KO	7	30	27	+	+				+				
LC	11	30	ND	+	+	+							
RM	8	35	35	+	+	+	+	+	+	+	+		
Lam	22	27	28	+	+	+	+						
MB	7	35	36	+	dysplasie	+							

Tableau 6: Données cliniques et paracliniques des 18 patientes AIC étudiées.

L'âge correspond à celui de l'entrée dans l'étude. V3 : 3ème ventricule, DO : disque optique, NO : nerf optique, CIA : communication interauriculaire. La dysmorphie cranio-faciale correspond à celle décrite récemment par Sutton et al. [28] et rapportée dans l'introduction de ce manuscrit. L'âge moyen des filles AIC est de 13,6 ans (médiane 10,5 ans). L'âge moyen de conception des mères est de 30,2 ans (médiane 30 ans) et celui des pères 30,2 ans (médiane 29). Les patientes 1, 6, 13, 14 et 18 (numéro en gras) présentent les critères étendus du diagnostic de syndrome AIC selon Aicardi et al [30].

2. Analyses cytogénétiques

Les 18 patientes AIC ont bénéficié d'une analyse cytogénétique constitutionnelle (caryotype lymphocytaire). Aucune anomalie de nombre et/ou de structure n'a été détectée au seuil de résolution de 550 bandes (bandes G et R).

3. Etudes de l'inactivation du chromosome X

Concernant les maladies génétiques de transmission dominante liée à l'X avec létalité chez les garçons, une inactivation de l'X biaisée d'un des chromosomes X chez les personnes porteuses de la mutation est classiquement observée. Les données de la littérature ne semblent pas révéler de biais significatif d'inactivation de l'X chez les patientes AIC [48]. Nous avons effectué une étude de l'inactivation du chromosome X sur l'ADN extrait de leucocytes de 18 patientes AIC et de leur mère respective au locus du gène codant le récepteur aux androgènes (HUMARA, [216]). Dans les cas où le locus HUMARA n'était pas informatif, l'étude a été réalisée au locus *FMR1* (fragile X mental retardation protein1). Nous avons considéré que l'inactivation était biaisée lorsqu'un des 2 était inactif dans plus de 70% de cellules. Le profil d'inactivation de l'X révèle chez les patients AIC une distribution identique à celle de la population générale sans biais majeur. Quatre-vingt-six pour cent des patientes (12/14) montre une inactivation aléatoire, 2 patientes présentent une inactivation biaisée, dont une inactivation totalement biaisé. Quatre patientes n'ont pas pu être étudiées (non informativité aux locus *HUMARA* et *FMR1*). Le tableau 7 résume les résultats des études d'inactivation.

Patiente AIC	Inactivation de l'X	
	Fille AIC	Mère
PB	**HUMARA : 87-13% / FMR1 : 73-27%**	HUMARA : 41-59% / FMR1 : 40-60%
LA	HUMARA : 45-47%	HUMARA : 73-27%
CC	ND	ND
CJ	HUMARA : 67-33%	HUMARA : 89-11%
PS	HUMARA : 34-66%	HUMARA : 20-80%
PAL	HUMARA : 33-67%	HUMARA : 65-35%
RC	ND	ND
TE	**HUMARA : 0 – 100%**	HUMARA : 48-52% / FMR1 : 76-24%
BM	ND	ND
AM	HUMARA : 50-50% / FMR1 : 53-47%	FMR1 : 61-39%
DE	HUMARA : 60-40%	HUMARA : 90-10%
HK	HUMARA : 35% - 65%	HUMARA : 75% - 25%
KA	HUMARA : 69-31%	HUMARA : 69-31% / FMR1 : 75-25%
KO	ND	ND
LC	**HUMARA : 77-23%**	HUMARA : 15-85%
RM	HUMARA : 93-7% / FMR1 : 66-34%	HUMARA : 93-7%
Lam	ND	ND
MB	ND	ND

Tableau 7 : Etude de l'inactivation de l' X chez 18 patientes AIC.

HUMARA : étude d'inactivation réalisée aux locus HUMARA et FMR1 : étude d'inactivation réalisée au locus FMR1. ND : non déterminé. Trois patientes PB, TE, LC (en gras) présentent une inactivation biaisée de leur chromosome X.

4. Recherche de microremaniements du chromosome X par puce génomique

4.1. Résultats des puces génomiques

Un criblage de l'ADN génomique à la recherche de microdélétions et microduplications a été réalisé chez 18 filles Aicardi. Les puces génomiques utilisées à cet effet ont été fournies par le laboratoire du Flanders Interuniversity Institute of Human Genetics [212]. Ces puces spécifiques du chromosome X sont constituées de 1875 BAC et possèdent une résolution

théorique de 82 kb. Les critères de qualité que nous avons utilisés sont l'efficacité d'hybridation (au moins 97%), la déviation standard (DS) des ratios d'intensité (DS <0 ,096) et une intensité du signal pour les clones au moins deux fois supérieure au bruit de fond. Un clone est considéré comme déviant, c'est-à-dire délèté ou dupliqué, lorsque son ratio d'intensité se trouve au-delà du seuil de 4 DS et que ces conditions sont retrouvées en dye swap. En effet, on considère que la distribution des ratios d'intensité normalisés suit une loi de Gauss. En utilisant un seuil à $\log_2(1) \pm 4$ DS, 99, 994% des clones seront dans l'intervalle normal et ceci conduit, pour une puce de 2000 clones, à environ un faux positif toutes les 8 analyses (0,12). Un second point est à prendre en compte lors de l'analyse des résultats. Les duplications ont un $\log_2$ ratio plus proche de la normalité que les délétions et sont donc plus difficiles à détecter. De ce fait, 4DS doit être placé sous la limite de détection d'une duplication : 4 DS ≤ log2 (3/2)-2 DS, soit DS ≤ 0.096. Un second seuil est ainsi défini à $\log_2$ (3/2) -2 DS. Avec ce seuil plus élevé, 97,72% des clones dupliqués et environ 100% des clones délétés sont ainsi détectés.

Dans cette étude avons travaillé dans l'optique de tirer un maximum de profit de nos données de puces génomiques. Pour ne pas passer à coté d'une variation, nous avons volontairement sélectionné des clones dans les limites des seuils fixés, augmentant ainsi volontairement le nombre de faux positifs puisque la technique de Q-PCR nous permettait par la suite de les éliminer. L'ADN d'une patiente portant une délétion des exons 18 à 44 du gène DMD (Duchenne Muscular Dystrophy) a été utilisé comme témoin. Ainsi 5 BAC, RP11-142J18 (162kb), RP5-1147O16 (132kb), RP4-556A22 (109kb), RP4-639D23 (95kb) et RP11-64I1 (156kb), ont été systématiquement identifiés comme délétés pour cet ADN. Ce témoin a été utilisé car le locus DMD n'est pas associé au syndrome d'Aicardi. De plus, ce sont deux maladies bien distinctes sans chevauchement phénotypique possible. Dans le cas ou le locus DMD aurait néanmoins été touché chez une fille AIC les profils des 5 clones déviants auraient été différents et donc l'anomalie aurait été détectée.

Nous avons observé pour nos 17 expériences une efficacité d'hybridation d'au moins 98% et une déviation standard comprise dans l'intervalle [0,056 ; 0,096]. Pour 3 expériences la déviation standard obtenue était supérieur à 0,096. Malgré un seuil de DS fixé à 0,096, les 5 clones servant de témoins positifs étaient bien détectés comme déviants et les expériences d'hybridation inverse répondaient aux critères de qualités, nous n'avons donc pas exclu ces cas de l'analyse.

Numéro de lame	Patientes	Puce Cy5[$]		Puce Cy3[§]		Clones déviants
		Eff, Hyb,*	SD	Eff, Hyb,*	SD	
1	PB	98,29%	0,096	97,98%	0,079	
2	LA	98,43%	0,072	98,47%	0,075	
3	CC	97,47%	0,091	98,34%	0,086	
4	CJ	98,38%	0,091	98,61%	0,069	
5	PS	98,34%	0,070	98,56%	0,084	
6	PAL	98,25%	0,088	98,29%	0,089	
7	RC	98,47%	0,109	98,52%	0,055	
8	TE	98,34%	0,072	98,43%	0,069	
9	BM	98,47%	0,080	98,47%	0,099	
10	AM	98,34%	0,097	98.38%	0.093	
11	DE	98,38%	0,063	98.16%	0.075	
12	HK	98,47%	0,093	ND	ND	
13	**KA**	98,43%	0,060	98.52%	0.075	RP11-388L20;RP5-1178I21; RP11-441L6; RP11-23N11; CTD-2511C7
14	**KO**	98,47%	0,072	98.47%	0.069	RP11-142K4; RP5-1000K24
15	LC	98,52%	0,089	98.52%	0.073	
16	**RM**	96,52%	0,065	96.27%	0,065	RP11-97N5
17	Lam	98,47%	0,060	97,71	0,065	
18	**MB**	98,34%	0,099	98.34	0.095	RP11-66N11;RP11-54I20; CTD-2511C7

Tableau 8 : Résultats de l'étude des puces génomiques sur 18 patientes AIC

* Efficacité d'hybridation. $ Puce sur laquelle l'ADN de la patiente a été hybridé en Cy5.
§ Puce sur laquelle l'ADN de la patiente a été hybridé en Cy3. ND : non déterminé

Quatre expériences (13, 14, 16, 18) présentent un total de 10 clones déviants. Quatre clones présentent des log_2ratio dépassant le seuil supérieur log_2 (3/2)-2 DS, 2 chez la patiente Ka (RP11-388L20, RP5-1178I21) apparaissent dupliqués et 2 chez la patiente MB (RP11-66N11, RP11-54I20) apparaissent délétés. Parmi les six clones restant qui appartiennent à l'intervalle [4DS ; log2 (3/2)-2 DS], RP11-441L6, RP11-23N11 apparaissent comme dupliqués chez la patiente KA, RP5-1000K24 et RP11-142K4 (clones colinéaires) apparaissent dupliqués chez la patiente KO, RP11-97N5 apparait délété chez la patiente RM et le clone CTD-2511C7 apparaît délété chez les deux patientes KA et MB.

Parmi ces 10 clones déviants, RP11-23N11 est définit comme clone polymorphe par le logiciel d'analyse des puces. Cinquante cinq clones ont en effet été déterminés comme des variants génomiques connus. Dans nos analyses, chaque clone sélectionné a été recherché dans la base de données Database of Genomic Variants (http ://projects.tcag.ca/variation/)[217] pour

savoir si il était indexé comme polymorphe ou non. Ce site répertorie les variations structurales du génome humain. Un CNV est ainsi défini par un gain ou une perte d'une séquence d'ADN > 1kb de longueur.

Trois clones (RP11-97N5, RP11-388L20, RP11-441L6) sont répertoriés comme des CNV polymorphes par la base de données mais ils ont néanmoins été testés par la technique de validation utilisée. Tous ces clones (sauf RP11-23N11) ont fait l'objet d'une étude par PCR quantitative (qPCR) afin d'éliminer les faux positifs et/ou valider les CNV chez les patientes AIC.

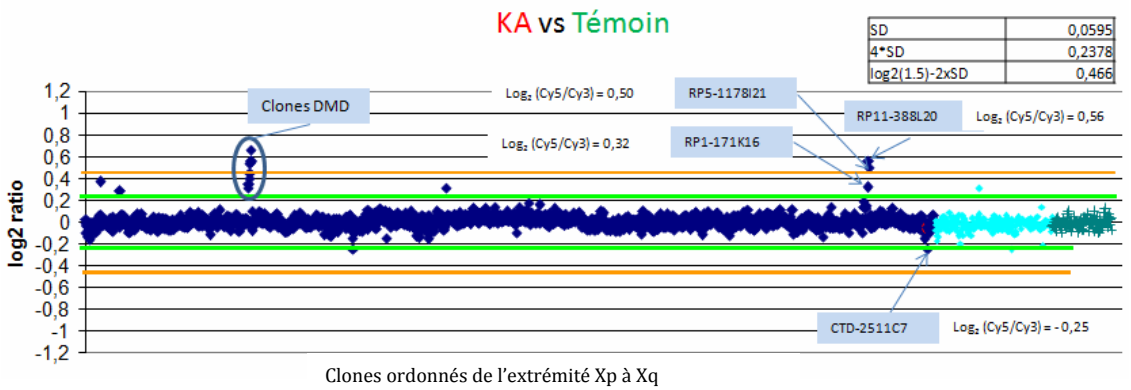
4.2. Validation des résultats obtenus par puces génomiques à l'aide de la Q-PCR

La Q-PCR a été effectuée avec au moins 3 couples d'amorces couvrant de façon homogène la séquence des clones déviants et situés dans des régions uniques. La technique de (q-PCR) pour la validation des résultats de puces génomiques n'était pas utilisée au laboratoire. Une première étape a donc consisté à mettre en place la technique à l'aide d'ADN de témoins mutés. Le gène *WWOX* cartographié sur le chromosome 16 a servi de référence interne pour les calculs de ratios lors des quantifications relatives (méthode des $\Delta\Delta$Ct). Un témoin féminin (ratio de 1) est systématiquement ajouté dans chaque série d'expérience de Q-PCR. Une mise au point de la technique a consisté d'une part à tester 3 délétions connues du gène *MECP2* : délétion de l'exon 4b (ratio obtenu en qPCR = 0,67), délétion promoteur exon1 (ratio = 0.56), délétion totale (ratio = 0,42), et 3 duplications connues du gène MECP2 : duplication exon 4b, duplication promoteur exon 1 (1,44) et duplication totale (1,51). Nous avons d'autre part testé deux duplications connues du gène DMD : duplication des exons 51 à 55 (ratio = 1,80), duplication des exons 42 à 54 (ratio= 1,30).Toutes les délétions et les duplications ont été identifiées par la technique de Q-PCR et donc la technique a ainsi été validée.

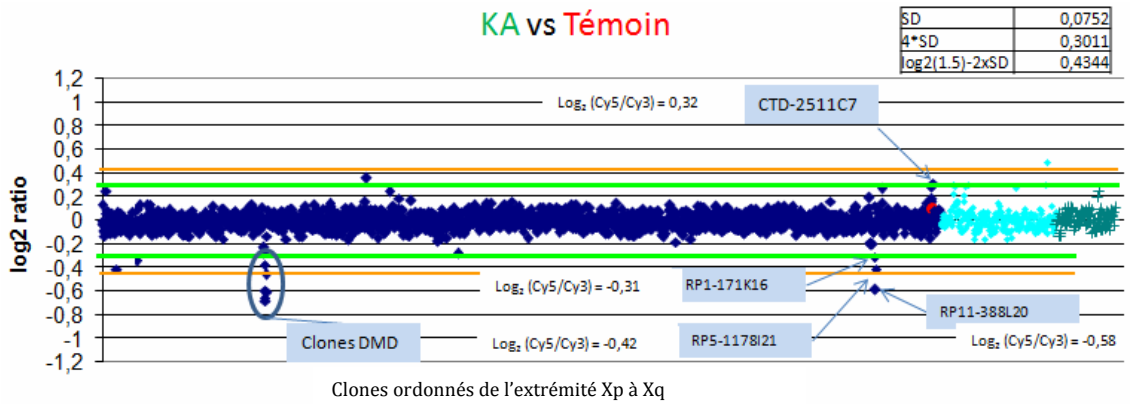
L'ensemble des clones testés par qPCR (RP11-388L20, RP5-1178I21, RP11-66N11, RP11-54I20, RP11-441L6, RP11-142K4, RP5-1000K24, RP11-97N5, CTD-1125C7) donnent des ratios

dans l'intervalle de la normalité [0,80 et 1,2]. Les résultats observés en puces génomiques ne sont donc pas confirmés et ces clones sont considérés comme des faux positifs. Ces résultats ne sont pas très étonnants puisque nous nous étions placés à des seuils limites. De plus, un seul clone déviant a été suffisant pour la sélection alors que le protocole d'analyse des puces recommande au moins deux clones chevauchants déviants dans l'intervalle entre les deux seuils pour considérer le microremaniement. Comme expliqué précédemment nous voulions minimiser les faux négatifs en sachant que le nombre de faux positifs serait dans le même temps supérieur. La technique de q-PCR a été utilisée pour écarter ces faux positifs. Parmi les clones sélectionnés, CTD-2511C7 apparaissant délété chez 2 patientes (KA et MB) a retenu notre attention (figure 1). En effet, ce clone contient le gène de la Filamine A *(FLNA)* qui avait précédemment été évoqué comme gène candidat pour le syndrome AIC [76]. Ce clone de 143 kb a été contrôlé comme normal par qPCR avec 3 couples d'amorces différents dont un placé dans la séquence du gène de la filamine A, pour les deux patientes Aicardi.

A



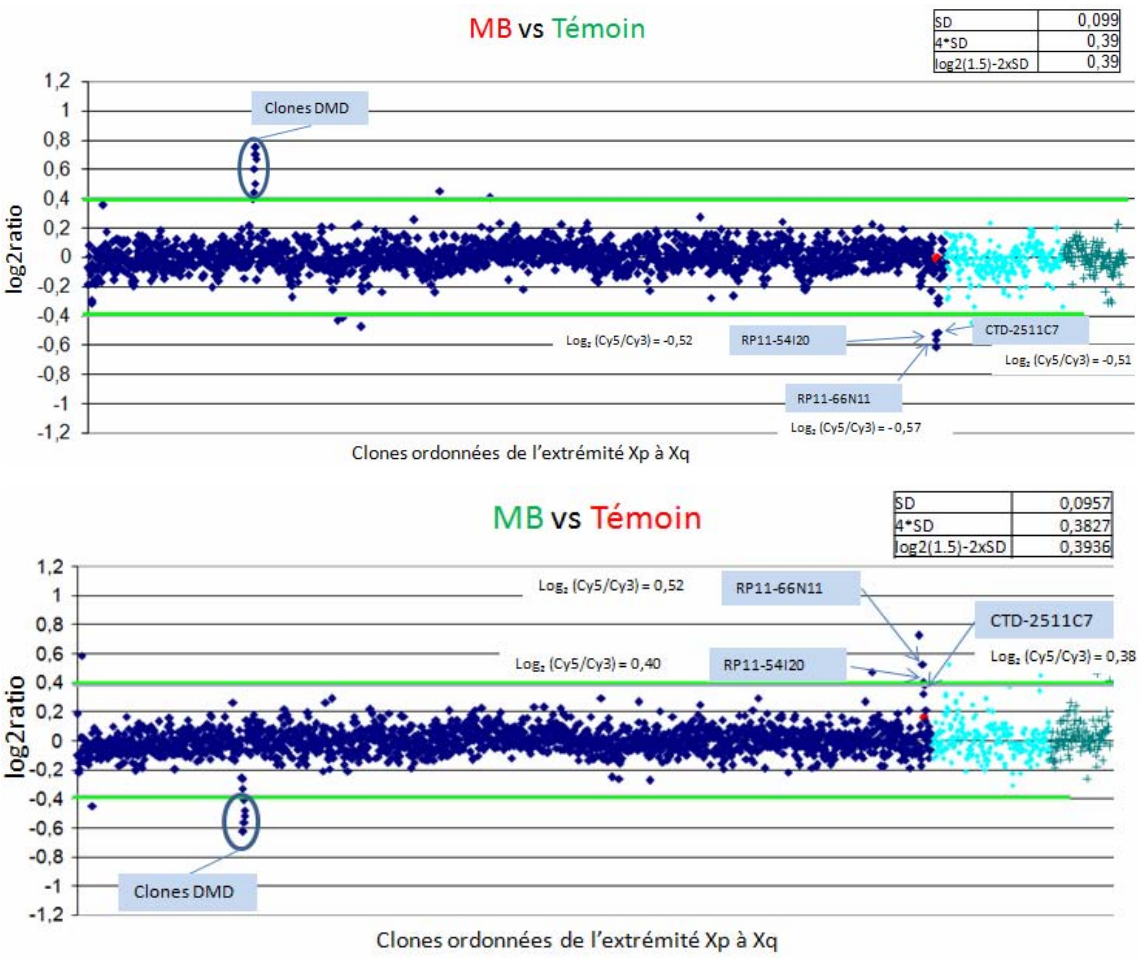
Clones ordonnés de l'extrémité Xp à Xq



Clones ordonnés de l'extrémité Xp à Xq

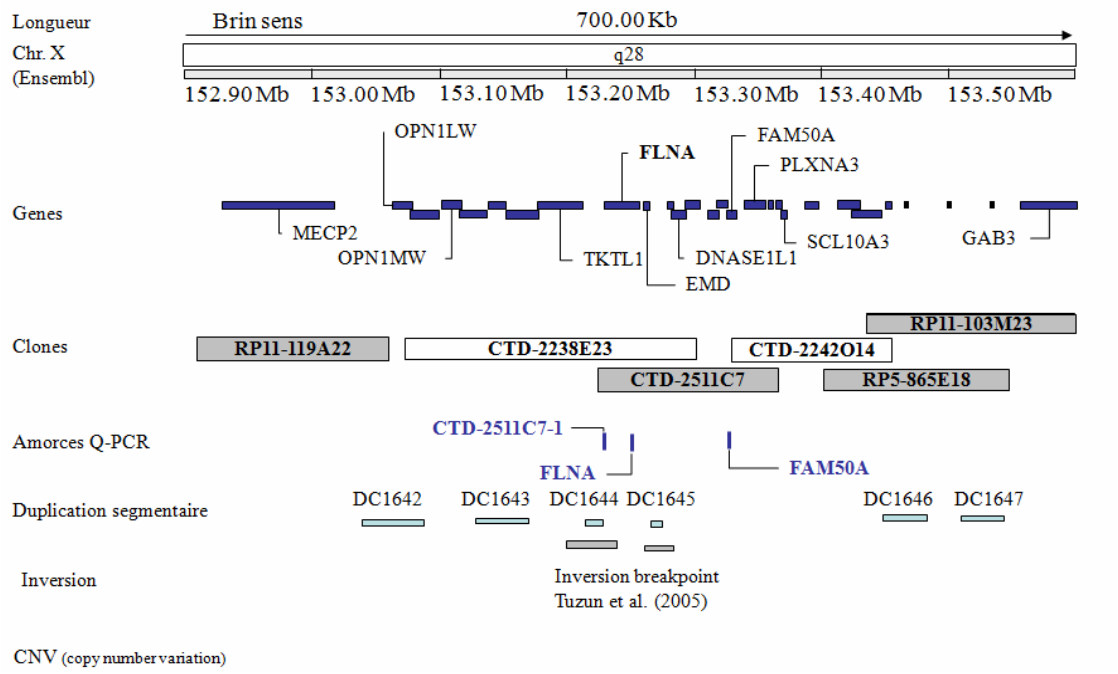
(Suite de cette figure 25, page suivante)

B



(Suite de cette figure 25, page suivante)

C



Ensemble release 42-Dec2006

Figure 25 : Résultat des puces génomiques pour les patientes KA (A) et MB (B).

Les clones de la puce ordonnés du télomère Xp vers le télomère Xq sont représentés sur l'axe des abscisses selon la cartographie de Ensembl (version 42). L'axe des ordonnées représente les $\log_2$ des ratios d'intensités. La ligne verte indique le seuil de ±4 SD et la ligne rouge le seuil de ($\pm\log_2$ (3/2)-2 SD). Les clones localisés sur les autosomes et servant de témoin sont représentés par des points bleus clairs et les contrôles négatifs (clones vides) en vert. Les 5 clones du gène DMD sont entourés. Le clone CTD-2511C7 et les clones déviants sont signalés par une flèche. Tous ces clones sont confirmés en hybridation inverse pour chaque expérience. Les autres clones qui apparaissent comme déviant sont normaux en hybridation inverse. Au-delà cette carte représente le positionnement des clones BAC, des gènes, des amorces de Q-PCR et des variations du génome dans la région du clone CTD-2511C7 selon Ensembl en décembre 2006.

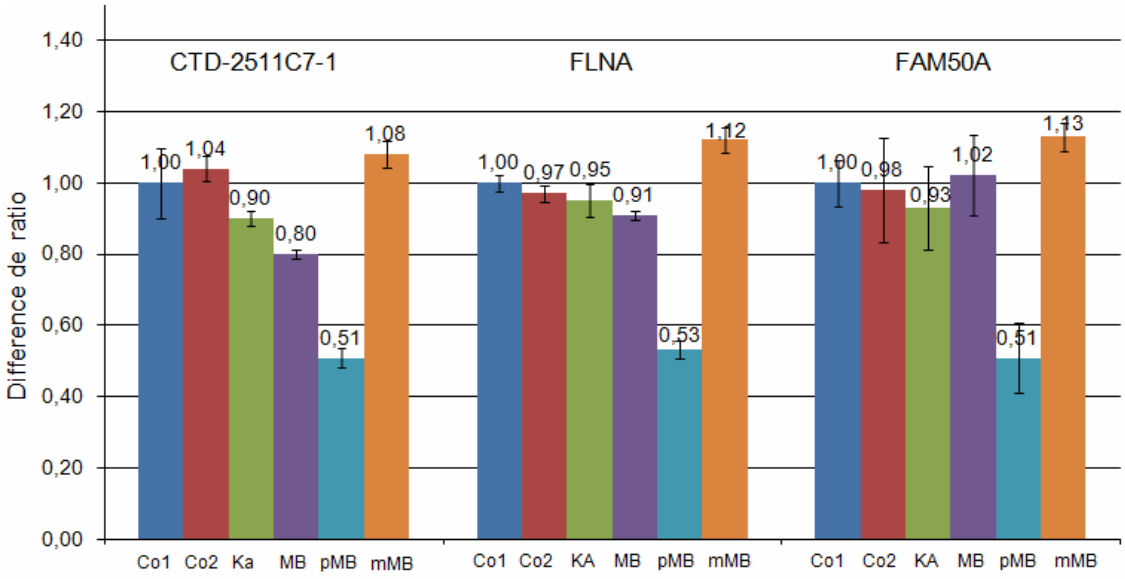


Figure 26 : Résultats de la Q-PCR quantitative aux niveaux de trois loci dans le clone CTD-2511C7.

CTD-2511C7-1, *FLNA* and *FAM50A* ne présente pas de variation significative en nombre de copies entre la patiente contrôle 1(Co1), la patiente contrôle 2 (Co2), les patientes Ka et MB. pMB et mMB sont respectivement le père et la mère de la patiente MB. Les essais ont été répétés trois fois. Des amorces situées au niveau des régions uniques du gène *WWOX* (cartographié sur le chromosome 16) ont été utilisées pour la référence interne dans la méthode des $\Delta\Delta$Ct.

L'analyse de l'ADN de 18 filles Aicardi par une puce à CGH possédant 1840 clones spécifiques du chromosome X, n'a révélé aucune anomalie délétère au seuil de résolution théorique de 82 kb.

Ce travail à fait l'objet d'une publication dans la revue European Journal of Medical Genetics ci- jointe.

Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
European Journal of Medical Genetics xx (2007) 1–6
EUROPEAN JOURNAL OF MEDICAL GENETICS
http://www.elsevier.com/locate/ejmg

Case report

Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH

Saliha Yilmaz [a], Hervé Fontaine [a], Karène Brochet [a], Marie-José Grégoire [a], Marie-Dominique Devignes [b], Jean-Luc Schaff [c], Christophe Philippe [a], Christophe Nemos [a], John Louis McGregor [d], Philippe Jonveaux [a,*]

[a] *Laboratoire de génétique, EA 4002-IFR111, Nancy-Université University Hospital (CHU) of Nancy-Brabois, Rue du Morvan, 54511 Vandoeuvre-les-Nancy, France*
[b] *CNRS-UMR 7503, LORIA, 54506 Vandoeuvre-les-Nancy, France*
[c] *Service de Neurologie, CHU Nancy, France*
[d] *INSERM Unité 689, Hôpital Lariboisière, Paris, France*

Received 9 February 2007; accepted 21 May 2007

Abstract

Aicardi syndrome (AIC) is an uncommon neurodevelopmental disorder affecting almost exclusively females. Chief features include infantile spasms, corpus callosal agenesis, and chorioretinal abnormalities. AIC is a sporadic disorder and hypothesized to be caused by heterozygous mutations in an X-linked gene but up to now without any defined candidate region on the X chromosome. Array based comparative genomic hybridisation (array-CGH) has become the method of choice for the detection of microdeletions and microduplications at high resolution. In this study, for the first time, 18 AIC patients were analyzed with a full coverage X chromosomal BAC arrays at a theoretical resolution of 82 kb. Copy number changes were validated by real-time quantitation (qPCR). No disease associated aberrations were identified. For such conditions as AIC, in which there are no familial cases, additional patients should be studied in order to

* Corresponding author. Tel.: +33 3 83 15 37 71; fax: +33 3 83 15 37 72.
E-mail address: p.jonveaux@chu-nancy.fr (P. Jonveaux).

1769-7212/$ - see front matter
doi:10.1016/j.ejmg.2007.05.006

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

identify rare cases with submicroscopic abnormalities, and to pursue a positional candidate gene approach.

Keywords: Aicardi syndrome; Array-CGH; X chromosome

1. Introduction

Aicardi syndrome (MIN 304050) was initially described as a congenital abnormality with a triad of total or partial agenesis of the corpus callosum, typical chorioretinal lacunae, and infantile spasms [1]. However, the spectrum of AIC seems broader than previously defined with a small proportion of the affected girls with moderate retardation. Additional features of the condition related to developmental defects of the central nervous system, eye, and skeletal system have been reported [2]. The brain malformation is complex with cortical polymicrogyria and migration abnormalities, often cystic formations and choroid plexus papillomas; the eye anomalies, often feature coloboma in addition to the lacunae, and focal seizures rather spasms, are common. In addition, it has recently been reported that AIC has a distinctive facial phenotype including a prominent premaxilla, upturned nasal tip, decreased angle of the nasal bridge, and sparse lateral eyebrows [3]. All Aicardi syndrome cases known to date are sporadic, and except for one isolated pair of sisters [4], no familial occurrence has been described. AIC presents only in females except three reported males with a confirmed diagnosis and a 47,XXY karyotype [5,6]. It is believed that AIC is an X-linked dominant disorder in females with early embryonic lethality in the hemizygous males [7]. As the Microphthalmia with Linear Skin defects syndrome (MLS) is associated with an Xp22.31 deletion and shares some symptoms with AIC, exhaustive sequencing and deletion studies were performed in AIC patients. No genomic abnormality was identified in AIC patients [8]. Many X-linked mental retardation genes have been identified by mapping chromosomal aberrations such as inversions, deletions, and translocations. Array based comparative genomic hybridisation (CGH) is a new powerful approach to detect submicroscopic chromosomal abnormalities too subtle for traditional cytogenetic techniques. The current study describes the analysis of a set of 18 AIC patients screened with a full coverage array for deletions and duplications on the X chromosome.

2. Patients and methods

Eighteen AIC patients were included in this study through the Association AAL-Syndrome Aicardi (http://www.aicardi.info). The diagnosis on all individuals was confirmed by review of medical records, neuroimaging studies and ophthalmological examination. Thirteen patients exhibited the classic features of AIC, including agenesis of the corpus callosum, chorioretinal lacunae, infantile spasms, severe physical and mental delay, and the remainder met criteria for diagnosis based on the expanded diagnostic criteria [6]. The ages ranged from 6 years to 25 years (mean 13.5 years; median 10.5 years). For each family the index patient showed a normal high resolution, G-banded peripheral lymphocyte chromosome analysis. Genomic DNA from AIC patients and their parents was isolated from blood samples according to standard procedures, and was purified using a QIAamp kit (QIAgen) following the supplier's instructions. The control sample (Co1) consisted of the DNA from a heterozygous woman for a partial dystrophin

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

gene deletion spanning exon 18 to exon 44. Blood samples (patients and control) were obtained according to our institutional ethical committee and after appropriate informed consent.

We used a full coverage X chromosome array provided by the Flanders Interuniversity Institute of Human Genetics [9]. This array contains 1875 validated X clones with a theoretical resolution of 82 kb. Labelling and hybridisation were performed essentially as described [9]. In brief genomic DNA was labelled with the Bioprime DNA Labelling System (Invitrogen, Carlsbad, CA) using Cy3- and Cy5-labelled dCTP's (Amersham Biosciences) as described by the manufacturer. The concentrations and labelling efficiencies were measured with the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies, Rockland, DE). For each hybridisation, 200 pmol of Cy5 and Cy3 probe each was mixed together with 100 μg Cot-1 DNA and probe preparation, pre-blocking and washing of the slide were performed as described previously [9]. Arrays were scanned with an Axon 4000B scanner (Axon Instruments, Burlingame, CA) and the acquired images were analyzed using the GenePix Pro 5.0 software (Axon Instruments). Only spots with signal intensities at least twofold above background signal intensities were included in the analysis. For each clone, a ratio of Cy5 to Cy3 fluorescent intensity was calculated. Data normalisation was performed over the mean of the spot ratios of all clones. Finally, the normalised ratio values of the duplicates were averaged and a $\log_2$ value was calculated. Color-flip hybridisations were always conducted.

Subsequent confirmation experiments for the presence of copy number changes were done by qPCR using the SYBR-green method as described elsewhere [10]. The ΔΔCT method was used for data analysis and the *WWOX* gene (located on chromosome 16) as a control gene. A standard curve was performed for each primer pair. Serial fivefold dilutions of each target (in duplicate) from 100 ng to 0.01 ng per experiment served as standard quantitation curves. Efficiency of each reaction was checked for all primer pairs and triplicate PCR amplifications were performed for each sample. Primers were selected within regions of unique sequence and designed with the Primer-Express V2.0 software (Applied Biosystems): a unique sequence at the beginning at the clone, CTD-2511C7-1f: 5′-CTGGGCACCAGTTCCGACTA-3′; CTD-2511C7r: 5′-GGACCGAAACGTGAAGTCGT-3′; in exon 3 of the *FLNA* gene, FLNAf: 5′-GTGAAC TCTGCCCGCTTCTT-3′; FLNAr: 5′-CCCTGCCAGGCATCG-3′; in exon 5 of the *FAM50A* gene, FAM50Af: 5′-GGCGTCGGCCATATCAAA-3′; FAM50Ar: 5′-TTGTGTCAAC GTCTGGGTTCT-3′. The *WWOX* gene (WWOXf: 5′-TTAACATTTCTCGGGTGAACACA-3′; WWOXr: 5′-GCC ATGAGGTGATGCCCTAA-3′) was used for normalisation.

3. Results

Using the quality criteria and threshold values as previously described [9,11], we regularly identified the deletion of the five BAC clones (RP11-142J18, RP5-1147O16, RP4-556A22, RP4-639D23 and RP11-64I1) encompassing exon 18 to exon 44 of the dystrophin gene in the DNA of the control female (Fig. 1). Four clones (RP11-388L20, RP5-1178I21 and RP11-66N11, RP11-54I20) fell beyond the 4SD threshold for patient 2 and patient 6. Six clones (RP11-441L6, RP11-23N11, CTD-2511C7, RP5-1000K24, RP11-142K4, RP11-97N5) fell between the 4SD and $\log_2(3/2) - 2X$SD thresholds for four patients. RP11-23N11 was known as a polymorphic clone filtered by the data analysis software. Three clones (RP11-97N5, RP11-388L20, RP11-441L6) were previously listed as copy number polymorphism by database of genomic variants (http://projects.tcag.ca/variation/). Eight clones (RP11-388L20, RP5-1178I21, RP11-66N11, RP11-54I20, RP11-441L6, RP11-142K4, RP5-1000K24, RP11-97N5) were controlled as normal by qPCR using probes from regions of unique sequence within the clones.

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

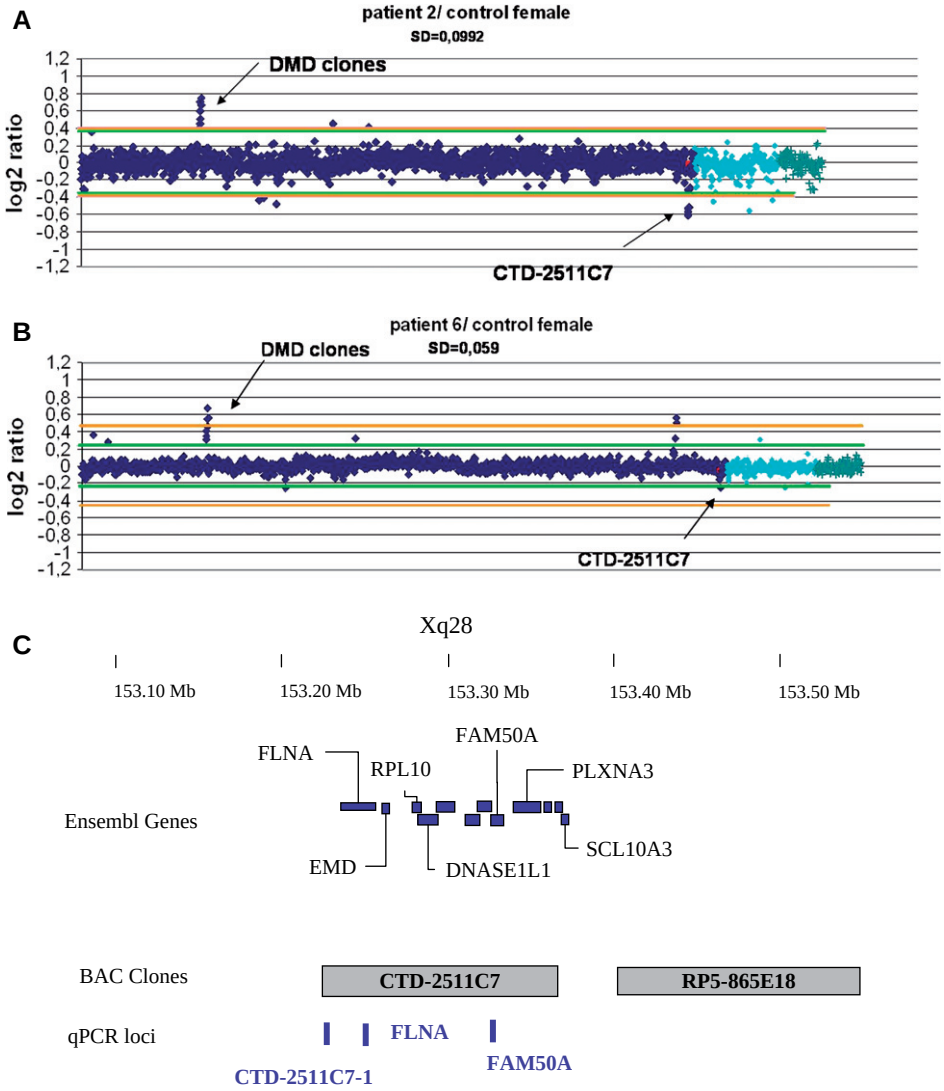


Fig. 1. Results of X-array-CGH for patients 2 (A) and 6 (B). For each panel, the *x*-axis represents the clones ordered from the Xp telomere to the Xq telomere according to their clone position in the November 2004 Ensembl freeze. The *y*-axis marks the hybridisation ratio plotted on a $\log_2$ scale. The green lines indicate the lower threshold for clone deletion or duplication ($\pm$4SD) and the red lines the upper threshold ($\log_2$ (3/2) − 2*X*SD); 98.34% of the spots were included in the analysis. The blue dots represent the autosomal control clones and the green ones represent the empty clones (C) The position of the BAC clones present on the X-array, the genes and qPCR primers is given according to the December 2006 Ensembl Release 42. The deletion of the five BAC clones within the dystrophin (DMD) gene at Xp21.1 can be seen as clones with ratios significantly higher than the normal interval since the graph was plotted as patient female to control female. Apparent deletion of the clone CTD-2511C7 is indicated by the arrow. Other clones that fell beyond the cut-off lines were not confirmed by the dye swap experiment or were controlled as normal by qPCR. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

These results indicate that the regions harbour numerous repeat sequences, like segmental duplications, leading to variance in copy numbers between individuals. Apparent deletion of a single clone CTD-2511C7 was noticed in patients 2 and 6 (Fig. 1). The intermediate $\log_2$ ratios of the clone (−0.25 in patient 2 and −0.51 in patient 6) might indicate that it was partially deleted. However, we did not confirm by qPCR the deletion using three different loci within the BAC clone, and particularly the Filamin A gene (*FLNA*) (Fig. 2). Finally, no disease associated copy number changes on the X chromosome were detected in the 18 AIC patients.

4. Discussion

Aicardi syndrome is a rare disorder that affects primarily females and is hypothesized to be caused by heterozygous mutations in an X-linked gene. The identification of the gene responsible for sporadic X-linked dominant disorders is challenging because no classical genetic mapping techniques can be applied. To date, there is no defined candidate region on the X chromosome for AIC. Previous reports have shown that array-CGH can be used efficiently to screen an entire chromosome for the presence of deletions and duplications that cannot be detected by cytogenetic analysis. This study reports for the first time the use of a full coverage X chromosome array to screen for imbalances in AIC patients. We did not find any deleterious copy number changes. Particular attention was given to the apparent deletion of the clone CTD-2511C7 in Xq28 including the *FLNA* gene. Indeed, previous functional candidate gene approach starting from a thorough clinicopathologic examination of post-mortem brain led the *FLNA* gene to a potential candidate gene. Sequencing of *FLNA* did not reveal any mutations in AIC patients [12]. Here we excluded a deletion of the *FLNA* gene in AIC. Finally, as previously suspected the CTD-2511C7 clone, with a 35 kb tandem repeat in its sequence, can be considered as a polymorphic clone [9]. However, we cannot exclude the possibility of mutations in an autosomal gene with sex-limited expression of the disorder. Further studies, in which more patients will be investigated using genome-wide tiling resolution arrays, could provide more insight in the AIC pathogenesis.

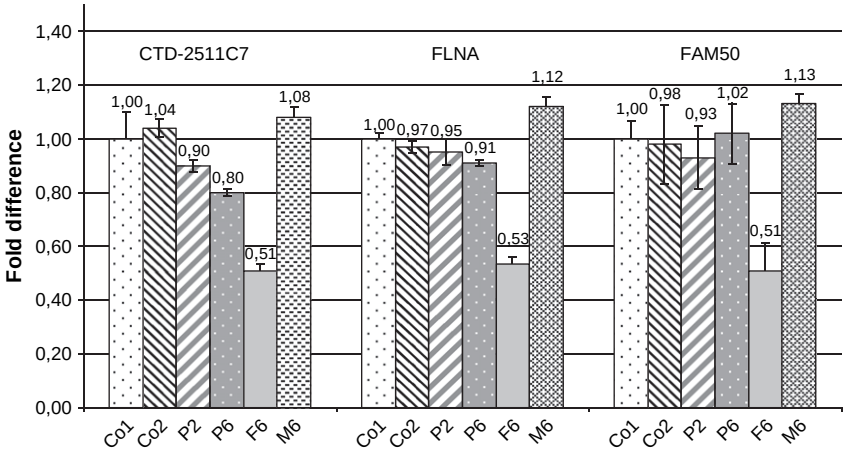


Fig. 2. Results of real-time quantitation for three different loci within the BAC clone CTD-2511C7, i.e., CTD-2511C7-1, FLNA and FAM50A showing no difference between control sample 1 (Co1), patients 2 (P2), 6 (P6), and a second control sample (Co2) corresponding to a normal female. Patients (F6) and (M6) are, respectively, the father and the mother of the patient 6.

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

Acknowledgments

We warmly thank all the patients and their families who participated in this study and the medical staff involved in the diagnosis of AIC patients. We thank Dr Marijke Bauters (Flanders Interuniversity Institute of Human Genetics) for helpful discussions and Dr Francis Martin (INRA, Champenoux). We are also grateful to Conseil Régional de Lorraine, Communauté Urbaine du Grand Nancy, Dr Bihain (JE 2482, Laboratoire médecine et thérapeutique moléculaire) and the Association AAL-Syndrome d'Aicardi for their valuable support and contribution.

References

[1] J. Aicardi, J. Levebre, A. Lerique-Koechlin, A new syndrome: spasms in flexion, callosal agenesis, ocular abnormalities, Electroencephalogr. Clin. Neurophysiol. 19 (1965) 609–610.
[2] J. Aicardi, Aicardi syndrome, Brain Dev. 27 (2005) 164–171.
[3] V.R. Sutton, B.J. Hopkins, T.N. Eble, N. Gambhir, R.A. Lewis, I.B. Van den Veyver, Facial and physical features of Aicardi syndrome: infants to teenagers, Am. J. Med. Genet. 138A (2005) 254–258.
[4] J.A. Molina, F. Mateos, M. Merino, J.L. Epifanio, M. Gorrono, Aicardi syndrome in two sisters, J. Pediatr. 115 (1989) 282–283.
[5] I.J. Hopkins, I. Humphrey, C.G. Keith, M. Susman, G.C. Webb, E.K. Turner, The Aicardi syndrome in a 47,XXY male, Aust. Paediatr. J. 15 (1979) 278–280.
[6] J. Aicardi, Aicardi syndrome: old and new findings, Int. Pediatr. 14 (1999) 5–8.
[7] I.B. Van den Veyver, Microphthalmia with linear skin defects (MLS), Aicardi, and Goltz syndrome: are they related X-linked dominant male lethal disorders? Cytogenet. Genome Res. 99 (2002) 289–296.
[8] S.K. Prakash, R. Paylor, S. Jenna, N. Lamarche-Vane, D.L. Armstrong, B. Xu, M.A. Mancini, H.Y. Zoghbi, Functional analysis of ARHGAP6, a novel GTPase-activating protein for RhoA, Hum. Mol. Genet. 9 (2000) 477–488.
[9] M. Bauters, H. van Esch, P. Marynen, G. Froyen, X chromosome array-CGH for the identification of novel X-linked mental retardation genes, Eur. J. Med. Genet. 48 (2005) 263–275.
[10] H. Van Esch, M. Bauters, J. Ignatius, M. Jansen, M. Raynaud, K. Hollanders, D. Lugtenberg, T. Bienvenu, L.R. Jensen, J. Gecz, C. Moraine, P. Marynen, J.P. Fryns, G. Froyen, Duplication of the MECP2 region is a frequent cause of severe mental retardation and progressive neurological symptoms in males, Am. J. Hum. Genet. 77 (2005) 442–453.
[11] J.R. Vermeesch, C. Melotte, G. Froyen, S. van Vooren, B. Dutta, N. Maas, S. Vermeulen, B. Menten, F. Speleman, B. de Moor, P. van Hummelen, P. Marynen, J.P. Fryns, K. Devriendt, Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis, J. Histochem. Cytochem. 53 (2005) 413–422.
[12] I.B. Van den Veyver, P.P. Panichkul, B.A. Antalffy, Y. Sun, J.V. Hunter, D.D. Armstrong, Presence of filamin in the astrocytic inclusions of Aicardi syndrome, Pediatr. Neurol. 30 (2004) 7–15.

Please cite this article in press as: S. Yilmaz et al., Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, Eur. J. Med. Genet. (2007), doi:10.1016/j.ejmg.2007.05.006

Bien que pour la majorité des pathologies génétiques, la proportion de microdélétions et microduplications de la totalité du gène reste faible par rapport aux variations délétéres d'un ou quelques nucléotides, la recherche de micro remaniements reste une méthode efficace de clonage positionnel, puisqu'elle peut réduire considérablement la région candidate. Les puces génomiques n'ont pas révélé de perte, ou de gain de matériel génétique au seuil de résolution théorique de 82Kb. Le succès de cette approche classique dépend finalement du nombre de cas étudiés et de la probabilité de trouver un cas avec une délétion (ou duplication détectable au seuil de résolution de la technique utilisé). Nous avons décidé d'enrichir les stratégies de clonage de gènes candidats par d'autres méthodologies. Une approche fonctionnelle d'étude du transcriptome a donc été envisagée. Cette approche vise à sélectionner les gènes dont l'expression diffère entre les filles Aicardi et des témoins. Outre la sélection de gènes candidats impliqués dans le syndrome, cette approche est surtout vouée à l'identification des fonctions biologiques dérégulées chez les patientes Aicardi.

5. Etude du transcriptome

5.1. Plan d'étude mis en œuvre

Notre étude a été conduite en fonction de la disponibilité du matériel biologique. Une première étude (AS1) a porté sur 3 familles à partir d'ARN extraits de lignées lymphoblastiques initialement disponibles au laboratoire de génétique. Puis une seconde étude transcriptomique (AS2) a porté sur 10 familles Aicardi et cette fois sur des ARN extraits de sang total. Un plan expérimental avec une référence commune nous a semblé le plus adapté à notre étude et donc a été utilisé pour AS1 et AS2. Les expériences des puces transcriptomiques (AS1 et AS2) ont été réalisées sous la direction de John Louis Mcgregor,

directeur de l'équipe génomique au Thrombosis Research Institute (http ://www.tri-london.ac.uk/). Les données de l'étude AS1 ont été soumises à la base de données ArrayExpress (http ://www.ebi.ac.uk/microarray-as/aer/#ae-main[0]) par le standard d'annotation MIAME (Minimum Information About a Microarray Experiment) conformément aux recommandations de MGED (Microarray Gene Expression Data Society). MGED est une organisation internationale composée de biologistes et bio-informaticiens dont le but est la standardisation et le partage des données issues des expériences de génomique fonctionnelle et de protéomique. Les données de notre étude AS1 sont consultables depuis le premier janvier 2007 sous le numéro d'accession E-MEXP-725 et le nom 'genetics-LBAS1'.

Les résultats des études de puces transcriptomiques seront présentées de la façon suivante : Avant de débuter les expériences d'analyse globale du transcriptome, nous allons au préalable poser les hypothèses de travail (paragraphe 5.2). Une première étude sera tout d'abord présentée (paragraphe 5.3). Cette étude pilote a été réalisée sur des ARN extraits de 3 lignées lymphoblastiques (AS1) et a conduit à la sélection des gènes signatures (sur/sous régulés) sur le chromosome X (paragraphe 5.3.1), à leur validation par Q-PCR (paragraphe 5.3.2) et au séquençage des 6 gènes les plus intéressants (paragraphe 5.3.3). Une nouvelle version (V3.0) du **logiciel Luminator Rosetta** a par la suite permis une comparaison directe des transcriptomes des filles AIC avec celui de leur mère. Ainsi, les fonctions biologiques dérégulées ont été analysées (paragraphe 5.3.4) par le logiciel **EASE (Expression Analysis Systematic Explorer)**. Une seconde étude (AS2) a été menée sur 9 couples fille versus mère (paragraphe 5.4). Une étude des fonctions biologiques dérégulées dans chacun des couples fille versus mère (étude par famille) sera d'abord présentée (paragraphe 5.4.1). Une analyse des résultats d'expression par **ANOVA (Analyse Of Variance)** a également été effectuée (paragraphe 5.4.2). Les premiers résultats d'une étude *in silico* à l'aide du prototype **ACGR (Approach for Candidate Gene Retrieval)** termineront cette session (paragraphe 5.5).
Ainsi, durant ces études transcriptomiques, 4 logiciels (ou méthodes d'analyse) ont été utilisés. Le logiciel luminator a servi à la sélection des gènes différentiellement exprimés

(gènes signatures), EASE a été utilisé pour effectuer des regroupements en classes fonctionnelles des gènes signatures. ANOVA, est un outil statistique implémenté dans le logiciel Luminator, a permis la sélection des gènes signatures influencés par le facteur maladie dans l'étude AS2. Le prototype ACGR est un logiciel mise en place au cours de mon travail de thèse pour la recherche de gènes candidats par une méthode *in silico.* On définit ainsi plusieurs valeurs de significativité. Une valeur p de Luminator. Cette valeur définit la significativité d'une variation de ratio (« fold change ») d'un gène sur une lame. Un indicateur est également calculé par le logiciel EASE (score EASE) il indique la significativité de sur représentation pour de chaque catégorie.

5.2. Hypothèses de travail

Nous avons émis l'hypothèse que les patientes AIC auraient une mutation sur un gène à l'origine d'une perte d'expression de l'allèle délétère (haploinsuffisance) avec d'éventuelles conséquences sur la régulation des voies biologiques auxquelles il contribue. Les mères de phénotype normal, seraient indemnes de cette mutation. Il a été récemment décrit [218] que le niveau de base d'expression de nombreux gènes est une caractéristique héritée. Nous avons donc envisagé de comparer le transcriptome des patientes AIC avec celui de leur mère afin de mettre en évidence d'éventuelles différences d'expression en rapport avec la maladie. Ces gènes signatures pourraient nous guider soit directement aux gènes candidats soit indirectement via des voies biologiques dérégulées et partagées entre les patientes AIC.

Les manifestions cliniques du syndrome d'Aicardi sont essentiellement le reflet d'atteintes cérébrale et rétinienne, cependant n'ayant pas accès à ces tissus, nous avons formulé l'hypothèse que le gène responsable est exprimé dans les cellules sanguines et/ou que les voies biologiques dérégulées par les mutations du gène candidat sont identifiables dans les cellules sanguines. L'étude du transcriptome a ainsi été réalisée sur des lymphocytes. Il est cependant possible que l'expression du gène soit tissu spécifique. Le gène *MECP2* responsable du syndrome de Rett (encéphalopathie dominante liée à l'X) a toutefois une expression ubiquitaire.

Par ailleurs, les gènes candidats doivent être présents sur la puce utilisée et la modification d'expression détectable au seuil de résolution de la technique.

5.3. Etudes sur des lignées lymphoblastiques (AS1)

Cette étude concerne trois patientes Aicardi (CC, BM et HK) et leur mère respective (FM, RL, HAK) par rapport à un pool de 10 témoins femmes. Une puce deux couleurs Agilent Human1A(V2) a été utilisée. Elle contient environ 18 000 gènes (22 000 oligonucléotides). Six hybridations ont été réalisées (3 pour les filles AIC, 3 pour leur mère), en hybridation compétitive avec le pool d'ARN de 10 témoins marqué en Cy5. Compte tenu du faible nombre d'échantillons (3 patientes AIC), les outils statistiques ne sont pas utilisables pour la sélection des gènes signatures (sous/sur-régulés). Nous avons donc sélectionnés les gènes signatures par un seuil de valeur p (« p- value ») de 1,0E-02 et à une variation de ratio (« fold change ») de 1, 5.

Deux axes majeurs ont été suivis. Le premier a pour but de sélectionner directement les candidats potentiels du syndrome d'Aicardi (paragraphe 5.3.1) tandis que le but du second axe (analyse fonctionnelle) est de mieux cerner les mécanismes et fonctions biologiques impliquées dans la maladie (paragraphe 5.3.2). Le schéma 27 de la page suivante résume les étapes de l'analyse de l'étude AS1.

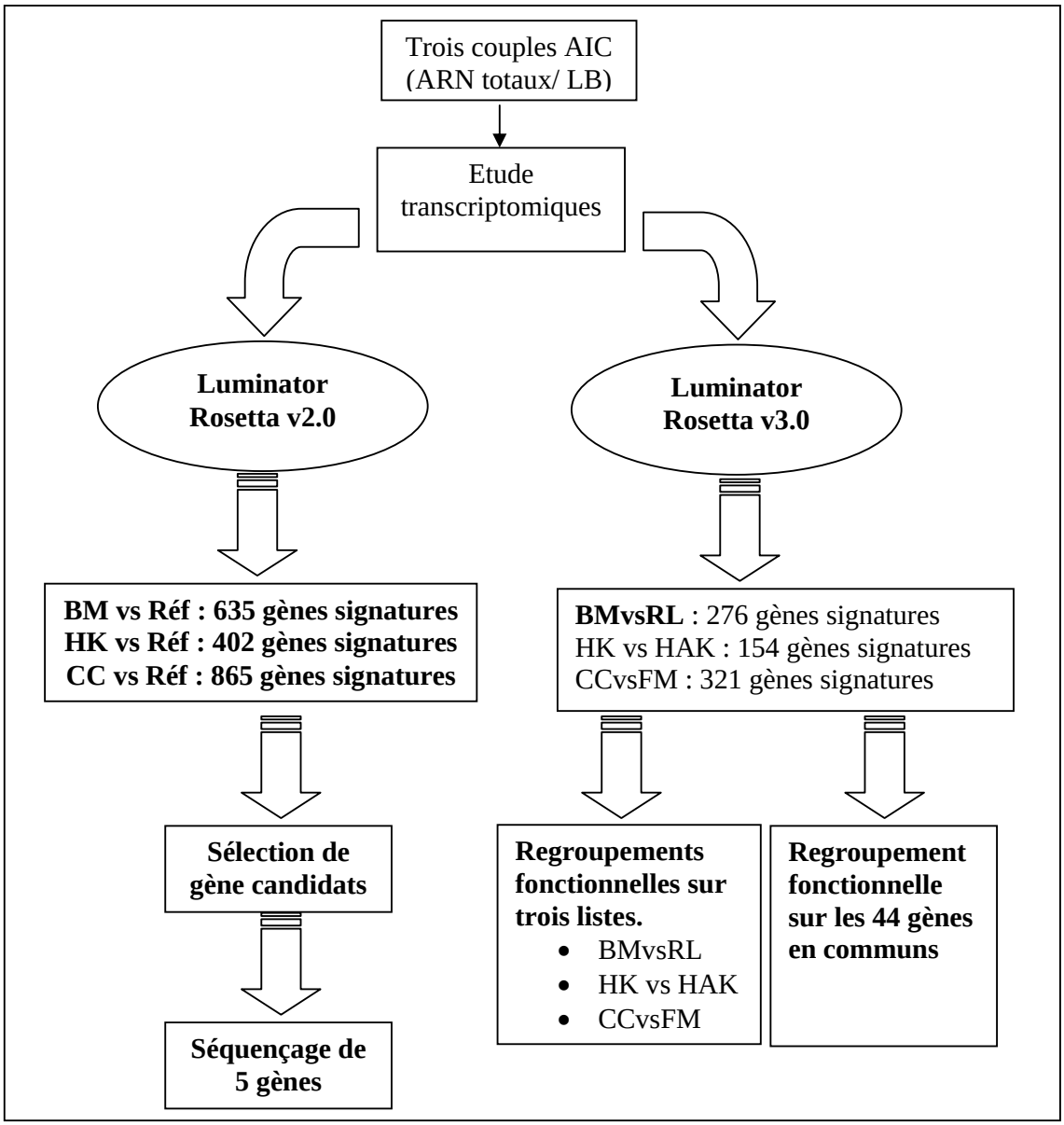


Figure 27 : Schémas récapitulatif de l'étude AS1

Le schéma décrit les différentes étapes de l'analyse de l'étude AS1. L'étude a été effectuée chez trois patientes AIC, sur des ARN totaux extraits de lignées lymphoblastiques. L'étude à été menée à l'aide des puce transcriptomique A1(V2) de 22K. La version V2.0 du logiciel Luminator Rosetta a permis la sélection de 3 listes de gènes signatures pour les trois filles vs Ref. A partir de ces 3 listes, les gènes signatures cartographiés sur le chromosome X ont été sélectionnés (paragraphe 5.3.1). Parmi les gènes variant dans leur expression (vérifié par Q-PCR), 6 gènes ont fait l'objet d'un séquençage.
La version V3.0, avec la fonctionnalité « re-ratio » permet une comparaison directe fille vs mère. Ainsi, 3 listes de gènes signatures ont été sélectionnées. Une analyse fonctionnelle à l'aide du logiciel EASE a été entreprise. Une première analyse a été effectuée indépendante sur chacune des trois listes : analyse fonctionnelle fille vs mère (paragraphe 5.3.2.1). Puis une seconde analyse fonctionnelle a consisté au regroupement fonctionnel des gènes signatures en commun dans AS1 (paragraphe 5.3.2.2).

5.3.1. Les gènes signatures du chromosome X

5.3.1.1. Choix des gènes candidats

Le gène candidat « idéal » recherché serait dérégulé chez les trois filles Aicardi et cartographié sur le chromosome X. Au cours des premières analyses il ne nous a pas été possible de faire une comparaison directe des filles Aicardi par rapport à leur mère. En effect, la version V2.0 de luminator ne dispose pas de la fonction « re-ratio » : cette fonction permet à partir de deux lames (fille cy3 versus référence cy5) et (Mère cy3 versus référence cy5), de calculer les ratios d'expression pour une lame virtuelle fille versus mère. Les premières analyses sont donc orientées vers une comparaison entre les filles versus la référence. La sélection des gènes signatures a été réalisée en fixant un seuil de significativité p à 1,0E-02. Cette valeur p indique la probabilité que la différence entre 2 valeurs ne soit pas significative.

Nous avons recherché les gènes cartographiés sur l'X et différentiellement exprimés chez les filles Aicardi par rapport à la référence. Trente six gènes cartographiés sur l'X et différentiellement exprimés ont ainsi été identifiés (tableau 9).

Certains gènes ont ainsi été sélectionnés à partir des résultats d'une seule puce d'expression (pour une seule fille). Nous sommes partis de l'hypothèse que, selon la mutation, l'ARNm pouvait être plus ou moins déstabilisé. Ainsi, la différence d'expression ne serait pas forcément visible pour toutes les filles. Les gènes sur et sous régulés ont été sélectionnés par le logiciel luminator. Le seuil de p est fixé à =1,0E-02. Un filtre basé sur la cartographie physique est appliqué pour ne garder que les gènes de l'X ; et, d'un point de vue pratique le nombre de gènes sur l'X restant gérable, un filtre plus stringent n'a pas été nécessaire.

	CC		BM		HK	
Nom du gène	**Amplitude des ratios**	**Valeur p**	**Amplitude des ratios**	**Valeur p**	**Amplitude des ratios**	**Valeur p**
ABCB7; ATP-binding cassette, sub-family B (MDR/TAP), member 7			1,6	1,00E-02		
ASMT; acetylserotonin O-methyltransferase	-4,4	1,90E-04	-2,4	2,00E-06	-4,4	3,50E-05
BTK; Bruton agammaglobulinemia tyrosine kinase	1,9	8,70E-04				
CD99 ; CD99 antigen	1,4	1,00E-02			1,6	9,90E-04
CD99L2 ; CD99 antigen-like 2					-1,5	3,90E-03
CTAG1B ; cancer/testis antigen 1B			13,3	7,70E-24		
CXCR3; chemokine (C-X-C motif) receptor 3	-1,9	3,60E-05				
EBP; emopamil binding protein (sterol isomerase)	1,6	3,30E-06	1,7	1,90E-04		
F9; coagulation factor IX (plasma thromboplastic component, hemophilia B)	4,3	3,00E-13				
FAM3C; family with sequence similarity 3, member C					-1,6	3,80E-05
H2AFB3; H2A histone family, member B3	-2	5,40E-07	-1,9	3,10E-06		
IL3RA ; interleukin 3 receptor, alpha (low affinity)	1,7	5,90E-04	1,8	3,40E-04	-2,6	1,30E-04
ITM2A; integral membrane protein 2A	-1,8	2,10E-03				
KIAA1280 ; KIAA1280 protein	-2	1,10E-04				
LOC340529 ; hypothetical protein LOC340529	-1,6	5,30E-04	-1,4	1,00E-02		
MGC27005 ; hypothetical protein MGC27005					11,3	7,60E-07
MPP1 ; membrane protein, palmitoylated 1, 55kDa	2,1	4,20E-06				
MSL3L1; male-specific lethal 3-like 1 (Drosophila)	-2,1	1,50E-04	-1,7	1,00E-02	1,9	3,00E-05
MST4; Mst3 and SOK1-related kinase	1,5	1,00E-02	1,5	1,20E-10		
MTCP1 ; mature T-cell proliferation 1	1,5	1,20E-04	1,8	2,80E-04		
NGFRAP1; nerve growth factor receptor (TNFRSF16) associated protein 1	2,8	1,90E-10	2,6	1,40E-08		
NSBP1 ; nucleosomal binding protein 1	1,8	5,10E-09	1,4	9,90E-04	1,3	1,00E-02
PCDH11Y ; protocadherin 11 Y-linked			-2	1,00E-02		
PIM2 ; pim-2 oncogene	-4	1,10E-14	-2,2	4,20E-07		
PIR; pirin (iron-binding nuclear protein)					-1,5	1,00E-02
PLP2; proteolipid protein 2 (colonic epithelium-enriched)	1,5	1,00E-02	1,6	3,50E-03		
PLXNB3 ; plexin B3	-1,7	7,80E-05	-1,5	1,00E-02		
PORCN ; porcupine homolog (Drosophila)	-1,6	7,30E-04	-1,5	7,24E-04		
PRDX4 ; peroxiredoxin 4	-1,5	9,50E-03			-1,5	1,00E-02
RPS4X; ribosomal protein S4, X-linked	1,5	1,00E-02	1,5	5,00E-03		
SAT ; spermidine/spermine N1-acetyltransferase			1,6	2,40E-04		
SMS ; spermine synthase	1,5	6,20E-08	1,5	4,00E-03		
SSR4 ; signal sequence receptor, delta (translocon-associated protein delta)	-1,6	9,50E-04				
STAG2 ; stromal antigen 2	2	1,70E-04				
SYN1 ; synapsin I	-2,3	1,30E-08	-2	8,90E-07		
TCEAL4; transcription elongation factor A (SII)-like 4			1,6	9,80E-04		

Tableau 9 : Gènes signatures de l'étude AS1 cartographiés sur le chromosome X.

5.3.1.2. Validation des gènes candidats de l'X

La PCR Quantitative (Q-PCR) est connue pour être quantitativement plus précise que les puces transcriptomiques bien que beaucoup plus onéreuse par gène. Elle est utilisée en complément des puces transcriptomiques. Les puces servent au criblage et la PCR quantitative à la validation des résultats. Ainsi, lorsque les résultats de Q-PCR vont dans le même sens que ceux de la puce, ils confirment les résultats de cette dernière, permettant ainsi leur validation. Ainsi, les gènes révélés comme effectivement dérégulés par Q-PCR et situés sur l'X, deviennent des candidats pour le séquençage systématique à la recherche de mutations délétères chez les patientes AIC.

Cette sélection a été guidée par une appréciation des connaissances biologiques disponibles dans les bases de données publiques et nous avons principalement utilisé le NCBI (http ://www.ncbi.nlm.nih.gov/). Le gène est il déjà impliqué dans une maladie ? (OMIM). Quelle sont les fonctions biologiques dans lesquelles il intervient ? (GENE). Quels sont les tissus où il est exprimé ? (UniGene). Nous avons ainsi réduit la liste des gènes à étudier par PCR quantitative à 15 gènes candidats : *ASMT, CTAG1B, CXCR3, ITM2A, KIAA1280, MST4, NGFRAP1, NSBP1, PIR, PLXNB3, PORCN, RPS4X, SSR4, SYN, TCEAL4.*

Par exemple, le gène *ASMT* est cartographié en Xp22.3. La protéine ASMT à une fonction méthyltransférase. C'est une hormone neuropeptidique impliquée dans la dernière réaction du métabolisme du tryptophane (acide aminé essentiel). Le gène est exprimé dans le cerveau et la rétine. D'après les résultats des puces à ADN, c'est le seul gène sous régulé chez les trois filles Aicardi.

	Q-PCR			Puces à ADN					
	Amplitude des ratios de Q-PCR			CC		BM		HK	
Nom du gène	**CC**	**HK**	**BM**	**Amplitude de ratio**	**p**	**Amplitude de ratio**	**p**	**Amplitude de ratio**	**p**
ASMT	***-1,4***	***-7,4***	***-2,2***	***-4,4***	***1,9E-04***	***-2,4***	***2,0E-06***	***-4,4***	***3,5E-05***
CTAG1B	0	0	0			13,3	7,70E-24		
CXCR	1,2	2,1	1,3	-1,9	3,6E-05				
ITM2A	1,3	-3,0	-1,2	-1,8	2,10E-03				
KIAA1280 (WWC3)	**-5,6**	**-1,8**	**-1,7**	**-2,0**	**1,1E-04**				
MST4		**2,4**	**1,2**	**1,5**	**1,0E-02**	**1,5**	**1,20E-10**		
NGFRAP1	**2,9**		**1,6**	**2,8**	**1,9E-10**	**2,6**	**1,4E-08**		
NSBP1	**9,4**	**1,4**	**5,1**	**1,8**	**5,1E-09**	**1,4**	**9,9E-04**	**1,3**	**1,0E-02**
PIR	4,2	-1,8	1,2					-1,5	1,0E-02
PLXNB3		**-1,4**	**-4,7**	**-1,7**	**7,8E-05**	**-1,5**	**1,00E-02**		
PORCN	6,0	-3,0	1,5	-1,6	7,3E-04	-1,5	7,24E-04		
RPS4X	**4,5**		**2,8**	**1,5**	**1,00E-02**	**1,5**	**5,00E-03**		
SSR4	**-1,7**	**-5,9**	**-1,5**	**-1,6**	**9,5E-04**				
SYN1	3,9	-2,0	-10,9	-2,3	1,3E-08	-2,0	8,9E-07		
TCEAL4	7,0	-1,6	1,1			1,6	9,8E-04		

Tableau 10 : comparaison des résultats de RTQ-PCR et puces à ADN.

Quinze gènes cartographiés sur le chromosome X et dérégulés d'après les analyses du transcriptome par puces chez au moins une fille Aicardi sont testés par Q-PCR. Le gène CTAG1B n'est pas retrouvé exprimé par Q-PCR.

Les gènes en gras montrent une concordance entre les résultats des puces à ADN et la Q-PCR chez les patientes.

Pour 70% des gènes (17/24), les deux technologies donnent des résultats concordants. Les résultats sont en accord avec les publications sont rapportées une concordance de l'ordre de 70 à 75 % [219]. Les gènes *ASMT, KIAA1280, NGFRAP1, NSBP1, PLXNB3, RPS4X et SSR4* sont dérégulés chez trois filles Aicardi. Ils sont tous exprimés dans le cerveau et la rétine. Ils sont donc des bons candidats. Nous pouvions supposer que cette liste constitue une série de gènes intervenant probablement à différents endroits de la (des) voie biologique(s) impliquée(s) dans la pathologie, restait à savoir si elle contenait le gène responsable lorsqu'il est muté du syndrome d'Aicardi.

5.3.1.3. Séquençage des gènes candidats

Une liste de gènes candidats à séquencer a été dressée. La classification de ces gènes a été mise en place grâce aux informations fonctionnelles recueillies par le criblage des bases de données biologiques publiques pour chacun des gènes candidats. Les gènes *ASMT*, *MST4*, *NSBP1*, *PLXNB3* et *SYN1* se sont révélés être les plus intéressants pour les raisons suivantes. *ASMT* est le seul gène sous régulé chez toutes les filles AIC. Il a été présenté comme candidat pour l'épilepsie [220]. *MST4* a un rôle dans l'activation de la voie des MAPK (mitogen-activated protein kinase) notamment durant le réarrangement du cytosquelette. D'après les résultats issus des puces, *NSBP1* est sur-régulé chez les trois filles Aicardi. La protéine intervient dans la régulation de la transcription. Plus exactement elle est un activateur de la transcription de part sa capacité de fixation à la chromatine. *PLXNB3* code une proéine avec une activité de récepteur pour la semaphorine 5A qui, elle est impliquée dans la neurogenèse. Le gène *SYN1* est impliqué dans la fixation de l'actine, dans des processus de sécrétion de neurotransmetteur et dans la transmission synaptique. Il a clairement été impliqué dans des formes d'épilepsies familiales [221] donc il a été testé bien que les résultats de la Q-PCR n'étaient pas cohérents entre les trois filles.

Nous avons donc réalisé le séquençage des parties codantes de ces gènes et des jonctions intron/exon qui comprennent les régions intervenant dans l'épissage. Les analyses de séquençage effectuées sur les 19 exons de *ASMT* révèlent un polymorphisme intronique (c.625-133A>C) chez les deux patientes HK et BM. Ces variants sont également retrouvés sur l'ADN de leur mères HAK et FM respectivement. *PLXNB3* est constitué de 36 exons. Le polymorphisme c.2670-59_2670-60ins28 qui est présent chez CC l'est aussi chez son père. Le variant nucléotidique c.4653T>C décrit dans les bases de données comme variant normal (rs5987155) est présent chez les trois filles. Aucune mutation délétère n'a été trouvée pour les gènes candidats testés chez les filles AIC.

Ces gènes sont *a priori* écartés en tant que candidats du syndrome d'Aicardi en l'absence de mutations pathogènes dans la séquence codante sur l'ADN des filles Aicardi. Notons bien que nous avons effectué un criblage en estimant que la majorité des mutations délétères se situe dans les régions codantes des gènes. Ainsi, les promoteurs et régions régulatrices n'ont pas été criblés. Cependant, nous ne pouvons pas exclure que des mutations interviennent dans ces régions. De plus, les expériences de séquençage ne détectent pas les grands réarrangements touchant un ou plusieurs exons et donc nous ne pouvons pas écarter non plus cette possibilité. Notons que nous avons menée cette stratégie jusqu'au

séquençage avant d'avoir effectué les analyses AS2 et donc nous n'avons pas attendue l'obtention de leurs résultats pour les confronter à celles obtenue avec l'analyse AS1. Ceci explique le séquençage de gènes candidats non retrouvées dans l'analyse transcriptomique AS2.

Ce premier axe de recherche des puces à ADN visait l'identification d'un candidat idéal du syndrome d'Aicardi. Les résultats ne révèlent pas le gène candidat du syndrome d'Aicardi. Cependant, 8 gènes (*ASMT, KIAA1280, NGFRAP1, NSBP1, PLXNB3, RPS4X* et *SSR4*) on été identifiés comme dérégulés chez les filles Aicardi et ceci aussi bien par les puces à ADN que la RTQ-PCR. Parmi ceux-ci, le criblage des bases de données biologiques à la recherche d'informations sur ces gènes nous a permis de sélectionner les gènes *ASMT, MST4, NSBP1, PLXNB3 et SYN1 qui* ont fait l'objet d'un séquençage de leur partie codante. Cependant, aucun variant délétère n'a été détécté chez les filles AIC testées.

Le but du second axe d'analyse des résultats obtenus par puces transcriptomiques (AS1) est de comprendre les mécanismes et fonctions biologiques impliqués dans la maladie. Une analyse fonctionnelle des résultats des puces a donc été entreprise. Cette étude est réalisée sur des gènes signatures et avec le logiciel EASE.

5.3.2. Analyse par regroupement fonctionnel dans l'étude AS1

D'une manière générale un groupe de gènes est dit « sur-représenté » en une fonction biologique si la proportion de gènes dans ce groupe, connus pour être impliqués dans cette fonction biologique, excède le nombre attendu par le hasard. Ainsi, si pour 23 gènes signatures à étudier 20 font partie de la catégorie « Chromatine », la fobs = 20/23 = 87%. Sur l'ensemble des gènes de la puce A1v2 on sait que 171 gènes sur 5675 appartiennent à la catégorie fonctionnelle étudiée (fref = 171/5675 = 3%). Nous aurons donc dans ce cas une sur représentation de la catégorie fonctionnelle « Chromatine ». EASE calcule la probabilité

d'observer une fréquence fobs par le hasard (compte tenu de la valeur de fref). Plus cette probabilité est faible, plus le groupe de gènes signatures est «fonctionnellement enrichi ». Nous parlerons dans ce cas de catégorie sur-représentée. Les gènes signatures sont soumis au logiciel. Ils sont classés dans des catégories fonctionnelles des trois systèmes composant Gene Ontology que sont : Processus Biologique (BP : Biological Process), Fonction Moléculaire (MF : Molecular Fonction) et Composant Cellulaire (CC : Cellular Component), Gene Ontology (GO) étant un vocabulaire contrôlé et structuré. Un score est affecté à chaque catégorie sur représentée. Ce score est calculé par rapport à une liste référence. La liste référence que nous avons utilisée est dans le cas des études pour AS1 la liste de gènes issue de la puce A1V2 et pour l'étude AS2, la liste des gènes issue de la puce « whole human genome ».

Dans le cadre de notre étude, il est intéressant de connaitre les catégories sur-représentées c'est-à-dire affectées chez les filles AIC.

Avec la version 3.0 de luminator, il est possible de comparer directement les filles AIC à leur mère respective. La sélection des gènes candidats par le logiciel Luminator s'est faîte avec un seuil d'amplitude de ratio égal à 2 et une valeur p égal à 1,0E-02. Nous avons effectué deux sortes d'analyse. La première consiste à faire une étude fonctionnelle des gènes signatures à l'intérieur de chaque couple fille/mère (ce que nous appelons analyse par famille) et par la suite de comparer les fonctions biologiques sur-représentées entre chaque fille AIC. L'analyse fonctionnelle se fait à l'aide du logiciel EASE. Ainsi, si les gènes signatures n'ont pas été sélectionnés par un outil statistique, leur classement fonctionnel par contre a été réalisé par une approche statistique. La seconde approche consiste à rechercher les gènes signatures communs aux filles Aicardi (paragraphe 5.3.2.2). Un regroupement fonctionnel de ces gènes signatures permet de mettre en évidence les fonctions biologiques dérégulées dans les listes de gènes signatures par rapport à la totalité des gènes portés par la puce A1V2.

5.3.2.1. Analyse par famille (fille/mère) dans l'étude AS1

Rappelons que le couple BMvsRL possède 276 gènes signatures, le couple HK vs HAK 154 signatures et le couple CCvsFM 321 signatures. Rappelons également que les gènes signatures ont été sélectionnés par le logiciel luminator sur la base d'un seuil de p à 1,0E-2 et une amplitude de ratio à 2 en valeur absolue. Pour chaque couple fille *vs* mère ces gènes

signatures ont été soumis au logiciel EASE. Nous avons procédé à un regroupement fonctionnel de ces gènes signatures.

Le tableau ci-dessous est un récapitulatif du **nombre de catégorie** en fonction du score EASE dans les trois systèmes d'ontologie processus biologique, fonction moléculaire et composant cellulaire. Il résume pour chaque couple fille *vs* mère le nombre de catégories donné par le logiciel pour les gènes signatures soumis. Ce tableau permet d'avoir un aperçu du nombre de réponses pour chaque analyse puisque pour la suite des résultats seul seront représentés les catégories ayant un score EASE les plus failes (les plus significatifs). Rapellons que EASE ne donne pas les catégories pour lesquelles il n'y a pas de gène et donc me nombre de catégorie total (indicateur EASE ≤1) est variable d'une analyse à l'autre.

	Score EASE	BMvsRL Nbr Cat	HKvsHAK Nbr Cat	CCvsFM Nbr Cat
Processus Biologique	≤1	539	376	518
	<1,0E-01	275	159	231
	<1,0E-02	56	30	36
Composant cellulaire	≤1	99	90	106
	<1,0E-01	50	40	47
	<1,0E-02	9	9	10
Fonction Moléculaire	≤1	284	174	246
	<1,0E-01	115	56	90
	<1,0E-02	13	3	5

Tableau 11: Nombre de catégories en fonction du score EASE pour les couples BM vs RL, HKvsHAK et CCvsFM.

Pour le couple BM *vs* RL les 276 gènes signatures sont impliqués dans 539 catégories du système processus biologique. Parmi ces catégories seules 56 ont un score de significativité EASE inférieur à 1,0E-02.

Nbr Cat : Nombre de catégories.

Les trois tableaux suivant présentent les catégories sur représentées pour les trois couples de l'étude AS1.

Distribution fonctionnelle des gènes signatures pour le couple BM vs RL

Puce A1V2		Gènes signature		
Système	**Catégorie (gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (5901 gènes)	response to external stimulus (1025)	43	107	3.92e-008
	immune response (597)	30		3.47e-007
	response to biotic stimulus (698)	32		8.92e-007
	defense response (652)	30		2.20e-006
	response to stress (656)	26		1.78e-004
Composant Cellulaire (5675 gènes)	cytosol (269)	16	104	8.77e-005
	soluble fraction (166)	10		3.00e-003
	sarcolemma (7)	3		6.46e-003
	obsolete cellular component (262)	11		1.99e-002
	extracellular (937)	26		2.64e-002
Fonction Moléculaire (5799 gènes)	antigen binding (56)	7	105	4.52e-004
	defense/immunity protein activity (66)	7		1.09e-003
	cytoskeletal protein binding (202)	10		9.94e-003
	binding (3956)	81		3.20e-002
	phosphopyruvate hydratase activity (2)	2		3.55e-002

Tableau 12 : Classification fonctionnelle des gènes signatures pour le couple BM vs RL.

Dans l'étude AS1, la puce transcriptomique A1V2 utilisée contient 5901 gènes annotés dans le système processus biologique. Parmi ceux-ci 1025 (17%) sont inclus dans la catégorie « réponse à un stimuli externe ». Pour ce couple BM *vs* RL, 107 des 276 gènes signatures appartiennent au système processus biologique et 43 (40%) participent à la catégorie « réponse à un stimuli externe ». cette catégorie est bien sur représentée, le logiciel lui attribue le score le plus significatif de la liste (3.92e-008).

Ainsi, les scores EASE les plus significatifs sont obtenus pour la catégorie fonctionnelle « réponse aux stimuli » suivi par ses termes fils de l'arborescence de GO (figure 27) : « réponse à un stimuli externe », « réponse immunitaire », « réponse à un stimulus biologique », « réponse de défense », « réponse au stress ».

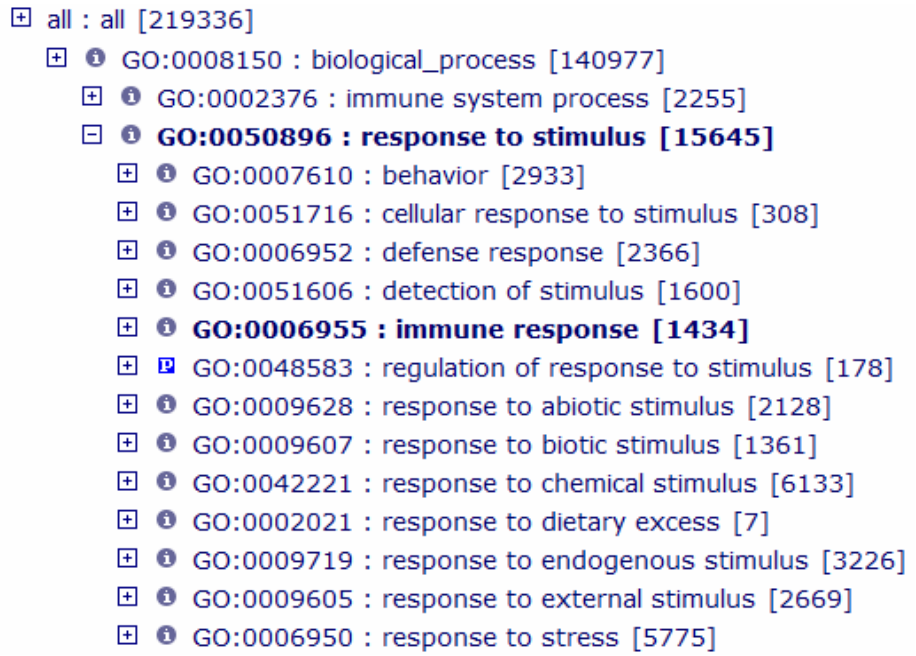


Figure 28 : Extrait du graphe de Gene Ontology représentant la catégorie (ou terme) « réponse aux stimuli » et ses termes fils.

Les gènes annotés par ces catégories inclus IGKC (immunoglobulin kappa constant), AIF1 (allograft inflammatory factor 1), ANXA1 (annexin A1), UBD (ubiquitin D), IL27 (interleukin 27).

Dans le système **composant cellulaire**, parmi les 99 catégories, 50 ont un score EASE inférieur à 1,0e-01 (tableau 11). Parmi les plus représentés figurent les catégories « cytosol », « fraction soluble » et « sarcolemme ». Dans le système **fonction moléculaire** la fonction « liaison à un antigène » semble la plus significative.

Distribution fonctionnelle des gènes signatures pour le couple HKvsHAK

Puce A1V2		Gènes signature		
Système	**Catégorie (gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (5901 gènes)	immune response (597)	20	67	1.69e-005
	defense response (652)	20		5.85e-005
	response to biotic stimulus (698)	20		1.48e-004
	humoral defense mechanism (sensu Vertebrata) (95)	7		6.20e-004
	response to external stimulus (1025)	23		1.19e-003
Composant Cellulaire (5675 gènes)	extracellular (937)	23	66	4.60e-004
	nucleosome (59)	5		4.45e-003
	multivesicular body (11)	3		6.65e-003
	lysosome (103)	5		2.98e-002
	lytic vacuole (108)	5		3.46e-002
Fonction Moléculaire (5799 gènes)	antigen binding (56)	10	63	5.13e-009
	defense/immunity protein activity (66)	10		2.30e-008
	MHC class II receptor activity (15)	3		1.08e-002
	sugar binding (61)	3		1.38e-001
	carbohydrate binding (65)	3		1.53e-001

Tableau 13 : Classification fonctionnelle des gènes signatures pour le couple HKvsHAK.

Les catégories les plus représentées du système processus biologique semblent être les mêmes que pour le couple précédemment étudié. La catégorie « réponse immunitaire » comprend chez cette patiente, les gènes AIF1 (allograft inflammatory factor 1), CXCL13 (chemokine (C-X-C motif) ligand 13 (B-cell chemoattractant)), IGKC (immunoglobulin kappa constant), TRA1 (tumor rejection antigen (gp96) 1). Notons également que pour les composants cellulaires, les gènes signatures annotés par la catégorie « nucléosome » semblent sur représentés. Cette catégorie comprend les 5 histones HIST1H2AJ (histone 1, H2aj), HIST1H4H (histone 1, H4h), HIST1H4L (histone 1, H4l), HIST1H4B (histone 1, H4b), HIST1H3B (histone 1, H3b). Ces gènes sont tous sous-régulés d'après les puces transcriptomiques pour ce couple.

Distribution fonctionnelle des gènes signatures pour le couple CCvsFM

Puce A1V2		Gènes signature		
Système	**Catégorie (gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (5901 gènes)	nucleosome assembly (68)	7	92	5.79e-004
	chromatin assembly/disassembly (72)	7		7.86e-004
	nucleobase, nucleoside, nucleotide and nucleic acid metabolism (1164)	32		8.55e-004
	chromosome organization and biogenesis (sensu Eukaryota) (79)	7		1.28e-003
	DNA replication and chromosome cycle (154)	9		2.44e-003
Composant Cellulaire (5675 gènes)	nucleosome (59)	7	95	3.91e-004
	chromosome (171)	11		4.81e-004
	chromatin (114)	9		5.48e-004
	nucleus (1782)	45		1.16e-003
	obsolete cellular component (262)	11		1.09e-002
Fonction Moléculaire (5799 gènes)	nucleic acid binding (1328)	34	93	3.10e-003
	DNA binding (1022)	28		3.67e-003
	transcription regulator activity (830)	21		3.38e-002
	transcription factor activity (602)	16		5.05e-002
	opioid receptor activity (7)	2		1.06e-001

Tableau 14 Classification fonctionnelle des gènes signatures pour le couple CCvsFM.

Dans le système processus biologique 36 termes ont un score EASE inférieur à 1,0 e-01. Ces termes peuvent être regroupés en plusieurs catégories dont la plus importante est « organisation du chromosome et biogenèse ». Il est l'ancêtre commun des termes « assemblage ou désassemblage de la chromatine », « assemblage du nucléosome ». Ces catégories comprennent les mêmes gènes histones que la catégorie « nucléosome » (composant cellulaire) : HIST1H2BK (histone 1, H2bk), HIST1H3H (histone 1, H3h), HIST2H2AA (histone 2, H2aa), HIST1H4I (histone 1, H4i), HIST1H4L (histone 1, H4l), HIST1H1E (histone 1, H1e), HIST1H2AH (histone 1, H2ah). D'après les puces génomiques ces gènes seraient sur-exprimés chez CC par rapport à sa mère FM et donc ce résultat est contraire à celui obtenue pour le couple HK vs HAK.

Distribution fonctionnelle des gènes signatures : bilan des trois couples de l'étude AS1

Les études précédentes étaient focalisées sur une analyse par famille (fille *vs* mère) sans tenir compte des autres couples. Pour effectuer un bilan et comparer les catégories sur représentées chez les trois couples, nous avons placé un seuil de score EASE = 1,0e⁻01 puis classé les catégories en fonction de leur score. Les résultats sont présentés dans le tableau suivant.

Processus biologique	HKvsHAK	CCvsFM	BMvsRL
response to pest/pathogen/parasite	**1,7E-03**	**9,3E-02**	**5,0E-04**
chromosome segregation	**3,9E-03**	**1,2E-02**	**8,6E-02**
response to stress	**1,2E-02**	**4,1E-02**	**1,8E-04**
immune response	1,7E-05	1,3E-01	3,5E-07
defense response	5,9E-05	2,0E-01	2,2E-06
response to biotic stimulus	1,5E-04	1,1E-01	8,9E-07
response to external stimulus	1,2E-03	2,2E-01	3,9E-08
chromatin assembly or disassembly	6,1E-03	3,3E-03	1,2E-01
nucleosome assembly	6,8E-03	5,8E-04	1,2E-01
chromosome organization and biogenesis (sensu Eukaryota)	1,1E-02	1,3E-03	1,7E-01
physiological process	4,5E-02	4,0E-01	2,0E-02
establishment and/or maintenance of chromatin architecture	6,1E-02	1,6E-02	4,3E-01
DNA packaging	7,4E-02	2,2E-02	2,5E-01
response to wounding	7,8E-02	2,3E-01	5,1E-02
chromosome organization and biogenesis (sensu Eukarya)	8,6E-02	2,8E-02	2,8E-01
DNA replication and chromosome cycle	9,3E-02	2,4E-03	1,4E-01
cell proliferation	1,9E-01	1,9E-02	3,5E-02
cell cycle	2,4E-01	4,9E-03	2,3E-02

Tableau 15 : Bilan des regroupements fonctionnels pour les trois couples de l'étude AS1.

Pour chacun des trois systèmes d'ontologie (processus biologique, composant cellulaire et fonction moléculaire), les catégories (termes GO) en commun aux trois filles sont retenues. Les chiffres indiquent les scores EASE pour chaque étude fille versus mère. Seize catégories du système processus biologiques sont en commun à au moins deux filles sur trois lorsque l'on applique un seuil de score EASE à 1,0E-02. Parmi elles, on retrouve le terme « réponse à un stimulus biologique » et les termes fils à celui-ci « réponse à un stimulus externe », « réponse immunitaire », «réponse de défense ». Le terme « organisation du chromosome et biogenèse » avec des termes fils tel que « assemblage et désassemblage de la

chromatine » sont des catégories de processus biologiques partagées par deux filles Aicardi sur trois.

Les catégories «réponse aux pathogènes/parasites », « ségrégation du chromosome » et « réponse au stress » sont des processus biologiques dérégulés communs aux trois filles. Les catégories « chromatine » et « nucléosome » sont les constituants cellulaires qui paraissent affectées chez les trois filles AIC.
La catégorie « Réponse aux pathogènes/parasites » inclut deux gènes : AIF1 (allograft inflammatory factor 1) et CLEC2B (C-type lectin domain family 2, member B) commun aux couples BMvsRL et HKvsHAK. Dans la catégorie « ségrégation du chromsome », le gène histone HIST1H4L (histone 1, H4l) est commun aux trois couples et HIST1H4I (histone 1, H4i) aux couples HK *vs* KAH et CC *vs* FM. Aucun de ces 4 gènes n'est localisé sur le chromosome X.

Chaque patiente AIC paraît avoir des catégories dérégulées spécifiques. Pourtant, en comparant les résultats avec différents seuils, on s'aperçoit que les catégories dérégulées sont partagées.

5.3.2.2. Regroupement fonctionnel des gènes signatures en commun dans AS1

Une analyse similaire a également été réalisée sur les gènes signatures communs aux trois couples AIC de l'étude AS1. Rappelons que, le couple BMvsRL possède 276 gènes signatures, le couple HKvsHAK 154 signatures et le couple CCvsFM 321. Cinquante six oligonuclétides (44 gènes) sont en commun à au moins deux des trois expériences. N'oublions également pas que cette comparaison entre les trois couples a été réalisée avec la version V3.0 de luminator sur les couples fille vs mère. L'étude précédente réalisée avec la version V2.0 (paragraphe 5.3.1.1) avait porté sur les études fille vs référence (pool témoins femmes).
Un gène *HIST1H4L* (histone cluster 1, H4l) est commun au trois filles Aicardi. *MSL3L1* (male-specific lethal 3-like 1) est le seul gène parmi les 44 gènes communs cartographié sur le chromosome X. Ce gène avait précédemment été séquencé à la recherche de variants délétères chez des patientes AIC. Aucune mutation délétère n'avait été détectée et donc les auteurs avait écarté *a priori* ce gène des candidats pour le syndrome AIC [76].

Un regroupement fonctionnel par EASE de ces 44 gènes (par rapport à la liste des 20000 clones de la puce Agilent Human1AV2) a été effectué. Les résultats sont présentés dans le tableau 16.

Puce A1V2		Gènes signature		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (5901 gènes)	**immune response (597)**	**8**	23	**4,60E-03**
	defense response (652)	8		7,50E-03
	response to external stimulus (1025)	10		7,98E-03
	response to biotic stimulus (698)	8		1,08E-02
	chromatin assembly or disassembly (66)	3		2,46E-02
	nucleosome assembly (68)	3		2,61E-02
	chromatin assembly/disassembly (72)	3		2,90E-02
	chromosome organization and biogenesis (79)	**3**		**3,44E-02**
	female meiosis chromosome segregation (17)	2		6,16E-02
	meiotic chromosome segregation (18)	2		6,51E-02
Composant Cellulaire (5675 gènes)	Extracellular (937)	10	23	5,66E-03
	obsolete cellular component (262)	5		1,68E-02
	Nucleosome (59)	**3**		**2,15E-02**
	Chromatin (114)	**3**		**7,12E-02**
	endoplasmic reticulum (253)	4		7,00E-02
	Nucleus (1782)	11		1,18E-01
	Chromosome (171)	3		1,41E-01
	Synapse (87)	2		2,89E-01
	extracellular matrix (144)	2		4,32E-01
	Intracellular (3688)	15		6,46E-01
Fonction Moléculaire (5799 gènes)	**antigen binding (56)**	**4**	23	**1,15E-03**
	defense/immunity protein activity (66)	**4**		**1,86E-03**
	Binding (3956)	19		1,24E-01
	sugar binding (61)	2		2,08E-01
	carbohydrate binding (65)	2		2,20E-01
	calcium ion binding (308)	3		3,28E-01
	enzyme inhibitor activity (127)	2		3,86E-01
	transcription corepressor activity (133)	2		4,00E-01
	metal ion binding (892)	5		4,45E-01
	DNA binding (1022)	5		5,60E-01

Tableau 16 : Analyse fonctionnelle de gènes signatures en commun à au moins deux filles Aicardi.

Seuls les 10 premiers scores EASE de chaque système sont indiqués. Dans le système **processus biologique**, la catégorie « réponse immunitaire » inclut les catégories « réponse de défense » et «réponse à un stimulus biologique », «réponse à un stimuli externe » ; la catégorie « organisation du chromosome et biogenèse » est le terme père de « assemblage

ou désassemblage de la chromatine » et «assemblage du nucleosome ». Dans le système **Composant Cellulaire** « nucléosome » et « chromatine » sont les termes fils de la catégorie « organisation du chromosome et biogenèse ». Dans le système **fonction moléculaire**, « ligation à un antigène » est le terme le plus significatif, il est en outre père du terme « ligation ».

Les termes « gene ontology » (GO) sur-représentés sont des indicateurs des voies biologiques dérégulées chez les filles Aicardi par rapport à leur mère saine. Il existe naturellement plusieurs causes pouvant expliquer les différences entre les filles et leur mère. Notons que cette variabilité peut avoir des causes autres que la maladie, telle que la différence d'âge, les médicaments et l'influence du génome paternel.

Cette première étude sur trois filles Aicardi a permis de mettre en évidence des gènes dérégulés chez les filles Aicardi par rapport à leur mère. Ces gènes sont : ASMT, KIAA1280, NGFRAP1, NSBP1, PLXNB3, RPS4X et SSR4. D'autre part, il apparait après analyses fonctionnelles que la réponse immunitaire, l'organisation du chromosome et la biogenèse soient des voies biologiques touchées chez les filles Aicardi par rapport à leur mère.

Cette première étude étant basée seulement sur trois patientes Aicardi, les résultats présentent des limites : nombre d'échantillon trop faible et ARN extraits de lignées lymphoblastiques. De plus, si les lignées nous ont permis de mettre en place cette approche sur le transcriptome, les conditions de culture cellulaire et l'immortalisation par le virus EBV modifient l'image « physiologique » du transcriptome. Nous avons donc souhaité étendre cette étude sur des échantillons d'ARN issus de prélèvements sanguins et non plus de lignées. (Etude AS2) Par ailleurs l'étude AS2 est enrichie en patientes (10 familles). Ceci permettra d'utiliser des outils statistiques, une option qui ne nous a pas été disponible dans « l'étude pilote ». Enfin, les puces utilisées dans cette seconde analyse transcriptomique contiennent cette fois 44 000 clones (Agilent human whole genome) permettant ainsi un criblage a priori exhaustif du transcriptome.

5.4. Etude transcriptomique sur ARN extraits de sang (AS2)

Selon les hypothèses de travail précédemment présentées (Paragraphe.5.1), cette étude s'adresse à 10 couples fille-mère différents. Les hybridations ont été réalisées à partir d'ARN issus de prélèvements sanguins. Un plan expérimental avec référence commune (pool de femmes saines) a été adopté.

Les données ont été analysées selon deux approches. La première consiste à sélectionner les gènes signatures pour 9 couples (fille *vs* mère) puis de faire une classification fonctionnelle de chacune des listes de gènes signatures. Une patiente AIC (PS) n'a pas pu être comparée à sa mère car l'ARN extrait n'était pas de qualité convenable. Cependant, les résultats pour cette patientes PS ont été inclus dans une seconde analyse. Cette seconde approche utilise les logiciels d'analyses statistiques pour la sélection des gènes signatures communs à l'ensemble des patientes Aicardi. Une approche ANOVA (Analyse Of Variance) implantée dans le logiciel Luminator a ainsi été utilisée. Les analyses pour l'étude AS2 diffèrent de celles réalisées pendant l'étude AS1 seulement pour cette seconde approche.

Le schéma 28 de la page suivante présente les étapes de l'analyse de l'étude AS2.

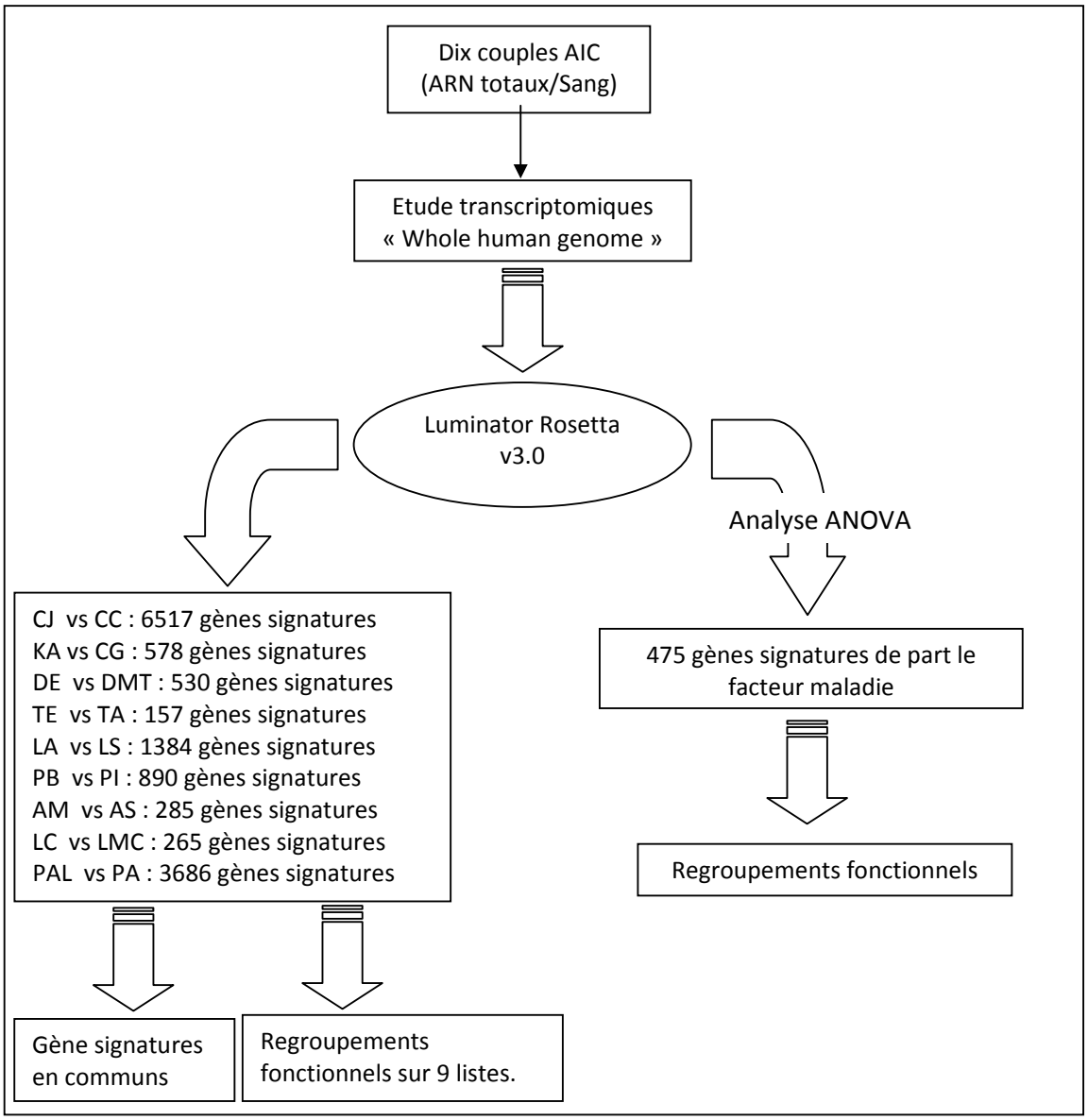


Figure 29 : Schémas récapitulatif de l'étude AS2

Le schéma décrit les différentes étapes de l'analyse de l'étude AS2. L'étude a été effectuée chez 10 patientes AIC, sur des ARN totaux extraits de sang. L'étude à été menée à l'aide des puce transcriptomique « whole human genome ». La version V3.0 du logiciel Luminator Rosetta a permis la sélection de 9 listes de gènes signatures pour les 9 couples filles vs mère disponibles. Une analyse fonctionnelle à l'aide du logiciel EASE a été entreprise. Une première analyse a consisté à rechercher les gènes sigantures en commun à un maximum de fille AIC (paragraphe 5.4.1.1.). Puis, nous avons effectuée un regroupement fonctionnel sur chacune des 9 listes : analyse fonctionnelle fille vs mère (paragraphe 5.4.1.2.). Une dernière analyse par la méthode ANOVA (sur 10 filles AIC vs réf) a permis la sélection de 475 gènes signatures. Une analyse fonctionnelle de ces gènes a été effectuée (paragraphe 5.4.2).

5.4.1. Analyses par famille (fille *vs* mère) dans l'étude AS2

La version 3.0 de luminator, permettant de faire des comparaisons directes entre le transcriptome des filles contre celui de leur mère, a été utilisée. Pour chacun des 9 couples fille *vs* mère, un seuil de l'amplitude de ratio de 2 et une valeur p inférieure à 1,0E-02 ont dicté la sélection des gènes signatures. Ainsi, 9 listes de gènes signatures sont analysées de façon indépendante (filles *vs* mère) par le logiciel EASE (de la même façon que pour l'étude AS1). L'objectif étant d'identifier les fonctions biologiques sur-représentées dans les listes de gènes différentiellement exprimés chez les filles Aicardi par rapport à leur mère.

Fille Aicardi (âge) vs Mère Aicardi	gènes signatures	Numéro de série
CJ (20) vs CC	6517	II & II
KA(20) vs CG	578	IV & IV
DE (23) vs DMT	530	II & IV
TE (25) vs TA	157	III & II
LA (6) vs LS	1384	I & I
PB (6) vs PI	890	I & III
AM (10) vs AS	285	IV & III
LC (11) vs LMC	265	III & I
PAL (16) vs PA	3686	II & I

Tableau 17 : Nombre de gènes signatures et séries d'hybridation pour chaque couple fille versus mère.

CJ(20) est une patiente Aicardi de 20 ans dont la mère est CC. L'hybridation de la lame correspondant à la patiente CJ a été effectuée dans la série II et celle correspondant à la sa mère CC dans la série II. L'analyse du transcriptome chez ce couple CJ (20) vs CC, révèle 6517 gènes signatures. Les hybridations sont divisées en deux groupes. Un groupe de patientes pubères constituée des couples CJ (20) vs CC, KA(20) vs CG, DE (23) vs DMT, TE (25) vs TA, et un groupe de patientes pré-pubères constitué des couples LA (6) vs LS, PB (6) vs PI, AM (10) vs AS, LC (11) vs LMC, PAL (16) vs AP. Les hybridations des 20 lames ont été effectuées en 5 séries de quatre lames. Les ARN à tester ont été répartis de façon aléatoire.

Numéro de série : numéro de la série d'hybridation. Ainsi, les hybridations CJ, CC, DE et TA ont été effectués dans la série II.

5.4.1.1. Les signatures communs dans AS2

Une des premières questions est de savoir si parmi les 9 listes de signatures définies au-delà, des gènes sont partagés par toutes les filles. Il n'y a pas de gènes communs aux 9 listes signatures correspondant aux analyses fille *vs* mère. Nous avons donc cherché les gènes partagés par un maximum de patientes Aicardi. Le gène est sélectionné lorsqu'il apparait dans au moins 5 expériences (tableau 18).

Symbole du gène ; nom complet	**Nbr exp**
IFI27 ; interferon, alpha-inducible protein 27	7
FLJ38973, hypothetical protein LOC205327	6
SARA1; SAR1 gene homolog A (S. cerevisiae)	6
NUDT4; nudix (nucleoside diphosphate linked moiety X)-type motif 4	6
FLJ11292 ; hypothetical protein FLJ11292	5
AGRP; agouti related protein homolog (mouse)	5
ZNF683 ; zinc finger protein 683	5
SNRPG, small nuclear ribonucleoprotein polypeptide G	5
RPS7 ; ribosomal protein S7	5
LOC441073; similar to 60S ribosomal protein L26 (Silica-induced gene 20 protein) (SIG-20)	5
FLJ31153, chromosome 16 open reading frame 63	5
CTNNB1, catenin (cadherin-associated protein), beta 1, 88kDa	5
CD47, CD47 molecule	5
BCLAF1, BCL2-associated transcription factor 1	5

Tableau 18 : Gènes signatures les plus représentés dans l'étude AS2 :

Analyse par couple fille vs mère
Nbr exp : Nombre d'expérience où le gène figure dans la liste de gènes signatures.

IFI27 (interferon, alpha-inducible protein 27) est le gène le plus représenté dans les listes signatures. Ce gène est sous régulé chez les patientes DE (variation de ratio =-3,0 ; p= 4,10E-04), PAL (variation de ratio =-3,1 ; p= 9.2E-36), KA (-2,4 ; 2.1E-21), AM (-2,2 ; 4,89E-06), CJ (-2,0 ; 7E-05). Ce gène est également dérégulé dans l'étude AS1 chez la patiente HK (-8,2 ; 2.1E-21).

5.4.1.2. Regroupement fonctionnel des résultats (AS2)

Pour chaque couple, fille versus mère, nous avons procédé à une analyse fonctionnelle des gènes signatures. Les gènes signatures ont été sélectionnés par le logiciel luminator sur la base d'un seuil de p à 1,0E-02 et une amplitude des ratios à 2 en valeur absolue.
Les couples fille *vs* mère ont été étudiés par EASE de façon indépendante afin de déterminer les fonctions, processus et composant cellulaires qui différencient le transcriptome des filles de celui leur mère. Dans une seconde étape, les catégories sur-représentées sur un maximum de patientes seront recherchées.

Le tableau 19 suivant permet d'avoir un aperçu du nombre de réponses pour chaque couple analysé. Ainsi, pour le couple AM *vs* AS les 245 gènes signatures (tableau) sont définis par 685 termes GO. Dans les résultats qui suivent, pour chaque couple fille *vs* mère étudié, nous ne présenterons les résultats que pour les catégories les plus significatives (scores EASE les plus faibles).

	Score EASE	AMvsAS Nbr Cat	TevsTA Nbr Cat	KavsCG Nbr Cat	DEvsDMT Nbr Cat	LCvsLMC Nbr Cat	CJvsCC Nbr Cat	PBvsPI Nbr Cat	PALvsPA Nbr Cat	LavsLS Nbr Cat
Processus Biologique	≤1	387	437	646	727	450	2775	836	1651	987
	<1,0E-1	172	217	270	342	203	1428	401	974	518
	<1,0E-2	12	50	21	36	14	138	38	107	89
Composant Cellulaire	≤1	91	114	132	160	110	424	206	345	235
	<1,0E-1	49	45	58	70	52	276	112	216	113
	<1,0E-1	5	10	5	2	1	28	25	34	30
Fonction Moléculaire	≤1	207	173	357	375	229	1438	397	928	497
	<1,0E-1	85	63	148	149	89	804	176	477	231
	<1,0E-2	2	4	5	7	6	56	26	43	43

Tableau 19: Nombre de catégories en fonction du score EASE pour les couples de l'étude AS2

Nbr Cat : Nombre de catégorie contenant les gènes signatures. Neuf couples fille *vs* mère sont représentés. Le couple AM vs AS possède 285 gènes signatures (tableau 17) annotés par 387 termes GO dans le système processus biologique. Ce système processus biologique inclut donc 387 termes dont 172 ont un score EASE inférieur à 1,0E-1.

A partir de ce point et jusqu'au paragraphe 5.4.2 les résultats seront présentés de façon séquentielle pour les 9 couples étudiés. Un tableau représentant la distribution fonctionnelle des gènes signatures pour chacun d'eux sera ainsi utilisé.

La liste de référence utilisée dans EASE est la liste des 41 000 clones de la puce « whole human genome » (Agilent). La puce « human whole genome » est annotée par 9929 gènes appartenant à processus biologiques, 9559 gènes annotés par des termes du système composant cellulaire et 9763 gènes annotés par des termes de fonction moléculaire.

Distribution fonctionnelle des gènes signatures pour le couple AM vs AS

Puce « whole human genome »		**Gènes signatures**		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 gènes)	**nucleosome assembly**	**9**	96	**2,73E-06**
	chromatin assembly/disassembly	**9**		**5,88E-06**
	chromosome organization and biogenesis (sensu Eukaryota)	8		5,86E-05
	chromosome organization and biogenesis (sensu Eukarya)	11		7,91E-05
	nuclear organization and biogenesis	11		1,55E-04
Composant Cellulaire (9559 gènes)	**nucleosome**	**9**	83	**3,30E-07**
	chromatin	9		7,35E-05
	chromosome	9		1,31E-03
	obsolete cellular component	10		8,90E-03
	nuclear chromatin	3		7,32E-02
Fonction Moléculaire (9763 gènes)	defense/immunity protein activity	6	91	1,48E-02
	antigen binding	5		4,11E-02
	oxygen transporter activity	2		1,05E-01
	GTP binding	5		1,82E-01
	guanyl nucleotide binding	5		1,86E-01

Tableau 20 : Classification fonctionnelle des gènes signatures pour le couple AM *vs* AS

Le tableau présente les résultats pour les 5 termes les plus représentés dans chacun des systèmes de gène ontologie. Le composant cellulaire « nucléosome » est la catégorie la plus sur représentée. Cette catégorie contient les gènes des histones également impliqués dans les catégories telles que « chromatine », « chromosome » et les catégories du système processus biologique « assemblage du nucléosome » et « assemblage/ désassemblage de la chromatine ».

Distribution fonctionnelle des gènes signatures pour le couple TE vs TA

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 gènes)	**chromosome organization and biogenesis (sensu Eukaryota)**	**9**	67	**3,84E-07**
	chromatin assembly or disassembly	8		1,72E-06
	chromosome segregation	8		1,72E-06
	nucleosome assembly	8		2,50E-06
	chromatin assembly/disassembly	8		4,96E-06
Composant Cellulaire (9559 gènes)	**nucleosome**	**8**	63	**6,86E-07**
	chromatin	8		8,24E-05
	chromosome	9		1,95E-04
	obsolete cellular component	10		1,32E-03
	major (U2-dependent) spliceosome	3		7,30E-03
Fonction Moléculaire (9763 gènes)	defense/immunity protein activity	6	65	3,59E-03
	antigen binding	5		1,36E-02
	poly-pyrimidine tract binding	2		8,80E-02
	pre-mRNA splicing factor activity	3		9,05E-02
	photoreceptor activity	2		1,12E-01

Tableau 21 : Classification fonctionnelle des gènes signatures pour le couple TE *vs* TA

Le tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 157 gènes signatures du couple TE*vs*TA (tableau 17). Ces gènes signatures sont annotés par 387 termes dans Processus Biologique, 91 termes dans Composant cellulaire et 207 termes dans le système Fonction Moléculaire (tableau 19). Comme pour le couple AM vs AS, les catégories fonctionnelles les plus sur représentées sont « nucléosome », « chromatine », « chromosome » et les catégories du processus biologique « assemblage du nucléosome » et « assemblage/ désassemblage de la chromatine ».

Distribution fonctionnelle des gènes signatures pour le couple KavsCG

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 gènes)	immune response	56	197	4,12E-13
	defense response	57		3,80E-12
	response to biotic stimulus	59		6,75E-12
	response to external stimulus	64		5,28E-07
	response to pest/pathogen/parasite	31		7,25E-07
Composant Cellulaire (9559 gènes)	ribosome	20	190	6,81E-08
	ribonucleoprotein complex	23		3,83E-06
	cytosolic large ribosomal subunit (sensu Eukaryota)	11		2,23E-05
	large ribosomal subunit	8		5,86E-05
	prefoldin complex	2		7,68E-02
Fonction Moléculaire (9763 gènes)	structural constituent of ribosome	22	191	1,33E-08
	GTPase activity	11		4,60E-03
	GTP binding	12		7,43E-03
	guanyl nucleotide binding	12		7,89E-03
	hydrolase activity\, acting on acid anhydrides	13		9,77E-03

Tableau 22 : Classification fonctionnelle des gènes signatures pour le couple KA *vs* CG

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 578 gènes signatures du couple KA *vs* CG (tableau 17). Ces gènes signatures sont annotés par 646 termes dans Processus Biologique, 132 termes dans Composant cellulaire et 357 termes dans le système Fonction Moléculaire (tableau 19). Le processus biologique « réponse immunitaire » semble être le terme GO père le plus sur représenté pour ce couple KA *vs* CG. Cette catégorie ainsi que d'autres catégories du système processus biologiques avaient été retrouvées dans l'étude AS1. Par contre, les composants cellulaires et fonctions moléculaires touchés pour cette fille KA par rapport à sa mère CG ne sont pas retrouvés pour les autres couples.

Distribution fonctionnelle des gènes signatures pour le couple DE vs DMT

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 gènes)	response to wounding	24	175	1,19E-06
	response to biotic stimulus	42		6,67E-06
	defense response	39		1,68E-05
	immune response	**36**		**3,68E-05**
	response to external stimulus	52		9,98E-05
Composant Cellulaire (9559 gènes)	hemoglobin complex	4	163	1,52E-03
	cytosol	16		6,25E-03
	cytosolic small ribosomal subunit (sensu Eukaryota)	4		1,05E-01
	plasma membrane	43		1,12E-01
	mitochondrion	19		1,38E-01
Fonction Moléculaire (9763 gènes)	oxygen transporter activity	4	170	9,99E-04
	transmembrane receptor activity	23		2,19E-02
	structural constituent of ribosome	**10**		**2,84E-02**
	GTPase activity	8		5,15E-02
	laminin receptor activity	2		8,36E-02

Tableau 23 : Classification fonctionnelle des gènes signatures pour le couple DE *vs* DMT

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 530 gènes signatures du couple DEvsDMT (voir tableau 17). Ces gènes signatures sont annotés par 646 termes dans Processus Biologique, 132 termes dans Composant cellulaire et 357 termes dans le système Fonction Moléculaire (tableau 19). Le processus biologique « réponse immunitaire » semble être le terme GO père le plus sur représenté pour ce couple DE *vs* DMT. Cette catégorie ainsi que autres catégories du système processus biologique avaient été retrouvées dans l'étude AS1 et pour le couple KA *vs* CG. La catégorie « réponse à une blessure » est un terme GO fils de « réponse à un stimulus externe » et donc également fils du terme réponse immunitaire.

Nous pouvions nous attendre à retrouver le terme GO « constituant structural du ribosome » dans le système composant cellulaire plutôt que dans le système fonction moléculaire. Cependant, la définition de ce terme GO fait également référence à une

fonction moléculaire « Action d'une molécule qui contribue à l'intégrité structurale du ribosome » (http ://www.geneontology.org/). Cette catégorie inclut des gènes ribosomaux et donc peut être rapprochée en termes de contenu en gènes des catégories « ribosomes » et «complexes ribonucléoprotéiques », catégories également présentes comme dérégulées chez la patiente KA par rapport à sa mère CG.

Distribution fonctionnelle des gènes signatures pour le couple LC vs LMC

Puce « whole human genome »		**Gènes signatures**		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	response to external stimulus	27	**95**	1,07E-02
	response to biotic stimulus	20		1,12E-02
	defense response	18		2,32E-02
	antimicrobial humoral response (sensu Vertebrata)	5		3,07E-02
	immune response	**16**		**4,40E-02**
Composant Cellulaire (9559 **gènes**)	cell	90	96	3,55E-02
	endomembrane system	7		1,05E-01
	endoplasmic reticulum	8		1,09E-01
	nucleus	36		1,33E-01
	Golgi apparatus	8		1,34E-01
Fonction Moléculaire (9763 **gènes**)	chemokine activity	3	94	5,54E-02
	chemoattractant activity	3		5,54E-02
	chemokine receptor binding	3		5,54E-02
	sugar binding	4		7,50E-02
	G-protein-coupled receptor binding	3		7,64E-02

Tableau 24 : Classification fonctionnelle des gènes signatures pour le couple LC *vs* LMC

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 265 gènes signatures du couple LC *vs* LMC (tableau 17). Ces gènes signatures sont annotés par 450 termes dans Processus Biologique, 110 termes dans Composant cellulaire et 229 termes dans le système Fonction Moléculaire (tableau 19). Cette patiente AIC présente les mêmes processus biologiques dérégulés que les patientes KA, DE, BM (AS1) et HK (AS1). Les catégories des systèmes composant cellulaires et fonction moléculaire ne sont par contre pas partagés avec d'autres patientes.

Distribution fonctionnelle des gènes signatures pour le couple CJ vs CC

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	defense response	359	2190	1,40E-19
	response to biotic stimulus	379		4,22E-19
	immune response	331		2,65E-18
	response to pest/pathogen/parasite	198		1,15E-12
	cellular defense response	80		1,47E-12
Composant Cellulaire (9559 **gènes**)	hemoglobin complex	12	2081	9,51E-06
	ribonucleoprotein complex	105		2,19E-03
	T-cell receptor complex	6		2,39E-03
	integral to plasma membrane	335		6,56E-03
	mitochondrion	204		8,47E-03
Fonction Moléculaire (9763 **gènes**)	antigen binding	77	2150	9,18E-17
	defense/immunity protein activity	80		3,90E-15
	oxygen transporter activity	10		1,36E-04
	endogenous peptide antigen binding	8		1,61E-04
	coreceptor activity	12		7,33E-04

Tableau 25 : Classification fonctionnelle des gènes signatures pour le couple CJ *vs* CC

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 6517 gènes signatures du couple CJ *vs* CC (tableau 17). Ces gènes signatures sont annotés par 2775 termes dans Processus Biologique, 424 termes dans Composant cellulaire et 1438 termes dans le système Fonction Moléculaire (tableau 19).Cette patiente AIC présente les même processus biologiques dérégulés que les patientes KA, DE, LC, BM (AS1) et HK (AS1). Les catégories des systèmes composant cellulaire et fonction moléculaire ne sont par contre pas partagés avec d'autres patientes.

Distribution fonctionnelle des gènes signatures pour le couple PB vs PI

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	**protein biosynthesis**	**68**	287	**2,36E-22**
	biosynthesis	83		1,57E-16
	macromolecule biosynthesis	73		3,08E-15
	protein metabolism	126		2,94E-13
	metabolism	194		1,69E-07
Composant Cellulaire (9559 **gènes**)	**ribonucleoprotein complex**	**48**	285	**1,05E-17**
	ribosome	**37**		**3,07E-17**
	cytosolic small ribosomal subunit (sensu Eukaryota)	20		3,55E-14
	intracellular	235		1,63E-10
	cytoplasme	159		1,27E-06
Fonction Moléculaire (9763 **gènes**)	structural constituent of ribosome	51	283	9,10E-29
	RNA binding	56		1,37E-16
	structural molecule activity	66		1,85E-15
	nucleic acid binding	103		8,94E-07
	translation elongation factor activity	8		4,29E-05

Tableau 26 : Classification fonctionnelle des gènes signatures pour le couple CJ *vs* CC

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 890 gènes signatures du couple PB vs PI (tableau 17). Ces gènes signatures sont annotés par 836 termes dans Processus Biologique, 206 termes dans Composant cellulaire et 397 termes dans le système fonction moléculaire (tableau 19). La catégorie « biosynthèse protéique » ainsi que la fonction moléculaire « composant structural du ribosome » sont de loin les catégories les plus sur représentées. Elles contiennent entre autre les gènes ribosomaux.

Distribution fonctionnelle des gènes signatures pour le couple PAL vs PA

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	**protein biosynthesis**	**161**	1115	**1,27E-26**
	macromolecule biosynthesis	195		6,68E-19
	biosynthesis	213		1,36E-16
	protein metabolism	374		2,00E-13
	ribosome biogenesis and assembly	35		2,35E-08
Composant Cellulaire (9559 **gènes**)	**ribonucleoprotein complex**	**125**	1082	**3,65E-31**
	ribosome	**85**		**1,31E-25**
	cytosolic small ribosomal subunit (sensu Eukaryota)	38		1,84E-18
	intracellular	792		1,15E-08
	spliceosome complexe	27		1,15E-08
Fonction Moléculaire (9763)	structural constituent of ribosome	102	1090	3,79E-34
	RNA binding	142		4,43E-23
	nucleic acid binding	338		1,74E-09
	mRNA binding	37		3,70E-08
	structural molecule activity	135		9,68E-08

Tableau 27 : Classification fonctionnelle des gènes signatures pour le couple PAL *vs* PA

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 3686 gènes signatures du couple PAL *vs* PA (tableau 17). Ces gènes signatures sont annotés par 1651 termes dans Processus Biologique, 345 termes dans Composant cellulaire et 928 termes dans le système Fonction Moléculaire (tableau 19). Cette patiente AIC (par rapport à sa mère PA) présente pour les trois systèmes GO, les mêmes catégories dérégulées que la patiente PB (par rapport à sa mère PI).

Distribution fonctionnelle des gènes signatures pour le couple LA vs LS

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	**protein biosynthesis**	**98**	443	**7,72E-30**
	macromolecule biosynthesis	106		8,51E-20
	biosynthesis	114		3,49E-18
	ribosome biogenesis and assembly	30		3,72E-15
	protein metabolism	181		6,20E-15
Composant Cellulaire (9559 **gènes**)	**ribonucleoprotein complex**	**90**	442	**8,65E-41**
	ribosome	**65**		**1,96E-34**
	cytosolic small ribosomal subunit (sensu Eukaryota)	31		6,55E-23
	intracellular	343		1,79E-08
	cytoplasme	235		6,78E-07
Fonction Moléculaire (9763 **gènes**)	structural constituent of ribosome	84	434	8,88E-52
	RNA binding	105		1,24E-39
	structural molecule activity	92		1,86E-18
	nucleic acid binding	164		1,26E-11
	mRNA binding	20		2,03E-06

Tableau 28 : Classification fonctionnelle des gènes signatures pour le couple LA *vs* LS

Ce tableau présente les 5 résultats les plus significatifs de la classification fonctionnelle des 1384 gènes signatures du couple LavsLS (tableau 17). Ces gènes signatures sont annotés par 987 termes dans Processus Biologique, 235 termes dans Composant cellulaire et 497 termes dans le système Fonction Moléculaire (tableau 19). Cette patiente AIC (par rapport à sa mère PA) présente pour les trois systèmes GO, les mêmes catégories dérégulées que les patientes PB (par rapport à sa mère PI) et PAL (par rapport à sa mère PA).

Les résultats de classifications fonctionnelles des gènes signatures des filles Aicardi par rapport à leur mère met en évidence trois groupes de patientes. EASE identifie les catégories sur-représentées « organisation du chromosome » et « nucléosome » chez les patientes AM (10ans) et TE (25ans). Pour les patientes KA (20 ans), DE (23ans), LC (11ans), CJ (20ans) la catégorie la plus représentée semble être la « réponse immunitaire» incluant ses termes fils tel que « réponse à un stimulus biologique », « réponse de défense », « réponse au stress », réponse à un stimuli externe ». Enfin, pour le troisième groupe constitué des patientes PB (6 ans), PAL (16 ans) et LA (6 ans) la catégorie identifiée par EASE est la « synthèse protéique ».

Distribution fonctionnelle des gènes signatures : bilan des 9 couples de l'étude AS2

Par rapport à l'étude AS1, le nombre de gènes signatures ainsi que les annotations fonctionnelles sont plus nombreuses dans la puce « whole human genome » utilisée pour l'étude AS2. Il en résulte des scores EASE plus élevés par rapport à ceux de l'étude AS1. On se place donc à un seuil de score EASE plus stringent (1,0e⁻ 03) pour la comparaison des patientes de l'étude AS2.

	Gene	**fille vs mère**
Processus Biologique	protein biosynthesis	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	protein metabolism	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	RNA metabolism	LavsLS ;PbvsPI ;PALvsPA ;CJvsCC
	physiological process	LavsLS ;PbvsPI ;KavsCG ;CJvsCC
	defense response	TevsTA ;KavsCG ;CJvsCC ;DEvsDMT
	immune response	TevsTA ;KavsCG ;CJvsCC ;DEvsDMT
	response to external stimulus	TevsTA ;KavsCG ;CJvsCC ;DEvsDMT
	response to biotic stimulus	TevsTA ;KavsCG ;CJvsCC ;DEvsDMT
Composant cellulaire	ribonucleoprotein complex	LavsLS ;PbvsPI ;PALvsPA ;KavsCG ;CJvsCC
	Ribosome	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	cytosolic large ribosomal subunit (sensu Eukaryota)	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	large ribosomal subunit	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	Mitochondrion	LavsLS ;PbvsPI ;PALvsPA ;CJvsCC
	Cytosol	LavsLS ;PbvsPI ;PALvsPA ;DEvsDMT
Fonction Moléculaire	structural constituent of ribosome	LavsLS ;PbvsPI ;PALvsPA ;KavsCG
	mRNA binding	LavsLS ;PbvsPI ;PALvsPA ;CJvsCC

Tableau 29 : Bilan des regroupements fonctionnels pour les 9 couples de l'étude AS2

La catégorie de gènes la plus partagée est le composant cellulaire « complexe ribonucléoprotéique ». Il est à noter que ce composant cellulaire est impliqué dans le processus biologique « synthèse protéique », et que la fonction du produit des gènes de ce composant cellulaire est de se lier à ARNm. Parmi les gènes impliqués dans ces catégories et signatures chez les patientes citons *HNRPDL* (heterogeneous nuclear ribonucleoprotein D-like). Sa protéine appartient aux ribonucléoprotéines nucléaires qui ont pour rôle l'épissage et l'export nucléaire des ARNm. La catégorie « complexe ribonucléoprotéique » inclut aussi les protéines ribosomiques telles que RPL21 (ribosomal protein L21), RPL26 (ribosomal protein L26), RPL31 (ribosomal protein L31), et RPL34 (ribosomal protein L34).

Dans le paragraphe .5.4.1, les gènes signatures ont été sélectionnés avec un seuil d'amplitude de ratio de 2 en valeur absolue et une valeur p inférieure à 0,01. Des gènes signatures ont été identifiés, cependant une des difficultés de l'étude provient de la différence d'âge des patientes par rapport à leur mère. Le transcriptome en est inévitablement affecté et donc cette différence d'âge constitue un biais dans notre recherche des gènes dérégulés de part la maladie. Nous avons donc voulu analyser les facteurs âge et maladie chez les 10 patientes Aicardi en nous appuyant cette fois sur un outil statistique qui est ANOVA. En effet, lorsque de multiples facteurs sont à analyser et que ces facteurs sont des valeurs discrètes, le logiciel évalue si les moyennes d'un ou plusieurs groupes d'échantillons sont significativement différentes et si un ou plusieurs facteurs affectent les mesures. Si la variance inter groupes est significativement supérieure à la variance à l'intérieur des groupes, alors l'hypothèse nulle selon laquelle la moyenne des groupes est identique est rejetée.

5.4.2. Analyse des résultats de l'étude AS2 par ANOVA

L'outil ANOVA utilisé est implanté dans le logiciel Luminator et utilise le modèle d'erreur spécifique aux puces utilisées. Les paramètres par défaut ont été utilisés (p =1,0E-02). Ainsi, les facteurs âge et maladie ont été analysés chez 10 patientes Aicardi par une analyse ANOVA 2 facteurs. Pour le facteur âge, le premier groupe comprend les 5 filles prè-pubères (LA, PB, AM, LC, PAL) et le second les 14 patientes pubères (5 filles +9 mères). Pour le facteur

maladie, le première groupe inclut 10 malades (toutes les patientes Aicardi) et le second 9 mères.

Il apparait d'après ANOVA que 279 gènes seraient dérégulés de part le facteur âge, **475 gènes seraient dérégulés de part le facteur maladie** et 239 seraient influencés par ces deux facteurs. Nous nous focaliserons donc pour la suite sur l'analyse des **475 gènes** dérégulés par le facteur maladie.

Je présenterai tout d'abord parmi ces gènes ceux qui sont cartographiés sur le chromosome X qui constituent, des gènes potentiellement candidats pour le syndrome AIC. Dans un second temps, un regroupement fonctionnel de ces 475 gènes par le logiciel EASE a permis la caractérisation des catégories GO sur représentées (pour les gènes dérégulés par le facteur âge) sur l'ensemble des filles AIC.

D'après ANOVA, 8 gènes cartographiés sur le chromosome X sont influencés par le facteur maladie.

CRSP2 (cofactor required for Sp1 transcriptional activation, subunit 2,) localisé en Xp11.4, code une sous-unité du complexe *CRSP* (cofactor required for SP1 activation). Ce complexe est requis pour l'activation par le facteur de transcription SP1 (Sp1 transcription factor). *ARMCX3* (armadillo repeat containing, X-linked 3) localisé en Xq21.33-q22.2, code un membre de la famille des protéines ALEX qui jouent un rôle dans la suppression tumorale. *ACRC* (acidic repeat containing) en Xq13.1 a été séquencé chez des patients avec dystonie associée à un syndrome de parkinson lié à l'X mais aucune mutation n'a été détectée dans la séquence codante. D'après les auteurs ce gène pourrait jouer un rôle dans la structure de la chromatine [4]. *DMD* (dystrophin (muscular dystrophy, Duchenne and Becker types) en Xp21.2, est impliqué dans la dystrophie musculaire de Duchenne. *TCTE1L* (dynein, light chain, Tctex-type 3) en Xp21 est un gène dont la fonction n'est pas connue. RAB9B (membre de la famille des oncogènes RAS) en Xq22.1-q22.3, appartient à une famille de protéines G qui régule le trafic vésiculaire. *WBP5* (WW domain binding protein 5) en Xq22.1-q22.2 code une protéine composé avec un domaine globulaire WW impliqué dans les interactions protéine-protéine. *GAGE7* (G antigen 7) en Xp11.4 code un antigène exprimé dans les tumeurs.

Une analyse fonctionnelle par EASE de la liste des 475 gènes dérégulés par la maladie a été effectuée.

Puce « whole human genome »		Gènes signatures		
Système	**Catégorie (Gènes)**	**Gènes/Cat**	**Gènes/ Sys**	**Score EASE**
Processus Biologique (9929 **gènes**)	mitotic sister chromatid segregation	3	125	6.52e-003
	DNA replication and chromosome cycle	3		6.52e-003
	DNA repair	6		3.68e-002
	cell cycle	7		3.84e-002
	mitosis	6		3.87e-002
	hemopoiesis	5		7.11e-002
	cytokinesis	4		1.02e-001
	sister chromatid cohesion	2		1.07e-001
	I-kappaB kinase/NF-kappaB cascade	3		1.12e-001
	microtubule nucleation	2		1.29e-001
Composant Cellulaire (9559 **gènes**)	**chromosome, pericentric region**	**4**	115	**8.26e-004**
	hemidesmosome	2		8.06e-002
	nucleus	41		8.25e-002
	nucleoplasm	4		9.47e-002
	mediator complex	2		2.04e-001
	spindle pole body	2		2.14e-001
	extracellular	11		2.14e-001
	endoplasmic reticulum	7		2.28e-001
	kinetochore	2		2.86e-001
	cellular_component unknown	19		3.59e-001
Fonction Moléculaire (9763 **gènes**)	dynein binding	2	120	3.61e-002
	molecular_function unknown	20		4.36e-002
	chromatin binding	4		6.14e-002
	zinc ion binding	15		8.63e-002
	specific RNA polymerase II transcription factor activity	5		1.15e-001
	axon guidance receptor activity	2		1.68e-001
	transcription factor activity	16		2.03e-001
	RNA polymerase II transcription mediator activity	2		2.08e-001
	structural constituent of cytoskeleton	5		2.12e-001
	GTPase activity	5		2.12e-001

Tableau 30 : classsification fonctionnelle des gènes signatures issues de la méthode ANOVA

Parmi les 9559 gènes annotés dans le système composant cellulaire, 16 sont annotés par le terme chromosome, région péricentrique. Quatre de ces gènes font partie des gènes signatures issus d'ANOVA. *CBX5* (chromobox homolog 5 (HP1 alpha homolog, Drosophila)) code la protéine HP1 (Heterochromatin protein-1) qui est une protéine qui se lie aux méthyl-lysines de l'hétérochromatines et participer ainsi à l'inhibition de l'expression des gènes a ces sites. Elle agit avec la protéine SUV39H1 (suppressor of variegation 3-9 homolog 1

(Drosophila)) qui elle méthyle l'histone H3. CDCA1 (NUF2, NDC80 kinetochore complex component, homolog (S. cerevisiae)), ZNF92 (zinc finger protein 92) qui est un facteur de transcription et CENPA (centromere protein A) qui possède un domaine similaire à celui des histones H3.De plus, nous retrouvons par cette analyse ANOVA l'implication des protéines Histones.

Les études d'expression ne révèlent pas de gènes candidats dérégulés chez toutes les filles Aicardi et cartographiés sur le chromosome X.

Cependant, l'analyse fonctionnelle aussi bien dans l'étude AS1 que l'étude AS2 montre que les gènes différentiellements exprimés chez les filles Aicardi appartiennent à des catégories de gènes communes codant les protéines :

- d'une part les gènes histones semblent être impliqué directement (CC, HK, BM, AM, TE) ou indirectement (étude ANOVA),
- d'autre part, les gènes ribosomaux (PB, PAL, LA).

Enfin la réponse au stimulus biologique et réponse immunitaire apparait comme significative dans les études couple par couple (AS1 et AS2).

L'accumulation des données biologiques notamment depuis l'initiative du projet de séquençage du génome humain a fait des bases de données biologiques une mine d'informations très impressionnante. Les bases de données dédiées aux animaux modèles, comme la drosophile et la souris, sont très complètes d'un point de vue annotation. Leur génome est complètement séquencé et les relations génotype-phénotype sont très détaillées. La conservation des séquences entre les espèces permet ainsi d'inférer des fonctions aux protéines humaines. Par exemple, la connaissance de la fonction d'un gène chez la drosophile, permet d'émettre une hypothèse quant à la fonction de son homologue

humain. L'annotation des gènes par une ontologie (vocabulaire contrôlé) permet d'étiqueter et donc de retrouver les gènes dans les bases de données. Le vocabulaire hiérarchisé GeneOntology (GO) constitue à l'heure actuelle le vocabulaire de référence pour l'annotation des gènes. Les gènes répertoriés dans la base de donnée Entrez GENE du NCBI (National Center for Biotechnology Information), MGI (Mouse Genome Informatics) et Flybase (Database of Drosophila Genes & Genomes) sont annotés par des termes GO. Cette annotation commune permet ainsi d'homogénéiser et de créer des relations entre les contenus des différentes bases de données. La comparaison et l'extraction d'informations sont ainsi facilitées. L'identification de gènes impliqués dans les maladies génétiques est aujourd'hui un sujet d'étude très actif en bioinformatique. Un nombre impressionnant de logiciels sont déjà disponibles alors que les premières approches datent seulement de 2001[193]. Les approches de recherche de gène candidats en bioinformatique peuvent être divisées en trois stratégies. Les systèmes généralistes essaient de prédire des gènes candidats en se basant sur les propriétés ou interactions de ceux-ci. Un second groupe utilise l'hypothèse selon laquelle les maladies phénotypiquement proches sont causées par des gènes fonctionnellement similaires. Et la dernière catégorie de logiciels propose une approche multicritère et multi-hypothèse pour la recherche de gènes candidats. Le prototype ACGR (« Approach for Candidate Gene Retrieval ») s'inscrit dans cette troisième catégorie de système. Cette partie, *in silico* est un projet mené en collaboration avec l'équipe Orpailleur et plus précisément avec Marie-Dominique Devignes et Malika Smaïl-Tabbone (LORIA Campus science-BP239-F-54506 Vandoeuvre les Nancy).

6. Approche *in silico* pour la recherche de gènes candidats pour le syndrome d'Aicardi

Une publication soumise donne le détail des analyses et des résultats de façon détaillée et donc cette partie aura pour vocation de présenter la logique et le déroulement de cette approche *in silico*.

6.1. Présentation du système ACGR (« Approach for Candidate Gene Retrieval »)

Initialement développé pour le syndrome d'Aicardi, le prototype avait été nommé ACGR pour Aicardi Candidate Gene Retrieval. Cependant la stratégie que nous avons mise en place s'est révélée applicable à d'autres maladies génétiques rares. Nous avons donc renommé le prototype pour un nom plus général.

Le but est de cribler les bases de données biologiques à la recherche de gènes candidats pour le syndrome d'Aicardi. Le prototype ACGR est organisé autour d'une base de données développée pour stocker des informations sur les gènes susceptibles d'être candidats, telles que l'espèce animale d'origine, la localisation chromosomique, les termes GO (Gene Ontology) associés, les gènes orthologues dans d'autres espèces. Cette base de données a été developpée à partir d'un modèle de données concu pour organiser toutes les informations nécessaires en vue de pouvoir répondre à diverses définitions de gène candidat (voir article). Elle est alimentée de façon semi automatique par l'utilisateur qui peut à partir de l'interface du logiciel lancer plusieurs types d'interrogation des bases de données publiques, la collecte des données étant effectuée grâce au logiciel Xcollect (Devignes *et al.* 2005 : http ://www.nettab.org/2005/docs/NETTAB2005_DevignesOral.pdf) développé au Loria.

La base de données complétée, il est par la suite nécessaire de définir les critères qui font d'un gène quelconque un gène intéressant pour la maladie étudiée. Il est possible de décrire un gène candidat par diverses définitions. Les définitions sont dépendantes de la connaissance que nous avons de la maladie étudiée et des données expérimentales dont nous disposons. Pour le syndrome d'Aicardi, nous disposons de données de puces transcriptomiques et nous savons que le gène candidat est *a priori* cartographié sur le chromosome X. Ainsi, une définition un peu simpliste du gène candidat pourra être : le gène candidat pour le syndrome d'Aicardi est un gène dérégulé chez les patientes et cartographié sur le chromosome X. D'autres définitions pourront faire intervenir un gène intermédiaire entre le gène candidate et la maladie. On dira par exemple que le gène candidat est une gène situé sur le chromosome X et qui interagit avec un gène dérégulé chez les patientes. Ces définitions peuvent être traduites sous la forme de vues (« view ») sur la BD. Ultimement ce sont des requêtes SQL (« Structured query language ») qui interrogent la

base de données ACGRdb pour construire des ensembles de gènes candidats répondant aux définitions.

6.2. Les sources de données intégrées dans ACGRdb

Avant d'interroger la base de données avec les différentes définitions de gènes candidats, une première étape de la stratégie a consisté à collecter les informations dans ACGRdb. Les bases de données GO, Entrez Gene, MGD, Flybase et le système HomoloGene ont ainsi été utilisées pour alimenter ACGRdb. Par ailleurs, les données issues des puces transcriptomiques ont également été intégrées dans ACGRdb. La consultation des BD ainsi que l'intégration des données expérimentales se fait systématiquement par l'interface d'ACGR.

- Collecte des données expérimentales

Parmi les 475 gènes signatures (dérégulés) issue de l'analyse ANOVA de l'étude AS2 (Résultats, chapitre 5), 300 gènes connus ont été intégrés dans ACGRdb. Parmi les 175 restant figuraient des identifiants ENSTnnnnnnnnnn (Ensembl), THCnnnnnnn (TIGR : The Institute for Genomic Research), des identifiants Genbank etc… Ces séquences n'ont pas été retenues par ACGR car elles n'ont pas de correspondance en symbole officiel de gène.

- Description de la maladie

Une seconde étape a consisté à rechercher comment décrire la maladie afin de pouvoir classer les gènes candidats par ordre de pertinence vis-àvis de la maladie. Il n'existe pas à ce jour d'ontologie qui permet d'indexer ou de faire le lien entre les phénotypes de maladies et les gènes. Le thésaurus MeSH permet d'annoter les premiers et GO les seconds. Dans GO, l'ontologie « processus biologique » a pu être utilisée pour sélectionner un certain nombre de termes reflétant au mieux le phénotype de la maladie. Ainsi, le terme « développement du corps calleux » représentait bien un processus biologique touché puisqu'une agénésie du corps calleux est retrouvée chez les patientes Aicardi. L'ensemble des termes reflétant le phénotype de la maladie constitue le jeu de termes AICARDI DS-GO (« Aicardi disease-

specific GO terms») (voir la publication). La sélection des termes GO se fait en interrogeant le site Amigo avec des mots clés tel que « corps calleux ». Les termes GO pertinents sont par la suite sélectionnés. Cette sélection étant une phase déterminante nous avons opté pour l'intervention d'un expert ayant une connaissance de la clinique du syndrome (B. Leheup, K. AngoÏ). Les termes GO du set AICARDI DS-GO sont utilisés pour collecter tous les gènes qui sont annotés par au moins un de ces termes ou de leurs enfants dans la hiérarchie GO. La collecte se fait à partir de BD Entrez Gene pour les gènes humains, MGD pour les gènes murins et Flybase pour les gènes de drosophile.

Un outil de mesure de similarité entre les annotations GO a été introduit dans le système. Ainsi **un score est affecté à chacun des gènes** en fonction de **la similarité des termes GO annotant ce gène avec le jeu AICARDI DS-GO.** La méthode utilisée pour calculer la similarité a été empruntée à l'outil GO Family du système GOTOOL box [222]. Plus il existe de termes communs entre les annotations du gène et le jeu de référence, plus le score est élevé.

- Collecte des données

Des étapes de collecte de données ont permis de receuillir lorsqu'elles sont disponibles, les informations relatives à chaque gène de la BD : symbole du gène, nom complet, cartographie physique, numéro OMIM, phénotype associé, orthologues humains des gènes collectés à partir des animaux modèles et interactants des gènes ou de leur produit.

La BD ACGR une fois complétée, des requêtes décrivant les gènes candidats ont pu être définies sous forme de vues. Les interactions peuvent se faire aussi bien entre protéines qu'entre une séquence et une protéine. Par la suite, pour ne pas alourdir le texte j'utiliserai la notion de gènes en interaction pour désigner les interactants des gènes aussi bien que les interactants des produits des gènes.

6.3. Les vues définies sur ACGRdb

Quatre vues ont été définies :

- Vue1 : les gènes de la BD classés en fonction de leur similarité au set AICARDI DS-GO. L'hypothèse est que la mutation dans ces gènes affecterait les processus biologiques

dans lesquelles ils interviennent. Un gène dont le score est élevé représente ainsi un gène dont les mutations entraineraient un phénotype proche de la maladie.

- Vue2 : les gènes de souris et de drosophile de la vue1 enrichis par leurs orthologues humains et la localisation cytogénétique de ces derniers. Cette vue permet de tenir compte de la richesse des annotations chez les animaux modèles. Elle permet, le cas échéant, de bien classer l'orthologue humain d'un gène de souris dans la mesure où ce dernier présente un score élevé même si le score du gène humain est faible (cas du syndrome CHARGE, voir article).

- Vue3 : les interactants des gènes ou produits des gènes de la vue1. Cette hypothèse prévoit que le gène candidat serait un interactant d'un gène bien annoté (dont les termes GO sont proches du set de référence). On peut imaginer qu'un facteur de transcription responsable d'une maladie soit en interaction avec un ou plusieurs gènes reflétant bien les processus biologiques touchés.

- Vue4 : les gènes de souris et de drosophile de la vue3 enrichis par les orthologues humains des gènes. en interaction. Cette vue permet de mettre en évidence des gènes candidats grâce à une interaction qui n'aurait été décrite que chez un organisme modèle (cas du syndrome de Goltz, voir article)

Étant donné que des données expérimentales sont disponibles pour le syndrome d'Aicardi, elles ont été intégrées dans les définitions de gène candidat. Ainsi, la vue1 devient vue1EXP : gènes de la base de données dérégulés, classés selon leur score de similarité En ce qui concerne le syndrome d'Aicardi pour lequel nous ne considérons que les gènes candidats cartographiés sur le chromosome X, cette définition fournit une liste de 9 gènes présentant de faibles scores. Ce résultat décevant justifie l'exploitation des autres définitions de gènes candidats.

En revenant à la vue 1, c'est-à-dire sans tenir compte des résultats expérimentaux, on trouve en première position comme gène candidat le gène de la plexin B3 (*PLXNB3*) avec un score de 56% de similarité avec le set AICARDI DS-GO. Le gène est cartographié en Xq28.

Une définition plus élaborée du gène candidat implique l'interaction entre les gènes dérégulés et le gène candidat (Vue3Exp). Le gène *DLX5* (distal-less homeobox 5) dérégulé chez les filles AIC, cartographié en 7q22 et affecté du score 50, interagit avec le gène *MAGED1* (melanoma antigen family D, 1) cartographié en Xp11.23. *MAGED1* interagit donc avec un gène « bien annoté » alors que son propre score est de 3.
Une autre possibilité d'exploiter la vue3 est de rechercher les gènes de l'X qui sont reliés au le plus grand nombre d'interactants dans ACGRdb. C'est la perturbation de voies biologiques qui conduit au phénotype observé et on peut supposer que le gène candidat fasse partie avec ses interactants rd'un telle voie. En d'autres mots, le gène candidat peut être déifini comme un gène en interaction avec un maximum de gènes dérégulés ou dont les termes GO sont proches du set AICARDI DS-GO. Le gène *SUV39H1* (suppressor of variegation 3-9 homolog 1) cartographié en Xp11.23 est en interaction avec 10 gènes de ACGRdb parmi lesquels, les gènes *CBX5* (chromobox homolog 5) et *PCM1* (pericentriolar material 1) font partie des gènes dérégulés.

Ce travail à fait l'objet d'une publication soumise dans la revue Bioinformatics ci-jointe.

A Database Approach for Candidate Gene Retrieval based on Model-driven Data Integration and Expert View Definition

Yilmaz[1] S, Jonveaux[1] P, Bicep C, Pierron L, Smaïl-Tabbone M and Devignes MD[*]
LORIA UMR7503, CNRS, INRIA, Nancy University, BP239, 54506 Vandoeuvre-les-Nancy cedex
[1]Laboratory for Human Genetics, Nancy Medical Faculty, rue du Morvan, 54500 Vandoeuvre-les-Nancy cedex.

ABSTRACT

Motivation: Understanding genetic diseases relies on the discovery of the genetic defects that are responsible for the observed disorder. Computational methods are nowadays widely used to discover candidate genes by exploiting the mass of data accumulated in public genomic databases and crossing these data with private data. In most methods a similarity measure is computed between annotations of candidate genes and either known disease genes or some disease description. However the relation between a disease gene and a disease may reveal more complex, especially in the case of rare diseases presenting mixed phenotypes. Exploiting prior knowledge about orthology relationships or interactions as well as data from multiple species is then required.

Results: We propose a model-driven approach for retrieving disease-specific candidate genes based on an integrated genomic information system. Knowledge embedded in expert's description of candidate gene is used to enrich the definition of precise relationships between genes and diseases by expliciting the role of intermediary genes (orthologous, interacting…). These definitions are used for guiding data modeling and are converted into views on the data which ultimately lead to retrieval of sets of candidate genes. Implementation using a relational data model involves an automated process of data collection from both public and private sources. Three case-studies are presented including a rare disease for which responsible gene is still unknown.

Availability: ACGR sources are freely available upon request to the contact author.

Contact: devignes@loria.fr

Supplementary information: See Folders "ACGR_views" and "ACGR_Xcollect_scenarios" on http://bioinfo.loria.fr/projects/acgr .

1 INTRODUCTION

Identifying genes responsible for human diseases is a complex and challenging task, especially today when most monogenic genetic diseases become better and better understood so that we are left with complex pleiotropic syndromes possibly relying on more than one gene and sometimes displaying very rare distribution all over the world. About 3700 phenotypes described in OMIM (Online Mendelian Inheritance in Man) are not yet associated with any responsible gene. About 1300 of these are designated as "syndromes". Understanding the molecular basis of a disease ultimately means correlating disease symptoms with altered gene function(s). This is obviously more difficult for syndromes that group together several symptoms.

Gene discovery for such complex syndromes rely on integrative genomics approaches as reported in the case of human cytochrome c oxydase deficiency (Leigh syndrome, OMIM #256000, Mootha *et al.*, 2003) or infantile hepatic mitochondrial DNA depletion (OMIM #251880, Spinazzola *et al.*, 2006). In general disease gene identification can be divided into two phases. A first exploration phase produces lists of candidate genes, ranked from the most to the less promising one according to user criteria and with respect to studied disease. Then comes the validation phase where each candidate gene is intensively studied to experimentally demonstrate its involvement in causing the disease.

This paper deals with the first exploration phase (candidate gene discovery) in the case of rare diseases (prevalence less than 5 per 10000) for which limited genetic/linkage data are available. This situation hinders precise delimitation of the genomic locus associated with the disease. To compensate for the lack of genetic data, the approach we describe here relies on integrating relevant gene annotations retrieved from public genomic databases and crossing them with experimental data such as transcriptomic results.

Available computer-based systems for identifying disease genes can be classified into three groups. Generalist systems try to predict disease genes according to their properties or interactions (Lopez-Bigas and Ouzounis, 2004; Adie *et al.*, 2005; Lopez-Bigas *et al.*, 2006; Oti *et al.*, 2006; Tu *et al.*, 2006; Xu and Li, 2006; Calvo *et al.*, 2007). Interesting tendencies are thus detected among the ~1600 disease genes listed in OMIM morbid map and used for these studies. Disease genes tend to be longer, are composed of more exons, show a higher degree of interspecies conservation, display more interactions, etc. However these approaches are unable to establish correspondence between a given disease and sets of candidate genes.

The second group of strategies relies on the hypothesis that similar diseases are caused by similar genes. These strategies are often called prioritization methods since they aim at ranking genes from a list with respect to their chance of being the responsible gene (Freudenberg and Propping, 2002; Perez-Iratxeta *et al.*, 2002; Turner *et al.*, 2003; Masseroli *et al.*, 2004 and 2005; Perez-Iratxeta *et al.*, 2005; Adie *et al.*, 2006; Aerts *et al.*, 2006; George *et al.*, 2006; Rossi *et al.*, 2006). Alternative strategies based on the same similarity hypothesis aim at clustering rather than ranking genes from a list (Masseroli *et al.*, 2004 and 2005; Barillot *et al.*, 2004; Chiang *et al.*, 2006; Franke *et al.*, 2006; Sun *et al.*, 2006). Prioritization methods may also be applied to single diseases for which several chromosomal loci are known in order to find additional responsible genes. Various statistical methods are used ton compute similarity measures that take into account variable gene features. Such features are particularly well covered in the Endeavour system (Aerts *et al.*, 2006): sequence similarity, domain composition, tissue expression, GO annotation, interspecies conservation, protein-protein interactions and involved pathways. However this type of strategy implies proposing one or several model genes for the disease.

Finally a third group of methods proposes integrated systems enabling users to formulate complex multi-criteria queries in order to retrieve appropriate collections of relevant genes. This is the case of the GeneSeeker system (Van Driel et al., 2003 and 2005) and the GeneSorter functionality proposed by UCSC Genome Browser (Kent et al., 2005). Such systems enable the expert to test various hypotheses about the criteria that can relate a disease to its candidate genes. An example of that is found in Tiffin et al. (2005) who developed a strategy to identify genes expressed in the tissue affected by the disease: candidate genes were selected if their annotation with a controlled vocabulary (eVOC, used in Ensembl EST annotation) matched the disease annotation. Relevant eVOC annotation for the disease was derived from PubMed abstracts by a text-mining approach.

The approach presented in this paper is inspired by this last group of methods. We propose to guide candidate gene retrieval by formalizing expert knowledge concerning the relationships that may exist between a gene and the disease it causes. Several possible definitions of candidate gene are expressed and exploited first for populating a dedicated customized database with meaningful data extracted from web resources and second for defining views which can be applied to the database in order to retrieve sets of candidate genes. The method allows integrating experimental results. It is tested with three complex syndromes, one of which (AICARDI syndrome, OMIM %304050) remains unexplained so far.

2 SYSTEMS AND METHODS

2.1 Explicit definitions of a candidate gene

A first very broad definition of a candidate gene is: "a gene that can be related to a disease". In order to efficiently guide candidate gene retrieval, this definition has to be refined into more specific definitions reflecting available information and knowledge pertaining from various public databases or wet-lab experiments.

The most obvious relationship between candidate genes and disease is their co-localisation on human chromosomes. This has been guiding positional cloning for a long time. This relationship can be expressed as "is_co-localized_with" (Rel1). Mapping resolution is highly variable depending on the disease. Mapping data retrieved from factual databases can be enriched by experimental data thanks to recent techniques such as CGH array (Shaw-Smith *et al.*, 2004; Vissers *et al.*, 2005; Vermeesch *et al.*, 2007).

Another direct relationship is tissue or developmental co-expression of both genes and disease features. This hypothesis has been used in various prioritization methods (Tiffin *et al.*, 2005). A variant of this relationship, the "is_dysregulated_in" relationship (Rel2), considers the dysregulation of candidate genes in patient samples.

Functional annotation of genes is improving in most available databases and can be related to disease description. The "has_similar_functional_annotation_with" relationship (Rel3) requires a similarity measure between functional annotations of gene and disease. A major problem here is that both annotations are not expressed with the same vocabularies. Therefore most prioritization methods restrict their similarity measure to a comparison between test genes and training genes considered as relevant disease genes (see for example Aerts *et al.*, 2006). The similarity measurement is then performed between the GO annotations of test and training genes according to one of the various algorithms published so far (Khatri and Draghici, 2005). Alternatively, we propose here to directly compute a gene-disease similarity thanks to a functional description of the disease of interest with a gene annotation vocabulary (such as GO).

A key-point of our approach is that the relationship between candidate gene and disease may also involve an intermediate gene which will itself satisfy some relationship with the disease. We explored two types of intermediate genes (Int): orthologous genes (Int1) and interacting genes (Int2). It can be observed that mapping relationship (Rel1) can only be directly applied to the candidate gene itself, whereas dysregulation (Rel2) or functional similarity (Rel3) relationships can be applied to intermediate genes. Complex definitions can thus be constructed such as: "a candidate gene is a gene co-localized with the disease and orthologous to a gene that has similar functional annotation with the disease" which combines the Rel1 and Rel3 types of relationships and the Int1 type of intermediary gene, or "a candidate gene is a gene co-localized with the disease that interacts with a gene that is dysregulated in patients with the disease" which combines the Rel1 and Rel2 types of relationships and the Int2 type of intermediate gene. Further complex definitions can be imagined such as "a candidate gene is a gene co-localized with the disease and interacting with a gene which is orthologous to a gene which has similar functional annotation with the disease", etc.

Retrieving from the mass of data present in biological data sources sets of candidate genes matching such complex definitions is the challenge taken up by the ACGR approach.

2.2 Functional presentation of the ACGR approach

We propose the following generic *in silico* methodology for complex disease candidate gene retrieval. Inputs to the system are a functional description of the disease using GO terms and available experimental datasets. The system then collects data from public resources: genes sharing common GO annotation with the disease, either in human or in model organisms, relevant annotations for these genes such as cytogenetic localization, functional annotation, interacting genes and human orthologs to genes from model organisms. All retrieved genes are then assigned values computed on the basis of their annotation similarity with the studied disease. Finally it is possible to build sets of candidate genes that correspond to various definitions.

To achieve these functions, we chose to use an architecture centered on a database (DB). Following section presents the advantages of using a DataBase Management System (DBMS) and an appropriate conceptual framework compared to more conventional data processing systems.

2.3 Database management system and conceptual framework

There are three main features of a DBMS that make it attractive to use: *centralized data management*, *data independence*, and *systems integration*. This contrasts with conventional data processing systems where each application program has direct access to the data it reads or manipulates. These programs are usually based on considerable prior knowledge about data structure and format. In such environment any change in data structure or format requires appropriate changes in the application programs. In DBMS, all data are integrated into one system thus reducing redundancies and inconsistencies and making data management more efficient. Better service is provided to users. Information availability and up-to-dateness are likely improved since data are shared. The ability to quickly obtain new and combined information is crucial for biologists exploring an "open hypothesis space". Finally, a major advantage of setting up a database system is the requirement that an overall data model for the domain be built ensuring the global data coherence.

The most commonly used conceptual framework for a DBMS is the three-level architecture suggested by the ANSI/SPARC committee (American National Standards Institute/Standards Planning and Requirements Committee) (ANSI/X3/SPARC, 1975). The three levels can be considered as three different views on the data: (i) External level - individual user view; (ii) Conceptual level- community user view; (iii) Internal level- physical or storage view. The three-level database architecture allows a clear separation of the information meaning (conceptual view) from the physical data structure layout. A database system that is able to separate these modeling levels is likely to be flexible and adaptable. The external level is the view that an individual user of the database has. This view is often a restricted view on the data and the same database may provide a number of different views for different categories of users or needs. In our study, we consider that the candidate gene definitions defined in section 2.1 constitute external views on data collected about genes and diseases. The conceptual level consists in the information model of the domain of interest without any concern for physical implementation. It constitutes the overall community view on the data and it includes all the information that is going to be represented in the database. This level is more stable than the two others since it may be desirable to introduce changes at the internal (physical) level in order to improve system performance without changing anything at the conceptual level of the database. In this paper we will substitute to the physical level the so-called *logical level* as introduced by almost all design methods (Teorey et al., 2006). The *logical* data model is close to the physical one but is not dependent of a specific commercial DBMS.

3 ALGORITHM

In this section, we give the major specifications for achieving ACGR functions. We first present the ACGR DB design according to the three levels described above (section 3.1). In section 3.2 we outline the specification of wrappers aimed at populating ACGR DB. Finally, the construction of sets of candidate genes is explained (section 3.3).

3.1 Database Design

At the external level, the individual user views include the various definitions of candidate genes presented in section 2.1. In-depth analysis of these definitions leads to specify the various types of data relevant for the study. The resulting conceptual data model expressed as an Entity-Relationship (ER) model is presented in Fig. 1.

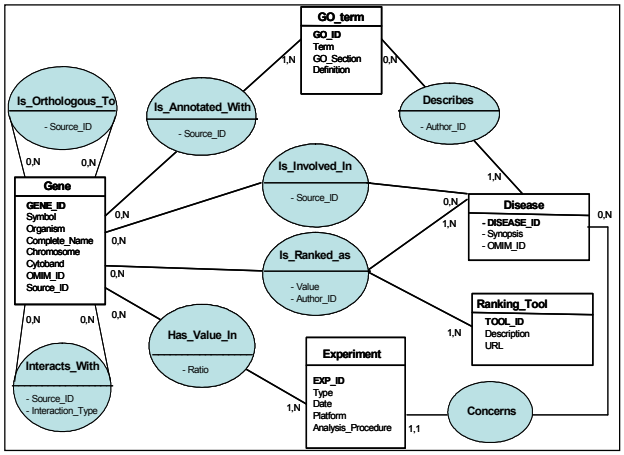


Fig. 1. Conceptual data model for ACGRD database. An entity type, represented as a box, correspond to a class of concrete or abstract objects. A relationship type, represented as an ellipse, correspond to an association between entity types. Two cardinalities are attributed to each participation of an entity to a relationship for representing the minimal and maximal number of times each occurrence of the entity can participate in the relationship. For instance, a Gene occurrence may be involved in no disease (minimum equals 0) or in several diseases (maximum equals N).

This modelling ensures that queries corresponding to any candidate gene definition can be addressed to a database constructed according to this model. For example the definition of a "candidate gene as a gene co-localized with the disease and orthologous to a gene that has similar functional annotation with the disease" can be represented thanks to the 'Gene', 'Disease', 'GO_term' and 'Ranking_Tool' entities and to the 'Is_Orthologous_To' and 'Is_Ranked _As' relationships.

The relational data model derived from the conceptual model shown in Fig. 1 is presented in Table 1.

3.2 Database populating

On the basis of the relational data model, it is possible to explicit the initialisation steps of ACGR DB. Entering a disease description consists in inserting one tuple into the 'Disease' table and several tuples in the 'GO_Term' and 'Disease_GO_Term' tables. Disease description is found in OMIM entries with links to MedLine abstracts that are indexed with MeSH terms. Since the correspondence between MeSH and GO terms is not yet well established (Bodenreider 2004; Marquet et al., 2003), disease-specific GO term sets (pertaining from the Biological Process branch of GO) have to be selected by the expert for expressing at best the specific features of the disease.

When available, experimental data will be entered by inserting one tuple into the 'Experiment' table (for each performed experiment), several tuples in the 'Gene' and 'Gene_Experiment' tables corresponding to all signature genes with their dysregulation ratio (only gene symbols are inserted into the 'Gene' table at this stage).

Table 1. Relational data model for ACGR database. It consists in a set of abbreviated table schemas. Each table contains a set of attributes including a primary key (in bold face) and one or several foreign keys (in italic).

Table name	Attribute set
Gene	**Gene_ID**, Symbol, Organism, Complete_name, Chromosome, Cytoband, OMIM_ID,Source_ID
GO_Term	**GO_ID**, Term, GO_section, Definition
Gene_GO_Term	***Gene_ID*, *GO_ID***, Source_ID)
Orthology	***Gene_ID1*, *Gene_ID2***, Source_ID
Interaction	***Gene_ID1*, *Gene_ID2***, Source_ID, Interaction_Type
Disease	**Disease_ID**, Synopsis, OMIM_ID
Disease_GO_Term	***Disease_ID*, *GO_ID*, Author_ID**
Involvement	***Gene_ID*, *Disease_ID*,** Source_ID
Ranking_Tool	**Tool_ID**, Description, URL
Gene_Disease_Rank	***Gene_ID*, *Disease_ID*, *Author_ID*, *Tool_ID*,** Value
Experiment	**Exp_ID**, Type, Date, Platform, Analysis_procedure, *Disease_ID*
Gene_Experiment	**Gene_ID, Exp_ID**, Ratio

The data collecting function then consists in retrieving from public databases the genes that are annotated by at least one GO term associated with the disease of interest and inserting them into the Gene table. Similarly, orthologous and interacting genes are retrieved and inserted into the Gene, Orthology and Interaction tables. Data wrapper specification implies selecting appropriate databases (see below: implementation) and mapping manually the relevant fields onto ACGR relational data model.

In addition specific wrappers must be designed for plugging available ranking tools in order to compute values aimed at ranking genes with respect to disease description. Such wrapper inserts tuples into the Gene_Disease_Rank table: one tuple per gene and per ranking tool.

3.3 Building sets of candidate genes

Our objective here is to build queries aimed at producing sets of annotated candidate genes corresponding to the various explicit definitions presented in section 2.1.

Views can be defined in Standard Query Language (SQL) at the logical level of our conceptual framework to express the candidate gene definitions. A view in SQL associates an SQL query with a view name leading to the creation of a virtual table. We selected here four basic definitions leading to the four views described below. The corresponding SQL queries can be found in the supplementary files.

View1: *Genes ranked according to their similarity with a given set of disease-specific GO terms*

The first view retrieves symbols, species, cytogenetic localization and rank of the ACGR DB genes, sorted according to decreasing rank. Human, mouse and fly genes are thus inter-classified according to their rank. Eventually, mouse genes can be better ranked

than their human orthologs because of richer annotation in the model organism. The better ranked the gene in View1, the stronger the relationship with the disease.

View2: *Human orthologs of genes from model organisms ranked according to their similarity with a given set of disease-specific GO terms*

This second view displays all features of View1 for genes retrieved from model organisms (here mouse and fly) together with the gene symbol, cytogenetic localization and rank of their human orthologs. Good ranking of a mouse gene can drag its human ortholog up to the top of View2 when it was at bottom in View1 because of poor GO annotation in human (see below the case of the *CHD7* gene responsible for CHARGE syndrome).

View3: *Genes interacting with the genes from View1*

For each gene in View1, the symbol, cytogenetic localization and rank of the genes reported as interacting with it (mostly via the gene products but other types of interactions are not excluded) are displayed. The source of information concerning these interactions is also displayed. Only intraspecies interactants are listed here (i.e. human interactants of human genes, mouse interactants of mouse genes, and fly interactants for fly genes). Candidate genes displaying proper cytogenetic mapping but poor ranking may reveal here interactions with well-ranked genes mapped anywhere else in the genome.

View4: *Human orthologs of genes from model organisms which interact with genes ranked according to their similarity with a given set of disease-specific GO terms.*

View4 is intended to retrieve candidate genes which are human orthologs of genes retrieved from model organisms and displaying interactions with well-ranked genes.

When experimental data are available, they can be included into each view described above producing View1Exp to View4Exp. An example is shown below in the AICARDI case-study.

Further queries on the four views defined above can then refine the four basic definitions with more specific hypotheses. Defining the sets of annotated candidate genes as SQL views allows biologists to benefit of the numerous advantages of this powerful mechanism. Besides the simplification of the query writing, the automatic updating of the views after data refreshment (adding new genes, new annotations or new experimental data) is also relevant. Defining views also contributes to DB security since end-users can receive tuned privileges on views but not necessarily on seminal data tables.

4 IMPLEMENTATION

The technical choices performed in this work for implementation are not mandatory since other techniques are envisageable depending on the deployment environment.

Wrappers for retrieving and integrating data from single data sources were implemented using the Xcollect software (Devignes et al., 2005). Xcollect scenarios are configured to automatically formulate queries, send them to the remote web resource, parse the returned HTML document and store the desired data in an XML document. Capturing the date of last database update is included in each scenario for ensuring data quality tracking. Specific Xcollect scenarios used here are available in the supplementary files.

Data sources were selected by the expert according to his preferences with respect to updating frequencies, annotation quality and coverage. Hence GO terms corresponding to keywords describing the disease were retrieved from AMIGO database, gene symbols of genes annotated with selected GO terms were retrieved from Entrez-Gene at NCBI for human genes and from MGI and Flybase databases for mouse and drosophila genes respectively, gene annotations concerning associated GO terms, genome localization, variant phenotypes and interactants were retrieved from Entrez-Gene at NCBI. Symbols of orthologous genes were retrieved from Entrez-HomoloGene.

Storage of collected data in ACGR DB was performed thanks to XSL transformations designed for converting each Xcollect session document into appropriate SQL commands.

Besides Xcollect wrappers, we developped a wrapper to invoke the GO-Family program available in the GOToolBox (Martin et al., 2004). The program was slightly modified since a list of GO-terms rather than reference gene symbols is provided as input to the system together with the list of genes to be ranked. Briefly the program fetches all GO-terms annotating the candidate genes, as well as their parent terms. It also fetches all parents of the disease-specific GO terms. It then computes a similarity percentage taking into account identical and non identical terms between the set of GO terms associated with each candidate gene and the set of disease-specific GO terms.

The EasyPHP package including a web server (Apache), a DBMS (mySQL) and a script language (PHP) was used for data management and user interface development.

5 RESULTS AND DISCUSSION

5.1 Three case-studies

Three complex and rare syndromes (CHARGE, GOLTZ and AICARDI) have been selected from the literature to illustrate the ACGR approach. Responsible genes have been reported for the first two syndromes (CHARGE and GOLTZ). However, since the ACGR approach does not make use of known disease-gene associations when building sets of candidate genes, it is quite relevant to test it on already elucidated diseases.

CHARGE syndrome (OMIM #214800) associates coloboma (defect of the iris of the eye, visible as a hole, split, or cleft in the iris) or choanal atresia (skull defect leading to the closure of one or both posterior nasal cavities) with retardation in growth and development and various other typical malformations affecting heart, ear and genitor-urinary system. Birth incidence is about 1 in 12000. Up to 2004 neither the gene responsible, nor the exact chromosomal locus of CHARGE syndrome were known. The comparative genomic hybridization (CGH) array method then revealed copy number variation (CNV) at 8q12 that concerned a total of 9 genes. Sequencing of all 9 genes in patients lead to identification of the *CHD7* gene as the responsible gene (Vissers et al., 2004). This gene encodes chromodomain helicase DNA-binding protein 7 and is composed of 38 exons.

GOLTZ syndrome (OMIM #305600) or focal dermal hypoplasia is inherited as X-linked dominant with *in utero* lethality in males. The features include atrophy and linear pigmentation of the skin, herniation of fat through the dermal defects, and multiple papillo-

mas of the mucous membranes or skin. In addition, digital anomalies consist of syndactyly, polydactyly, camptodactyly, and absence deformities. One responsible gene has recently been identified thanks to CGH array method (Grzeschik *et al.*, 2007; Wang et al., 2007). The Porcupine gene (*PORCN*) maps at Xp11.23, it belongs to an evolutionarily conserved gene family that encodes endoplasmic reticulum proteins with multiple transmembrane domains. Porcupine proteins are involved in the processing of Wnt (*wingless* and *int* drosophila homologues) proteins. The gene is composed of 15 exons and multiple alternatively spliced transcript variants encoding distinct isoforms have been observed.

AICARDI syndrome (OMIM %304050) is the initial motivation for the ACGR approach since experimental investigations are currently carried out in the laboratory (Yilmaz *et al.*, 2007). This neuro-developmental disorder is characterized by a typical triad of anomalies that include complete or partial agenesis of the corpus callosum, typical chorioretinal lacunae, and severe early-onset seizures, starting as infantile spasms (Aicardi *et al.*, 1969). However, the complete triad is not absolutely needed to diagnose the disease. Additional common features are related to developmental defects of the central nervous system, eye and skeletal systems (Aicardi, 2005). Mental retardation is most of the time severe although a recent publication in 2007 reports a girl with normal intellectual abilities (Grosso *et al.*, 2007). This underlines the extremely broad phenotype variations of this syndrome.

Concerning AICARDI syndrome genetics, all cases known to date are sporadic except for one isolated pair of sisters (Molina *et al.*, 1989). The phenotypic expression of the disease is restricted to individuals carrying two X chromosomes (Aicardi, 1999; Aicardi, 2005; Hopkins *et al.*, 1979; Wettke-Schafer and Kantner, 1983) and AICARDI syndrome is therefore considered as an X-linked dominant disorder caused by a de novo heterozygous mutation of a gene that is subject to X-chromosomal inactivation (Sutton *et al.*, 2005; Van den Veyver *et al.*, 2004). From the few dozens of patients investigated so far, no consensus caryotypic anomalies have yet been detected that unequivocally associate with Aicardi syndrome, neither translocations, nor micro-deletions, nor copy-number variations (Yilmaz *et al.*, 2007). Thus, the whole X chromosome is candidate. Various genes have been considered already as candidate genes for Aicardi syndrome but sequencing studies never confirmed their involvement (Nielsen *et al.*, 1991; Prakash *et al.*, 2002; Van den Veyver *et al.*, 2004). Differential microarray hybridization was carried out in the laboratory for studying genome-wide gene expression levels in patients compared to their mother (Yilmaz, 1977). A set of 10 mother-daughter pairs was used and ANOVA method was applied to the data leading to the identification of about 300 significantly dysregulated genes. These genes constitute the experimental dataset that will be crossed with gene annotations retrieved by the ACGR system in this study.

5.2 Expressing disease description as a set of GO terms

Table 2 summarizes for the three case-studies the correspondence between disease phenotypes, keywords and GO terms. In the case of AICARDI syndrome, the third symptom (infantile spasms) did not yield any specific GO term but according to the expert, it is covered by "Forebrain development" GO term.

Table 2. List of GO terms defined by the expert on August 31, 2007. Phenotypes were selected from OMIM notices as determinant for the diagnosis. Keywords (not shown, see supplementary file) were chosen as reflecting at best each phenotype. For a given keyword, GO terms were selected at the relevant level of the GO hierarchy. A GO term is included when all its children are relevant. The number of genes annotated by a GO term is indicated between brackets.

Syndrome	Phenotype	GO term (Biological Process section)
CHARGE	Coloboma	Camera-type eye morphogenesis [47]
	Choanal atresia	Nose development [2] Embryonic cranial skeleton morphogenesis [16]
	Ear abnormality Deafness	Ear development [155] Sensory perception of sound [203]
	Heart anomaly	Heart morphogenesis [69]
GOLTZ	Skin defects	Skin development [22]
	Digital anomalies	Embryonic digit morphogenesis [28]
	Skeletal defects	Embryonic skeletal morphogenesis [25]
AICARDI	Corpus callosum agenesis	Forebrain development [191] Corpus callosum development [0] Corpus callosum morphogenesis [0] Neuron migration [139] Neural plate development [117]
	Chorioretinal lacunae	Camera-type eye morphogenesis [47]

Table 3 provides quantitative monitoring of ACGR DB in terms of collected genes for the three case-studies (see section 4 for data origin). The *#GO* column displays the number of GO terms specific of the disease. The #f*ly*, *#mouse*, *#human* columns indicate the number of genes annotated by at least one of these GO terms for each organism. The column *#dysregulated* indicates the number of dysregulated genes entered in ACGR DB when experimental data are available. The last column provides the total number of genes after the retrieval of orthologuous and interacting genes.

Table 3. Monitoring collected genes in ACGR DB for the 3 case-studies.

Disease	#GO	#fly	#mouse	#human	#dysregulated	#genes
CHARGE	6	29	172	223	0	1410
GOLTZ	3	0	55	272	0	1583
AICARDI	6	2	182	166	300	2218

5.3 Building sets of annotated candidate genes

Views 1 to 4 were constructed for each case-study as described in section 3.3 for enabling queries reflecting expert hypotheses about candidate genes.

Table 4 displays the first four tuples from CHARGE View2. The human *CHD7* gene responsible for the disease appears at second position as orthologous to the mouse Chd7 gene which is well

ranked (48 %). It is worth noting that the low similarity of human *CHD7* gene annotation with the CHARGE GO terms (4%) relegates it at the bottom of View1. Selecting human genes with a cytogenetic localization at 8q12 in CHARGE View2 yields as first-ranked candidate gene the *CHD7* gene.

Table 4. The first four tuples from CHARGE View2. Orthol_Symbol, Orthol_Cytoband and Orthol_Value columns concern the human ortholog of considered mouse genes

Symbol	Organism	Cytoband	Value (%)	Orthol_ Symbol	Orthol_ Cytoband	Orthol_ Value
Tmie	mouse	9 64.0 cM	62	*TMIE*	3p21	62
Chd7	mouse	4 1.0 cM	48	*CHD7*	8q12.2	4
Gjb6	mouse	14 22.5 cM	48	*GJB6*	13q12	45
Dfna5h	mouse	6 B2.3	47	*DFNA5*	7p15	49

This case-study shows that the ACGR approach would have been able to designate the *CHD7* gene as the best candidate gene among the nine genes located at 8q12 thus prioritizing its sequencing. One can notice that although the association *CHD7*-CHARGE syndrome has been established for 3 years the Entrez-Gene anGO annotation of the gene still does not reflect it, probably due to the lack of evidence for precise genotype-phenotype association.

Table 5 shows the first five tuples from GOLTZ View4. In fact, an interaction is reported in the Entrez-Gene database between the mouse Wnt7a and Porcn genes but not between their human orthologs. The human *PORCN* gene is poorly ranked (7 %). However it appears well ranked in GOLTZ View4. Thus, the ACGR approach could have pointed to the *PORCN* gene before the mapping refinement provided by the CGH array experiment (Grzeschik *et al.*, 2007; Wang et al., 2007).

In the case of AICARDI syndrome Views 1Exp to 4Exp were constructed including transcriptomic data. A first query on View1Exp retrieved 71 genes located on human chromosome X. Table 6 displays the first four genes of this list. The best-ranked *PLXNA3* gene appears an interesting candidate. Its annotation is significantly similar to AICARDI GO terms (56%). So far, it has not been associated with any human disease. This 15-kb long gene with 26 exons mapping at Xq28 is an interesting but challenging candidate for sequence analysis aimed at discovering mutations in patients' genomes. The following *ARX* and *SOX3* genes are both responsible for diseases involving mental retardation: MRX54 (OMIM #300419) and MRGH (OMIM #300123) respectively. The following *DCX* gene is a good internal control since it is responsible for X-linked lissencephaly (LISX, OMIM #300067), a disease involving agenesis of corpus callosum and multiple heterotopia.

An interesting conclusion drawn from View1Exp analysis is that the best-ranked X-located genes for AICARDI disease are not dysregulated in our transcriptomic experiments ("Ratio" column at 0 in Table 6). AICARDI View2Exp was used to verify that the poor ranking of dysregulated human genes does not reflect lack of annotation in human genes when compared with model organisms. Actually all mouse or fly orthologs of the dysregulated X-located genes of View2Exp have a ranking below 10% except for mouse Dmd gene (17 %). However human *DMD* gene is already associated to the Duchenne muscular dystrophy and cannot be considered as a valuable candidate gene for AICARDI syndrome. This shows that in this case-study and for this transcriptomic dataset, dysregulated genes are not good candidate genes.

Further queries were applied to AICARDI View3Exp to explore possible interactions between dysregulated genes and candidate genes. Table 7 shows four candidate genes ("Interac_Symbol" column) from View3Exp, located on chromosome X and interacting with the four best-ranked dysregulated genes ("Symbol" column). The *MAGED1* gene interacts with the *DLX5* gene which is dysregulated in our transcriptomic experiments and its GO annotation displays 50 % similarity with the AICARDI specific GO terms. The interaction between these two gene products is based on in vivo experiments [Masuda et al., 2001].

The *UBQLN2* is also an interesting candidate. It interacts with the *UBE3A* gene which has been associated with Angelmann syndrome (OMIM #105830) that involves among others mental retardation and seizure symptoms. The other genes retrieved by the query are more difficult to interpret as candidate genes.

Further exploitation of View3Exp consisted in ranking the interacting genes located on chromosome X according to the decreasing number of their interactants. The *IKBKG* gene located at Xq28 was thus found interacting with 28 genes and the *DMD* gene with 17 genes. These two genes are already associated with diseases showing no significant overlap with AICARDI syndrome. The following *SUV39H1* gene has 10 interactants among which the *CBX5* and *PCM1* genes that are found dysregulated in our transcriptomic dataset. The next two genes, *RNF12* and *MAGED1*, have 8 and 7 interactants respectively among which one is dysregulated. Apart from the already mentioned *MAGED1* gene, this observation points to the *SUV39H1* and *RNF12* genes as candidate genes for AICARDI syndrome.

Table 5. The first five tuples from GOLTZ View4. Orthol_Symbol, Orthol_Cytoband and Orthol_Value columns concern the human orthologs of considered mouse interacting genes.

Symbol	Organism	Cytoband	Value (%)	Interac_ Symbol	Source	Interac_ Cytoband	Interac_Value (%)	Orthol_ Symbol	Orthol_ Cytoband	Orthol_ Value(%)
Gna12	Mouse	5 82.0 cM	35	Ppp5c	BIND	7 4.0 cM	4	*PPP5C*	19q13.3	4
Col5a2	Mouse	1 C1	32	Smad2	BIND	18 48.0 cM	15	*SMAD2*	18q21.1	17
Col5a2	Mouse	1 C1	32	Smad7	BIND	18 unknown	5	*SMAD7*	18q21.1	5
Col5a2	Mouse	1 C1	32	Samd3	BIND	9 unknown	15	*SMAD3*	15q22.33	16
Wnt7a	mouse	6 39.5 cM	31	Porcn	BIND	X 2.15 cM	5	***PORCN***	**Xp11.23**	7

All these new results have motivated the launching of a new sequencing campaign for AICARDI candidate genes as well as new CGH array experiments aimed at investigating the Xp11.23 and Xq28 regions at high resolution.

Table 6. The first four human genes mapping to chromosome X from AICARDI View1Exp

Symbol	Organism	Value (%)	Cytoband	Ratio
PLXNA3	human	56	Xq28	0
ARX	human	40	Xp21	0
SOX3	human	34	Xq27.1	0
DCX	human	26	Xq22.3-q23	0

5.4 Discussion

In summary, the ACGR approach has yielded very satisfying results in the CHARGE and GOLD case-studies. We have shown that in both cases the responsible gene has been retrieved by the system at an appropriate rank. Thus, the ACGR approach would have been useful at the time of their discovery to avoid unnecessary sequencing.

AICARDI case-study is a particularly difficult one due to the small number of recruited patients. However the ACGR approach has now provided several meaningful and promising candidate genes. For instance the *MAGED1* gene displays several features associated with disease genes (Tu et al., 2006): it is 99.3 kb long due to a giant intron (91 kb) separating the first exon from the 12 other exons which are grouped over the remaining 8 kb. Interestingly three of the retrieved candidate genes (*MAGED1, SUV39H1, UBQLN2*) are located in the same cytogenetic band (Xp11.23) which is known to be correlated with several neuro-psychiatric disorders.

Beyond the factual results, these case-studies have revealed satisfying user feedback since the ACGR approach enables them expressing various candidate gene definitions and interpreting the retrieval and ranking of each candidate gene.

The main difference between ACGR and other systems aimed at identifying candidate genes is its flexibility in candidate gene definition. Most computational approaches use a standard "similar disease, similar gene" hypothesis. Some systems (Endeavour, GeneLibrarian, GFSST, UCSC Genesorter) were tested for the three diseases studied in this paper but results were deceiving. In fact their rationale does not cope with indirect definitions in which the candidate gene is "a gene orthologous to a gene similar to another disease gene" or "that interacts with a gene similar to another disease gene". In addition these systems do not allow integrating private experimental data.

Nevertheless it can be envisaged to re-use some of the published procedures for disease-gene studies in the ACGR approach. For instance, data about interactions networks could be retrieved from the high-confidence protein complexes curated by Lage et al. (2007). The GO-Family algorithm used for gene ranking in this study could be replaced by another software computing similarity measurement between GO terms (Khatri and Draghici, 2005; Lord *et al.*, 2003; Zhang *et al.*, 2006, Wang et al., 2007). Our approach can also be extended to other ranking criteria such as similarity between eVOC terms annotating sites of gene expression and affected tissues (Hide *et al.*, 2003; Tiffin et al., 2005). In that case the data model would have to be adapted and the eVOC annotation collected just as the GO annotation so far. Finally, multi-criteria ranking methods could be integrated assuming that the proper data are collected. This would enlarge the field of investigation and circumvent the limits of GO annotations.

From a computer science point of view, the ACGR approach presented in this paper is a model-driven integration solution enabling complex querying. The ACGR methodology includes the rigorous construction of a database in which collected data are available for processing (ranking according to functional similarity with the disease) and querying tasks. Views corresponding to several candidate gene definitions are constructed to facilitate query writing. Origin of collected data is fully traced. The effort of carefully designing a database is valuable because it renders various exploitations possible such as data mining procedures for searching hidden patterns in the data.

The ACGR approach was motivated by the search for candidate genes for rare diseases. However our methodology can be adapted to other problems in which multiple complex queries are to be formulated on a set of integrated heterogeneous up-to-date data.

ACKNOWLEDGEMENTS

This work was funded by the Contrat de Plan Etat-Région Lorraine (PRST Intelligence Logicielle). We thank Sylvain Lambermont for its contribution at early stage of the work and Dr Leheupp for helping in selecting disease specific GO terms. S.Y. was supported by the AAL association and Region Lorraine

Table 7. The first four human tuples from AICARDI View3Exp. Symbol to Ratio columns refer to the dysregulated genes and Interac_Symbol to Interac_Ratio columns refer to the interacting candidate genes. The Source column indicates the database where the interaction is documented.

Symbol	Cytoband	Value (%)	Ratio	Interac_Symbol	Interac_Cytoband	Interac_Value (%)	Interac_Ratio	Source
DLX5	7q22	50	1	*MAGED1*	Xp11.23	3	0	HPRD
UBE3A	15q11-q13	22	1	*UBQLN2*	Xp11.23-p11.1	8	0	HPRD
CXCL10	4q21	21	1	*CXCR3*	Xq13	10	0	HPRD
IGF1	12q22-q23	21	1	*IGSF1*	Xq25	6	0	BIND

REFERENCES

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization, *BMC bioinformatics*, **6**, 55.

Adie, E.A., Adams, R.R., Evans, K.L., Porteous, D.J. and Pickard, B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates, *Bioinformatics (Oxford, England)*, **22**, 773-774.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P. and Moreau, Y. (2006) Gene prioritization through genomic data fusion, *Nature biotechnology*, **24**, 537-544.

Aicardi, J. (1999) Aicardi syndrome: Old and new findings, *Int. Pediatr*, **14**, 5-8.

Aicardi, J. (2005) Aicardi syndrome, *Brain & development*, **27**, 164-171.

Aicardi, J., Chevrie, J.J. and Rousselie, F. (1969) [Spasma-in-flexion syndrome, callosal agenesis, chorioretinal abnormalities], *Archives francaises de pediatrie*, **26**, 1103-1120.

ANSI/X3/SPARC Study Group on Data Base Management Systems, Interim Report, FDT 7 No. 2, ACM, New York, 1975.

Barillot, R., Poix, J., Groppi, A., Barre, A., Goffard, N., Sherman, D., Dutour I., and de Daruvar, A. (2004): New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucl. Acids Res.* **32**(12), 3581-3589.

Bodenreider, O. (2004): The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucl. Acids Res.* **32** (Database-Issue), 267-270.

Calvo, B., Lopez-Bigas, N., Furney, S.J., Larranaga, P. and Lozano, J.A. (2007) A partially supervised classification approach to dominant and recessive human disease gene prediction, *Computer methods and programs in biomedicine*, **85**, 229-237.

Chiang, J.H., Shin, J.W., Liu, H.H. and Chin, C.L. (2006) GeneLibrarian: an effective gene-information summarization and visualization system, *BMC bioinformatics*, **7**, 392.

Codd, E.F. (1970) A relational Model of Data for Large Shared Data Banks, *Communications of the ACM,* **13**, 377-387.

Devignes, M.D., de Palma, H., Pierron, L., Domenjoud, L. and Smaïl-Tabbone, M. (2005) User-designed web services to support heterogeneous biological data retrieval. *NETTAB workshop on Workflows management: new abilities for the biological information overflow*, http://www.nettab.org/2005/ progr.html.

Franke, A., Wollstein, A., Teuber, M., Wittig, M., Lu, T., Hoffmann, K., Nurnberg, P., Krawczak, M., Schreiber, S. and Hampe, J. (2006) GENOMIZER: an integrated analysis system for genome-wide association data, *Hum Mutat*, **27**, 583-588.

Freudenberg, J. and Propping, P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes, *Bioinformatics (Oxford, England)*, **18 Suppl 2**, S110-115.

George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D. and Wouters, M.A. (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction, *Nucleic acids research*, **34**, e130.

Grosso, S., Lasorella, G., Russo, A., Galluzzi, P., Morgese, G. and Balestri, P. (2007) Aicardi syndrome with favorable outcome: Case report and review, *Brain & development*, **29**, 443-446.

Grzeschik, K.H., Bornholdt, D., Oeffner, F., Konig, A., Del Carmen Boente, M., Enders, H., Fritz, B., Hertl, M., Grasshoff, U., Hofling, K., Oji, V., Paradisi, M., Schuchardt, C., Szalai, Z., Tadini, G., Traupe, H. and Happle, R. (2007) Deficiency of PORCN, a regulator of Wnt signaling, is associated with focal dermal hypoplasia, *Nat Genet*, **39**, 833-835.

Hide, W., Smedley, D., McCarthy, M. and Kelso, J. (2003) Application of eVOC: controlled vocabularies for unifying gene expression data, *Comptes rendus biologies*, **326**, 1089-1096.

Hopkins, I.J., Humphrey, I., Keith, C.G., Susman, M., Webb, G.C. and Turner, E.K. (1979) The Aicardi syndrome in a 47, XXY male, *Australian paediatric journal*, **15**, 278-280.

Kent, W.J., Hsu, F., Karolchik, D., Kuhn, R.M., Clawson, H., Trumbower, H. and Haussler, D. (2005) Exploring relationships and mining data with the UCSC Gene Sorter, *Genome research*, **15**, 737-741.

Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems, *Bioinformatics (Oxford, England)*, **21**, 3587-3595.

Lage, K., Karlberg, E.O., Størling, Z.M., Ólason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y. and Brunak, S. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology,* **285**, 309-316.

Lopez-Bigas, N., Blencowe, B.J. and Ouzounis, C.A. (2006) Highly consistent patterns for inherited human diseases at the molecular level, *Bioinformatics (Oxford, England)*, **22**, 269-277.

Lopez-Bigas, N. and Ouzounis, C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease, *Nucleic Acids Research*, **32**, 3108-3114.

Lord, P.W., Stevens, R.D., Brass, A. and Goble, C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation, *Bioinformatics (Oxford, England)*, **19**, 1275-1283.

Marquet, G., Burgun, A., Moussouni, F., Guerin, E., Le Duff, F. and Loreal, O. (2003) BioMeKe: an ontology-based biomedical knowledge extraction system devoted to transcriptome analysis, *Studies in health technology and informatics*, **95**, 80-85.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D. and Jacq, B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology, *Genome biology*, **5**, R101.

Masseroli, M., Galati, O. and Pinciroli, F. (2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists, *Nucleic acids research*, **33**, W717-723.

Masseroli, M., Martucci, D. and Pinciroli, F. (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining, *Nucleic acids research*, **32**, W293-300.

Molina, J.A., Mateos, F., Merino, M., Epifanio, J.L. and Gorrono, M. (1989) Aicardi syndrome in two sisters, *The Journal of pediatrics*, **115**, 282-283.

Mootha, V.K., Lepage, P., Miller, K., Bunkenborg, J., Reich, M., Hjerrild, M., Delmonte, T., Villeneuve, A., Sladek, R., Xu, F., Mitchell, G.A., Morin, C., Mann, M., Hudson, T.J., Robinson, B., Rioux, J.D. and Lander, E.S. (2003) Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 605-610.

Nielsen, K.B., Anvret, M., Flodmark, O., Furuskog, P. and Bohman-Valis, K. (1991) Aicardi syndrome: early neuroradiological manifestations and results of DNA studies in one patient, *American journal of medical genetics*, **38**, 65-68.

Oti, M., Snel, B., Huynen, M.A. and Brunner, H.G. (2006) Predicting disease genes using protein-protein interactions, *J Med Genet*, **43**, 691-698.

Perez-Iratxeta, C., Bork, P. and Andrade, M.A. (2002) Association of genes to genetically inherited diseases using data mining, *Nature genetics*, **31**, 316-319.

Perez-Iratxeta, C., Wjst, M., Bork, P. and Andrade, M.A. (2005) G2D: a tool for mining genes associated with disease, *BMC genetics*, **6**, 45.

Prakash, S.K., Cormier, T.A., McCall, A.E., Garcia, J.J., Sierra, R., Haupt, B., Zoghbi, H.Y. and Van Den Veyver, I.B. (2002) Loss of holocytochrome c-type synthetase causes the male lethality of X-linked dominant microphthalmia with linear skin defects (MLS) syndrome, *Human molecular genetics*, **11**, 3237-3248.

Rossi, S., Masotti, D., Nardini, C., Bonora, E., Romeo, G., Macii, E., Benini, L. and Volinia, S. (2006) TOM: a web-based integrated approach for identification of candidate disease genes, *Nucleic acids research*, **34**, W285-292.

Shaw-Smith, C., Redon, R., Rickman, L., Rio, M., Willatt, L., Fiegler, H., Firth, H., Sanlaville, D., Winter, R., Colleaux, L., Bobrow, M. and Carter, N.P. (2004) Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features, *J Med Genet*, **41**, 241-248.

Spinazzola, A., Viscomi, C., Fernandez-Vizarra, E., Carrara, F., D'Adamo, P., Calvo, S., Marsano, R.M., Donnini, C., Weiher, H., Strisciuglio, P., Parini, R., Sarzi, E., Chan, A., DiMauro, S., Rotig, A., Gasparini, P., Ferrero, I., Mootha, V.K., Tiranti, V. and Zeviani, M. (2006) MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion, *Nature genetics*, **38**, 570-575.

Sun, H., Fang, H., Chen, T., Perkins, R. and Tong, W. (2006) GOFFA: Gene Ontology For Functional Analysis - A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data, *BMC bioinformatics*, **7 Suppl 2**, S23.

Sutton, V.R., Hopkins, B.J., Eble, T.N., Gambhir, N., Lewis, R.A. and Van den Veyver, I.B. (2005) Facial and physical features of Aicardi syndrome: infants to teenagers, *American journal of medical genetics*, **138**, 254-258.

Teorey, T.J., Lightstone, S.S. and Nadeau, T. (2006) *Database Modeling and Design: Logical Design.* The Morgan Kaufmann.

Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA.(2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates, *Nucl. Acids Res.*, **33**,1544-52.

Tu, Z., Wang, L., Xu, M., Zhou, X., Chen, T. and Sun, F. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes, *BMC genomics*, **7**, 31.

Turner, F.S., Clutterbuck, D.R. and Semple, C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes, *Genome biology*, **4**, R75.

Van den Veyver, I.B., Panichkul, P.P., Antalffy, B.A., Sun, Y., Hunter, J.V. and Armstrong, D.D. (2004) Presence of filamin in the astrocytic inclusions of Aicardi syndrome, *Pediatric neurology*, **30**, 7-15.

van Driel, M.A., Cuelenaere, K., Kemmeren, P.P., Leunissen, J.A., Brunner, H.G. and Vriend, G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases, *Nucl. Acids Res.*, **33**, W758-761.

Vermeesch, J.R., Fiegler, H., de Leeuw, N., Szuhai, K., Schoumans, J., Ciccone, R., Speleman, F., Rauch, A., Clayton-Smith, J., Van Ravenswaaij, C., Sanlaville, D., Patsalis, P.C., Firth, H., Devriendt, K. and Zuffardi, O. (2007) Guidelines for molecular karyotyping in constitutional genetic diagnosis, *Eur J Hum Genet*.

Vissers, L.E., van Ravenswaaij, C.M., Admiraal, R., Hurst, J.A., de Vries, B.B., Janssen, I.M., van der Vliet, W.A., Huys, E.H., de Jong, P.J., Hamel, B.C., Schoenmakers, E.F., Brunner, H.G., Veltman, J.A. and van Kessel, A.G. (2004) Mutations in a new member of the chromodomain gene family cause CHARGE syndrome, *Nat Genet*, **36**, 955-957.

Vissers, L.E., Veltman, J.A., van Kessel, A.G. and Brunner, H.G. (2005) Identification of disease genes by whole genome CGH arrays, *Human molecular genetics*, **14 Spec No. 2**, R215-223.

Wang, X., Reid Sutton, V., Omar Peraza-Llanes, J., Yu, Z., Rosetta, R., Kou, Y.C., Eble, T.N., Patel, A., Thaller, C., Fang, P. and Van den Veyver, I.B. (2007) Mutations in X-linked PORCN, a putative regulator of Wnt signaling, cause focal dermal hypoplasia, *Nat Genet*, **39**, 836-838.

Wettke-Schafer, R. and Kantner, G. (1983) X-linked dominant inherited diseases with lethality in hemizygous males, *Hum Genet*, **64**, 1-23.

Xu, J. and Li, Y. (2006) Discovering disease-genes by topological features in human protein-protein interaction network, *Bioinformatics (Oxford, England)*, **22**, 2800-2805.

Yilmaz, S. (2007) Recherche de gènes candidats responsables du Syndrome d'Aicardi: complémentarité des approches expérimentales et bioinformatiques. PhD Thesis, Nancy University Henri Poincaré, France.

Yilmaz, S., Fontaine, H., Brochet, K., Gregoire, M.J., Devignes, M.D., Schaff, J.L., Philippe, C., Nemos, C., McGregor, J.L. and Jonveaux, P. (2007) Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH, *Eur J Med Genet*.

Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. and Chen, C.-F. (2007) A New Method to Measure the Semantic Similarity of GO Terms, *Bioinformatics (Oxford, England)*, **23**, 1274-1281.

Zhang, P., Zhang, J., Sheng, H., Russo, J.J., Osborne, B. and Buetow, K. (2006) Gene functional similarity search tool (GFSST), *BMC bioinformatics*, **7**, 135.

Discussion

Le syndrome d'Aicardi est une affection génétique rare et de survenue sporadique. Il s'agit probablement d'une pathologie dominante liée à l'X et létale chez les garçons hémizygotes. Cependant, cette hypothèse ne peut pas être testée puisque les patientes Aicardi qui présentent un lourd handicap et une faible espérance de vie n'ont pas de descendance. Les études d'inactivation de l'X ne révèlent aucune corrélation entre le phénotype clinique et le profil d'inactivation chez les patientes AIC. Nous n'avons relevé aucune élévation de l'âge moyen des mères ou des pères au moment de la procréation d'une fille AIC comme c'est par exemple le cas avec l'achondroplasie où il existe des néomuations d'origine paternelle dans le gène *FGFR3* (fibroblast growth factor receptor 3). Seule la caractérisation du gène candidat pourrait aider à tester cette hypothèse. Je discuterai dans cette session des résultats de l'analyse du caryotype moléculaire à l'aide de puces génomiques puis des résultats des puces transcriptomiques, et je terminerai par les études *in silico*.

1. Caryotype moléculaire des patientes AIC

L'efficacité des puces génomiques pour la détection d'anomalies quantitatives délétères n'est plus à prouver [223-229]. Les premières études utilisaient des puces génomiques souvent spécifiques d'un chromosome et d'une résolution de l'ordre de 1 Mb. Avec l'amélioration de la technologie, des puces à oligonucléotides ont permis de détecter un nombre plus important de CNV [103, 230]. Un consensus d'utilisation de la technique n'a pas encore été établi pour le caryotype moléculaire. Cependant, un certain nombre de points importants sont clairement définis et sont à contrôler avant de passer à l'interprétation des données pour établir une corrélation entre la variation quantitative et le phénotype observé [231, 232].

1.1. La détection des clones variants

1.1.1. Paramètres pouvant influencer la détection des clones variants

1.1.1.1. Les séquences répétées dans les sondes

Une des difficultés de l'utilisation des puces génomiques constituées de clones BAC est liée à la variabilité en taille et en composition de bases des sondes. La quantité d'ADN Cot-1 à ajouter pour masquer les séquences répétées n'est pas facile à déterminer. Il en résulte des variations dans les ratios d'intensité. Tout le problème est ensuite de déterminer si ces variations sont issues de ce biais ou si effectivement ce sont des variants quantitatifs. C'est le cas notamment du clone RP-11-761E20. Lors de la sélection des amorces pour la Q-PCR pour ce clone, une étape consiste à cacher les séquences répétées du clone à tester. Plus de 75 % (144 035pb sur 189 726pb) du clone correspond à des répétitions. Il est clair que pour ce BAC les variations observées sont dues aux séquences répétées et donc ce BAC n'a pas été retenu. Un autre clone (RP5-1178I21, variant pour la patiente KA) possède environ 50% de séquences répétées (30943pb/62267pb). Et enfin le clone RP11-66N11 (variant chez la patiente KA) environ 25 % (44 955pb/180 442pb).

1.1.1.2. Seuil de détection

Un seuil pour le ratio d'intensités permettant de déterminer si un clone est variant ou non doit être choisi. Il doit permettre d'éliminer les faux positifs tout en écartant le minimum de faux négatifs. Certaines études utilisent des seuils fixes. Dans ce cas la principale difficulté est la variabilité des ratios d'intensités des différents clones pour une même expérience. Pour des seuils fixés à $\log_2$ ratio = 0,3 (ou/et 0,5), le nombre de faux positifs dépendra de la variation des ratios d'intensités. Pour un seul fixe, lors de l'hybridation d'un même échantillon, une expérience avec des variations d'intensités élevées donnera plus de faux positif. Ainsi, l'utilisation des seuils variables a été proposée[232]. S'agissant des expériences

effectuées au laboratoire sur des puces génomiques spécifiques de l'X provenant du Flanders Interuniversity Institute of Human Genetics, nous avons donc utilisé les seuils variables et les critères de validité définis par les créateurs et utilisateurs de cette puce. Ainsi, des critères de qualité de la puce sont définis par un minimum de 95% d'hybridation, une intensité de signal au moins 2 fois supérieure à celle du bruit de fond et une déviation standard maximale de 0,096, seuil au-delà duquel les duplications ne sont plus détectées. Deux valeurs seuils pour la sélection des clones déviants (±4 DS et $\pm\log_2$ (3/2)-2 DS) ont été utilisées. Ces seuils prévoient un clone faux positif toutes les 4 hybridations pour une puce contenant 3000 clones. Sachant que nous avons utilisé les puces génomiques avec 1875 clones nous sommes donc dans des conditions idéales de détection.

1.1.2. Comment repérer les clones réellement variants

Les hybridations ont été effectuées et seuls les clones déviants lors des deux hybridations inverses Cy3/Cy5 et Cy5/Cy3 ont été sélectionnés. Lors des expériences de puce génomique il est également conseillé d'utiliser un témoin (interne au laboratoire ou commercial) [231]. Nous avions initialement opté pour un témoin femme présentant une délétion complète du gène *MECP2* (plus de 85kb), testant par la même occasion les limites de détection des puces utilisées. Pour 3 hybridations, le clone RP11-119A22 contenant *MECP2* n'apparaît pas délété (ratio de 0,80). Un autre témoin interne a été utilisé avec une délétion partielle du gène *DMD* (exons 18 à 44 couverts par 5 clones) et dans toutes les expériences de CGH réalisées, la délétion des 5 clones a été identifiée. L'irrégularité dans la détection de la délétion du gène *MECP2* chez le témoin a été expliquée par la mise en évidence ultérieure d'une mosaïque somatique pour la délétion chez le témoin (70% des cellules délétées). Cette hypothèse a été émise à la suite les expériences de Q-PCR chez cette patiente. Le ratio se situait dans un intervalle entre 0,7 et 0,8. A partir d'un nouveau prélèvement sanguin, nous avons confirmé par FISH et à l'aide d'une sonde spécifique du gène *MECP2* la présence une délétion dans 70% des cellules chez ce témoin.

1.2. CNV et maladies

Les variations du nombre de copies d'un segment d'ADN peuvent influencer l'expression des gènes de diverses façons et donc être impliquées dans des maladies. Le rôle des CNV dans les pathologies humaines est connu depuis l'élucidation de l'étiologie moléculaire de la maladie de Charcot-Marie-Tooth. La majorité des patients présentent une duplication du gène *PMP22* (peripheral myelin protein 22)[233]. Cette duplication est favorisée par une recombinaison homologue non allélique entre deux duplicons encadrant ce gène. D'une manière générale, et comme expliqué précédemment, les CNV peuvent contribuer à expliquer la variabilité phénotypique entre patients et aussi certaines différences inter-individuelles entre sujets témoins. Les CNV ont aussi une part importante dans l'étiologie des retards mentaux [103, 123, 212, 234-236]. Ces variations quantitatives peuvent ainsi être une réponse à la question des auteurs de la publication «Nonsyndromic X-linked mental retardation : where are the missing mutations ?» [237]. On peut également imaginer que les CNV expliqueraient en partie le décalage du sexe ratio chez les patients avec RM entre les garçons hémizygotes pour le chromosome X et les filles. On peut donc comprendre l'intérêt de l'utilisation des puces génomiques dans la recherche de causes moléculaires de RMLX. C'est dans cette optique que nous nous sommes intéressés à une puce spécifique de l'X pour la recherche de variations quantitatives chez les patientes AIC [212]. Cinq études ont été publiées sur des puces spécifiques de l'X [212, 226, 236, 238, 239]. Une de ces études décrit la mise en place de la puce spécifique de l'X que nous avons utilisée dans nos expériences pour analyser le génome de 18 filles AIC [212, 239]. Une seconde l'étude de 40 familles avec RMLX non syndromique a permis d'identifier 7 variants parmi lesquels 3 duplications ont été impliquées dans le phénotype observé [236]. Une autre étude menée sur 108 patients avec un RM a révélé 15 variants chez 14 patients (13%), chez 5 patients (4,6%), les auteurs ont pu démontré que les CNV étaient responsables du phénotype observé [226].

Dans notre étude avec des puces génomiques, nous avons volontairement sélectionné des clones donnant des ratios d'intensités se situant dans les limites des seuils choisis pour nous donner toutes les chances de détecter toutes les variations quantitatives chez les patientes AIC. Nous nous attendions donc à des faux positifs pour minimiser le nombre de faux

négatifs. Nous avons par la suite étudié tous les clones validés en hybridation inverse par Q-PCR. Sur 18 patientes, les puces génomiques ont détecté 4 clones à un seuil de 4 DS et 6 clones entre les seuils 4DS et log2 (3/2)-2XSD. Les résultats de la technique Q-PCR effectuée sur plusieurs amplicons par clone n'ont pas conduit à impliquer ces clones dans la pathologie.

Le clone CTD-2511C7 a présenté un intérêt particulier de par la présence du gène de la filamine A. Le gène de la filamine A contenu dans cette région est borné par deux duplications segmentaires et est associé à des inversions retrouvées dans la population normale [105, 240]. Des mutations dans ce gène ont été impliquées dans des hétérotopies périventriculaires familiales dominantes liées à l'X [241]. La recherche de mutations ponctuelles dans la séquence codante du gène *FLNA* s'étant révélée négative chez des patientes AIC dans une étude antérieure [76], nous avons ainsi voulu tester si un grand réarrangement du gène ou de la région pouvait intervenir dans la maladie. Cependant, la délétion détectée par les puces génomiques n'a pas été confirmée par les expériences de Q-PCR. On peut exclure chez ses 18 patientes de notre étude, les grands réarrangements touchant la totalité du gène de la *FLNA* comme causes moléculaires syndrome d'Aicardi. Les résultats des puces transcriptomiques et les données de la base ACGR n'ont pas orienté nos recherches vers ce gène qui peut donc être raisonnablement écarté en tant que candidat dans le syndrome d'Aicardi.

Notons cependant que le seuil de résolution de 80 kb des puces est clairement limitant et les anomalies quantitatives des autosomes n'ont pas été testées. L'hypothèse d'une pathologie autosomique reste cepandant possible. Dans ce cas, il faudrait imaginer que l'expression du gène est différente entre les deux sexes et donc qu'une mutation du gène n'affecte pas les garçons. Dans cette hypothèse, des puces pan-génomiques seraient plus adéquates, l'utilisation de telles puces est envisagée à court terme. Des puces à oligonucléotides, couvrant l'ensemble du génome sont aujourd'hui disponibles. Une plate-forme d'analyse du caryotype moléculaire conjuguant les efforts du laboratoire de génétique et les compétences informatiques du LORIA (Laboratoire Lorrain d'Informatique est ses Applications) sera opérationnelle d'ici la fin de l'année 2007. Ces puces très résolutives permettront de cribler l'ensemble du génome des patientes Aicardi (résolution de quelques kilobases) multipliant ainsi les chances de trouver une patiente avec une anomalie quantitative telle qu'une délétion partielle du gène AIC.

2. Etudes du transcriptome des patientes AIC

Les études du transcriptome sont riches en informations mais difficiles d'interprétation. Comme pour la partie puce génomique, la discussion sera tout d'abord orientée vers les consensus d'analyse établis puis j'aborderai le sens biologique des données présentées dans la session résultats. Les catégories fonctionnelles sur-représentées chez les patientes AIC seront discutées à la lumière de données de la littérature.

2.1. Hypothèses de travail et étude du profil d'expression du syndrome d'Aicardi

Plusieurs études globales du transcriptome ont été menées sur des maladies neurologiques et ont montré qu'un profil d'expression spécifique de la maladie était retrouvé dans le sang total [242-247]. Pour la maladie de Huntington (HD), des marqueurs spécifiques de la maladie ont été décrits, rendant possible la distinction grâce au profil d'expression du tissu sanguin, des patients témoins, des patients HD pré-symptomatiques et des patients HD symptomatiques [248]. Etant donné que les tissus cérébraux sont très difficiles à obtenir, seules les études sur le sang total sont donc réalisables pour ce type de maladies neurologiques. Les manifestions cliniques du syndrome d'Aicardi étant essentiellement le reflet d'une atteinte cérébrale et rétinienne, et n'ayant pas accès à ces types de tissu, l'étude d'expression a été réalisée sur des lymphocytes. Nous avons donc fait l'hypothèse que le gène responsable est exprimé dans le sang et / ou que les conséquences de la dérégulation de ce gène sont identifiables dans le transcriptome des lymphocytes.

2.2. Analyse des résultats d'expression : les consensus établis

Toujours en plein développement, l'analyse ou la comparaison des résultats obtenus avec les puces d'expression pose de sérieux problèmes d'homogénéité entre les études [138]. Je

justifierai dans ce qui suit les choix que nous avons effectués pour le plan expérimental (paragraphe 2.2.1) et l'analyse des données expérimentales (paragraphe 2.2.2). Je préciserai notamment le choix de la référence commune et celui des témoins. S'agissant des étapes de traitement de l'image et de normalisation, nous avons utilisé les paramètres par défaut établis pour les puces d'expression Agilent par le logiciel Feature extraction.

2.2.1. Choix du plan expérimental lors des études de puces transcriptomiques

Un plan expérimental avec une référence commune nous a paru être le plus adapté. Les ADN des patientes AIC et de leur mère ont pu ainsi être marqués par le même flurochrome, ce qui facilite les analyses statistiques sur l'ensemble des lames et rend les lames comparables entre elles.

La référence a été construite avec un pool de 10 témoins de sexe féminin. Les ARN commerciaux tels que la référence universelle de Stratagene (Universal Human Reference RNA) provenant du mélange d'ARN de lignées cellulaires différentes sont certes intéressants pour les comparaisons inter plates-formes et entre les laboratoires mais ne présentent pas d'intérêt dans notre cas d'étude. Notre étude n'a pas de raison d'être vérifiée ou reproduite ailleurs. D'autre part, l'homogénéité entre les échantillons nous a paru constituer un point important. Dans une hybridation entre les ARN du tissu sanguin d'une patiente et d'un mélange de lignées cellulaires pour la référence, la sélection des gènes différentiellement exprimés risque d'être biaisée de par la nature différente des deux extraits.

Une autre question importante a été le choix des **témoins**. Dans l'idéal, une sœur du même âge pour chaque patientes AIC aurait du être choisie. Cependant, le nombre de familles Aicardi est très faible et les fratries sont de petites tailles. Des filles témoins du même âge auraient pu être sélectionnées. Il est cependant très difficile d'obtenir des prélèvements sanguin de filles non majeures et en bonne santé. En outre, des polymorphismes d'expression existent entre les individus. Une étude sur le sang frais de 75 volontaires sains a révélé qu'il existait des différences interindividuelles dans le transcriptome de personnes saines. Les auteurs proposent que ces différences seraient soit causées par les différences dans les génotypes soit le reflet d'effets épigénétiques ou de l'environnement [249]. Ces

facteurs génotypiques peuvent vraisemblablement être les CNV. Nous avons donc choisi de comparer une fille Aicardi à sa mère plutôt qu'à un individu non apparenté, certes du même âge mais avec un transcriptome très différent.

2.2.2. Sélection des gènes signatures lors d'une expérience de puce transcriptomique

Dans les premières études globales du transcriptomie, la sélection des gènes signatures était basée sur un seuil de variation de ratio (« fold change »). Il est maintenant admis que la sélection de gènes par cette méthode est insuffisante à elle seule [138]. Dans nos études, nous avons voulu comparer les différences d'expression entre une fille AIC et sa mère. C'est ce que nous appelons l'analyse par famille. Nous avons utilisé une sélection par l'amplitude de ratio et une valeur p (p-value) inférieure à 1.E-02. Cependant, notre analyse ne s'est pas limitée à cette sélection puisque nous avons effectué un regroupement fonctionnel avec une approche statistique rendant donc notre méthode plus robuste. La seconde analyse des résultats a été effectuée par une approche ANOVA, outil statistique qui se révèle être particulièrement adapté aux données d'expression.

2.3. Les gènes signatures sélectionnés dans l'étude AS1 sont des candidats potentiels pour les RMLX

Rappelons que l'étude AS1 a été menée jusqu'au séquençage de gènes candidats avant d'avoir effectué les analyses avec la version 2.0 du logiciel luminator et donc nous n'avons pas attendu l'obtention des résultats de cette analyse informatique avant la sélection de gènes candidats pour le syndrome d'Aicardi. Ceci explique le séquençage de gènes candidats non retrouvés dans l'analyse AS2. Néanmois, nous avons noté que les gènes sélectionnés comme candidats (session résultats paragraphe .5.3.1) restent de bons candidats pour les RMLX. Comme le signal H.H Ropers dans une revue de 2002 intitulée « *X-linked mental retardation : many genes for a complex disorder »* [21], Les retards mentaux liés à l'X forment un groupe hétérogène résultant de mutations dans de nombreux gènes. On ne

s'étonnera donc pas que des voies biologiques communes soient retrouvées. Durant la recherche de gènes candidats pour le syndrome d'Aicardi, certains gènes signatures sélectionnés avaient déjà été impliqués dans des maladies avec retard mental. La liste des gènes candidats pour le syndrome d'Aicardi présente donc des candidats potentiels pour d'autres retards mentaux liés au chromosome X. Nous pouvons raisonnablement ajouter que ces gènes peuvent faire partie d'une voie biologique également impliquée dans le syndrome d'Aicardi et donc que ces gènes contribuent à leur mesure au phénotype des filles AIC. Trois gènes principaux, *SYN1*, *ASMT* et *PORC* ont été sélectionnés à la suite des expériences AS1 et validés par Q-PCR. *ASMT* et *SYN1* ont fait l'objet d'un séquençage des régions codantes et aucune mutation délétère n'a été détectée chez les filles AIC étudiées. Le gène *SYN1* (Synapsin 1) cartographiée en Xp11.23 code une phosphoprotéine associée à la surface des vésicules synaptiques. La famille des synapsines est impliquée dans la synaptogenèse et la modulation des neurotransmetteurs suggérant un rôle dans les maladies neurologiques sévères. Le gène *SYN1* a été impliqué dans une forme familiale de RMLX récessif avec de individus masculins touchés par une épilepsie variable. Cette protéine est également présentée comme candidat potentiel pour les RMLX par l'équipe de Laumonier et al. de par son appartenance au protéome post-synaptique (PSP) [22]. Le gène *ASMT* code la dernière enzyme de la voie de synthèse de la mélatonine, il est le seul gène dérégulé commun aux trois filles étudiées dans AS1. Une étude de 2007 a rapporté deux mutations délétères de ce gène, présente chez deux familles avec troubles autistiques. Les analyses biochimiques ont révélé une diminution importante de l'activité *ASMT* corrélée à une forte diminution du niveau de mélatonine chez les patients atteints [250].

2.4. Sens biologique des données d'expression

Donner un sens biologique aux résultats des puces transcriptomiques revient à imaginer les sources de variabilité possibles entre les échantillons. Nous ne discuterons pas ici de la variabilité technique issue des expériences mais uniquement de la variabilité biologique. Notre but étant d'écarter toutes les causes de variabilité du transcriptome non liées au syndrome d'Aicardi (biais) pour ne garder que l'effet de la maladie et comprendre par la suite son impact sur le transcriptome des filles AIC. Dans notre étude, parmi les causes

possibles de la variabilité du transcriptome chez les filles Aicardi par rapport à leur mère nous pouvons citer : Le facteur âge (paragraphe 2.3.1), le facteur heure du prélèvement (paragraphe 2.3.2),

2.4.1. Influence du facteur « âge » sur les transcriptomes des filles AIC

Le facteur âge est nécessairement un facteur d'impact fort sur les résultats. L'analyse ANOVA tient compte de ce biais par contre les analyses par famille, où l'on compare chaque fille à sa mère présentent nécessairement ce biais. Cinq filles sont incluses dans un premier groupe pré-pubère (PB : 6 ans, PAL : 16 ans, et LA : 6 ans, AM : 10 ans, LC : 11 ans et quatre filles de 20 à 25 ans composent le second groupe pubère (TE : 25ans, KA : 20 ans, DE : 23ans, CJ : 20ans). Les mères ont en moyenne 41 ans. L'impact de l'âge dans le premier groupe est sans doute plus important que pour le second. Cependant, vu l'intervalle qui sépare les filles pubères de leur mère une influence de l'âge existe aussi dans le second groupe. L'équipe de Whitney à l'université de Standford fut la première à montrer que les profils spécifiques du sang total sont corrélés à l'âge et au sexe chez 75 volontaires [249]. Leur étude montre une corrélation négative entre l'expression des gènes des immunoglobulines et le facteur âge. Une seconde étude sur le sang total montre également cette corrélation entre les gènes associés aux immunoglobulines et l'âge des patients [251]. Dans les études d'expression que nous avons menées (AS1 et AS2), le processus biologique « réponse immunitaire » et la catégorie fonction moléculaire « défense et activité des protéines immunes » contiennent les gènes des immunoglobulines. Toutes les analyses par famille (fille *vs* mère) révèlent une sur-représentation de ces catégories sans distinction claire entre les deux groupes pubères et pré pubères. Nous pouvons donc raisonnablement écarter cette catégorie comme impliquée dans le syndrome d'Aicardi.

2.4.2. Influence du facteur « heure du prélèvement » sur les transcriptomes des filles AIC

Un autre facteur de variabilité est vraisemblablement **l'heure du prélèvement**. Dans l'étude menée par Witney *et al.*, l'expression des gènes dans le sang varie de façon significative avec l'heure du prélèvement. Un groupe de gènes parmi lesquels 35 % codent des protéines ribosomales a ainsi été identifié [249]. Une étude sur le cycle circadien à partir de tissus (cœur et foie) chez la souris, a montré également des variations d'expression des protéines ribosomales [252]. Une autre étude du cycle circadien chez la souris a identifié le large groupe fonctionnel comprenant la synthèse protéique, les protéines ribosomales et les gènes impliqués dans le transport et la dégradation protéique [253]. Le contrôle de la synthèse et de la dégradation protéique pourrait refléter le cycle circadien des cellules en croissance (et/ou la présence de nutriments) et peut être fondamental dans le cycle diurne des mammifères [249]. Les études chez la souris citent notamment le gène RPS4X comme signature dans le cycle circadien. Dans l'étude menée par l'équipe de Witney sur le sang de témoins volontaires, *RPS4X* est également retrouvé comme dérégulé par le facteur heure du prélèvement. Ce gène avait été sélectionné par les expériences AS1 comme variable chez deux des trois filles et confirmé par Q-PCR. Les données issues de ces publications tendraient donc à exclure ce gène des candidats pour le syndrome AIC. Toutefois, il ne faut pas tirer des conclusions trop hâtives. Il est important de tenir compte des effets croisés des différents facteurs. Une publication récente de Laumonnier et al. [22] sur les gènes du protéome post-synaptique montre qu'un certain nombre d'entre eux sont responsables de RMLX. En effet, des mutations dans 6 des 7 gènes du NRC/MASC (85%) et 19 des 39 gènes du PSD (49%) ont été rapportées dans des familles avec RMLX. Les autres gènes non encore impliqués dans des maladies et liés au complexe PSP sont des candidats potentiels pour les RMLX. *RPSX4* fait partie de ce complexe. Il est cartographié en Xq13.1 et n'est impliqué à ce jour dans aucune maladie.

Dans nos études, les prélèvements sanguins extraits dans des tubes PAXgene™(Quiagen) nous ont été envoyés séquentiellement par les familles. Les heures de prélèvement sont nécessairement variables entre les filles et leur mère. Dans les expériences AS2, les résultats de EASE pour les études par famille révèlent que les catégories « ribosome », « complexe

ribonucleoproteique » (composant cellulaire) et « biosynthèse protéique », « biosynthèse des macromolécules » (processus biologiques) contiennent effectivement les même gènes ribosomaux et ce pour 6 des 8 couples étudiés. Ces catégories ne sont pas représentées dans l'étude AS1 et il peut être intéressant de noter que les expériences AS1 ont été effectué à partir de lignées lymphoblastiques. Notons que le clustering fonctionnel des 475 gènes issu des analyses par ANOVA n'identifit pas de sur-représentation de ces catégories probablement parce que cette approche n'identifie que les gènes dérégulés chez une majorité de filles AIC.

2.4.3. Influence du facteur environnement sur les transcriptomes des filles AIC

Les autres facteurs de variabilité comprennent les **médicaments, les toxines et les infections** [246].Ce sont des facteurs d'impact important de la variabilité du transcriptome. Or les filles AIC que nous avons étudiées sont toutes soumises à des traitements anti-épileptiques multiples et lourds.

Les mécanismes moléculaires de l'épilepsie et les voies impliquées dans ce groupe d'affections neurologiques ne sont pas encore bien connus et les quelques études globales du transcriptome ne sont pas concordantes. Une étude du transcriptome sur des tissus cérébraux de 12 patients souffrant d'épilepsie depuis plus de 10 ans implique des gènes du système GABA (gamma-aminobutyric acid), du métabolisme des lipides et de la cascade des MAPK (Mitogen Activated Potein Kinase)[254]. Ces voies et complexes protéiques n'ont pas été identifiés dans notre étude comme des catégories sur-représentées.

D'autres études ont tenté de trouver des voies dérégulées chez des patientes épileptiques. Ainsi l'une d'entre elles rapporte une corrélation forte entre l'administration d'acide valproique et l'inhibition des histones déacétylases *(HDAC)* [255]. L'acide valproïque est un traitement anti-épileptique et l'inhibition des *HDAC* est certainement une conséquence du traitement et non de la maladie, cependant la prise du médicament stoppe les crises et donc touche vraisemblablement la voie biologique impliquée dans l'épilepsie. Une autre étude sur des cellules HEK (Human Embryonic Kidney) traitées par de l'acide valproïque non seulement confirme son implication dans les modifications des histones mais aussi son rôle dans la

modification épigénique de l'ADN influençant ainsi l'expression de certains gènes [256]. Dans nos expériences, cette catégorie correspondrait dans les études AS1 aussi bien que AS2 à la catégorie « nucléosome » sur-représentée chez les patientes AIC. Les catégories « nucléosome » et « assemblage/ désassemblage de la chromatine » incluent les gènes. AM et TE (AS2) et CC (AS1) sont les trois patientes qui présentent la catégorie « assemblage du nucléosome » comme surreprésentée dans les analyses par famille.

2.4.4. Influence du facteur variabilité naturelle sur les transcriptomes des filles AIC

Il existe des gènes pour lesquels il existe une variabilité inter-individuelle physiologique. Parmi ces gènes, on peut citer les gènes dérégulés par les interférons [246]. Dans nos études d'expression, le gène *IFI27* fait partie de cette catégorie. Ce gène avait particulièrement attiré notre attention. Il est ainsi sous régulé chez les couples AS1, chez AM vs As, CJ vs CC, DE vs DMT, Ka vs CG, Pal vs PA et l'étude par ANOVA. Les amplitudes de ratio sont comprises entre -2 et -28. Le gène n'est pas cartographié sur le chromosome X et donc ne constitue pas un bon candidat pour le syndrome AIC cependant il avait attiré notre attention en raison de sa fréquence d'apparition dans les résultats. D'après la littérature, nous pouvons raisonnablement penser que la variabilité de ce gène chez les patientes AIC serait le reflet d'une variabilité naturelle.

En résumé, les facteurs âge, différence d'heure du prélèvement, et médicaments représentent les facteurs de variabilité dominant du transcriptome. Les impacts de ces facteurs sont visibles lors des analyses par famille (couples fille versus mère). L'effet de la maladie sur le transcriptome se trouve dilué par les profils d'expressions non liés au syndrome d'Aicardi. Ainsi, pour une analyse par famille (fille vs mère) le facteur maladie semble moins bien représenté que les facteurs âge, heure du prélèvement. L'analyse statistique par ANOVA tente quant à elle de trouver des variations communes à l'ensemble des filles AIC et non des fluctuations de sens et d'amplitude variables. Les facteurs maladie

et médicaments sont probablement ceux qui différencient le plus les filles AIC de leur mère. Le caractère dilué de la réponse du transcriptome au phénotype AIC pourrait notamment expliquer la raison pour laquelle l'analyse fonctionnelle par EASE des résultats d'ANOVA donne des scores très faibles, reflétant ainsi la faible « réponse commune » (le petit nombre de gène dérégulés) chez l'ensemble des filles AIC.

2.4.5. Que représentent les variabilités transcriptomiques restantes ?

Les études de regroupement fonctionnel révèlent trois catégories surreprésentées. Comme nous l'avons démontré précédemment, les catégories « synthèse protéique » et « réponse immunitaire » sont à relier avec les facteurs de variabilité du transcriptome (âge et heure de prélèvement) indépendant de la maladie. Notons que des effets croisés de plusieurs facteurs sont néanmoins possibles. Ainsi, la dégradation protéique est une catégorie fonctionnelle présentée comme pouvant être impliquée dans les RMLX [257]. La dernière catégorie représentée par le terme nucléosome et assemblage du nucléosome (BP) et ses termes parents serait le reflet des traitements par les médicaments anti-épileptiques. Notons cependant que le facteur épilepsie (et donc son traitement) peut révéler des voies impliquées dans la maladie puisque ce signe clinique fait partie de la triade définissant cette pathologie. Cette catégorie reste donc potentiellement à explorer. Des maladies impliquant des gènes annotés par ce terme sont connues. Le gène CHD7 (chromodomain helicase DNA binding protein 7) annoté par des termes tels que « assemblage/désassemblage de la chromatine », « modification de la chromatine » est impliqué par exemple dans le syndrome de CHARGE. Il existe d'autres pathologies génétiques dans l'espèce humaine causées par des mutations dans des gènes impiqués dans la modification dynamique de la structure de la chromatine (tableau 31) [258].

Disease	OMIM	Gene	Type of protein	Phenotypes
ATR-X syndrome	301040	ATRX	SNF2-type helicase?	a-Thalassemia, mental retardation, facial and skeletal abnormalities, urogenital abnormalities, microcephaly
Juberg-Marsidi syndrome	309590			
Sutherland-Hann syndrome	309470			
Smith-Fineman-Myers syndrome	309580			
ICF syndrome	242860	DNMT3B	*De novo* DNA methyltransferase	Immunodeficiency, instability of pericentric heterochomatin in lymphocytes, facial anomalies, mild mental retardation, developemental delay
RTT (only in females and males with Klinefelter's syndrome)	312750	MeCP2	Methyl-CpG binding protein	Progressive encephalopathy, loss of purposeful hand use, stereotyped hand movements, severe breathing dysfunction (apnea), apraxia, mental retardation
Male RTT mutation (MRM) syndrome (RTT-causing mutation in hemizygous males)	N/A	MeCP2	Methyl-CpG binding protein	Severe neonatal-onset microcephaly, hypotonia, severe breathing dysfunction (apnea)
Rubilstein-Taybi syndrome	180849	CBP	CREB binding protein, histone (?) acetyltransferase	Cardiac anomalies, broad thumbs and big toes, characteristic facies (beaked nose), mental retardation, microcephaly
Coffin-Lowry syndrome (CLS)	303600	RSK2	Histone H3 (?) serine/ threonine kinase	Facial dysmorphism, progressive skeletal deformation and abnormal digits, mental retardation, hypotonia

Tableau 31 : Chromatin diseases, genes mutated, corresponding proteins and phenotypes

La plupart des gènes impliqués dans les pathologies de la chromatine ont un lien avec la modification des nucléosomes. Parmi les résultats de nos études, *H2AFB3* (H2A histone family, member B3) carographié en Xq28 est présenté comme sous régulé dans les expériences AS1 mais non retrouvé dans les expériences AS2 et les analyses ANOVA. Nous ne retiendrons donc pas ce gène comme candidat prioritaire. D'autre part, le gène *SUV39H3* (suppressor of variegation 3-9 homolog 1) n'est pas détecté par les études d'expression (valeur p non significatifs). Cependant, ce gène localisé sur l'X a été ciblé de façon indirecte lors des analyses ANOVA. En effet, les 475 gènes signatures issue des analyses ANOVA ont été groupés en classes fonctionnelles par le logiciel EASE. La catégorie « chromosome, région péricentrique » (score EASE=8,26E-004) est la plus représentée. Elle contient quatre gènes non cartographiés sur le chromosome X : *CBX5* (chromobox homolog 5 (HP1 alpha homolog, Drosophila)), ZNF92 (zinc finger protein 92), *CDCA1* (NUF2, NDC80 kinetochore complex component, homolog (S. cerevisiae) et *CENPA* (centromere protein A). *SUV39H3* un intéractant de *CBX5* (chromobox homolog 5) est cartographié en Xp11.23. Ce gène

intervient dans la méthylation de l'histone H3. Il n'a été impliqué dans aucune maladie mais est candidat au RMLX [11]. D'autre part, *SUV39H3* est également un gène candidat d'après les résultats d'ACGR car il est l'interactant cartographié sur le chromosome X de deux gènes *CBX5* et *PCM1* présentés comme dérégulées par ANOVA. *SUV39H3* est donc considéré comme candidat pour le syndrome d'Aicardi. Le séquençage de ce gène chez les patientes AIC fait partie des perspectives à court terme.

3. Etudes *in silico* pour la recherche de gènes candidats

Les solutions actuelles de la bio informatique pour la recherche de gènes candidats par les approches bioinformatiques se fondent soit sur les propriétés des gènes de maladies (approches généralistes), soit sur la recherche de maladies proches d'un point de vue phénotypique (approches comparatives). Ces solutions sont insuffisantes dans le cas où il n'est pas possible de formuler un *a priori* sur le gène candidat. Des définitions plus complexes (logiciels multicritères) sont nécessaires. Pourtant dans le cas du syndrome d'Aicardi, les logiciels actuels donnent des résultats décevants. Le prototype ACGR est un système multicritère qui permet d'intégrer des données privées et publiques et permet d'effectuer des requêtes complexes dans un système non figé.

3.1. Apport et limites du prototype ACGR

Une base de données spécifique est un réel avantage voir une nécessité pour intégrer correctement, et surtout efficacement les données biologiques. Le stockage, l'intégration, et l'exploration des données issue de puces transcriptomiques sont nécessairement facilités par une BD dédiée au projet d'étude. La BD facilite la manipulation des données par rapport à une solution sur tableur Excel ou une solution commerciale telle que Luminator. L'exploitation des données sur Excel devient rapidement très difficile à gérer notamment

pour des analyses nécessitant des requêtes, même simples. Par exemple, la simple question : parmi toutes les patientes AIC lesquelles présentent une dérégulation du gène SUV39H1 ? est un réel problème. Il nécessite de rechercher le gène dans chacun des résultats des 9 analyses fille *vs* mère (9 tableaux Excel). Lorsque la question devient : quels sont les gènes significativement dérégulés chez un maximum de couples fille vs mère le problème devient encore plus complexe et l'utilisation de macro Excel devient alors inévitable. Ainsi, les 9 listes de gènes signatures sont listées par colonne et une macro est chargée de rechercher les identifiants par nombre d'apparition dans les 9 colonnes. La création de macro Excel n'est malheureusement pas à la portée de tous les biologistes. La solution commerciale quand à elle n'est pas flexible. Dans le cas de Luminator v3.0, les informations sur les variations de ratio et la valeur p ne deviennent plus disponibles lorsque l'on regroupe des gènes en commun à plusieurs expériences (« bioset »). Le logiciel ACGR permet de solutionner ces problèmes. Il présente également l'avantage d'intégrer les données publiques. Le système est non figé puisqu'il offre la possibilité d'intégrer des modules divers tels des méthodes de priorisation différentes de celle utilisée. Les BD publiques consultées peuvent varier en fonction des besoins puisque le prototype Xcollect permet de définir rapidement des scénarios permettant d'analyser les pages HTML des sites d'intérêts pour recueillir des informations pertinentes.

Une des limites du système est sa dépendance aux sites consultés. Puisque le logiciel se connecte directement à la source, un problème dans le site se répercutera forcément sur le logiciel. D'autre part, les changements des interfaces nécessitent également la mise à jour des scénarios Xcollect. Le système est également dépendant du choix des bases de données. Nous avons choisi les BD du NCBI pour extraire un maximum d'information à partir d'un minimum de sites consultés. Les BD MGD Flybase sont des références pour les données génétiques chez la souris et la drosophile respectivement. D'autre part, ces deux BD sont maintenues par des membres fondateurs du consortium GO, ce qui garantit la qualité de leurs annotations GO. La traçabilité des informations dans le système ACGR permet de vérifier la qualité des données. Les erreurs des BD et les données manquantes constituent une des limites de notre système aussi bien que des autres logiciels actuellement disponibles.

3.2. Annotations GO

Un point important dans la stratégie que nous avons élaborée est l'intervention de l'expert. Il constitue aussi bien une force qu'une limite de notre système. La sélection du set AICARDI DS-GO est un élément déterminant dans l'alimentation de ACGRdb par les gènes issus des trois espèces (Homo sapiens, Mus musculus, Drosophila melanogaster). L'expert, sera plus compétent qu'un programme dans la sélection de ces termes GO « reflétant le phénotype » des patientes AIC. Une des difficultés est que inversement, la sélection des termes GO pourra varier d'une personne à l'autre puisque cette sélection est subjective et est surtout très difficile. Bien qu'une vingtaine d'organisations travaille sur l'annotation des gènes par GO, il n'en demeure pas moins que certains gènes et notamment chez l'homme, restent pauvrement annotés et donc sont difficilement voir non identifiables par les termes sélectionnés par l'expert. A ceci s'ajoute un autre problème : les termes GO ne sont pas adaptés pour annoter les phénotypes décrivant les maladies. Le thésaurus MeSH est au contraire bien adapté pour décrire les phénotypes de maladies mais ne permet pas d'indexer les gènes. Des solutions comme G2D [196, 259] et eVOC [208] sont disponibles. Le premier utilise MeSH pour la description phénotypique et GO pour l'annotation des gènes. Un algorithme de similarité est utilisé pour passer des termes MeSH vers les termes GO. Cette solution présente l'avantage majeur de ne pas présenter une étape manuelle cependant elle présente nécessairement l'inconvénient d'une perte d'information lors du passage aux termes GO. Evoc est quand à lui une solution qui utilise les tissus affectés par la maladie. L'hypothèse est que le phénotype est représenté dans les tissus affecté et que le gène impliqué est dérégulé dans ces tissus.

Une autre limite du système concerne les séquences dérégulées pour lesquelles il n'y a pas de symbole de gène. Ainsi, parmi les 475 gènes décrits comme dérégulés par la méthode ANOVA, 175 n'ont pas pu être exploités. Une majorité de ces séquences sont des EST pour lesquels peu d'informations sont disponibles sur les bases de données. Cependant, des données sur leurs tissus d'expression peuvent être extraitesdes BD UniGene ou Ensembl. L'exploitation de ces données pourrait être possible en remplaçant, pour ces EST, les annotations GO non disponible par des annotations eVOC que l'on pourait mettre en correspondance par une mesure de similarité avec les sites d'expression du phénotype malade à l'instar de la méthode proposée par Tiffin et al. [208].

3.3. Relations entre gène candidat et maladies

Les définitions que l'utilisateur donne à un gène candidat peuvent être très variées. Ces définitions peuvent être fondées sur la fonction du gène, son lieu d'expression ou sa place dans un réseau biologique. Il est également possible de combiner des hypothèses multiples.

La définition simpliste disant qu'un gène candidat pour le syndrome d'Aicardi peut être un gène dont l'expression est dérégulée chez les patientes AIC par rapport à des patientes témoins et cartographié sur le chromosome X, n'a donné dans notre étude aucun gène candidat satisfaisant. Il est possible d'imaginer que les expériences de puces transcriptomiques n'ont pas permis de détecter la dérégulation de ce gène. Cette explication est plausible si le gène est tissus-spécifique ou si les puces transcriptomiques ne sont pas assez résolutives pour la détection d'une telle variation. Il est également possible que la méthode ANOVA n'ait pas permis de sélectionner le gène candidat.

On peut alors se tourner vers une définition alternative du gène candidats. Il peut également être un gène dont les annotations GO reflètent les voies biologiques dérégulées chez les patientes AIC. Une mutation dans ce gène affecterait les voies biologiques (décrites par les termes GO) et conduirait au phénotype observé. L'introduction d'une mesure de similarité entre les annotations GO des gènes et un jeu de termes GO de référence, spécifique de la maldie, correspond à cette définition. Cette approche a été utilisée pour la priorisation et classification des gènes. Mais ici se pose le problème des gènes pauvrement annotés, voire dépourvus d'annotation. Par ailleurs, il est possible que le gène candidat ne corresponde pas directement au phénotype de la maladie et donc ne possède pas un score élevé. C'est le cas pour les maladies dont le gène responsable possède une fonction générale telle que par exemple un facteur de transcription.

Et ce sont précisément sur ces points que notre système est performant. Des relations indirectes sont ainsi introduites dans les définitions de gène candidat. Ces relations indirectes entre le gène et la maladie peuvent se traduire soit par des notions d'interactions entre le gène candidat et un autre gène, soit par des relations d'orthologie entre le gène candidat et un gène orthoogue bien étudié dans une autre espèce, sachant que le gène interactant ou orthologue possède les propriétés attendues, définies par l'utilisateur. Des définitions récursives peuvent ainsi être formulées.

Typiquement le produit du gène *MAGED1* interagit avec le gène *DLX* qui est par ailleurs dérégulé dans les expériences transcriptomiques. Le gène *DLX* possède un score élevé (50) par rapport au jeu de termes AICARDI-DS-GO sélectionnés. D'autres hypothèses nous ont apporté d'autres listes de candidats également intéressantes, dont la majeure partie est présentée dans l'article. Cependant, l'exploitation du système n'est pas terminée et sa flexibilité permet d'envisager des évolutions. Une des perspectives sera l'exploitation d'ACGRdb, après l'intégration des résultats de puces génomiques 244 k. cette étude expérimentale sera effectuée sur l'ADN génomique de 24 patientes AIC. La détection d'une région contenant un CNV délétère pourra rapidement nous orienter vers la priorisation des gènes de la région candidate.

Conclusion

Conclusion

Jusqu'à présent toutes les études visant à identifier le gène impliqué dans le syndrome d'Aicardi se sont révélées infructueuses. Les approches classiques de clonage positionnel par cartographie génétique ne sont pas applicables, de plus aucune des anomalies chromosomiques détectées chez des patientes AIC n'a abouti à la délimitation d'une région candidate.La seule approche fonctionnelle suite à quoi, à mentionner n'a pas révélé de mutations dans le gène candidat potentiel de la filamine A. Bien que pour la majorité des pathologies génétiques, la proportion de microdélétions et microduplications au sein de l'ensemble des allèles délétères reste faible par rapport aux variations d'un ou quelques nucléotides, la recherche de microremaniements demeure une méthode efficace de clonage positionnel, puisqu'elle peut réduire considérablement la région candidate. Nous avons donc tout d'abord effectué l'analyse du caryotype constitutionnel chez 18 patientes AIC. Aucune anomalie du caryotype n'a été détectée au seuil de résolution de la technique (5 Mb). Une recherche de variants quantitatifs de l'ADN avec l'aide de microréseaux d'ADN spécifiques du chromosome X (résolution théorique de 82 kb) a également été effectuée. Bien qu'aucune variation quantitative délétère n'ait été associée au syndrome, cette étude a permis d'écarter l'hypothèse d'une délétion (ou d'une duplication) complète du gène de la filamine A dans la pathologie du syndrome d'Aicardi, hypothèse jusqu'alors non testée. Ces analyses de puces génomiques présentent cependant plusieurs limites. Tout d'abord, ce sont des puces dont les sondes sont constituées par des BAC avec une résolution ne permettant pas de détecter des aneusomies segmentaires de taille inférieure à 82 Kb. D'autre part, bien que cette hypothèse soit peu probable, on ne peut pas complètement exclure que le gène responsable, quand il est muté, du syndrome d'Aicardi soit localisé sur un autosome. Nous avons aujourd'hui à notre disposition une plate-forme permettant l'utilisation de puces à oligonucléotides de haute densité (244 K) couvrant la totalité du génome (résolution de 6,4 Kb). Un total de 24 filles AIC, comprenant les 18 filles étudiées au cours de mon travail de thèse et 6 nouvelles patientes, sera ainsi étudié à l'aide de puces haute densité. En effet, dans une approche visant à l'identification de microremaniements, plus le nombre de patientes analysées est important plus la probabilité de trouver un cas présentant une anomalie de structure est grande. Ceci montre bien l'importance du recrutement d'un

maximum de patientes AIC pour nos recherches. L'association AAL-Syndrome d'Aicardi, le réseau de neuropédiatres, généticiens et le serveur d'information sur les maladies rares Orphanet aussi bien que les présentations des travaux du laboratoire dans les congrès nationaux et internationaux sont autant d'intermédiaires et d'actions importants pour le recrutement des patientes Aicardi. Nous comptons sur ce réseau actif pour recruter un nombre toujours croissant de familles AIC et augmenter ainsi nos chances de trouver le gène candidat pour cette pathologie.

Il nous a alors été nécessaire d'enrichir les stratégies de recherche de gènes candidats par d'autres méthodologies. Une approche fonctionnelle basée sur l'étude du transcriptome a donc été envisagée dans le but de sélectionner des gènes dont l'expression diffère entre les filles Aicardi et des témoins. Outre la sélection du gène candidat impliqué dans le syndrome d'Aicardi, cette approche est surtout vouée à l'identification des fonctions biologiques dérégulées chez les patientes Aicardi. Une étude pilote à l'aide de puces à oligonucléotides de 20 k a été réalisée sur des lignées lymphoblastiques issues de trois patientes AIC. Nous avons par la suite complété notre étude transcriptomique par l'étude des ARN issus de sang total de 10 patientes AIC, grâce à une puce à oligonucléotides de 44 k couvrant théoriquement l'ensemble des transcrits. Les gènes sélectionnés par cette approche ont été séquençés. Aucune mutation délétère n'a été identifée et ces gènes ont été *a priori* exclus. Néanmois, ils pourraient tout à fait intervenir dans les voies biologiques impliquées dans la physiopathologie du syndrome d'Aicardi puisque leur dérégulation a été confirmée par RT Q-PCR. Par ailleurs les regroupements fonctionnelles des gènes signatures lors des comparaisons « fille vs mère » montrent que l'effet de la maladie sur le transcriptome semble dilué par des profils d'expression prépondérantes. Ainsi, les facteurs âge, heure du prélèvement, et médicaments représentent les facteurs de variabilité dominant du transcriptome par rapport à l'effet de la maladie dans notre étude.

Il est vrai que ces approches de puces à ADN (génomique ou transcriptomique) sont des technologies complexes dans leur utilisation. Le projet a débuté alors que nous ne disposions pas encore, au laboratoire de génétique à Nancy, d'une plate-forme pour l'analyse de puces génomique. Par ailleurs, rapidement il s'est avéré nécessaire d'inclure au projet de thèse, une approche bioinformatique pour l'analyse du grand nombre de données générées par des puces transcriptomiques. La bioinformatique est une discipline en plein développement et son utilisation nécessite un savoir faire bien particulier.

L'acquisition des compétences pour ces trois approches novatrices, complexes et en plein essor, a été rendue possible grâce à des collaborations étroites et précieuse avec des «équipes compétantes. Concernant l'utilisation des puces génomiques (D'une part, le laboratoire du Flanders Interuniversity Institute of Human Genetics avec l'aide précieuse de Marike Bauters pour les expériences de puces génomiques et d'autre part la mise en place des protocoles, des expériences et l'analyses des puces transcriptomiques grâce à l'équipe de génomique fonctionnelle de John MacGregor à Londres. C'est aussi à Londres et notamment avec Marion Leleu que j'ai débuté l'approche bioinformatique du projet. Le projet bioinformatique pour d'analyse des puces transcriptomiques est par la suite devenue une approche *in silico* de recherche de gène candidats ou « *gene discovery by integrative genomics* ». Ce projet ACGR a été mené sous la direction de Marie Dominique Devignes et Malika Smail de l'équipe Orpailleur (LORIA, Nancy). Ainsi, Le logiciel ACGR a permis d'intégrer les données expérimentales d'analyse du transcriptome ainsi que des informations issues des bases de données publiques pour la recherche de gènes candidats pour le syndrome d'Aicardi. Une base de données dédiée à la maladie (ACGRdb) et stockant les informations pertinentes a ainsi été créée. Des requêtes, correspondant à diverses définitions du gène candidat, permettent d'interroger la base de données et de sortir une liste ordonnée de gènes candidats. Des résultats intéressants ont déjà été obtenus et le séquençage de trois gènes identifiés par cette méthode de génomique intégrative (*PLXNA3*, *MADGED1, SUV39H1*) fait partie des perspectives à court terme.

Le logiciel ACGR est un outil qui s'inscrit dans les nouvelles méthodes bioinformatiques de recherche de gènes candidats pour des maladies génétiques. Pour le moment, seules les données transcriptomiques ont été intégrées dans ACGRdb, cependant la flexibilité du système prévoit la possibilité d'enrichir la base de données par d'autres résultats expérimentaux. Par exemple, dans le cas d'une détection d'un variant quantitatif délétère à l'aide d'une puce génomique, l'intégration de ces données dans ACGRdb permettra, d'une part d'identifier les candidats potentiels de la région en déséquilibre et d'autre part de prioriser ces gènes candidats permettant ainsi de cibler le gène candidat le plus pertinent.

Ces approches intégratives reflètent l'évolution de nos concepts de recherche passant de la génétique du retard mental à la génomique du retard mental en tenant compte de la multiplicité des réseaux d'interactions et de régulations qui doivent être exploités pour

mieux cibler et comprendre les mécanismes physiopathologiques à l'origine du retard mental.

Bibliographie

Bibliographie

1. **Manuel de Diagnostic et Statistiques des Troubles Mentaux - Quatrième Edition (DSM-IV)**. Washington DC: American Psychiatric Association, APA; 1994.
2. Wechsler: **Intelligence Scale for Children-Revised (WISC-R)**; 1974.
3. Stevenson RE: **Splitting and lumping in the nosology of XLMR**. *American journal of medical genetics* 2000, **97**(3):174-182.
4. Glass IA: **X linked mental retardation**. *Journal of medical genetics* 1991, **28**(6):361-371.
5. Leonard H, Wen X: **The epidemiology of mental retardation: challenges and opportunities in the new millennium**. *Mental retardation and developmental disabilities research reviews* 2002, **8**(3):117-134.
6. McLaren J, Bryson SE: **Review of recent epidemiological studies of mental retardation: prevalence, associated disorders, and etiology**. *Am J Ment Retard* 1987, **92**(3):243-254.
7. Hagberg B, Kyllerman M: **Epidemiology of mental retardation--a Swedish survey**. *Brain & development* 1983, **5**(5):441-449.
8. Flint J, Knight S: **The use of telomere probes to investigate submicroscopic rearrangements associated with mental retardation**. *Current opinion in genetics & development* 2003, **13**(3):310-316.
9. Chelly J, Khelfaoui M, Francis F, Cherif B, Bienvenu T: **Genetics and pathophysiology of mental retardation**. *Eur J Hum Genet* 2006, **14**(6):701-713.
10. Ropers HH, Hamel BC: **X-linked mental retardation**. *Nature reviews* 2005, **6**(1):46-57.
11. de Brouwer AP, Yntema HG, Kleefstra T, Lugtenberg D, Oudakker AR, de Vries BB, van Bokhoven H, Van Esch H, Frints SG, Froyen G *et al*: **Mutation frequencies of X-linked mental retardation genes in families from the EuroMRX consortium**. *Human mutation* 2007, **28**(2):207-208.
12. Lehrke R: **Theory of X-linkage of major intellectual traits**. *American journal of mental deficiency* 1972, **76**(6):611-619.
13. Fishburn J, Turner G, Daniel A, Brookwell R: **The diagnosis and frequency of X-linked conditions in a cohort of moderately retarded males with affected brothers**. *American journal of medical genetics* 1983, **14**(4):713-724.
14. Herbst DS, Miller JR: **Nonspecific X-linked mental retardation II: the frequency in British Columbia**. *American journal of medical genetics* 1980, **7**(4):461-469.
15. Raymond FL: **X linked mental retardation: a clinical guide**. *Journal of medical genetics* 2006, **43**(3):193-200.
16. Mandel JL, Chelly J: **Monogenic X-linked mental retardation: is it as frequent as currently estimated? The paradox of the ARX (Aristaless X) mutations**. *Eur J Hum Genet* 2004, **12**(9):689-693.
17. Neri G, Gurrieri F, Gal A, Lubs HA: **XLMR genes: update 1990**. *American journal of medical genetics* 1991, **38**(2-3):186-189.
18. Chiurazzi P, Tabolacci E, Neri G: **X-linked mental retardation (XLMR): from clinical conditions to cloned genes**. *Critical reviews in clinical laboratory sciences* 2004, **41**(2):117-158.

19. Yu YW, Tsai SJ, Hong CJ, Chen TJ, Yang CW: **Association analysis for MAOA gene polymorphism with long-latency auditory evoked potentials in healthy females**. *Neuropsychobiology* 2004, **50**(4):288-291.
20. Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, Taylor A, Poulton R: **Role of genotype in the cycle of violence in maltreated children**. *Science (New York, NY* 2002, **297**(5582):851-854.
21. Ropers HH: **X-linked mental retardation: many genes for a complex disorder**. *Current opinion in genetics & development* 2006, **16**(3):260-269.
22. Laumonnier F, Cuthbert PC, Grant SG: **The role of neuronal complexes in human X-linked brain diseases**. *American journal of human genetics* 2007, **80**(2):205-220.
23. McBride SM, Choi CH, Wang Y, Liebelt D, Braunstein E, Ferreiro D, Sehgal A, Siwicki KK, Dockendorff TC, Nguyen HT *et al*: **Pharmacological rescue of synaptic plasticity, courtship behavior, and mushroom body defects in a Drosophila model of fragile X syndrome**. *Neuron* 2005, **45**(5):753-764.
24. Aicardi J: *Electroencephalogr Clin Neurophysiol* 1965, **19**(Suppl):609-610.
25. Aicardi J, Chevrie JJ, Rousselie F: **[Spasma-in-flexion syndrome, callosal agenesis, chorioretinal abnormalities]**. *Archives francaises de pediatrie* 1969, **26**(10):1103-1120.
26. Donnenfeld AE, Packer RJ, Zackai EH, Chee CM, Sellinger B, Emanuel BS: **Clinical, cytogenetic, and pedigree findings in 18 cases of Aicardi syndrome**. *American journal of medical genetics* 1989, **32**(4):461-467.
27. Menezes AV, MacGregor DL, Buncic JR: **Aicardi syndrome: natural history and possible predictors of severity**. *Pediatric neurology* 1994, **11**(4):313-318.
28. Sutton VR, Hopkins BJ, Eble TN, Gambhir N, Lewis RA, Van den Veyver IB: **Facial and physical features of Aicardi syndrome: infants to teenagers**. *Am J Med Genet A* 2005, **138**(3):254-258.
29. Rosser TL, Acosta MT, Packer RJ: **Aicardi syndrome: spectrum of disease and long-term prognosis in 77 females**. *Pediatric neurology* 2002, **27**(5):343-346.
30. Aicardi J: **Aicardi syndrome**. *Brain & development* 2005, **27**(3):164-171.
31. Menezes AV, Enzenauer RW, Buncic JR: **Aicardi syndrome--the elusive mild case**. *The British journal of ophthalmology* 1994, **78**(6):494-496.
32. Chau V, Karvelas G, Jacob P, Carmant L: **Early treatment of Aicardi syndrome with vigabatrin can improve outcome**. *Neurology* 2004, **63**(9):1756-1757.
33. Grosso S, Lasorella G, Russo A, Galluzzi P, Morgese G, Balestri P: **Aicardi syndrome with favorable outcome: Case report and review**. *Brain & development* 2007, **29**(7):443-446.
34. Aicardi J: **Aicardi syndrome: Old and new findings**. *Int Pediatr* 1999, **14**:5-8.
35. Yacoub M, Missaoui N, Tabarli B, Ghorbel M, Tlili K, Selmi H, Essoussi A: **[Aicardi syndrome with favorable outcome]**. *Arch Pediatr* 2003, **10**(6):530-532.
36. Carney SH, Brodsky MC, Good WV, Glasier CM, Greibel ML, Cunniff C: **Aicardi syndrome: more than meets the eye**. *Survey of ophthalmology* 1993, **37**(6):419-424.
37. McPherson E, Jones SM: **Cleft lip and palate in Aicardi syndrome**. *American journal of medical genetics* 1990, **37**(3):318-319.
38. Umansky WS, Neidich JA, Schendel SA: **The association of cleft lip and palate with Aicardi syndrome**. *Plastic and reconstructive surgery* 1994, **93**(3):595-597.
39. Trifiletti RR, Incorpora G, Polizzi A, Cocuzza MD, Bolan EA, Parano E: **Aicardi syndrome with multiple tumors: a case report with literature review**. *Brain & development* 1995, **17**(4):283-285.
40. King AM, Bowen DI, Goulding P, Doran RM: **Aicardi syndrome**. *The British journal of ophthalmology* 1998, **82**(4):457.

41. Matlary A, Prescott T, Tvedt B, Lindberg K, Server A, Aicardi J, Stromme P: **Aicardi syndrome in a girl with mild developmental delay, absence of epilepsy and normal EEG**. *Clinical dysmorphology* 2004, **13**(4):257-260.
42. Lee SW, Kim KS, Cho SM, Lee SJ: **An atypical case of Aicardi syndrome with favorable outcome**. *Korean J Ophthalmol* 2004, **18**(1):79-83.
43. Prats Vinas JM, Martinez Gonzalez MJ, Garcia Ribes A, Martinez Gonzalez S, Martinez Fernandez R: **Callosal agenesis, chorioretinal lacunae, absence of infantile spasms, and normal development: Aicardi syndrome without epilepsy?** *Developmental medicine and child neurology* 2005, **47**(6):419-420; discussion 364.
44. Menezes AV, Lewis TL, Buncic JR: **Role of ocular involvement in the prediction of visual development and clinical prognosis in Aicardi syndrome**. *The British journal of ophthalmology* 1996, **80**(9):805-811.
45. Molina JA, Mateos F, Merino M, Epifanio JL, Gorrono M: **Aicardi syndrome in two sisters**. *The Journal of pediatrics* 1989, **115**(2):282-283.
46. Taggard DA, Menezes AH: **Three choroid plexus papillomas in a patient with Aicardi syndrome. A case report**. *Pediatric neurosurgery* 2000, **33**(4):219-223.
47. Hopkins IJ, Humphrey I, Keith CG, Susman M, Webb GC, Turner EK: **The Aicardi syndrome in a 47, XXY male**. *Australian paediatric journal* 1979, **15**(4):278-280.
48. Hoag HM, Taylor SA, Duncan AM, Khalifa MM: **Evidence that skewed X inactivation is not needed for the phenotypic expression of Aicardi syndrome**. *Human genetics* 1997, **100**(3-4):459-464.
49. Lyon MF: **Sex chromatin and gene action in the mammalian X-chromosome**. *American journal of human genetics* 1962, **14**:135-148.
50. Barr M HJ: **A quantitative study of certain morphological changes in spinal motor neurons during axon reaction**. *J Comp Neurol* 1948, **89**:93–121.
51. Brown CJ, Robinson WP: **The causes and consequences of random and non-random X chromosome inactivation in humans**. *Clinical genetics* 2000, **58**(5):353-363.
52. Brown CJ, Lafreniere RG, Powers VE, Sebastio G, Ballabio A, Pettigrew AL, Ledbetter DH, Levy E, Craig IW, Willard HF: **Localization of the X inactivation centre on the human X chromosome in Xq13**. *Nature* 1991, **349**(6304):82-84.
53. Brown SD: **XIST and the mapping of the X chromosome inactivation centre**. *Bioessays* 1991, **13**(11):607-612.
54. Penny GD, Kay GF, Sheardown SA, Rastan S, Brockdorff N: **Requirement for Xist in X chromosome inactivation**. *Nature* 1996, **379**(6561):131-137.
55. Panning B, Dausman J, Jaenisch R: **X chromosome inactivation is mediated by Xist RNA stabilization**. *Cell* 1997, **90**(5):907-916.
56. Xu N, Tsai CL, Lee JT: **Transient homologous chromosome pairing marks the onset of X inactivation**. *Science (New York, NY* 2006, **311**(5764):1149-1152.
57. Bacher CP, Guggiari M, Brors B, Augui S, Clerc P, Avner P, Eils R, Heard E: **Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation**. *Nature cell biology* 2006, **8**(3):293-299.
58. Carrel L, Willard HF: **Counting on Xist**. *Nat Genet* 1998, **19**(3):211-212.
59. Wettke-Schafer R, Kantner G: **X-linked dominant inherited diseases with lethality in hemizygous males**. *Human genetics* 1983, **64**(1):1-23.
60. Migeon BR, Axelman J, Jan de Beur S, Valle D, Mitchell GA, Rosenbaum KN: **Selection against lethal alleles in females heterozygous for incontinentia pigmenti**. *American journal of human genetics* 1989, **44**(1):100-106.
61. Wieacker P, Zimmer J, Ropers HH: **X inactivation patterns in two syndromes with probable X-linked dominant, male lethal inheritance**. *Clinical genetics* 1985, **28**(3):238-242.

62. Neidich JA, Nussbaum RL, Packer RJ, Emanuel BS, Puck JM: **Heterogeneity of clinical severity and molecular lesions in Aicardi syndrome**. *The Journal of pediatrics* 1990, **116**(6):911-917.
63. Collins FS: **Positional cloning: let's not call it reverse anymore**. *Nat Genet* 1992, **1**(1):3-6.
64. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millasseau P, Vaysseix G, Lathrop M: **A second-generation linkage map of the human genome**. *Nature* 1992, **359**(6398):794-801.
65. Ropers HH, Zuffardi O, Bianchi E, Tiepolo L: **Agenesis of corpus callosum, ocular, and skeletal anomalies (X-linked dominant Aicardi's syndrome) in a girl with balanced X/3 translocation**. *Human genetics* 1982, **61**(4):364-368.
66. Neidich JA, Nussbaum RL, Packer R, Graham JM, Jr., Donnenfeld AE, Emanuel BS, Puck JM: **Heterogeneity in clinical severity and molecular lesions in Aicardi syndrome**. *American journal of human genetics* 1988, **43**(A91).
67. Donnenfeld AE, Graham JM, Jr., Packer RJ, Aquino R, Berg SZ, Emanuel BS: **Microphthalmia and chorioretinal lesions in a girl with an Xp22.2-pter deletion and partial 3p trisomy: clinical observations relevant to Aicardi syndrome gene localization**. *American journal of medical genetics* 1990, **37**(2):182-186.
68. Naritomi K, Izumikawa Y, Nagataki S, Fukushima Y, Wakui K, Niikawa N, Hirayama K: **Combined Goltz and Aicardi syndromes in a terminal Xp deletion: are they a contiguous gene syndrome?** *American journal of medical genetics* 1992, **43**(5):839-843.
69. Van den Veyver IB: **Microphthalmia with linear skin defects (MLS), Aicardi, and Goltz syndromes: are they related X-linked dominant male-lethal disorders?** *Cytogenetic and genome research* 2002, **99**(1-4):289-296.
70. al-Gazali LI, Mueller RF, Caine A, Antoniou A, McCartney A, Fitchett M, Dennis NR: **Two 46,XX,t(X;Y) females with linear skin defects and congenital microphthalmia: a new syndrome at Xp22.3**. *Journal of medical genetics* 1990, **27**(1):59-63.
71. Temple IK, Hurst JA, Hing S, Butler L, Baraitser M: **De novo deletion of Xp22.2-pter in a female with linear skin lesions of the face and neck, microphthalmia, and anterior chamber eye anomalies**. *Journal of medical genetics* 1990, **27**(1):56-58.
72. Happle R, Daniels O, Koopman RJ: **MIDAS syndrome (microphthalmia, dermal aplasia, and sclerocornea): an X-linked phenotype distinct from Goltz syndrome**. *American journal of medical genetics* 1993, **47**(5):710-713.
73. Morleo M, Pramparo T, Perone L, Gregato G, Le Caignec C, Mueller RF, Ogata T, Raas-Rothschild A, de Blois MC, Wilson LC *et al*: **Microphthalmia with linear skin defects (MLS) syndrome: clinical, cytogenetic, and molecular characterization of 11 cases**. *Am J Med Genet A* 2005, **137**(2):190-198.
74. Prakash SK, Cormier TA, McCall AE, Garcia JJ, Sierra R, Haupt B, Zoghbi HY, Van Den Veyver IB: **Loss of holocytochrome c-type synthetase causes the male lethality of X-linked dominant microphthalmia with linear skin defects (MLS) syndrome**. *Human molecular genetics* 2002, **11**(25):3237-3248.
75. Nielsen KB, Anvret M, Flodmark O, Furuskog P, Bohman-Valis K: **Aicardi syndrome: early neuroradiological manifestations and results of DNA studies in one patient**. *American journal of medical genetics* 1991, **38**(1):65-68.
76. Van den Veyver IB, Panichkul PP, Antalffy BA, Sun Y, Hunter JV, Armstrong DD: **Presence of filamin in the astrocytic inclusions of Aicardi syndrome**. *Pediatric neurology* 2004, **30**(1):7-15.

77. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system**. *Genome research* 2004, **14**(5):942-950.
78. Wang X, Reid Sutton V, Omar Peraza-Llanes J, Yu Z, Rosetta R, Kou YC, Eble TN, Patel A, Thaller C, Fang P *et al*: **Mutations in X-linked PORCN, a putative regulator of Wnt signaling, cause focal dermal hypoplasia**. *Nat Genet* 2007, **39**(7):836-838.
79. Grzeschik KH, Bornholdt D, Oeffner F, Konig A, Del Carmen Boente M, Enders H, Fritz B, Hertl M, Grasshoff U, Hofling K *et al*: **Deficiency of PORCN, a regulator of Wnt signaling, is associated with focal dermal hypoplasia**. *Nat Genet* 2007, **39**(7):833-835.
80. Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y *et al*: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays**. *Nat Genet* 1998, **20**(2):207-211.
81. Pinkel D, Albertson DG: **Comparative genomic hybridization**. *Annual review of genomics and human genetics* 2005, **6**:331-354.
82. Schoumans J, Anderlid BM, Blennow E, Teh BT, Nordenskjold M: **The performance of CGH array for the detection of cryptic constitutional chromosome imbalances**. *Journal of medical genetics* 2004, **41**(3):198-202.
83. Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P: **Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances**. *Genes, chromosomes & cancer* 1997, **20**(4):399-407.
84. Pinkel D, Albertson DG: **Array comparative genomic hybridization and its applications in cancer**. *Nat Genet* 2005, **37 Suppl**:S11-17.
85. Albertson DG: **Profiling breast cancer by array CGH**. *Breast cancer research and treatment* 2003, **78**(3):289-298.
86. Jehee FS, Rosenberg C, Krepischi-Santos AC, Kok F, Knijnenburg J, Froyen G, Vianna-Morgante AM, Opitz JM, Passos-Bueno MR: **An Xq22.3 duplication detected by comparative genomic hybridization microarray (Array-CGH) defines a new locus (FGS5) for FG syndrome**. *Am J Med Genet A* 2005, **139**(3):221-226.
87. Bruder CE, Hirvela C, Tapia-Paez I, Fransson I, Segraves R, Hamilton G, Zhang XX, Evans DG, Wallace AJ, Baser ME *et al*: **High resolution deletion analysis of constitutional DNA from neurofibromatosis type 2 (NF2) patients using microarray-CGH**. *Human molecular genetics* 2001, **10**(3):271-282.
88. de Stahl TD, Hartmann C, de Bustos C, Piotrowski A, Benetkiewicz M, Mantripragada KK, Tykwinski T, von Deimling A, Dumanski JP: **Chromosome 22 tiling-path array-CGH analysis identifies germ-line- and tumor-specific aberrations in patients with glioblastoma multiforme**. *Genes, chromosomes & cancer* 2005, **44**(2):161-169.
89. Ishkanian AS, Malloff CA, Watson SK, DeLeeuw RJ, Chi B, Coe BP, Snijders A, Albertson DG, Pinkel D, Marra MA *et al*: **A tiling resolution DNA microarray with complete coverage of the human genome**. *Nat Genet* 2004, **36**(3):299-303.
90. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, Brown PO: **Genome-wide analysis of DNA copy-number changes using cDNA microarrays**. *Nat Genet* 1999, **23**(1):41-46.
91. Dhami P, Coffey AJ, Abbs S, Vermeesch JR, Dumanski JP, Woodward KJ, Andrews RM, Langford C, Vetrie D: **Exon array CGH: detection of copy-number changes at the resolution of individual exons in the human genome**. *American journal of human genetics* 2005, **76**(5):750-762.

92. Carvalho B, Ouwerkerk E, Meijer GA, Ylstra B: **High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides**. *Journal of clinical pathology* 2004, **57**(6):644-646.
93. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(20):12963-12968.
94. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR *et al*: **High-resolution analysis of DNA copy number using oligonucleotide microarrays**. *Genome research* 2004, **14**(2):287-295.
95. Okamoto T, Suzuki T, Yamamoto N: **Microarray fabrication with covalent attachment of DNA using bubble jet technology**. *Nature biotechnology* 2000, **18**(4):438-441.
96. Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR *et al*: **Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer**. *Nature biotechnology* 2001, **19**(4):342-347.
97. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE: **Recent segmental duplications in the human genome**. *Science (New York, NY* 2002, **297**(5583):1003-1007.
98. Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X: **Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements**. *Human molecular genetics* 2003, **12**(17):2201-2208.
99. Bailey JA, Baertsch R, Kent WJ, Haussler D, Eichler EE: **Hotspots of mammalian chromosomal evolution**. *Genome biology* 2004, **5**(4):R23.
100. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome**. *Nat Genet* 2004, **36**(9):949-951.
101. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ *et al*: **Common deletion polymorphisms in the human genome**. *Nat Genet* 2006, **38**(1):86-92.
102. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M *et al*: **Large-scale copy number polymorphism in the human genome**. *Science (New York, NY* 2004, **305**(5683):525-528.
103. de Vries BB, Pfundt R, Leisink M, Koolen DA, Vissers LE, Janssen IM, Reijmersdal S, Nillesen WM, Huys EH, Leeuw N *et al*: **Diagnostic genome profiling in mental retardation**. *American journal of human genetics* 2005, **77**(4):606-616.
104. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME *et al*: **Copy number variation: new insights in genome diversity**. *Genome research* 2006, **16**(8):949-961.
105. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D *et al*: **Fine-scale structural variation of the human genome**. *Nat Genet* 2005, **37**(7):727-732.
106. Aldred PM, Hollox EJ, Armour JA: **Copy number polymorphism and expression level variation of the human alpha-defensin genes DEFA1 and DEFA3**. *Human molecular genetics* 2005, **14**(14):2045-2052.
107. Sharp AJ, Cheng Z, Eichler EE: **Structural variation of the human genome**. *Annual review of genomics and human genetics* 2006, **7**:407-442.

108. Kleinjan DJ, van Heyningen V: **Position effect in human genetic disease**. *Human molecular genetics* 1998, **7**(10):1611-1618.
109. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE: **Positive selection of a gene family during the emergence of humans and African apes**. *Nature* 2001, **413**(6855):514-519.
110. Paulding CA, Ruvolo M, Haber DA: **The Tre2 (USP6) oncogene is a hominoid-specific gene**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(5):2507-2511.
111. Beckmann JS, Estivill X, Antonarakis SE: **Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability**. *Nature reviews* 2007, **8**(8):639-646.
112. Emanuel BS, Shaikh TH: **Segmental duplications: an 'expanding' role in genomic instability and disease**. *Nature reviews* 2001, **2**(10):791-800.
113. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R *et al*: **Segmental duplications and copy-number variation in the human genome**. *American journal of human genetics* 2005, **77**(1):78-88.
114. Zhang X, Snijders A, Segraves R, Zhang X, Niebuhr A, Albertson D, Yang H, Gray J, Niebuhr E, Bolund L *et al*: **High-resolution mapping of genotype-phenotype relationships in cri du chat syndrome using array comparative genomic hybridization**. *American journal of human genetics* 2005, **76**(2):312-326.
115. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.
116. Locke DP, Segraves R, Nicholls RD, Schwartz S, Pinkel D, Albertson DG, Eichler EE: **BAC microarray analysis of 15q11-q13 rearrangements and the impact of segmental duplications**. *Journal of medical genetics* 2004, **41**(3):175-182.
117. Shianna KV, Willard HF: **Human genomics: in search of normality**. *Nature* 2006, **444**(7118):428-429.
118. Carter NP: **Methods and strategies for analyzing copy number variation using DNA microarrays**. *NATURE GENETICS* 2007, **39**(s16):S16.
119. Bueno Filho JS, Gilmour SG, Rosa GJ: **Design of microarray experiments for genetical genomics studies**. *Genetics* 2006, **174**(2):945-957.
120. Vissers LE, van Ravenswaaij CM, Admiraal R, Hurst JA, de Vries BB, Janssen IM, van der Vliet WA, Huys EH, de Jong PJ, Hamel BC *et al*: **Mutations in a new member of the chromodomain gene family cause CHARGE syndrome**. *Nat Genet* 2004, **36**(9):955-957.
121. Solomon NM, Ross SA, Morgan T, Belsky JL, Hol FA, Karnes PS, Hopwood NJ, Myers SE, Tan AS, Warne GL *et al*: **Array comparative genomic hybridisation analysis of boys with X linked hypopituitarism identifies a 3.9 Mb duplicated critical region at Xq27 containing SOX3**. *Journal of medical genetics* 2004, **41**(9):669-678.
122. Dave BJ, Sanger WG: **Role of cytogenetics and molecular cytogenetics in the diagnosis of genetic imbalances**. *Seminars in pediatric neurology* 2007, **14**(1):2-6.
123. Shaw-Smith C, Redon R, Rickman L, Rio M, Willatt L, Fiegler H, Firth H, Sanlaville D, Winter R, Colleaux L *et al*: **Microarray based comparative genomic hybridisation (array-CGH) detects submicroscopic chromosomal deletions and duplications in patients with learning disability/mental retardation and dysmorphic features**. *Journal of medical genetics* 2004, **41**(4):241-248.
124. Murthy SK, Nygren AO, El Shakankiry HM, Schouten JP, Al Khayat AI, Ridha A, Al Ali MT: **Detection of a novel familial deletion of four genes between BP1 and BP2**

of the Prader-Willi/Angelman syndrome critical region by oligo-array CGH in a child with neurological disorder and speech impairment. *Cytogenetic and genome research* 2007, **116**(1-2):135-140.

125. Van Esch H, Jansen A, Bauters M, Froyen G, Fryns JP: **Encephalopathy and bilateral cataract in a boy with an interstitial deletion of Xp22 comprising the CDKL5 and NHS genes**. *Am J Med Genet A* 2007, **143**(4):364-369.
126. Toruner GA, Streck DL, Schwalb MN, Dermody JJ: **An oligonucleotide based array-CGH system for detection of genome wide copy number changes including subtelomeric regions for genetic evaluation of mental retardation**. *Am J Med Genet A* 2007, **143**(8):824-829.
127. Vissers LE, Stankiewicz P, Yatsenko SA, Crawford E, Creswick H, Proud VK, de Vries BB, Pfundt R, Marcelis CL, Zackowski J *et al*: **Complex chromosome 17p rearrangements associated with low-copy repeats in two patients with congenital anomalies**. *Human genetics* 2007.
128. Lugtenberg D, Yntema HG, Banning MJ, Oudakker AR, Firth HV, Willatt L, Raynaud M, Kleefstra T, Fryns JP, Ropers HH *et al*: **ZNF674: A New Kruppel-Associated Box-Containing Zinc-Finger Gene Involved in Nonsyndromic X-Linked Mental Retardation**. *American journal of human genetics* 2006, **78**(2):265-278.
129. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science (New York, NY* 1999, **286**(5439):531-537.
130. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ *et al*: **A gene-expression signature as a predictor of survival in breast cancer**. *The New England journal of medicine* 2002, **347**(25):1999-2009.
131. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS *et al*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications**. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(19):10869-10874.
132. Alizadeh AA, Staudt LM: **Genomic-scale gene expression profiling of normal and malignant immune cells**. *Current opinion in immunology* 2000, **12**(2):219-225.
133. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X *et al*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. *Nature* 2000, **403**(6769):503-511.
134. Churchill GA: **Fundamentals of experimental design for cDNA microarrays**. *Nat Genet* 2002, **32 Suppl**:490-495.
135. Yang YH, Speed T: **Design issues for cDNA microarray experiments**. *Nature reviews* 2002, **3**(8):579-588.
136. Pavlidis P, Li Q, Noble WS: **The effect of replication on gene expression microarray experiments**. *Bioinformatics (Oxford, England)* 2003, **19**(13):1620-1627.
137. Lee ML, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations**. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(18):9834-9839.
138. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nature reviews* 2006, **7**(1):55-65.
139. Leung YF, Cavalieri D: **Fundamentals of cDNA microarray data analysis**. *Trends Genet* 2003, **19**(11):649-659.

140. Dobbin K, Shih JH, Simon R: **Questions and answers on design of dual-label microarrays for identifying differentially expressed genes**. *Journal of the National Cancer Institute* 2003, **95**(18):1362-1369.
141. Miller RA, Galecki A, Shmookler-Reis RJ: **Interpretation, design, and analysis of gene array expression experiments**. *The journals of gerontology* 2001, **56**(2):B52-57.
142. Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes**. *Bioinformatics (Oxford, England)* 2001, **17**(6):509-519.
143. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data**. *J Comput Biol* 2000, **7**(6):819-837.
144. Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA: **Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays**. *Bioinformatics (Oxford, England)* 2003, **19**(11):1348-1359.
145. Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B *et al*: **A comparison of oligonucleotide and cDNA-based microarray systems**. *Physiological genomics* 2004, **16**(3):361-370.
146. Moreau Y, Aerts S, De Moor B, De Strooper B, Dabrowski M: **Comparison and meta-analysis of microarray data: from the bench to the computer desk**. *Trends Genet* 2003, **19**(10):570-577.
147. Draghici S, Khatri P, Eklund AC, Szallasi Z: **Reliability and reproducibility issues in DNA microarray measurements**. *Trends Genet* 2006, **22**(2):101-109.
148. Lock C, Hermans G, Pedotti R, Brendolan A, Schadt E, Garren H, Langer-Gould A, Strober S, Cannella B, Allard J *et al*: **Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis**. *Nature medicine* 2002, **8**(5):500-508.
149. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM: **Delineation of prognostic biomarkers in prostate cancer**. *Nature* 2001, **412**(6849):822-826.
150. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
151. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A *et al*: **Molecular classification of cutaneous malignant melanoma by gene expression profiling**. *Nature* 2000, **406**(6795):536-540.
152. Young AN, Amin MB, Moreno CS, Lim SD, Cohen C, Petros JA, Marshall FF, Neish AS: **Expression profiling of renal epithelial neoplasms: a method for tumor classification and discovery of diagnostic molecular markers**. *The American journal of pathology* 2001, **158**(5):1639-1651.
153. Abdullah-Sayani A, Bueno-de-Mesquita JM, van de Vijver MJ: **Technology Insight: tuning into the genetic orchestra using microarrays--limitations of DNA microarrays in clinical practice**. *Nature clinical practice* 2006, **3**(9):501-516.
154. Simon R, Radmacher MD, Dobbin K: **Design of studies using DNA microarrays**. *Genetic epidemiology* 2002, **23**(1):21-36.
155. Peyman GA, Kazi AA, Riazi-Esfahani M, Aydin E, Kivilcim M, Sanders DR: **The effect of combinations of flurbiprofen, low molecular weight heparin, and doxycycline on the inhibition of corneal neovascularization**. *Cornea* 2006, **25**(5):582-585.
156. Quackenbush J: **Genomics. Microarrays--guilt by association**. *Science (New York, NY* 2003, **302**(5643):240-241.

157. Balasubramanian R, LaFramboise T, Scholtens D, Gentleman R: **A graph-theoretic approach to testing associations between disparate sources of functional genomics data**. *Bioinformatics (Oxford, England)* 2004, **20**(18):3353-3362.
158. Giallourakis C, Henson C, Reich M, Xie X, Mootha VK: **Disease gene discovery through integrative genomics**. *Annual review of genomics and human genetics* 2005, **6**:381-406.
159. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F *et al*: **Identification of a gene causing human cytochrome c oxidase deficiency by integrative genomics**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(2):605-610.
160. Meng H, Vera I, Che N, Wang X, Wang SS, Ingram-Drake L, Schadt EE, Drake TA, Lusis AJ: **Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(11):4530-4535.
161. Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung CS: **The GenBank genetic sequence databank**. *Nucleic acids research* 1986, **14**(1):1-4.
162. Hamm GH, Cameron GN: **The EMBL data library**. *Nucleic acids research* 1986, **14**(1):5-9.
163. George DG, Barker WC, Hunt LT: **The protein identification resource (PIR)**. *Nucleic acids research* 1986, **14**(1):11-15.
164. Galperin MY: **The Molecular Biology Database Collection: 2007 update**. *Nucleic acids research* 2007, **35**(Database issue):D3-4.
165. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic acids research* 2007, **35**(Database issue):D26-31.
166. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE: **The Mouse Genome Database (MGD): updates and enhancements**. *Nucleic acids research* 2006, **34**(Database issue):D562-567.
167. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic acids research* 2007, **35**(Database issue):D5-12.
168. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, Frazer K, Haendel M, Howe DG, Mani P, Ramachandran S *et al*: **The Zebrafish Information Network: the zebrafish model organism database**. *Nucleic acids research* 2006, **34**(Database issue):D581-585.
169. Hirschman JE, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hong EL, Livstone MS, Nash R *et al*: **Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome**. *Nucleic acids research* 2006, **34**(Database issue):D442-445.
170. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al*: **The COG database: an updated version includes eukaryotes**. *BMC bioinformatics* 2003, **4**:41.
171. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM: **FlyBase: genomes by the dozen**. *Nucleic acids research* 2007, **35**(Database issue):D486-491.
172. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM**. *American journal of human genetics* 2007, **80**(4):588-604.
173. Stoeckert CJ, Jr., Causton HC, Ball CA: **Microarray databases: standards and ontologies**. *Nat Genet* 2002, **32 Suppl**:469-473.

174. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
175. Consortium" GO: **Creating the gene ontology resource: design and implementation**. *Genome research* 2001, **11**(8):1425-1433.
176. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R: **The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase**. *In silico biology* 2004, **4**(1):5-6.
177. Khatri P, Draghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems**. *Bioinformatics (Oxford, England)* 2005, **21**(18):3587-3595.
178. Pasquier C, Girardot F, Jevardat de Fombelle K, Christen R: **THEA: ontology-driven analysis of microarray data**. *Bioinformatics (Oxford, England)* 2004, **20**(16):2636-2643.
179. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S *et al*: **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome biology* 2003, **4**(4):R28.
180. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens RM, Bryant D, Burt SK *et al*: **High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID)**. *BMC bioinformatics* 2005, **6**:168.
181. Draghici S, Khatri P, Bhavsar P, Shah A, Krawetz SA, Tainsky MA: **Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate**. *Nucleic acids research* 2003, **31**(13):3775-3781.
182. Al-Shahrour F, Diaz-Uriarte R, Dopazo J: **FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes**. *Bioinformatics (Oxford, England)* 2004, **20**(4):578-580.
183. Dryja TP, McGee TL, Reichel E, Hahn LB, Cowley GS, Yandell DW, Sandberg MA, Berson EL: **A point mutation of the rhodopsin gene in one form of retinitis pigmentosa**. *Nature* 1990, **343**(6256):364-366.
184. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes**. *Nature* 2001, **409**(6822):853-855.
185. Lopez-Bigas N, Blencowe BJ, Ouzounis CA: **Highly consistent patterns for inherited human diseases at the molecular level**. *Bioinformatics (Oxford, England)* 2006, **22**(3):269-277.
186. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease**. *Nucleic acids research* 2004, **32**(10):3108-3114.
187. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization**. *BMC bioinformatics* 2005, **6**:55.
188. Tu Z, Wang L, Xu M, Zhou X, Chen T, Sun F: **Further understanding human disease genes by comparing with housekeeping genes and other genes**. *BMC genomics* 2006, **7**:31.
189. Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network**. *Bioinformatics (Oxford, England)* 2006, **22**(22):2800-2805.
190. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions**. *J Med Genet* 2006, **43**(8):691-698.

191. Calvo B, Lopez-Bigas N, Furney SJ, Larranaga P, Lozano JA: **A partially supervised classification approach to dominant and recessive human disease gene prediction**. *Computer methods and programs in biomedicine* 2007, **85**(3):229-237.
192. Chelly J, Mandel JL: **Monogenic causes of X-linked mental retardation**. *Nature reviews* 2001, **2**(9):669-680.
193. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes**. *Bioinformatics (Oxford, England)* 2002, **18 Suppl 2**:S110-115.
194. Turner FS, Clutterbuck DR, Semple CA: **POCUS: mining genomic sequence annotation to predict disease genes**. *Genome biology* 2003, **4**(11):R75.
195. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining**. *Nature genetics* 2002, **31**(3):316-319.
196. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA: **G2D: a tool for mining genes associated with disease**. *BMC genetics* 2005, **6**:45.
197. Masseroli M, Martucci D, Pinciroli F: **GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining**. *Nucleic acids research* 2004, **32**(Web Server issue):W293-300.
198. Masseroli M, Galati O, Pinciroli F: **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists**. *Nucleic acids research* 2005, **33**(Web Server issue):W717-723.
199. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates**. *Bioinformatics (Oxford, England)* 2006, **22**(6):773-774.
200. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S: **TOM: a web-based integrated approach for identification of candidate disease genes**. *Nucleic acids research* 2006, **34**(Web Server issue):W285-292.
201. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B *et al*: **Gene prioritization through genomic data fusion**. *Nature biotechnology* 2006, **24**(5):537-544.
202. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction**. *Nucleic acids research* 2006, **34**(19):e130.
203. Chiang JH, Shin JW, Liu HH, Chin CL: **GeneLibrarian: an effective gene-information summarization and visualization system**. *BMC bioinformatics* 2006, **7**:392.
204. Franke A, Wollstein A, Teuber M, Wittig M, Lu T, Hoffmann K, Nurnberg P, Krawczak M, Schreiber S, Hampe J: **GENOMIZER: an integrated analysis system for genome-wide association data**. *Hum Mutat* 2006, **27**(6):583-588.
205. Sun H, Fang H, Chen T, Perkins R, Tong W: **GOFFA: Gene Ontology For Functional Analysis - A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data**. *BMC bioinformatics* 2006, **7 Suppl 2**:S23.
206. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases**. *Nucleic acids research* 2005, **33**(Web Server issue):W758-761.
207. Kent WJ, Hsu F, Karolchik D, Kuhn RM, Clawson H, Trumbower H, Haussler D: **Exploring relationships and mining data with the UCSC Gene Sorter**. *Genome research* 2005, **15**(5):737-741.
208. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates**. *Nucleic acids research* 2005, **33**(5):1544-1552.

209. Philippe C: **Cartographie physique du chromosome X humain**. Nancy: Institut National Polytechnique de Lorraine; 1994.
210. Rainen L, Oelmueller U, Jurgensen S, Wyrich R, Ballas C, Schram J, Herdman C, Bankaitis-Davis D, Nicholls N, Trollinger D *et al*: **Stabilization of mRNA expression in whole blood samples**. *Clinical chemistry* 2002, **48**(11):1883-1890.
211. Chai V, Vassilakos A, Lee Y, Wright JA, Young AH: **Optimization of the PAXgene blood RNA extraction system for gene expression analysis of clinical samples**. *Journal of clinical laboratory analysis* 2005, **19**(5):182-188.
212. Bauters M, Van Esch H, Marynen P, Froyen G: **X chromosome array-CGH for the identification of novel X-linked mental retardation genes**. *European journal of medical genetics* 2005, **48**(3):263-275.
213. Nadon R, Shoemaker J: **Statistical issues with microarrays: processing and analysis**. *Trends Genet* 2002, **18**(5):265-271.
214. Kreil DP, Russell RR: **There is no silver bullet--a guide to low-level data transforms and normalisation methods for microarray data**. *Briefings in bioinformatics* 2005, **6**(1):86-97.
215. Quackenbush J: **Microarray data normalization and transformation**. *Nat Genet* 2002, **32 Suppl**:496-501.
216. Allen RC, Zoghbi HY, Moseley AB, Rosenblatt HM, Belmont JW: **Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation**. *American journal of human genetics* 1992, **51**(6):1229-1239.
217. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW: **Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome**. *Cytogenetic and genome research* 2006, **115**(3-4):205-214.
218. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG: **Genetic analysis of genome-wide variation in human gene expression**. *Nature* 2004, **430**(7001):743-747.
219. Mutch DM, Berger A, Mansourian R, Rytz A, Roberts MA: **The limit fold change model: a practical approach for selecting differentially expressed genes from microarray data**. *BMC bioinformatics* 2002, **3**:17.
220. Doherty MJ, Glass IA, Bennett CL, Cotter PD, Watson NF, Mitchell AL, Bird TD, Farrell DF: **An Xp; Yq translocation causing a novel contiguous gene syndrome in brothers with generalized epilepsy, ichthyosis, and attention deficits**. *Epilepsia* 2003, **44**(12):1529-1535.
221. Garcia CC, Blair HJ, Seager M, Coulthard A, Tennant S, Buddles M, Curtis A, Goodship JA: **Identification of a mutation in synapsin I, a synaptic vesicle protein, in a family with epilepsy**. *Journal of medical genetics* 2004, **41**(3):183-186.
222. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B: **GOToolBox: functional analysis of gene datasets based on Gene Ontology**. *Genome biology* 2004, **5**(12):R101.
223. Jehee FS, Bertola DR, Yelavarthi KK, Krepischi-Santos AC, Kim C, Vianna-Morgante AM, Vermeesch JR, Passos-Bueno MR: **An 11q11-q13.3 duplication, including FGF3 and FGF4 genes, in a patient with syndromic multiple craniosynostoses**. *Am J Med Genet A* 2007, **143**(16):1912-1918.
224. Maas ND, Van de Putte T, Melotte C, Francis A, Schrander-Stumpel CT, Sanlaville D, Genevieve D, Lyonnet S, Dimitrov B, Devriendt K *et al*: **The C20orf133 gene is disrupted in a patient with Kabuki syndrome**. *Journal of medical genetics* 2007.
225. Thienpont B, de Ravel T, Van Esch H, Van Schoubroeck D, Moerman P, Vermeesch JR, Fryns JP, Froyen G, Lacoste C, Badens C *et al*: **Partial duplications of the ATRX gene cause the ATR-X syndrome**. *Eur J Hum Genet* 2007.

226. Froyen G, Van Esch H, Bauters M, Hollanders K, Frints SG, Vermeesch JR, Devriendt K, Fryns JP, Marynen P: **Detection of genomic copy number changes in patients with idiopathic mental retardation by high-resolution X-array-CGH: important role for increased gene dosage of XLMR genes**. *Human mutation* 2007.
227. Thienpont B, Mertens L, de Ravel T, Eyskens B, Boshoff D, Maas N, Fryns JP, Gewillig M, Vermeesch JR, Devriendt K: **Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients**. *Eur Heart J* 2007.
228. Castermans D, Vermeesch JR, Fryns JP, Steyaert JG, Van de Ven WJ, Creemers JW, Devriendt K: **Identification and characterization of the TRIP8 and REEP3 genes on chromosome 10q21.3 as novel candidate genes for autism**. *Eur J Hum Genet* 2007, **15**(4):422-431.
229. Wozniak A, Sciot R, Guillou L, Pauwels P, Wasag B, Stul M, Vermeesch JR, Vandenberghe P, Limon J, Debiec-Rychter M: **Array CGH analysis in primary gastrointestinal stromal tumors: cytogenetic profile correlates with anatomic site and tumor aggressiveness, irrespective of mutational status**. *Genes, chromosomes & cancer* 2007, **46**(3):261-276.
230. Friedman JM, Baross A, Delaney AD, Ally A, Arbour L, Armstrong L, Asano J, Bailey DK, Barber S, Birch P *et al*: **Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation**. *American journal of human genetics* 2006, **79**(3):500-513.
231. Vermeesch JR, Fiegler H, de Leeuw N, Szuhai K, Schoumans J, Ciccone R, Speleman F, Rauch A, Clayton-Smith J, Van Ravenswaaij C *et al*: **Guidelines for molecular karyotyping in constitutional genetic diagnosis**. *Eur J Hum Genet* 2007.
232. Vermeesch JR, Melotte C, Froyen G, Van Vooren S, Dutta B, Maas N, Vermeulen S, Menten B, Speleman F, De Moor B *et al*: **Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis**. *J Histochem Cytochem* 2005, **53**(3):413-422.
233. Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ, Saucedo-Cardenas O, Barker DF, Killian JM, Garcia CA *et al*: **DNA duplication associated with Charcot-Marie-Tooth disease type 1A**. *Cell* 1991, **66**(2):219-232.
234. Menten B, Maas N, Thienpont B, Buysse K, Vandesompele J, Melotte C, de Ravel T, Van Vooren S, Balikova I, Backx L *et al*: **Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports**. *Journal of medical genetics* 2006, **43**(8):625-633.
235. Vissers LE, Veltman JA, van Kessel AG, Brunner HG: **Identification of disease genes by whole genome CGH arrays**. *Human molecular genetics* 2005, **14 Spec No. 2**:R215-223.
236. Lugtenberg D, de Brouwer AP, Kleefstra T, Oudakker AR, Frints SG, Schrander-Stumpel CT, Fryns JP, Jensen LR, Chelly J, Moraine C *et al*: **Chromosomal copy number changes in patients with non-syndromic X linked mental retardation detected by array CGH**. *Journal of medical genetics* 2006, **43**(4):362-370.
237. Ropers HH, Hoeltzenbein M, Kalscheuer V, Yntema H, Hamel B, Fryns JP, Chelly J, Partington M, Gecz J, Moraine C: **Nonsyndromic X-linked mental retardation: where are the missing mutations?** *Trends Genet* 2003, **19**(6):316-320.
238. Veltman JA, Yntema HG, Lugtenberg D, Arts H, Briault S, Huys EH, Osoegawa K, de Jong P, Brunner HG, Geurts van Kessel A *et al*: **High resolution profiling of X chromosomal aberrations by array comparative genomic hybridisation**. *Journal of medical genetics* 2004, **41**(6):425-432.

239. Yilmaz S, Fontaine H, Brochet K, Gregoire MJ, Devignes MD, Schaff JL, Philippe C, Nemos C, McGregor JL, Jonveaux P: **Screening of subtle copy number changes in Aicardi syndrome patients with a high resolution X chromosome array-CGH**. *European journal of medical genetics* 2007.
240. Gargiulo A, Auricchio R, Barone MV, Cotugno G, Reardon W, Milla PJ, Ballabio A, Ciccodicola A, Auricchio A: **Filamin A is mutated in X-linked chronic idiopathic intestinal pseudo-obstruction with central nervous system involvement**. *American journal of human genetics* 2007, **80**(4):751-758.
241. Sheen VL, Dixon PH, Fox JW, Hong SE, Kinton L, Sisodiya SM, Duncan JS, Dubeau F, Scheffer IE, Schachter SC *et al*: **Mutations in the X-linked filamin 1 gene cause periventricular nodular heterotopia in males as well as in females**. *Human molecular genetics* 2001, **10**(17):1775-1783.
242. Tang Y, Gilbert DL, Glauser TA, Hershey AD, Sharp FR: **Blood gene expression profiling of neurologic diseases: a pilot microarray study**. *Archives of neurology* 2005, **62**(2):210-215.
243. Hershey AD, Tang Y, Powers SW, Kabbouche MA, Gilbert DL, Glauser TA, Sharp FR: **Genomic abnormalities in patients with migraine and chronic migraine: preliminary blood gene expression suggests platelet abnormalities**. *Headache* 2004, **44**(10):994-1004.
244. Moore DF, Li H, Jeffries N, Wright V, Cooper RA, Jr., Elkahloun A, Gelderman MP, Zudaire E, Blevins G, Yu H *et al*: **Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation**. *Circulation* 2005, **111**(2):212-221.
245. Achiron A, Gurevich M, Friedman N, Kaminski N, Mandel M: **Blood transcriptional signatures of multiple sclerosis: unique gene expression of disease activity**. *Annals of neurology* 2004, **55**(3):410-417.
246. Sharp FR, Xu H, Lit L, Walker W, Apperson M, Gilbert DL, Glauser TA, Wong B, Hershey A, Liu DZ *et al*: **The future of genomic profiling of neurological diseases using blood**. *Archives of neurology* 2006, **63**(11):1529-1536.
247. Steinman L, Zamvil S: **Transcriptional analysis of targets in multiple sclerosis**. *Nat Rev Immunol* 2003, **3**(6):483-492.
248. Borovecki F, Lovrecic L, Zhou J, Jeong H, Then F, Rosas HD, Hersch SM, Hogarth P, Bouzou B, Jensen RV *et al*: **Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(31):11023-11028.
249. Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO: **Individuality and variation in gene expression patterns in human blood**. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(4):1896-1901.
250. Melke J, Goubran Botros H, Chaste P, Betancur C, Nygren G, Anckarsater H, Rastam M, Stahlberg O, Gillberg IC, Delorme R *et al*: **Abnormal melatonin synthesis in autism spectrum disorders**. *Mol Psychiatry* 2007.
251. Tang Y, Lu A, Ran R, Aronow BJ, Schorry EK, Hopkin RJ, Gilbert DL, Glauser TA, Hershey AD, Richtand NW *et al*: **Human blood genomics: distinct profiles for gender, age and neurofibromatosis type 1**. *Brain research* 2004, **132**(2):155-167.
252. Storch KF, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, Weitz CJ: **Extensive and divergent circadian gene expression in liver and heart**. *Nature* 2002, **417**(6884):78-83.
253. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB: **Coordinated transcription of key pathways in the mouse by the circadian clock**. *Cell* 2002, **109**(3):307-320.

254. Arion D, Sabatini M, Unger T, Pastor J, Alonso-Nanclares L, Ballesteros-Yanez I, Garcia Sola R, Munoz A, Mirnics K, DeFelipe J: **Correlation of transcriptome profile with electrical activity in temporal lobe epilepsy**. *Neurobiology of disease* 2006, **22**(2):374-387.
255. Gurvich N, Berman MG, Wittner BS, Gentleman RC, Klein PS, Green JB: **Association of valproate-induced teratogenesis with histone deacetylase inhibition in vivo**. *Faseb J* 2005, **19**(9):1166-1168.
256. Milutinovic S, D'Alessio AC, Detich N, Szyf M: **Valproate induces widespread epigenetic reprogramming which involves demethylation of specific genes**. *Carcinogenesis* 2007, **28**(3):560-571.
257. Bahi-Buisson N, Chelly J, des Portes V: **[Update on the genetics of X-linked mental retardation]**. *Revue neurologique* 2006, **162**(10):952-963.
258. Hendrich B, Bickmore W: **Human diseases with underlying defects in chromatin structure and modification**. *Human molecular genetics* 2001, **10**(20):2233-2242.
259. Perez-Iratxeta C, Bork P, Andrade-Navarro MA: **Update of the G2D tool for prioritization of gene candidates to inherited diseases**. *Nucleic acids research* 2007, **35**(Web Server issue):W212-216.

Liste des publications

Liste des publications

Les travaux de thèse ont fait l'objet des deux publications ci-dessous. Les articles sont présentés dans la session résultats de ce manuscrit.

- **Yilmaz S, Fontaine H, Brochet K, Grégoire MJ, Devignes MD, Schaff JL, Philippe C, Nemos C, McGregor JL, Jonveaux P.** Screening of subtle copy number changes in Aicardi Syndrome Patients with a high resolution X-chromosome array-CGH. European Journal of Medical Genetics Eur J Med Genet. 2007 Jun 7; [Epub ahead of print].

- **Yilmaz S, Jonveaux P, Bicep C, Pierron L, Smaïl-Tabbone M and Devignes MD.** A Database Approach for Candidate Gene Retrieval based on semantic Data Integration and Expert View Definition. *Submited to Bioinformatics*

Durant mon stage de maîtrise (master1), j'ai participé aux recherches sur le syndrome de Rett avec la mise en place d'un protocole de criblage des variants génomiques par DHPLC (Denaturing High Performance Liquid Chromatography). L'article est présenté ci-après.

- **Quenard A., Yilmaz S, Fontaine H, Bienvenu T, Moncla A, des Portes V, Rivier F, Mathieu M, Raux G, Jonveaux P and Philippe C**. Deleterious mutations in exon 1 of MECP2 in Rett syndrome. Eur J Med Genet, 2006 Jul-Aug;49(4):313-22.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Medical Genetics 49 (2006) 313–322
EUROPEAN JOURNAL OF MEDICAL GENETICS
http://france.elsevier.com/direct/ejmg

Original article

Deleterious mutations in exon 1 of *MECP2* in Rett syndrome

Aline Quenard [a], Saliha Yilmaz [a], Hervé Fontaine [a], Thierry Bienvenu [b], Anne Moncla [c], Vincent des Portes [d], François Rivier [e], Michèle Mathieu [f], Grégory Raux [g], Philippe Jonveaux [a], Christophe Philippe [a,*]

[a] *Laboratoire de génétique, EA 3441, CHU Brabois, avenue du Morvan, 54511 Vandoeuvre-les-Nancy cedex, France*
[b] *Institut Cochin et laboratoire génétique moléculaire, pavillon Cassini, CHU, hôpital Cochin, université Paris-V, 75679 Paris cedex 14, France*
[c] *Département de génétique médicale, hôpital d'enfants de la Timone, 13385 Marseille cedex 05, France*
[d] *Service de neurologie pédiatrique, centre hospitalier Lyon-sud, 165, chemin du grand Revoyet, 69495 Pierre-Benite cedex, France*
[e] *Service de neurologie pédiatrique, CHU, hôpital Gui-de-Chauliac, 80, avenue Augustin-Fliche, 34295 Montpellier cedex, France*
[f] *Unité de génétique clinique, département de pédiatrie, CHU, 80054 Amiens cedex 01, France*
[g] *Inserm EMI U614, faculté de médecine, 76183 Rouen cedex, France*

Available online 20 December 2005

Abstract

The *MECP2* gene is responsible for 80–85% of typical cases of Rett syndrome with deleterious mutations affecting exons 3 and 4. Recently, an alternate transcript including exon 1 was discovered with a new protein isoform (MeCP2_e1) much more abundant in brain. We screened exon 1 of *MECP2* for mutations and for large rearrangements in a panel of 212 typical cases of Rett syndrome and one family case with atypical Rett syndrome. We identified two deleterious mutations (c.48_55dup and c.62+2_62 +3del) and four large rearrangements encompassing exon 1 of *MECP2*. We also identified the c.16_21dup alteration formerly reported as c.3_4insGCCGCC and give additional support to classify this sequence variation as polymorphic. In our large panel of typical Rett, mutations affecting exon 1 of

☆ Databases: OMIM: 312750 (RTT), Genbank: NT_025913 and AY541280.1; RettBASE: http://mecp2.chw.edu.au.

* Corresponding author. Tel.: +33 3 83 15 37 78; fax: +33 3 83 15 37 72.
E-mail address: c.philippe@chu-nancy.fr (C. Philippe).

1769-7212/$ - see front matter
doi:10.1016/j.ejmg.2005.11.002

MECP2 represent 1% of the deleterious alleles. This study confirms that mutations in exon 1 of *MECP2* are a rare cause of Rett syndrome.

Keywords: Rett; Exon 1; MeCP2 isoforms; Deleterious mutation

1. Introduction

Rett syndrome (RTT, MIM # 312750) is a progressive devastating neurological developmental disorder affecting almost exclusively girls, it occurs in 1:15,000 female births and is the most common genetic cause of profound mental retardation in females [17]. Mutations in *MECP2* are a frequent cause of RTT [3,11,17]. The human *MECP2* gene was cloned in 1996 and described first as a three exon gene [6]. Comparative sequence analysis enabled the identification of a fourth 5′ exon initially considered as non coding [15] and until recently, a single *MECP2* transcript (*MECP2_e2*) with all four exons and a start codon in exon 2 was identified. Two studies combining bioinformatics and molecular biology reported a second alternative splice variant (MECP2_e1) with the start codon in exon 1 joined to exon 3 in mature mRNA [10,12]. The *MECP2_e1* open reading frame leads to different MeCP2 terminus with a distinctive 21 amino-acid peptide, MeCP2_e1 is the predominant protein isoform of 498 amino acids in adult human brain [12] where its level is 10 times higher compared to those of the MeCP2_2 isoform of 486 amino acids.

As exon 1 was thought to be non coding, it was therefore not screened for mutation, and deleterious alterations were reported exclusively in exons 3 and 4 ([3,11,17], RettBASE). Mutations in exon 2 present only in the transcript generating MeCP2_e2 were never found, this strongly supports a more important functional role for the MeCP2_e1 isoform and suggests that exon 1 has to be analysed in mutation negative RTT patients. So far, few mutations in exon 1 were found, almost exclusively in typical RTT [1,12,14,16]. In this study we assessed the mutation rate in exon 1 of *MECP2* in a large cohort of classical RTT by denaturating high pressure liquid chromatography (DHPLC) and quantitative multiplex PCR of short fluorescent fragments (QMPSF).

2. Methods

2.1. Patients and DNA samples

We studied 212 patients with classical sporadic Rett diagnosed according to the international criteria [2,8]. Blood samples were obtained from patients after informed consent. Total genomic DNA was extracted using the Nucleon BACC3 kit (Amersham Life Technologies) according to the manufacturer's protocol.

2.2. Mutation detection

2.2.1. PCR amplification

Exon 1, a 207 bp long GC-rich amplicon difficult to amplify, was PCR amplified with the Thermoprime Taq polymerase (ABgene) by using 7.5% DMSO, 1 mM $MgCl_2$ and 38 cycles with annealing at 56 °C. The primer pair used to PCR amplify exon 1 of MECP2 is reported in Table 1.

2.2.2. DHPLC analysis

Exon 1 derived PCR products were screened for heterozygous base changes with a Transgenomic Wave DNA fragment analysis system.

Heteroduplex formation was induced by heat denaturation of PCR products at 98 °C for 5 min, followed by gradual reannealing from 98 to 25 °C over 40 min. Before DHPLC analysis, the PCR products are controlled by electrophoresis on a 2% agarose gel. The buffer gradients (buffer A, 0.1 M TEAA; buffer B, 0.1 M TEAA/25% acetonitrile), start and end points of the linear gradient and the melting temperature predictions were determined using to the WAVEMAKER software 4.1 (Transgenomic). Analysis took 2.5 min per sample with the WAVE accelerator on a WAVE 3500HT. Column temperatures for each exon 1: 63, 66.5 °C. Heterozygous profiles were detected after visual inspection of the chromatograms and comparison with normal controls.

2.2.3. Detection of large heterozygous deletions or duplications

Two multiplex PCRs were performed to amplify the promoter, all four exons and part of the 3′ UTR region of MECP2. For the coding part of exon 4 which is 1084 bp long, we amplified four short fragments. A standard amplicon (exon 20 of BRCA1) was included in both multiplex PCRs for the subsequent analysis of fluorescent profiles. Primer sequences used in

Table 1
Primer sets for DHPLC and multiplex fluorescent PCRs
[a] For each amplicon, forward (F) and reverse (R) are equimolar. The sizes of fluorescent PCR products for the QMPSF include the hexadecamer sequences added on the 5′ side: 5′-CGTTAGATAG-3′ (forward primers) and 5′-GATAGGGTTA-3′ (reverse primers). For the QMPSF multiplex PCR 1 and 2, forward primers are 5′ labelled with the fluorescein 6-FAM dye.

Exon	Name	Sequence (5'>3')	Name	Sequence (5'>3')	Concentration (µM)	Size (bp)
PCR DHPLC						
1	E1F	GGAGAGAGGGCTGTGGTAAAAG	E1R	CATCCGCCAGCCGTGTCGTCCG	0,4	207
QMPSF [a]						
Multiplex 1						
1	E1F	GGAGAGAGGGCTGTGGTAAAAG	E1R	CATCCGCCAGCCGTGTCGTCCG	0,4	207
2	E2F	AAAGGTCGTGCAGCTCAATG	E2R	TTACCTGAGCCCTAACATCCC	1,5	201
4b	E4AF	AACCACCTAAGAAGCCCAAATC	E4AR	CTTCCCAGGACTTTTCTCCAG	1	166
4c	E4BF	CGGGAAAGGACTGAAGACCTG	E4BR	TGAGTGGTGGTGATGGTGGTG	0,5	144
4d	E4CF	GTTTCATCCTCCATGCCAAGG	E4CR	GGGAAGCTTTGTCAGAGCCC	0,3	190
3'UTR	3'UTR F	GGTTGGAGGGAAAGCCTTAAG	3'UTR R	GGAAAGCTGTCGGTCAATTCTG	1	180
BRCA1-exon20	E20F3	GCTCCACTTCCATTGAAGGAAG	E20R2	AGCGGCCCATCTCTGCAAAG	2,5	251
Multiplex 2						
promotor	PRO F	GCTCCAAATCAACTCTTCTAGG	PRO R	AAACTGACAACAAATCGCAACC	1,4	228
3	E3F	GAAAGAGGGCAAGCATGAGC	E3R	CAGGCAGGGTGGGGTCAT	0,6	211
4a	E4A1F	GTCACCACCATCCGCTCTGC	E4A1R	GTTTCTGCTCTCGCCGGGAG	0,3	243
BRCA1-exon20	E20F3	GCTCCACTTCCATTGAAGGAAG	E20R2	AGCGGCCCATCTCTGCAAAG	1,4	251

the QMPSF protocol are listed in Table 1. Primer design was performed as described previously in [5], except that all primers carried different 5′ extension of 10 bases (Table 1). Primers were HPLC purified (Applied Biosystems).

QMPSF reactions were performed in a 25 μl reaction mix containing 0.2 mM dNTPs, 2 and 1 mM $MgCl_2$ for multiplexes 1 and 2, respectively, 7.5% DMSO, 1.5 units Thermoprime plus DNA polymerase (Abgene, Epson, UK) and 300 ng of genomic DNA or less, however using the same amounts within each series of samples to be compared. Final concentrations for primer pairs are listed in Table 1, they were adjusted to obtain fluorescent peak levels within the same range. The PCR consisted of an initial denaturation at 95 °C for 7 min, followed by 23 cycles (95 °C for 30 s, 56 °C for 30 s, 72 °C for 30 s) and ended by a 7 min extension at 72 °C.

PCR products were separated by capillary electrophoresis for 24 min on an ABI prism 310 DNA sequencer (Applied Biosystems) at 15 kV using a 47 cm capillary and POP4 polymer. Usually, 1 μl of multiplex PCR product was combined with 12 μl formamide and 0.3 μl size standard ROX 400HD (Applied Biosystems), and a 2 min denaturation step at 95 °C followed by a quick cooling on ice was performed before electrophoresis. The total yield of multiplex PCR was sometimes estimated by agarose gel electrophoresis and ethidium bromide staining prior to capillary electrophoresis.

Data were analysed with the Genescan 2.1 software. Normalisation of fluorescence intensities was performed by adjusting the peak heights of the *BRCA1* exon 20 amplicons in both the control and the patient. A twofold reduction or a 1.5-fold increase of the peak area reveals a deletion or a duplication, respectively, of the corresponding exon of the *MECP2* gene.

2.2.4. Sequencing analysis

When a shift was detected by DHPLC, a second PCR reaction was performed, the product was purified using the Millipore PCR96 cleanup plates on a TECAN Genesis RSP 100. The purified PCR product was bidirectionally sequenced with the BigDye terminator v1.1 cycle sequencing kit (Applied Biosystems), in a 10 μl reaction. Cycle sequencing products were purified on DyeEx 2.0 spin kit (Qiagen) and sequenced on both strand on an ABI prism 3100 DNA sequencer (Applied Biosystems). Sequences were visually analysed with the sequencing analysis 3.7 software.

2.2.5. X-inactivation assay

X-chromosome inactivation pattern was determined as described previously in [4].

3. Results

We routinely applied a DHPLC based molecular diagnostic protocol for all four exons of the *MECP2* gene. For all RTT patients that did not show any deleterious mutation we searched for large rearrangements of the *MECP2* gene by QMPSF. We tested all four exons of the *MECP2* gene in a cohort of 212 typical RTT and one family case with atypical RTT. We found 148 deleterious mutations in exons 3 and 4 together with 27 heterozygous gross rearrangements of *MECP2* ranging from partial deletions of exon 4 to deletion of the whole *MECP2* locus ([3,13], and unpublished data). We found two deleterious variations in exon 1 of *MECP2* in typical RTT i.e. a 8 bp frameshifting duplication (c.48_55dup) and a 2 pb deletion affecting the donor splice site of intron 1 (c.62+2_62+3del) (Table 2). In addition, four out of

Table 2
Mutations and polymorphisms in *MECP2*
a: Sequence variations in exon 1 are numbered starting from the first base of the ATG start codon, numbering based on reference sequence AY541280.1.
b: Breakpoints were not sequenced for large rearrangements, the numbering is based on the genomic reference sequence AF030876.2.
c: Deduced but not experimentally verified on mRNAs from the patient.
d: Also reported by Evans et al. [2004].
N/A: non applicable.

Nucleotide change	Amino acid change	MeCP2 isoforms affected	Pathogenic / Polymorphism
Subtle sequence variations [a]			
c.16_21dup	**p.Ala6_Ala7dup**	**MeCP2_e1**	**Polymorphism** [d]
c.48_55dup	**p.Glu19AlafsX46**	**MeCP2_e1**	**Pathogenic**
c.62+2_62+3del	**altered splicing** [c]	**MeCP2_e1**	**Pathogenic**
Large rearrangements [b]			
Heterozygous deletion of the promotor and exon 1 **g.90033+?_88969-?**	**N/A**	**MeCP2_e1/_e2**	**Pathogenic**
Heterozygous deletion of exon 1 and exon 2 **g.89175+?_83658-?**	**N/A**	**MeCP2_e1/_e2**	**Pathogenic**
Heterozygous deletion of the promotor, exon 1 and exon 2 **g.90033+?_83658-?**	**N/A**	**MeCP2_e1/_e2**	**Pathogenic**
Heterozygous deletion of the complete MECP2 locus **90033+?_18465-?**	**N/A**	**MeCP2_e1/_e2**	**Pathogenic**

27 heterozygous deletions detected by QMPSF encompassed exon1 of *MECP2* (Table 2). In our series, the overall mutation detection rate was 83.5% for classical RTT (177/212). The vast majority of mutations was located in exons 3 and 4 (148/150). Mutations in exon 1 and large rearrangements accounted for 1.12% and 15.2% of all deleterious mutations found in *MECP2*, respectively. Mutations in exon 1 were found in 0.95% of typical RTT (2/212).

3.1. c.48_55dup

The out of frame 8 pb duplication in exon 1 was identified in a 5-year-old girl with typical RTT (Fig. 1). After a normal prenatal and perinatal period, her early development was normal until 12 months when a regression was diagnosed with deceleration of head growth. When seen at the age of 3 years and 3 months, her head circumference was 45 cm (< 3rd percentile) and she had lost hand skills. She was noted to have stereotypic hand movements. She sat at 6 months but was unable to walk. She was able to utter one word and social interactions were disturbed with autistic features. The mutation is de novo and associated with a random pattern of X-chromosome inactivation on DNA extracted from lymphocytes (data not shown).

3.2. c.62+2_62+3del

This atypical Rett patient was a female born at 40 weeks of gestation after a normal pregnancy. Her head circumference at birth was 33 cm (between 3rd and 10th percentile), and her

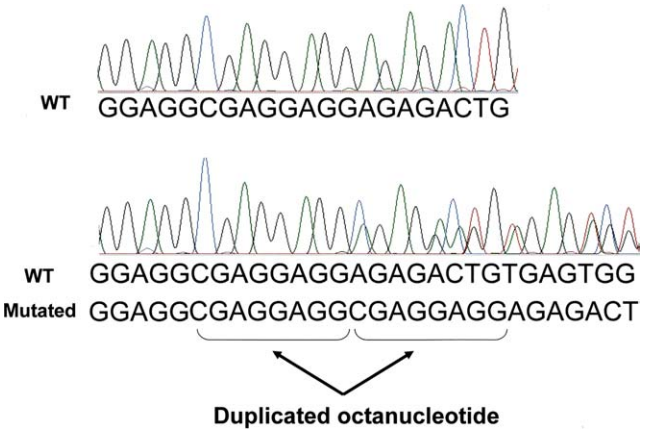


Fig. 1. Out of frame duplication in exon 1 of *MECP2B* splice variant.
Forward sequences from a control and the patient of the PCR products from genomic DNA using primers E1F and E1R. Top, forward wild-type (WT) sequence; bottom, forward mixed WT and mutated sequences from patient with the tandem duplication (48_55dupCGAGGAGG) of an octanucleotide.

birth weight was 3.510 g. Neonatal hypotonia was present. She presented to the paediatric neurologist at the age of 3 years with severe development delay. Her weight was 9.100 g and her head circumference was 44.5 cm (< 3rd percentile). She had never been able to crawl or walk and had never developed speech. She had some hand use without stereotypic movements. She developed myoclonic seizures at 4 years and displayed persistent severe hypotonia. A 2 bp deletion affecting the donor splice site of intron 1 (c.62+2_62+3del) was detected (Fig. 2). It was not possible to have access to a new blood sample from the affected girl (she died when she was six and a half before we identified the 2 bp deletion in intron 1), excluding the possibility to confirm the effect of the sequence variation on the splicing of the MECP2_e1 pre-mRNA.

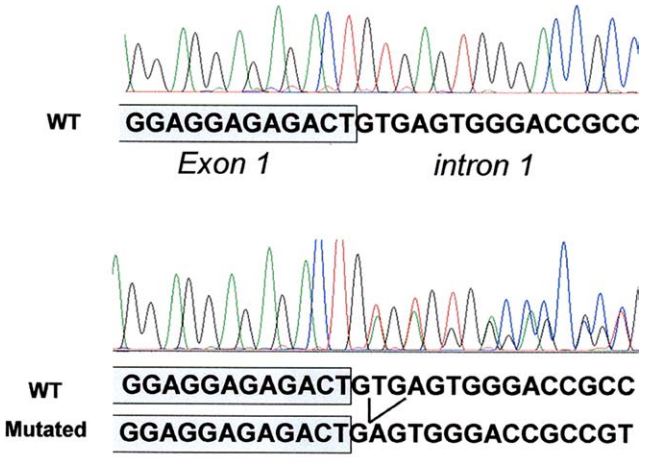


Fig. 2. Mutation affecting the donor splice site of intron 1 of *MECP2*.
Forward sequences from a control and the patient of the PCR products from genomic DNA using primers E1F and E1R. Top, forward WT sequence; bottom, forward mixed WT and mutated sequences from patient with the 2 bp deletion affecting donor splice site of intron 1. Grey box: 3′ end of exon 1.

After the death of their affected daughter the parents divorced and did not wish to participate to the study, it was therefore not possible to distinguish between a de novo and an inherited sequence variation. The pattern of X-inactivation in the affected girl was random on DNA extracted from leucocytes (data not shown).

3.3. c.16_21dup

We also investigated the *MECP2* gene in two sisters with a similar phenotype including intellectual disability and seizures, one sister being more severely affected with autistic traits. The mother presented with seizures during adolescence as the only clinical sign. A 6 bp duplication in the trinucleotide repeat tract at the start of exon 1 was present in the more severely affected sister and her mother. This sequence variation was absent in the sister with a mild phenotype, suggesting that it is a non pathogenic variant. The pattern of X-inactivation was random on DNA extracted from leucocytes in the sisters and their mother (data not shown).

3.4. Large rearrangements of MECP2

We implemented a QMPSF protocol for the detection of large rearrangements of *MECP2* and found four heterozygous deletions encompassing exon 1 of *MECP2* in typical RTT patients (Fig. 3 and Table 2). We detected three partial deletions of the *MECP2* locus involving exon 1 and one deletion removing the complete *MECP2* locus. The partial deletions were confirmed by using the Multiple Ligation-dependant Probe Amplification (MLPA, MRC Holland b.v.) kit (data not shown). We confirmed the heterozygous complete deletion of the *MECP2* locus by replacing the control amplicon of the QMPSF assay with a different internal control (exon 3 of the C1 inhibitor gene *SERPING1*) (data not shown). The X-inactivation patterns on leucocytes were random except for the partial deletion involving the promoter and exon 1 which is associated with a skewed X-chromosome inactivation (90–10%) (data not shown). All patients with large rearrangements of the MECP2 locus presented with a classical form of RTT, including the girl with the heterozygous deletion of the whole locus.

4. Discussion

Recent studies indicated that mutations in exon 1 are not a common cause of RTT [1,7,16]. To our knowledge, before our study a total of five deleterious mutations affecting exclusively the synthesis of MeCP2_e1 were reported in exon 1 of *MECP2*. The c.38_48del/c.47_57del deletions was reported three times in typical [12,14,16] and atypical RTT [1]. Symmetric elements predispose DNA sequences to meiotic microdeletions [9]. The hotspot in exon 1 of *MECP2* consists of symmetric elements at nucleotides c.38_c46 and c.49_c.57 (*GAGGAG GAG*GC*GAGGAGGAG*). The c.38_48del encompasses the c38_c.46 element with the adjacent GC dinucleotide whereas the c.47_57del starts from the GC dinucleotide and extends to the 3′ symmetric element. We believe that the c.38_48del/c.47_57del deletions should be regarded as the result of a single meiotic deletion hotspot in exon 1 of *MECP2*. Apart from this recurrent 11 bp deletion, a 5 bp duplication (c.23_25dup) and a 2 bp deletion affecting the donor splice site of exon 1 (c.62+1_62+2del) were found in typical RTT [1,14].

In a cohort of 212 typical RTT, we found two deleterious mutations in exon 1 and four large rearrangements encompassing exon 1 of *MECP2*. The 8 bp duplication (c.48_55dup) re-

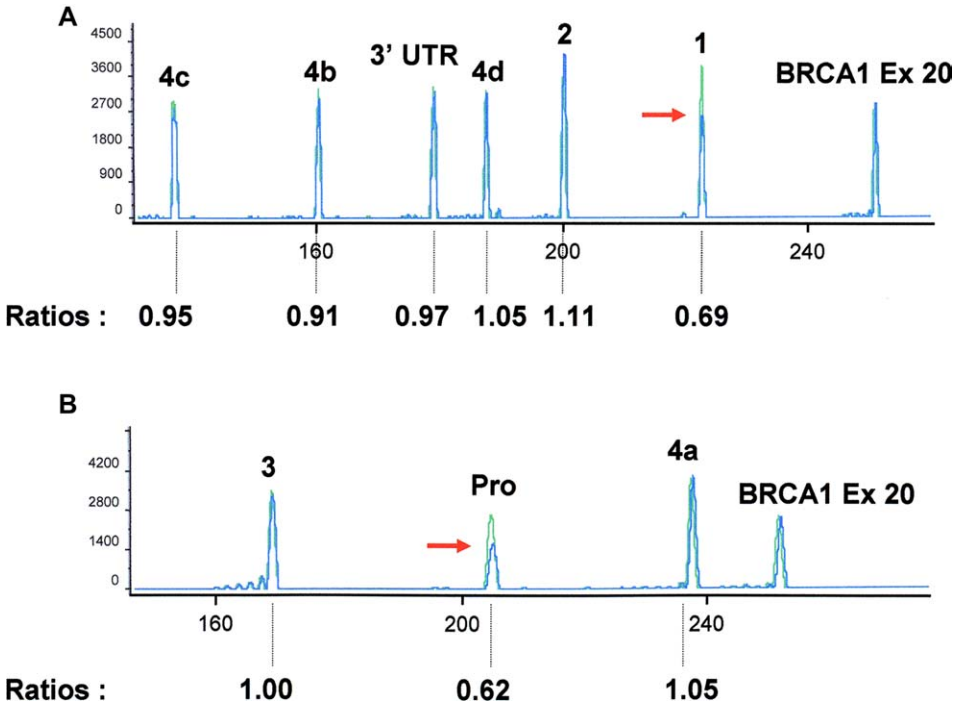


Fig. 3. Example of a large rearrangement involving exon 1 of *MECP2* detected by QMPSF.
A and B correspond to QMPSF Nos. 1 and 2, respectively, peak numbers refer to MECP2 amplicons. The horizontal scale is in bp, the vertical scale is in arbitrary units of fluorescence. The profile obtained with the RTT patient (blue) was superimposed with the one of a normal control (green) by adjusting to the same level the height of the BRCA1 exon 20 control peaks. For each amplicon, the ratio between the peak areas is indicated below the electrophoregrams. The arrows pinpoint amplicons with a twofold reduction of the intensity of the corresponding peaks indicative of a heterozygous deletion of the promoter and exon 1 of *MECP2* in the patient.

sults in a frame shift in the *MECP2_e1* open reading frame and may lead to a severely truncated MeCP2_e1 protein (p.Glu19AlafsX46). This case confirms that an intact MeCP2_e2 protein is not capable to compensate for the loss of MeCP2_e1 [1]. Disruption of MeCP2_e1 seems sufficient to cause a classic RTT phenotype although the duplication could also affect the 5′ UTR of *MECP2_e2* and modify its transcription and/or translation efficiencies. Recently, it has been shown that translation but not transcription of the MeCP2_e2 isoform is ablated by the 11 nucleotide deletion in exon 1 (c.47_57del), 103 nucleotides upstream of the MeCP2_e2 translation start site [16]. The 2 bp deletion affecting the donor splice site of intron 1 (c.62+2_62+3del) is very likely pathogenic. Even though we could not provide additional evidence supporting that it generates an alternate transcript with a deleterious effect on the production of the MeCP2_e1 isoform as RNA was not available for this patient. Amir et al. [1] reported another deletion affecting the donor splice site of intron 1 of *MECP2* (c.62+1_62 +2del) in a classic RTT patient. By RT-PCR, Amir and collaborators failed to detect abnormally spliced transcripts that are probably subject to the nonsense mediated mRNA decay and therefore highly unstable. For the patient with the 8 bp duplication (c.48_55dup), it would be of interest to test the impact of this duplication on translational efficiency of the *MECP2_e2* transcript by immunocytochemical analysis on uncultured lymphocytes.

For the four large rearrangements involving exon 1 of *MECP2* detected by QMPSF, we can assume that these gross rearrangements result in the absence of synthesis of both protein isoforms from the mutated allele.

The low prevalence of deleterious mutation in exon 1 may be due to the fact that the exon 1 coding sequence (62 nucleotides) is quite short as compared to that of exons 3 and 4 (1435 nucleotides). Less likely, the low prevalence of deleterious mutations in exon 1 of *MECP2* could be due to the more severe phenotype, compatible with survival only if the pattern of X-inactivation is skewed at least in brain. We found a non random pattern of inactivation in one out of six deleterious mutations affecting exon 1. However the studies were performed on DNA extracted from leucocytes and the results might be different in the brain where a skewing of X-inactivation could modulate the severity of the phenotype.

We also detected the c.16_21dup sequence variation in a family case with epilepsy and intellectual disability, leading to the insertion of an extra two alanine residues in the polyalanine sequence of MeCP2_e1. This variant was already reported by Evans et al. [7] in a girl with a classical RTT and her unaffected mother and described as a 6 bp insertion (c.3_4insGCCGCC). In fact c.3_4insGCCGCC and c.16_21GCCGCC correspond to the same sequence variation (see the nomenclature for the description of sequence variations, http://www.hgvs.org/mutnomen/). Therefore, we can conclude that this sequence alteration is most likely a polymorphism because it did not cosegregate with the clinical phenotype in two independent family studies.

From our study, we conclude that mutations within exon 1 of *MECP2* are indeed not a frequent cause of RTT. However, we found a deleterious mutation in exon 1 or a large rearrangement encompassing exon 1 of *MECP2* in around 1% and 2.2% of classical RTT, respectively. Our mutation detection rate in exon 1 of MECP2 is in agreement with the data in the literature based on the screening of 209 RTT patients [16]. We confirm the importance of mutation screening of exon 1 for both alterations and large rearrangements in classical RTT.

Acknowledgements

This study was supported by grant from GIS-Institut des maladies rares. We wish to thank T. Frébourg and M. Tosi for help in implementing the QMPSF protocol for the detection of large rearrangements in the MECP2 gene.

References

[1] R.E. Amir, P. Fang, Z. Yu, D.G. Glaze, A.K. Percy, H.Y. Zoghbi, B.B. Roa, I.B. Van den Veyver, Mutations in exon 1 of MECP2 are a rare cause of Rett syndrome, J. Med. Genet. 42 (2005) e15.

[2] A. Arzimanoglou, P. Castelnau, Signes cliniques et diagnostiques, in: Rett AFdSd (ed): Le syndrome de Rett, une maladie génétique, 2004, pp. 12–25.

[3] V. Bourdon, C. Philippe, O. Labrune, D. Amsallem, C. Arnould, P. Jonveaux, A detailed analysis of the MECP2 gene: prevalence of recurrent mutations and gross DNA rearrangements in Rett syndrome patients, Hum. Genet. 108 (2001) 43–50.

[4] V. Bourdon, C. Philippe, D. Martin, A. Verloes, A. Grandemenge, P. Jonveaux, MECP2 mutations or polymorphisms in mentally retarded boys: diagnostic implications, Mol. Diagn. 7 (2003) 3–7.

[5] F. Casilli, Z.C. Di Rocco, S. Gad, I. Tournier, D. Stoppa-Lyonnet, T. Frebourg, M. Tosi, Rapid detection of novel BRCA1 rearrangements in high-risk breast–ovarian cancer families using multiplex PCR of short fluorescent fragments, Hum. Mutat. 20 (2002) 218–226.

[6] M. D'Esposito, N.A. Quaderi, A. Ciccodicola, P. Bruni, T. Esposito, M. D'Urso, S.D. Brown, Isolation, physical mapping, and northern analysis of the X-linked human gene encoding methyl CpG-binding protein, MECP2, Mamm. Genome 7 (1996) 533–535.
[7] J.C. Evans, H.L. Archer, S.D. Whatley, A. Kerr, A. Clarke, R. Butler, Variation in exon 1 coding region and promoter of MECP2 in Rett syndrome and controls, Eur. J. Hum. Genet. 13 (2005) 124–126.
[8] B. Hagberg, F. Hanefeld, A. Percy, O. Skjeldal, An update on clinically applicable diagnostic criteria in Rett syndrome. Comments to Rett Syndrome Clinical Criteria Consensus Panel Satellite to European Paediatric Neurology Society Meeting, Baden Baden, Germany, 11 September 2001, Eur. J. Paediatr. Neurol. 6 (2002) 293–297.
[9] M. Krawczak, D.N. Cooper, Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment, Hum. Genet. 86 (1991) 425–441.
[10] S. Kriaucionis, A. Bird, The major form of MeCP2 has a novel N-terminus generated by alternative splicing, Nucleic Acids Res. 32 (2004) 1818–1823.
[11] G. Miltenberger-Miltenyi, F. Laccone, Mutations and polymorphisms in the human methyl CpG-binding protein MECP2, Hum. Mutat. 22 (2003) 107–115.
[12] G.N. Mnatzakanian, H. Lohi, I. Munteanu, S.E. Alfred, T. Yamada, P.J. MacLeod, J.R. Jones, S.W. Scherer, N. C. Schanen, M.J. Friez, J.B. Vincent, B.A. Minassian, A previously unidentified MECP2 open reading frame defines a new protein isoform relevant to Rett syndrome, Nat. Genet. 36 (2004) 339–341.
[13] C. Philippe, L. Villard, N. De Roux, M. Raynaud, J.P. Bonnefont, L. Pasquier, G. Lesca, J. Mancini, P. Jonveaux, A. Moncla, J. Chelly, T. Bienvenu, Spectrum and distribution of MECP2 mutations in 424 Rett syndrome patients: a molecular update, Eur. J. Med. Genet. (in press).
[14] K. Ravn, J.B. Nielsen, M. Schwartz, Mutations found within exon 1 of MECP2 in Danish patients with Rett syndrome, Clin. Genet. 67 (2005) 532–533.
[15] K. Reichwald, J. Thiesen, T. Wiehe, J. Weitzel, W.A. Poustka, A. Rosenthal, M. Platzer, W.H. Stratling, P. Kioschis, Comparative sequence analysis of the MECP2-locus in human and mouse reveals new transcribed regions, Mamm. Genome 11 (2000) 182–190.
[16] A. Saxena, D. de Lagarde, H. Leonard, S. Williamson, V. Vasudevan, J. Christodoulou, E. Thompson, P. Macleod, D. Ravine, Lost in translation: translational interference from a recurrent mutation in exon 1 of MECP2, J. Med. Genet. (2005).
[17] L.S. Weaving, C.J. Ellaway, J. Gecz, J. Christodoulou, Rett syndrome: clinical review and genetic update, J. Med. Genet. 42 (2005) 1–7.

Liste des posters avec comité de lecture

- **Yilmaz S, Fontaine H, Brochet K, M. Grégoire, Devignes MD, Schaff JL, Philippe C, Nemos C, McGregor JL, Jonveaux P.**
  Screening of subtle copy number changes in Aicardi Syndrome Patients with a high resolution X-chromosome array-CGH. European Human genetics conference 2007, June 16-19 2007, Nice, France. Eur J Hum Genet, 15, Suppl 1, June 2007, P0290, p99

- **Yilmaz S, Bodelot S, Lambermont S, Fontaine H, Rousselin A, Philippe C, Smail-Tabbone M, Devignes MD, MacGregor J, Jonveaux P.**
  Approches génétiques du syndrome d'Aicardi. Assises de génétique humaine et médicale (Montpellier, 26-28 january 2006). Médecine Sciences, 2006, 22, 124.

Liste des autres posters

- **Yilmaz S, Bodelot S, Lambermont S, Fontaine H, Rousselin A, Philippe C, Smail-Tabbone M, Devignes MD, MacGregor JL, Jonveaux P.**
  Approches génétiques du syndrome d'Aicardi. 3rd international meeting on « medical engineering and therapy » (Nancy, 15-16 may 2006)

- **Yilmaz S, Rousselin A, Philippe C, Devignes MD, MacGregor J, Jonveaux P.** Génétique du syndrome d'Aicardi. Gene signature symposium 2005 Applied Biosystem, Paris,27 april 2005 (Maison de la chimie).

- **Yilmaz S, Rousselin A, Philippe C, Devignes MD, MacGregor J, Jonveaux P.** Génétique du syndrome d'Aicardi. Journée de recherche clinique du CHU de Nancy, 11 January 2005

- **<u>Yilmaz S</u>, Rousselin A, Philippe C, Devignes MD, MacGregor J, Jonveaux P.** Genetics of Aicardi syndrome. European conference on rare diseases Luxembourg, 2005, 21-22 June 2005

Références Internet

Références Internet

Nom	Adresse	Page
Geneimprint	http ://www.geneimprint.com	21
EuroMRX	http ://www.euromrx.com/	23
XLMR Genes Update	http ://xlmr.interfree.it/home.htm	25
Greenwood Genetic Center	http ://www.ggc.org/xlmr.htm	25
Prévalence maladies rares (Orphanet)	http ://www.orpha.net/orphacom/cahiers/docs/FR/Prevalence_des_maladies_rares.pdf	35
National Center for Biotechnology Information - NCBI	http ://www.ncbi.nlm.nih.gov/SNP/	45
Genome Browser	http ://genome.ucsc.edu/	54
Database of Genomic Variants	http ://projects.tcag.ca/variation/	54
Human Structural Variation Database	http://humanparalogy.gs.washington.edu/structuralvariation/	54
Wellcome Trust Sanger Institute	www.sanger.ac.uk/PostGenomics/decipher/	58
l'European Cytogenetics Association	www.ecaruca.net	58
Cyber-T	http://cybert.ics.uci.edu/	64
NIH working definition of bioinformatics and computational biology	http ://www.bisti.nih.gov/CompuBioDef.pdf	68
Entrez Gene	www.ncbi.nlm.nih.gov/entrez/query.fcgi ?db=gene	73
Aperçu de la page Entrez GENE concernant le gène *SUV39H1*	www.ncbi.nlm.nih.gov/sites/entrez ?Db=gene&Cmd=ShowDetailView&TermToSearch=6839&ordinalpos=1&itool=EntrezSystem2.Pentrez.Gene.Gene_ResultsPanel.Gene_RVDocSum	74
Aperçu d'un rapport détaillé MGD concernant le gène Maged1	http://www.informatics.jax.org/javawi2/servlet/WIFetch?page=markerDetail&key=65381	76
FlyBase	www.flybase.org	76
Aperçu d'un rapport détaillé Flybase concernant le gène kismet	www.flybase.org/reports/FBgn0001309.html	77
HomoloGene	www.ncbi.nlm.nih.gov/sites/entrez/query.fcgi ?db=homologene	78
Medline	www.nlm.nih.gov/cgi/mesh/2007/MB_cgi http ://www.ncbi.nlm.nih.gov/sites/entrez ?db=mesh	82
Gene Ontology - GO	www.geneontology.org	83
UniProt	http ://www.expasy.uniprot.org	85
Protein Data Bank - PDB	www.rcsb.org/pdb	86
Portail Entrez du NCBI	http://www.ncbi.nlm.nih.gov/sites/gquery	87
GeneOntology	www.geneontology.org	85
Genome Browser	http ://genome.ucsc.edu	107
ExonPrimer	http ://ihg.gsf.de/ihg/ExonPrimer.html	107
The Institute for Genomic Research - TIGR	http ://compbio.dfci.harvard.edu/tgi/cgi-bin/magic/r1.pl	121
L'association AAL- syndrome d'Aicardi	www.aicardi.info	123

Orphanet	www.orphanet.net	123
Database of Genomic Variants	http ://projects.tcag.ca/variation/	129
Thrombosis Research Institute	http ://www.tri-london.ac.uk/	136
ArrayExpress	http ://www.ebi.ac.uk/microarray-as/aer/#ae-main[0]	136
National Center for Biotechnology Information - NCBI	http ://www.ncbi.nlm.nih.gov/	142
Xcollect	http ://www.nettab.org/2005/docs/NETTAB2005_DevignesOral.pdf	176

Nancy-Université

Université Henri Poincaré

Nom : **Mademoiselle YILMAZ**

Prénom : **Saliha**

DOCTORAT DE L'UNIVERSITÉ HENRI POINCARÉ, NANCY 1

en **BIOLOGIE SANTÉ ENVIRONNEMENT**

Spécialité : ***GÉNOMIQUE***

VU, APPROUVÉ ET PERMIS D'IMPRIMER N°

Nancy, le 22 janvier 2008

Le Président de l'Université

Jean-Pierre FINANCE

UNIVERSITÉ HENRI POINCARÉ, NANCY 1 ★ LE PRESIDENT

24 - 30 rue Lionnois | BP 60120
54003 Nancy Cedex | FRANCE
Tél. : +33(0)3 83 68 20 00
Fax : +33(0)3 83 68 21 00

info@uhp-nancy.fr | www.uhp-nancy.fr

Résumé en anglais

Candidate gene retrieval for Aicardi Syndrome: Complementarities of the experimental and bioinformatics approaches

Aicardi syndrome (AIC) is a severe X-linked dominant neurodevelopmental disorder affecting almost exclusively females. Chief features include infantile spasms, corpus callosal agenesis, and chorioretinal abnormalities. Aicardi syndrome is a sporadic disorder and hypothesized to be caused by heterozygous mutations in an X linked-gene but up to now no defined candidate region on the X chromosome has been identified. Positional candidate gene approach is not possible because no familial case were reported. Eighteen AIC patients were analyzed with a full-coverage X-chromosomal BAC arrays. No disease-associated Copy Number Variant was identified and we excluded total deletion and duplication of *FLNA* gene wich had been previously pointed out as a functional candidate. To complete this approach, 2 microarrays studies were performed to compare gene expression between AIC patients and a pool of healthy patients. The first study, on RNA extracted from lymphoblastoid cell lines isolated between 3 AIC patients used 22k oligonucleotide microarray. For the screened patients, no deleterious mutations were found in the 6 selected candidate genes (*ASMT*, *PLXNB3*, *MST4*, *SYN1*, *SSR4*, and *NSBP1*). The second study was performed with 44k microarray, on RNA directly extracted from 10 AIC patients blood samples. Functional clustering analyses revealed the effects of the factors: age, time of blood sample extraction, and inter-individual gene expression variance. A group of gene annotated by “nucleosome” GO term seemed influenced by the factor “use of antiepileptic drugs”. In a last strategy, we proposed a knowledge-guided approach for retrieving disease-specific candidate genes named ACGR (Approach for Candidate Gene Retrieval). Knowledge embedded in expert’s definitions of candidate gene was expressed as relations between genes and the disease. These definitions were used for guiding data modelling and are converted into views on the data which ultimately led to retrieval of sets of candidate genes. Thus *PLXNB3*, *MADEG1* and *SUV39H3* were selected as candidate genes. The perspectives of our work will include sequencing analysis of these genes. These integrative approaches reflect the evolution of our concepts and allow, with the use of biological pathways, the transition between the genetics of mental retardation to the genomics of mental retardation.

Keywords: Aicardi syndrome, mental retardation, X chromosome, microarrays, bioinformatics

Le syndrome d'Aicardi (AIC) est caractérisé par la triade agénésie du corps calleux, spasmes infantiles et lacunes choriorétiniennes. Cette triade s'accompagne d'un retard mental souvent sévère. Le syndrome survient chez les filles de façon sporadique, selon un mode d'hérédité dominant lié au chromosome X. Une approche de clonage positionnel n'est donc pas possible puisque aucun cas de transmission familiale n'a été répertorié à ce jour. Une puce génomique spécifique de l'X (résolution théorique de 82 kb) a été utilisée pour cribler le génome de 18 filles AIC à la recherche de variations quantitatives délétères. Aucun variant en nombre de copie (CNV) n'a été impliqué dans la pathologie et nous avons exclu chez les 18 patientes de notre étude les grands réarrangements touchant la totalité du gène *FLNA,* gène évoqué antérieurement comme candidat fonctionnel. Nous avons alors complété cette stratégie par deux études transcriptomiques. Cette approche vise à sélectionner les gènes dont l'expression diffère entre les filles AIC et des témoins. Initialement à partir d'ARN de 3 lignées cellulaires et d'une puce 22 000 clones (22K) nous avons exclu, *a priori*, par séquençage 5 gènes candidats : *ASMT*, *MST4*, *NSBP1*, *PLXNB3* et *SYN1*. Une deuxième étape a été engagée sur des ARN de prélèvements sanguins de 10 couples fille-mère et une puce 44K afin d'enrichir les données et de pallier à l'influence des lignées cellulaires. Outre la sélection de gènes candidats impliqués dans le syndrome, cette approche est surtout vouée à l'identification des fonctions biologiques dérégulées chez les patientes Aicardi. Les groupements fonctionnels des gènes signatures chez les filles révèlent clairement les effets des facteurs âge, heure de prélèvement, variabilité inter-individuelle. Un groupe de gènes annotés par le terme GO « nucléosome » semble être influencé par le facteur « prise d'anti-épileptique ». Un logiciel baptisé ACGR (Approach for Candidate Gene Retrieval) a été conçu et prototypé. Le but est de cribler les bases de données biologiques en incluant des données privées (données des puces transcriptomiques) à la recherche des gènes qui, lorsqu'ils sont mutés donnent un phénotype de syndrome d'Aicardi. Par cette approche, les gènes *PLXNB3*, *MADEG1* et *SUV39H3* sont trois gènes candidats pour le Syndrome d'Aicardi. Le séquençage de ces trois gènes s'inscrit dans les perspectives à court terme. Ces approches intégratives reflètent l'évolution de nos concepts de recherche passant de la génétique du retard mental à la génomique du retard mental en tenant compte de la multiplicité des réseaux d'interactions et de régulations.

Mots-clés : Syndrome d'Aicardi, Retard mental, Chromosome X, Puces à ADN, Bioinfomatique