Analyse statistique d'une base de translocations réciproques et modélisation du risque de survenue d'un enfant malformé

Christine Cans

THESE

présentée par

Christine CANS

pour obtenir le titre de **docteur**
de l'Université Joseph Fourier - Grenoble I
(arrêtés ministériels du 5 juillet 1984 et du 23 Novembre 1988)

spécialité Génie Biologique et Médical

Analyse statistique d'une base de translocations réciproques, et modélisation du risque de survenue d'un enfant malformé.

Date de soutenance : 22 Avril 1994

Composition du jury :

J. DEMONGEOT	Président
J. FEINGOLD	Rapporteurs
M. GILLOIS	
P. JALBERT	Examinateurs
C. ROBERT	
C. LAVERGNE	
O. COHEN	

Thèse préparée au sein des laboratoires **TIM3 et Laboratoire de Génétique**
à l'Université Joseph Fourier - Grenoble I -Domaine de la Merci - 38706 La Tronche Cedex

Je tiens à remercier,

Monsieur Jacques Demongeot, qui me fait l'honneur de présider le jury de cette thèse et qui m'a accueilli dans son laboratoire d'informatique et de statistique médicale. Il m'a orienté vers les statistiques appliquées à la cytogénétique et au conseil génétique, et je le remercie de m'avoir fait connaître ce sujet, certes difficile, mais passionnant.

Messieurs José Feingold et Michel Gillois, qui ont bien voulu accepter d'être rapporteurs de cette thèse. Leurs remarques pertinentes sont le témoin de l'intérêt qu'ils ont porté à ce travail, et je les en remercie.

Monsieur Pierre Jalbert, qui m'a toujours accueilli de façon sympathique dans son laboratoire universitaire. Il a été l'instigateur de ce travail et a su me faire profiter de sa grande expérience sur les translocations réciproques.

Madame Claudine Robert, qui a su m'éclairer de ses conseils avisés et m'assurer un bon encadrement pour cette thèse, toujours avec une grande gentillesse.

Monsieur Christian Lavergne, qui m'a dirigé dans ce travail avec une grande rigueur scientifique. Je le remercie pour sa disponibilité et sa gaité d'humeur qui ont contribué à la réussite de ce travail.

Monsieur Olivier Cohen, qui m'a intégrée chaleureusement dans son équipe. Jour après jour il a su me faire partager sa passion pour le conseil génétique, en m'éclairant de son point de vue de clinicien.

Je tiens aussi à remercier :
- Marie-Ange Mermet pour son aide précieuse, sa convivialité et sa rigueur dans le travail;
- Hubert Roth pour son aide en informatique et sa grande disponibilité;
- Marie-Jo Bossan pour son efficacité à régler les problèmes logistiques quotidiens;
- et enfin Vincent Rialle et Bruno Crémilleux qui ont bien voulu m'initier à un de leurs thèmes de recherche.

Introduction

Les malformations congénitales et les maladies héréditaires sont devenues dans nos pays industrialisés une des premières causes de morbidité et de mortalité pendant l'enfance (avec 15 à 20 % de cet ensemble représenté par les anomalies chromosomiques). D'où la place de plus en plus importante du conseil génétique en matière de dépistage et de prévention de certaines anomalies congénitales. Lors du conseil génétique, la responsabilité du cytogénéticien est importante, puisque le couple qui consulte doit être clairement informé du risque d'anomalie foetale, du risque d'erreur de la méthode diagnostique et de l'éventualité d'une interruption de grossesse. Pour assurer ce conseil génétique auprès des familles, les généticiens peuvent faire appel à des connaissances déjà acquises (risque de récurrence bien connu pour certaines anomalies) et / ou à des méthodes plus ou moins invasives comme le diagnostic prénatal.

Grâce aux progrès à la fois en obstétrique (techniques de prélèvement) et en biologie (techniques de dosage et techniques cytogénétiques), des améliorations considérables ont été réalisées depuis 30 ans dans le diagnostic prénatal de certaines affections. Les premières amniocentèses ont été réalisées en 1960, et ce n'est que vers 1967-1968 que les premiers diagnostics prénatals avec critères biologiques ont été réalisés, comme par exemple le diagnostic des aberrations chromosomiques (Boué, 1989). Parmi les différentes techniques utilisées, le diagnostic prénatal chromosomique représente 90 % du diagnostic anténatal biologique.

Si tout le monde concorde sur l'intérêt du diagnostic prénatal (dans le cas du dépistage de la trisomie par exemple), les positions sont plus nuancées quant aux politiques de Santé Publique à adopter face à ces méthodes préventives. De même, sur un plan éthique, l'information qui peut-être connue en diagnostic prénatal n'est pas forcément 'bonne' à connaître dans tous les cas. Par exemple, la connaissance du sexe de l'enfant peut avoir des effets très positifs en cas de transmission de maladie liée au sexe, ou au contraire des effets très pervers lorsqu'il est finalisé par une modification du sex-ratio.

Les translocations réciproques font partie des anomalies pour lesquelles une prévention est parfois possible grâce au diagnostic prénatal. Celui-ci permet de prévoir la survenue d'un enfant sévèrement malformé impliquant une prise en charge lourde et coûteuse. Disposant à Grenoble d'un recensement très important de translocations (base de données la plus volumineuse au monde à notre connaissance), le but de ce travail est de fournir une information quant au risque de survenue d'un enfant malade, afin que familles et soignants puissent prendre la meilleure décision en terme de prévention.

Sommaire

Chapitre I.

Etat de la question

Le sujet de la recherche est présenté dans un premier temps, puis les moyens d'acquisition des données dans la base sont explicités. Un dernier sous-chapitre présente les travaux déjà réalisés sur le sujet par le laboratoire de Cytogénétique de Grenoble.

I. 1. Le sujet

En cytogénétique constitutionnelle, il existe plusieurs types d'anomalies chromosomiques. De façon simplifiée, on peut distinguer :

• les anomalies de nombre acquises (accidents dans la mécanique chromosomique lors de la méiose - exemple la trisomie 21)

• les anomalies de structure (avec cassures chromosomiques suivies de recollements anormaux déjà présentes chez le géniteur - exemple inversion péricentrique du 9). Parmi ces anomalies de structure, les translocations réciproques sont les plus fréquentes, devant les translocations robertsoniennes.

Nous exposerons d'abord ce qu'est une translocation réciproque ainsi que les données épidémiologiques de cette anomalie (un rappel sur les caractéristiques cytogénétiques des chromosomes et sur la division méiotique figure en annexe 1). Puis nous décrirons les études de la littérature portant sur les translocations réciproques ainsi que le conseil génétique actuellement proposé en cas de translocations. Enfin, les objectifs de la recherche seront détaillés, en insistant sur les différents axes d'utilisations possibles de ce travail.

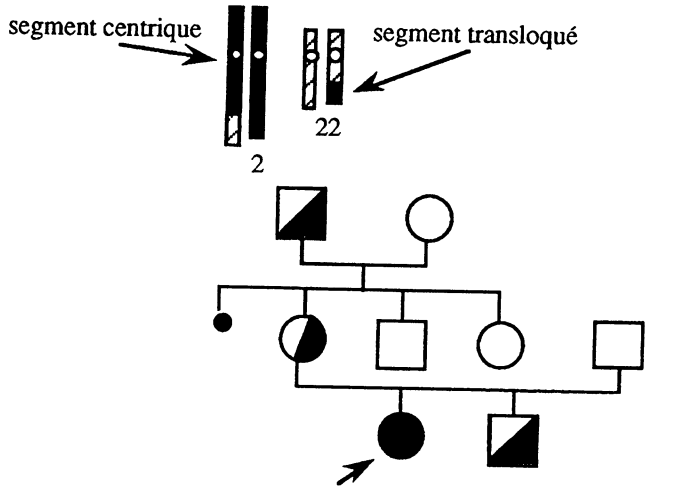
I. 1.1. Les translocations réciproques

I. 1.1.1. Définition

La translocation réciproque est un échange de segments chromosomiques entre deux chromosomes non homologues (c'est-à-dire n'appartenant pas à la même paire). Sur la figure 1 il y a eu cassure au niveau du bras long du chromome 2 et cassure au niveau du bras long du chromosome 22, l'extrémité du chromosome 2 venant se positionner à la place de celle du chromosome 22 et vice-versa. Il est important de remarquer que le matériel chromosomique est en quantité normale, même s'il y a eu échange de segments chromosomiques : les deux parties du chromosome 2 (segments centrique et transloqué) constituent à elles deux le matériel chromosomique d'un chromosome 2 entier et intact, et même chose pour le chromosome 22.

Cet arbre généalogique illustre le fait qu'un individu porteur d'une translocation peut donner naissance à des enfants normaux, à des enfants porteurs de la translocation, ou à des enfants "malades".

Fig. 1. Modèle d'un arbre généalogique



t(2;22)(q14.1;q11.1)

Légende

- parent porteur de la translocation (mère)
- enfant "atteint de la maladie", fille
- individu normal, père
- fausse couche spontanée

Les points de cassure sont théoriquement situés en n'importe quelle partie du chromosome, et sur n'importe lequel des 23 chromosomes. Il en résulte un nombre très important de différentes translocations possibles. De Arce souligne qu'avec un niveau de résolution* de 293 bandes, il existerait théoriquement 40 000 translocations différentes (De Arce, 1986). Dans la réalité, on en observe beaucoup moins (de l'ordre de 3 à 5 %). Les raisons en sont diverses : d'une part, certaines translocations n'ont pas encore été découvertes (en raison d'une expression clinique phénotypique très peu fréquente); d'autre part, certains points de cassure ou chromosomes ne sont quasiment jamais impliqués, car ils concernent une région probablement vitale du génome.

En ce qui concerne l'écriture d'une translocation réciproque, nous rappelons qu'elle comporte toujours le nombre de chromosomes de l'individu porteur, le sexe de cet individu, les 2 chromosomes impliqués dans la translocation, le bras et la bande du point de cassure du premier chromosome, et enfin le bras et la bande du point de cassure du deuxième chromosome. Une convention d'usage se traduit par le fait que le premier chromosome est toujours celui de plus grande taille et le deuxième celui de plus petite taille :

46 XY (2; 22)(q14.1; q11.1)

---

* Le niveau de résolution correspond au nombre de bandes visualisées sur les chromosomes pré-métaphasiques par des techniques de coloration. Avec des techniques de haute résolution on peut atteindre un niveau de 850 bandes.

I. 1.1.2. La méiose dans les translocations

La méiose est définie par les étapes successives qui conduisent les cellules à passer de l'état diploïde (2n chromosomes) à l'état haploïde (n chromosomes). Elle aboutit à la formation des gamètes. Sur les schémas ci-dessous seront montrés : 1) une méiose normale impliquant deux paires de chromosomes, 2) une méiose avec remaniement conduisant à la formation de gamètes remaniés ou normaux, 3) une méiose impliquant deux paires de chromosomes remaniés.

Le schéma 1 montre les 16 combinaisons possibles de chromatides, ces 16 combinaisons se réduisant en fait à seulement 4 combinaisons différentes.

Le schéma 2 ne montre pas un crossing over mais un remaniement entre deux segments chromsomiques non homologues. Il permet de visualiser la formation de gamètes porteurs d'un remaniement chromosomique (ou survenue d'une translocation "de novo"), ces gamètes pouvant être équilibrés ou non en matériel chromosomique.

Schéma 1. Méiose relative à deux paires de chromosomes homologues (chromosomes n°1 et n°2)

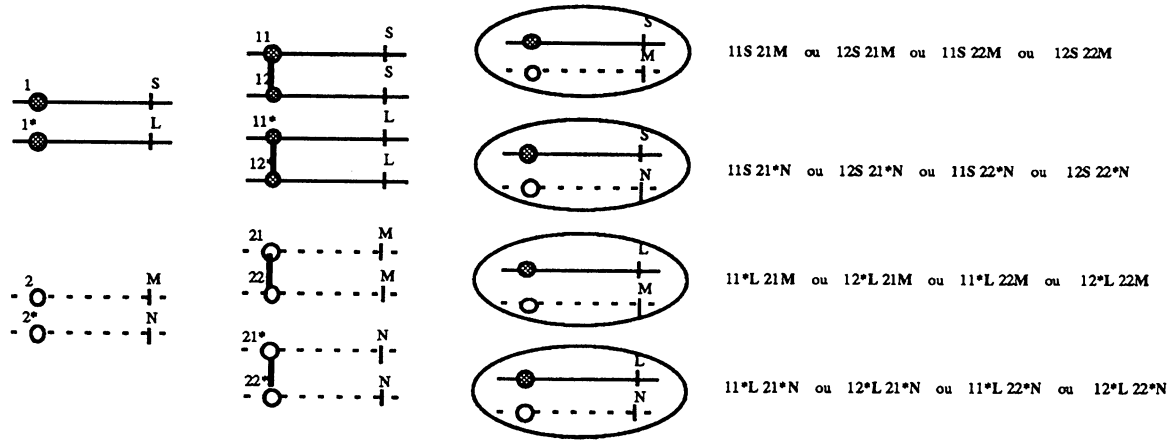


Schéma 2. Même méiose avec remaniement chromosomique entre ces deux paires de chromosomes

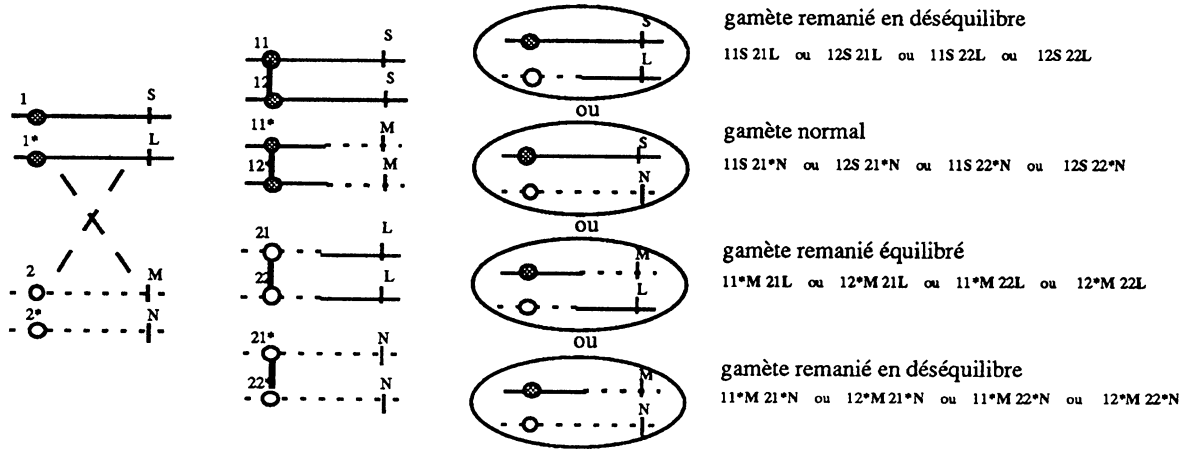
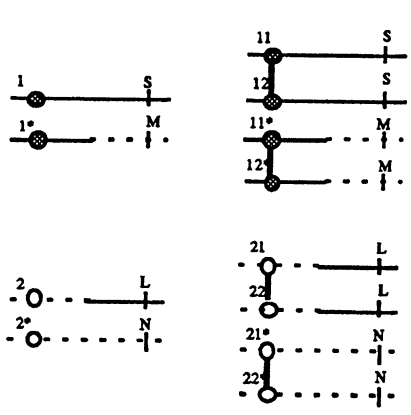
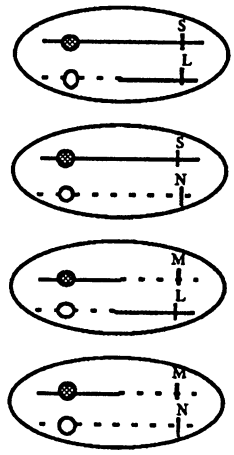


Schéma 3. Méiose relative à deux paires de chromosomes homologues remaniés (n°1 et n°2) t(1;2) sans crossing over





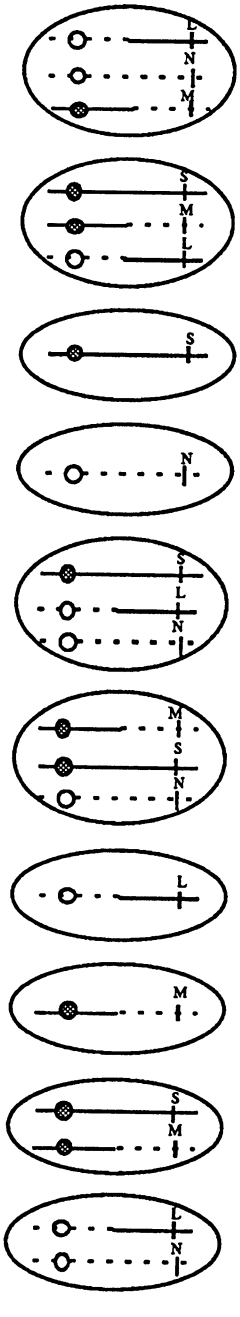
gamète remanié en déséquilibre adjacent1, der2
11S 21L ou 12S 21L ou 11S 22L ou 12S 22L

gamète normal
11S 21°N ou 12S 21°N ou 11S 22°N ou 12S 22°N

gamète remanié équilibré
11°M 21L ou 12°M 21L ou 11°M 22L ou 12°M 22L

gamète remanié en déséquilibre adjacent1, der1
11°M 21°N ou 12°M 21°N ou 11°M 22°N ou 12°M 22°N

En cas de non disjonction méiotique (entre 1 et 1*, ou entre 2 et 2*, c'est à dire entre chromosomes homologues et non pas entre centromères frères), on peut voir apparaître les gamètes suivants



3:1 échange, 24 Xmes, trisomie du 2
22L 21°N 12°M ou 21L 21°N 12°M ou 22L 21°N 11°M ou 21L 21°N 11°M
22L 22°N 12°M ou 21L 22°N 12°M ou 22L 22°N 11°M ou 21L 22°N 11°M

3:1 échange, 24 Xmes, trisomie du 1
11S 12°M 21L ou 12S 12°M 21L ou 11S 11°M 21L ou 12S 11°M 21L ou
11S 12°M 22L ou 12S 12°M 22L ou 11S 11°M 22L ou 12S 11°M 22L

3:1 échange, 22 Xmes, monosomie du 2
11S ou 12S

3:1 échange, 22 Xmes, monosomie du 1
21°N ou 22°N

3:1 tertiaire, 24 Xmes, trisomie composite du 2 et du 1
11S 22L 22°N ou 12S 22L 22°N ou 11S 21L 22°N ou 12S 21L 22°N ou
11S 22L 21°N ou 12S 22L 21°N ou 11S 21L 21°N ou 12S 21L 21°N

3:1 tertiaire, 24 Xmes, trisomie composite du 1 et du 2
11°M 12S 21°N ou 12°M 12S 21°N ou 11°M 12S 22°N ou 12°M 12S 22°N ou
11°M 11S 21°N ou 12°M 11S 21°N ou 11°M 11S 22°N ou 12°M 11S 22°N

3:1 tertiaire, 22 Xmes, monosomie composite du 2 et du 1
21L ou 22L

3:1 tertiaire, 22 Xmes, monosomie composite du 1 et du 2
12°M ou 11°M

gamète remanié en déséquilibre adjacent2, der1
11S 11°M ou 12S 11°M ou 11S 12°M ou 12S 12°M

gamète remanié en déséquilibre adjacent2, der2
21L 22°N ou 22L 22°N ou 21L 21°N ou 22L 21°N

Le schéma 3 montre, lorsque le déroulement de la méiose est "normal", les 16 combinaisons possibles de chromatides, se réduisant à 4 combinaisons d'allèles. Lorsque le déroulement de la méiose n'est pas "normal" on retrouve 48 combinaisons possibles de chromatides, se réduisant à 10 combinaisons d'allèles. Dans ces derniers cas il se produit une non disjonction entre chromosomes homologues alors que la disjonction entre chromatides et centromères s'effectue normalement. Le mauvais appariement entre chromosomes homologues, du fait du remaniement des segments chromosomiques, serait responsable de cette non disjonction. En annexe 2 figure ce même schéma de méiose impliquant deux paires de chromosomes remaniés avec survenue d'un crossing over entre deux chromosomes homologues.

I. 1.1.3. Statut chromosomique

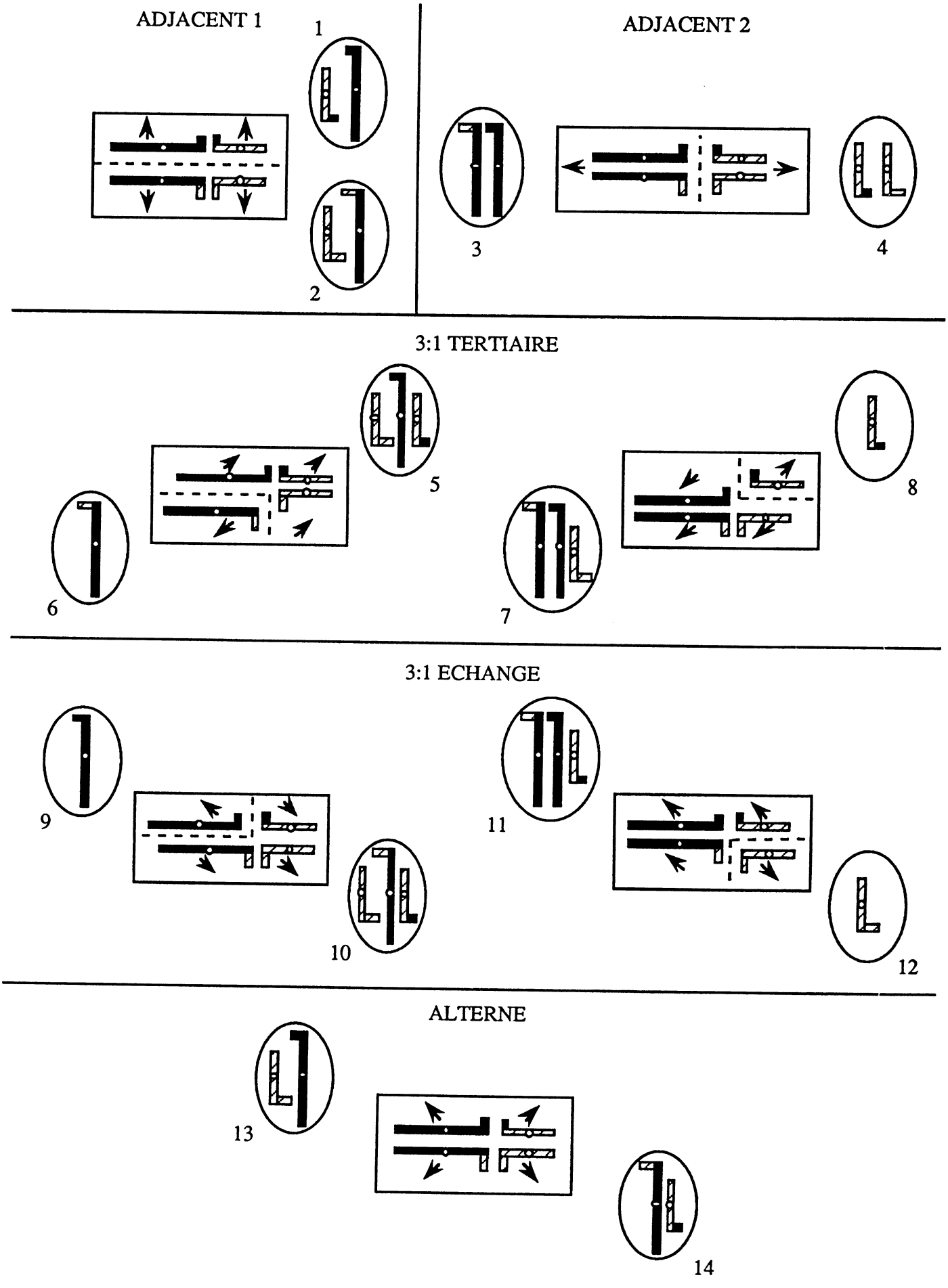
Un individu peut donc être porteur d'une translocation à l'état équilibré (sans perte ni gain de matériel chromosomique) ou à l'état déséquilibré. L'existence d'un déséquilibre est le reflet soit d'une méiose normale soit d'une méiose particulière. Les gamètes déséquilibrés présentent des segments de chromosomes en plus ou en moins par rapport au génome haploïde neutre*(ou autosomal*). La figure 2 illustre les différents modes de ségrégation possibles d'une translocation réciproque au cours de la méiose : les 12 premiers modes, pouvant être regroupés en 4 principaux, conduisent à un individu malformé ou à un avortement spontané, et les 2 autres modes, dits "modes de ségrégation alterne" conduisent à des individus "porteurs sains" ou normaux. Ces derniers (les individus normaux) ne présentent plus trace de la translocation réciproque.

Les 4 principaux modes de ségrégation donnant lieu à un déséquilibre chromosomique sont les modes adjacent 1, adjacent 2, 3:1 tertiaire et 3:1 échange. Seul le premier mode, adjacent 1, ne comporte pas de non disjonction méiotique. Au niveau gamétique on retrouve ces mêmes 4 principaux modes de déséquilibre en sachant qu'un mode adjacent 1 peut être le reflet parfois d'une ségrégation méiotique alterne avec crossing over (annexe 2). La quantité de matériel chromosomique en déséquilibre varie selon les chromosomes impliqués et selon le mode de ségrégation méiotique. Pour le mode adjacent 1, le déséquilibre porte sur les segments transloqués, alors que pour le mode adjacent 2 il porte sur les segments centriques. Pour le mode 3:1 tertiaire, le déséquilibre porte sur les segments centriques et transloqués, et pour le mode 3:1 échange c'est un chromosome entier qui est concerné par le déséquilibre (trisomie ou monosomie complète d'un des deux chromosomes). De manière générale on peut remarquer que la longueur des segments en déséquilibre est plus grande en cas de non disjonction méiotique que pour les adjacent 1 (mais ceci n'est pas toujours vrai : par exemple si les segments centriques sont plus petits que les segments transloqués, le mode de ségrégation adjacent 2 peut conduire à un déséquilibre moins important que le mode de ségrégation adjacent 1).

---

* Génome Haploïde Neutre (G.H.N) : ensemble du génome (autosomes plus le X)
Génome Haploïde Autosomal (G.H.A) : ensemble du génome non sexué (autosomes seulement)

Fig. 2. Les différents modes de ségrégation méiotique



I. 1.1.4. Expression phénotypique

Selon le mode de ségrégation, il existe différentes formes de descendance pour un individu porteur d'une translocation réciproque, avec des conséquences cliniques particulières à chacune.

• **mode de ségrégation alterne**

- enfant porteur sain de la translocation : il est phénotypiquement normal, mais peut être soumis à des risques plus élevés de pathologie ultérieure selon certains auteurs (Fryns, 1986).

- enfant normal ne portant pas trace de la translocation.

• **autres modes de ségrégation**

- aucune descendance, car l'individu porteur de la translocation présente une stérilité (masculine surtout) et ceci s'observe avec une plus grande fréquence lorsque des chromosomes acrocentriques sont impliqués dans la translocation (Luciani, 1987).

- avortements à répétition : le plus souvent les avortements sont précoces mais ils peuvent être tardifs conduisant alors à une Mort Foetale In Utero (MFIU). Soit il n'y a jamais obtention d'une descendance "saine" (échec de la repoduction), soit il y a alternance d'enfants normaux ou porteurs sains et de fausses couches.

- enfant malformé à terme : dans ce cas, l'enfant est phénotypiquement anormal, sa durée de vie est très variable, de quelques minutes à plusieurs années. Bien qu'il existe une grande variété de syndromes cliniques en fonction des segments chromosomiques impliqués dans ce déséquilibre, le plus souvent il y a coexistence d'un syndrome polymalformatif et d'un retard mental sévère. La pathologie de ces enfants est suffisamment grave pour qu'ils soient considérés comme des enfants polyhandicapés et que ne se soit jamais encore posé le problème de leur propre reproduction.

I. 1.1.4. Epidémiologie

En première approximation, environ 1 couple sur 600 est porteur d'une translocation réciproque. En terme de prévalence* ou d'incidence, les résultats disponibles dans la littérature montrent des taux homogènes et stables. Ils peuvent différer selon qu'il s'agit de mesure en population générale, parmi les naissances vivantes, ou bien encore parmi les avortements spontanés. Selon les études toutes les translocations ou seulement les translocations à l'état équilibré sont prises en compte.

• **Une étude de population,** réalisée en Hollande par Jacobs (Jacobs, 1970), étude caryotypique sur 4833 adultes, a montré une prévalence de translocation réciproque équilibrée de 2,5 ‰ [1,1-3,9]* porteurs. Crandall a trouvé un taux de 4,3 ‰ [1,4-7,2] porteurs de translocations réciproques parmi 1887 couples ayant eu recours à une amniocentèse en raison d'un

---

* La prévalence est définie par le nombre de cas malades, à une période donnée, rapporté à la population moyenne de cette période. L'incidence est définie par le nombre de nouveaux cas malades, survenus durant une période donnée, rapporté à la population moyenne de cette période.

* [ ] représente l'intervalle de confiance à 95 %

âge maternel ≥ 34 ans (Crandall, 1980). Boué donne un taux global plus faible de 1,6 ‰ [0-4,1] se répartissant en 2/3 de translocations transmises et 1/3 de translocations *de novo* (chez des couples ayant eu recours à l'amniocentèse pour un âge maternel ≥ 38 ans) (Boué, 1989).

• **Parmi 55 000 naissances vivantes**, Jacobs aux USA a noté un taux de 1,9 ‰ [1,5-2,3] translocations équilibrées (Jacobs, 1977). Boué, lui, signale un taux de 0,4 ‰ [0,2-0,6] pour les translocations équilibrées et déséquilibrées parmi les naissances vivantes (Boué, 1989).

• **Dans une étude portant sur 1500 fausses couches précoces** (fausses couches caryotypées de moins de 12 semaines d'aménorrhée) Boué notait, dans 3,8 % [2,8-4,8] des cas, l'existence d'une translocation réciproque (Boué, 1975). Des taux similaires sont retrouvés dans des études portant sur des couples ayant eu 2 ou plus avortements spontanés dans leurs antécédents : 4,4% [2,6-6,2] pour Sachs en 1985 et 4,2 % [2,5-5,9] pour Fortuny en 1988 sur 500 couples étudiés (Sachs, 1985; Fortuny, 1988).

I. 1.2. Situation actuelle

I. 1.2.1. Données de la littérature.

• **Dès 1970**, à partir d'études visant à estimer la fréquence des translocations en population générale, Jacobs en Angleterre remarque que le risque de survenue d'un déséquilibre à terme (c'est à dire à la naissance) n'est pas le même pour chaque translocation réciproque et qu'il semble varier avec le mode de détection de celle-ci : cas index (ou proposant) "malade" ou "non malade" (Jacobs, 1970).

• **Daniel aux USA** s'intéresse à la viabilité des déséquilibres issus d'une translocation, et émet en 1979 l'hypothèse de l'existence de seuils quantitatifs de viabilité du déséquilibre chromosomique (ces seuils sont définis par une ligne joignant les 4 % de trisomie et les 2 % de monosomie, annexe 7) (Daniel, 1979).

• **En 1980**, Jalbert en France constate que, parmi les modes de ségrégation non alterne, il en existe un mode préférentiel souvent unique, spécifique à une translocation donnée. En s'appuyant sur la configuration géométrique du diagramme du pachytène, il établit une méthode prédictive des différents modes de ségrégation non alterne (incluant des critères à la fois quantitatifs et qualitatifs chromosomiques) (Jalbert, 1980).

• **Dès 1975**, des systèmes non informatisés de recueil d'information sur les translocations réciproques se mettent en place à la fois au niveau des laboratoires de cytogénétique et des centres de diagnostic prénatal (Français et Européens).

• **En 1984**, en France, Boué observe, comme Jalbert, que les numéros des chromosomes impliqués dans la translocation jouent un rôle important dans la survenue d'un déséquilibre à terme, tout comme le rôle joué par la taille des segments chromosomiques (Boué, 1984).

• **Le nombre d'informations** sur les translocations réciproques augmentant, il devient possible pour certaines translocations (les plus fréquentes) d'estimer, de manière purement descriptive à partir des données observées, un risque de survenue d'un enfant déséquilibré à terme. C'est le travail réalisé en 1988 par Stengel-Rutkowski et Stène en Allemagne qui insistent sur le danger de considérer le mode de détection comme un facteur prédictif de ce risque. Leurs arguments sont les suivants : 1) ce mode de détection paraît soumis à l'interprétation (parfois subjective) du médecin qui a été contacté pour la première fois, 2) la distribution de ce mode de détection dans un échantillon de familles avec translocation réciproque est variable et probablement le reflet de différences de pratique en conseil prénatal ou en recherche étiologique pour infertilité. Ils suggèrent plutôt l'intérêt prédictif de certaines variables caractéristiques du parent porteur de la translocation réciproque comme le sexe et l'âge (Stengel-Rutkowski, 1988).

Une prédiction purement descriptive de ce risque ne satisfait pas parfaitement les généticiens dans la mesure où, compte tenu du très grand nombre des translocations, les résultats ne concernent que celles les plus fréquemment observées, le nombre d'observations devant être suffisant pour que la prédiction soit fiable. D'où le souci du laboratoire de cytogénétique de Grenoble de chercher à obtenir des méthodes prédictives par modélisation afin de pouvoir disposer d'une estimation du risque quelle que soit la translocation réciproque présentée.

I. 1.2.2. Le conseil génétique

Etant donné le risque de donner naissance à un enfant malformé, actuellement, devant tout parent porteur d'une translocation réciproque, en France et dans beaucoup de pays, le recours à un diagnostic prénatal est systématique : quand le foetus est porteur d'une translocation à l'état déséquilibré, une interruption de grossesse pour motif thérapeutique est alors proposée au couple. Le choix de la méthode diagnostique (amniocentèse ou choriocentèse) est le plus souvent fonction des habitudes de l'équipe médicale en place. On connaît bien les risques de chacune des techniques du diagnostic prénatal, leurs avantages et leurs inconvénients, mais on connaît mal les risques de survenue d'un enfant en déséquilibre, cet élément ne peut donc être pris en compte dans le choix de la méthode. Les inconvénients du diagnostic prénatal résident essentiellement dans le fait de pouvoir induire un avortement "iatrogène", alors que le conceptus est porteur sain ou normal pour la translocation. Le risque d'avoir un enfant "malade" n'est donné aux familles que de façon souvent imprécise avec une grande fourchette de valeurs (1 à 40 % par exemple), excepté pour quelques translocations très fréquentes.

Concernant les différentes méthodes de diagnostic prénatal, chacune possède ses propres avantages et inconvénients.

• **La choriocentèse** ou prélèvement de villosités choriales est la méthode de diagnostic prénatal pouvant être réalisée le plus précocément par rapport à la date de début de grossesse. Ce prélèvement peut s'effectuer à partir de la 10^{e} semaine d'aménorrhée (le plus souvent entre la 10^{e} et la 12^{e} semaine, optimum à 11 semaines). Il comporte des résultats faux positifs dans un nombre de cas non négligeable (sensibilité = 99,9), (spécificité = 99,0), les erreurs pouvant être d'autant

plus fréquentes que les segments transloqués sont petits (Simoni, 1987). De plus cette méthode présente un risque d'induire une fausse couche d'environ 2 à 5 %. Ce risque peut varier d'une équipe à l'autre ou d'un pays à l'autre (Boué, 1989) ; en terme de risque additionnel (excès de perte foetale après avoir exclu le risque intrinsèque de fausse couche durant le premier trimestre) on obtient un chiffre de 1 à 3 % (Gardner, 1989).

• **La méthode de référence pour le diagnostic prénatal est en fait l'amniocentèse** ou prélèvement de liquide amniotique. Elle s'effectue à partir de la 16^{e} semaine d'aménorrhée seulement (le plus souvent entre la 16^{e} et la 19^{e} semaine) et nécessite le recours à une culture cellulaire ce qui demande un délai d'environ 3 semaines pour les résultats. Le résultat du caryotype n'est pas connu avant la 20^{e} semaine de grossesse au minimum, par contre les résultats sont plus fiables que ceux de la choriocentèse (sensibilité = 99,99 ; spécificité = 99,9 %)(Lippman, 1992). Aussi le risque de fausse couche induite est nettement plus faible, autour de 0,7 % (entre 0,5 et 1 %, l'avortement survenant dans un délai de 3 à 6 semaines suivant la ponction).

• Enfin il existe aussi une autre méthode de diagnostic prénatal qui est un peu plus tardive : **le prélèvement de sang foetal**. Il ne peut être réalisé qu'à partir de la 18^{e} semaine d'aménorrhée. Ses résultats sont variables avec 0,3 à 3 % d'échec de la ponction (Boué, 1989) et un risque de fausse couche induite de l'ordre de 1 % pour des grossesses sans signe de souffrance foetale chronique. Cette méthode est très rarement utilisée dans le cas des translocations réciproques car elle conduit à une éventuelle interruption de grossesse trop tardive. Ses principales indications en sont plutôt des signes d'appel cliniques ou échographiques.

Le tableau 1, emprunté à Gardner (1989), montre le nombre respectif de choriocentèses ou d'amniocentèses effectuées pour translocation réciproque et leurs résultats. Les choriocentèses sont moins souvent effectuées (1 fois sur 10 seulement), mais le taux d'anomalies dépistées est presque 2 fois plus élevé que pour les amniocentèses. Ceci s'explique par une proportion non négligeable d'interruptions spontanées de grossesse entre la 10^{e} et la 16^{e} semaine de grossesse. Le choix de réaliser plus souvent des amniocentèses est aussi dicté par la meilleure fiabilité de cet examen.

Tableau 1. Diagnostic prénatal dans les translocations réciproques
Etude collaborative - source Mikkelsen et Aymé 1986

	nbre d'examens	nbre de déséquilibres	% de déséquilibre
choriocentèse	75	17	23
amniocentèse	609	71	12

I. 1.3. Les objectifs

• *Un couple se sachant porteur d'une translocation réciproque* demande un avis aux généticiens quant au risque pour sa descendance. Cette demande auquel les généticiens doivent faire face peut s'effectuer dans des circonstances diverses.

1) Au cours d'une enquête familiale

Un des parents est porteur d'une translocation découverte de manière systématique et le couple cherche à obtenir les informations suivantes avant toute procréation :

- quel est le risque d'avoir un enfant "malade" ?
- quel est le risque d'avoir une fausse couche ?
- en cas de grossesse, quelle surveillance prénatale sera-t-il nécessaire d'effectuer ?

2) Au cours d'une consultation pour échecs de la reproduction

Au cours de cette consultation est découverte une translocation réciproque chez le père ou chez la mère. La demande formulée par le couple est alors la suivante :

- quel est le risque d'avoir un enfant "malade" si la prochaine grossesse se poursuit normalement ?
- faut-il envisager un diagnostic prénatal très précoce ?
- si le risque d'avoir un enfant "malade" est très élevé, serait-il justifié d'avoir recours à un don de gamètes pour procréer ?

3) Suite à la naissance d'un enfant polymalformé

Un couple ayant déjà eu un enfant "atteint" (translocation découverte lors d'un caryotype de cet enfant) désire connaître le risque pour les grossesses ultérieures : faut-il envisager un diagnostic prénatal et si oui lequel ?

4) Au cours d'un diagnostic prénatal effectué pour une raison indépendante

Selon l'état équilibré ou déséquilibré de cette translocation, la grossesse en cours est poursuivie ou interrompue. Ce couple se retrouve alors dans le cas des situations n°1 ou n°3 relativement aux interrogations pour les grossesses ultérieures. Cette population est très particulière dans la mesure où la raison qui a conduit à la pratique d'un diagnostic prénatal (et donc à la découverte de la translocation) est complètement indépendante des conséquences pathologiques des translocations réciproques. C'est un peu comme si ce groupe pouvait être rapproché d'une recherche systématique de translocations en population générale.

En résumé, quelle que soit la situation, les généticiens se trouvent face à deux questions auxquelles ils se doivent de répondre, mais pour lesquelles ils disposent de peu d'éléments objectifs (quelle méthode conseiller si l'on ignore le risque dû à cette translocation d'avoir un enfant "malade"?).

- **quel est le risque d'avoir un enfant "malade"?**
- **quel diagnostic prénatal recommander ?**

• *En pratique, les conséquences attendues pourraient être les suivantes :*

1) Si le risque de survenue d'un déséquilibre à terme est très faible (inférieur à 1 %) pour une translocation réciproque, aucune méthode particulière de diagnostic prénatal ne serait "théoriquement" recommandée. Cette grossesse produira très probablement, soit une fausse couche spontanée, soit la naissance d'un enfant normal. Le risque d'avortement induit par le diagnostic prénatal est important.

Le risque d'enfant malformé en population générale est de 2 à 3 % (Goujard, 1983), mais toutes les malformations ne conduisent pas à des enfants handicapés. On peut raisonnablement estimer à 1 % le risque maximum en population générale d'avoir un enfant porteur d'un handicap d'origine prénatale. Pour la trisomie 21, ce risque varie de 1 % chez les mères d'âge supérieur à 38 ans à 0,5 ‰ dans les autres cas.

Comme le risque d'avoir un enfant malformé à terme à cause d'une translocation réciproque n'est pas plus important que celui d'avoir un enfant malformé en population générale, il semblerait licite d'un point de vue Santé Publique de ne pas proposer de diagnostic prénatal particulier dans ce cas, mais juste une surveillance échographique classique. D'un point de vue individuel, il est clair que les choses ne sont pas aussi simples, et qu'entre le risque de moins de 1 % d'avoir un enfant malformé à terme et le risque de 1 % de faire une fausse couche, un couple peut préférer courir le deuxième risque (sauf s'il s'agit d'un couple ayant déjà eu de nombreuses fausses couches et pour lequel une grossesse ne s'arrêtant pas spontanément peut être considérée comme très précieuse).

2) Si le risque de survenue d'un déséquilibre est supérieur à 1 % et inférieur à 5 %, il est faible et l'amniocentèse sera préférée à la choriocentèse lors du diagnostic prénatal. Même si l'inconvénient de l'amniocentèse réside dans le fait d'être amené à réaliser tardivement une éventuelle interruption de grossesse, cette méthode a des résultats plus fiables et moins d'effets secondaires néfastes.

3) Si le risque de survenue d'un déséquilibre est supérieur à 5 %, c'est déjà un risque élevé. Comme le risque iatrogène de fausses couches dû à la choriocentèse lui est inférieur, cette méthode sera donc préférée pour le diagnostic prénatal, car elle a l'avantage de permettre une éventuelle interruption de grossesse beaucoup plus précoce que l'amniocentèse, même si ses résultats sont moins fiables. La question peut être posée de savoir si cette attitude est réellement pertinente, dans la mesure où le nombre de faux positifs n'est pas négligeable pour la choriocentèse et qu'il s'agit de couples avec fort risque de déséquilibre, qui ont pu déjà subir plusieurs interruptions de grossesse.

• *En ce qui concerne l'aide au diagnostic prénatal*, une amorce de réflexion peut être envisagée. Pour pouvoir y répondre complétement, il serait nécessaire de disposer des informations suivantes :

1) Pour chaque méthode de diagnostic prénatal :

- le risque de fausse couche iatrogène (indépendamment du risque de fausse couche dû à la translocation),
- le risque de faux négatifs,
- le risque de faux positifs,
- et le coût de l'examen (en terme économique et en terme de faisabilité par les laboratoires, ainsi qu'en terme de facteur de stress pendant la grossesse).

2) Pour chaque translocation :

- le risque de donner naissance à un enfant malformé,
- le risque de donner une fausse couche spontanée,
- et l'estimation du coût de la naissance d'un enfant malformé (en terme économique et en terme de souffrance humaine pour une famille) même si cela paraît extrêmement difficile à mesurer.

Enfin, il y a le couple concerné, avec sa sensibilité particulière, face au médecin généticien détenteur du savoir cité ci-dessus. En attendant que toutes ces informations puissent être disponibles, la connaissance du risque de survenue d'un enfant malformé pour chaque translocation pourrait influencer le diagnostic prénatal. Actuellement, il n'existe pas de politique de Santé Publique visant à réguler les méthodes de diagnostic prénatal pour la population ciblée des parents porteurs d'une translocation réciproque. On remarque une attitude de diagnostic prénatal systématique commune à tous les médecins, mais avec des méthodes variables d'un médecin à un autre.

• L'objectif de ce travail est donc d'apporter une aide pour répondre à la première question posée grâce à l'analyse statistique. Pour répondre à la deuxième question posée, il sera nécessaire de conduire une réflexion sur les différents critères pouvant intervenir dans le choix d'une méthode de diagnostic prénatal (critères à la fois économiques, éthiques et de santé publique). Un système d'aide au diagnostic prénatal pourrait faire l'objet d'un autre travail de recherche, en sachant que l'information sur le risque d'avoir un enfant malade ne serait qu'un des éléments parmi les nombreux autres à prendre en compte.

Ce travail devrait permettre aussi une meilleure connaissance des facteurs chromosomiques ou individuels qui influencent la descendance en cas de translocation réciproque : mécanismes de ségrégation méiotiques, viabilité potentielle des segments chromosomiques impliqués, influence de l'âge parental, etc.. Les connaissances dont on dispose actuellement ont été obtenues le plus souvent à partir de petits échantillons de translocations ne permettant pas l'étude de l'existence de phénomènes d'interactions sur les modes de ségrégation et le devenir à terme. Il importe donc de

définir quelles sont les variables les plus influentes (du parent porteur ou des chromosomes), et pour chacun de ces groupes lesquelles ?

Ce travail devrait aussi aider à préciser les variables indispensables à l'étude des translocations et les modalités d'enregistrement de celles-ci. Une base de données cytogénétiques est un investissement important, et l'on peut concevoir qu'elle puisse servir d'autres objectifs que le conseil génétique. Cet aspect ne sera abordé qu'en conclusion comme élément de réflexion sur les orientations futures à prendre pour enrichir la base de données existente.

I. 2. Acquisition des données

Nous exposerons tout d'abord comment la base a été construite, et nous décrirons les données disponibles de cette base. Dans un deuxième temps nous expliciterons comment ces données ont été exploitées pour construire de nouvelles variables pouvant servir à la modélisation du problème posé. La notion de viabilité des gamètes en déséquilibre issus d'une translocation sera introduite ici.

I. 2.1. La base informatique

En 1988, le laboratoire de Cytogénétique de Grenoble a élaboré une base de données de translocations réciproques sur MacIntosh en utilisant le logiciel 4^{e} Dimension (Latche, 1988). Cette base de données concerne des familles porteuses d'une translocation réciproque : individu ayant permis la découverte de la translocation (ou proposant) ainsi que ses proches et collatéraux. L'idée de construire une base de données est venue progressivement avec l'augmentation du nombre d'observations disponibles pour chacune des translocations observées. Depuis bientôt 20 ans (avec la mise en place de différents systèmes de recueil de données en France et en Europe), lorsqu'une translocation est découverte chez un individu, il est demandé aux membres de sa famille de bien vouloir effectuer un caryotype afin de pouvoir disposer de la généalogie la plus informative possible. L'enquête n'est pas toujours complète en raison, soit du refus des proches de se soumettre à l'examen, soit de parents ou collatéraux déjà décédés, soit même de négligence ou de difficultés pratiques (cas de familles très dispersées par exemple). Le nombre d'individus par famille est donc variable selon les résultats de l'enquête familiale ; on note, dans notre base, en moyenne 5 individus par famille (minimum 1, maximum 46).

Pour le recueil de données dans la base informatique, les critères d'inclusion et d'exclusion ont été les suivants.

Critères d'inclusion :

- translocation caryotypée avec points de cassure validés (selon le niveau de résolution de 400 bandes ISCN 1981) pour le proposant.

- dans une famille, quelle que soit la "position" du proposant, tous les individus ou Fausse Couche Spontanée (FCS) ou Interruption Thérapeutique de Grossesse (ITG) issus d'un parent porteur caryotypé sont enregistrés, que ces descendants aient été eux-mêmes caryotypés ou non. Le parent porteur caryotypé est également enregistré. Lorsque le caryotype des deux parents n'est pas connu, les descendants sont tout de même enregistrés et l'information "origine inconnue" est annotée pour les premiers descendants.

Deux exceptions : les individus présentant un phénotype anormal (enfant malformé) mais sans caryotype ne sont pas enregistrés (impossibilité de savoir quelle anomalie est responsable du syndrome malformatif), et les fausses couches non caryotypées ne sont pas non plus enregistrées. L'information sur les fausses couches non caryotypées est rapportée au parent porteur de la translocation dont elles sont issues (nombre de fausses couches et nombre de grossesses pour ce parent).

- translocations associées à une trisomie.

Critères d'exclusion :

- les translocations robertsoniennes. Dans le cas particulier où les points de cassure surviennent au niveau du centromère et que les deux chromosomes impliqués sont des chromosomes acrocentriques, il s'agit alors d'une translocation Robertsonienne et non d'une translocation réciproque. Les translocations Robertsoniennes sont mieux connues, un peu moins fréquentes que les translocations réciproques et surtout moins nombreuses (15 translocations différentes seulement). Le conseil génétique devant un parent porteur d'une translocation robertsonienne est déjà bien codifié, et celles-ci ne feront donc pas partie de notre étude.

- les translocations impliquant un chromosome sexuel seront aussi exclues. Elles doivent être étudiées à part en raison des caractéristiques spécifiques à chaque chromosome sexuel : possibilité d'inactivation pour le chromosome X, segment important de chromatine (matériel inerte) pour le chromosome Y (Gardner, 1989).

- le plus fréquemment les translocations sont transmises par un parent porteur mais elles peuvent parfois apparaître "de novo" (survenue de l'anomalie au cours de la gamétogénèse) si les deux parents ont un caryotype normal (1,8 % des translocations). Dans cette étude, ces translocations ne seront pas prises en compte, car la transmission de l'anomalie est non informative.

- les translocations multiples (plus de deux points de cassure), pour lesquelles les mécanismes de ségrégation méiotique sont plus complexes et difficilement comparables à ceux survenant en cas de translocation réciproque simple.

- les translocations associées à d'autres anomalies chromosomiques (inversion, insertion, délétion). Pour ces translocations il n'est pas possible actuellement d'attribuer de façon certaine le syndrome malformatif à l'une ou l'autre des anomalies.

Pour constituer la base de données, on dispose de trois sources différentes de données:

- la première est issue de 15 laboratoires* de cytogénétique français : le laboratoire de cytogénétique de Grenoble a recueilli cette information de 1975 à 1986 (fichier 1, 1267 individus soit 17 % des données).

- une deuxième source est constituée par des cas recueillis lors de consultations de diagnostic prénatal : A. Boué (à Paris) a enregistré ces cas à partir de 71 centres européens* de 1977 à 1981 (fichier 3, 2024 individus soit 27 % des données).

- la dernière source de données comporte les translocations ayant fait l'objet d'une publication dans la littérature internationale entre 1971 et 1990 (fichier 2, 4092 individus soit 56 % des données) (Betend, 1989).

Les individus communs à plusieurs de ces fichiers n'ont été enregistrés qu'une seule fois.

Au total, à l'heure actuelle, la base comporte 1590 familles (indépendantes d'un point de vue génétique), regroupant 7383 individus et 1379 équations différentes. Le nombre d'équations différentes n'est pas égal au nombre de familles, dans la mesure où deux familles différentes peuvent être porteuses d'une même translocation. Cette base constitue à notre connaissance la base la plus nombreuse au monde de translocations réciproques. D'autres bases de données d'anomalies chromosomiques existent, mais elles ne sont pas spécifiques des translocations réciproques et se contentent le plus souvent de recenser les différentes anomalies sans objectif explicatif, ni prédictif (Borgaonkar, 1990). On peut supposer que cette base est assez exhaustive des translocations connues à ce jour, compte tenu du gros travail de recherche bibliographique qui a été effectué et qui continue d'être effectué régulièrement.

Dès 1991, des procédures d'aide au conseil génétique à partir des données contenues dans cette base ont été mises en place. En présence d'une translocation réciproque, un généticien peut demander au laboratoire de cytogénétique de Grenoble des informations concernant :

- l'existence de translocations identiques ou voisines,
- les déséquilibres chromosomiques potentiellement engendrés par cette translocation,
- et la prédiction du mode de ségrégation méiotique le plus probable.

Il est prévu de rendre cette aide au conseil génétique (RECI-CONSEIL) prochainement accessible par voie télématique, en permettant une réponse en temps réel à toute demande de renseignements (Cohen, 1992).

En contrepartie, les données relatives à une famille porteuse d'une translocation et pour laquelle une demande de renseignements sera effectuée pourront être recueillies et venir ainsi incrémenter la base de données (que la translocation soit déjà connue ou nouvelle). Un enregistrement complémentaire des translocations publiées dans la littérature depuis 1991 est aussi prévu.

---

* La liste de ces laboratoires et centres de diagnostic prénatal figure en annexe 3

I. 2.2. Les données disponibles

Les informations enregistrées dans la base pour chaque translocation figurent à l'annexe 4 (modèle de fiche de saisie). On peut remarquer l'absence de données sur les différentes expressions cliniques des enfants "malades". Une étude de ces signes cliniques serait certainement très intéressante étant donné la grande diversité des segments chromosomiques en déséquilibre, mais elle ferait alors l'objet d'un autre travail de recherche. Les symptômes présentés par un enfant déséquilibré à terme sont toujours graves, et l'absence de précision les concernant n'a pas de conséquence particulière pour répondre aux objectifs cités ci-dessus.

Certaines rubriques de la base de données sont directement disponibles, comme le sexe du parent porteur, le mode de ségrégation méiotique observé ou les bras des points de cassure, alors que d'autres ne sont obtenues que secondairement par calculs cytogénétiques, comme la longueur des segments centriques et transloqués ou la viabilité des déséquilibres potentiels, ou encore par regroupement (chromosomes impliqués dans la translocation) (tableaux 2-3). Les données obtenues secondairement sont présentées au paragraphe suivant (exploitation des données).

1) mode de ségrégation méiotique observé

Chaque individu est issu d'un parent porteur selon un mode de ségrégation méiotique particulier. Le plus fréquemment il s'agit du mode de ségrégation alterne. Dans 20 % des cas environ il y a eu un mode de ségrégation non alterne, selon l'une des 4 modalités suivantes *adjacent 1*, *adjacent 2*, *3:1 tertiaire* ou *3:1 échange* (fig. 2).

2) histoire naturelle

Lorsqu'il y a eu une ségrégation non alterne, celle-ci peut se rapporter, soit à une fausse couche, soit à un mort-né, soit à un enfant malade vivant à terme, soit encore à un résultat de diagnostic prénatal. Il a été procédé à un enregistrement de ces caractéristiques concernant uniquement les individus avec mode de ségrégation non alterne.

Ces caractéristiques sont exprimées de la manière suivante :

- *Interruption Thérapeutique de Grossesse* (ITG). Après un diagnostic prénatal montrant une ségrégation non alterne, une interruption de la grossesse en cours est effectuée. Dans ce cas, l'évolution naturelle de la grossesse est "tronquée" par l'acte d'interruption médicale.

- *Fausse couche spontanée* (FCS), traduisant un mode de ségrégation non alterne qui conduit à un individu "non viable" (puisque seules les fausses couches caryotypées ont été enregistrées dans la base).

- *Mort Foetale In Utero* (MFIU) ou *mort-né*, caractéristique très proche de la précédente, mais le décès s'est produit soit in utero après le 180^{e} jour de vie intra-utérine, soit au moment de la naissance.

- *Diagnostic prénatal* (DPN), obtenu lors d'un prélèvement de diagnostic prénatal (aussi bien villosités choriales que liquide amniotique) traduisant le fait que l'information sur l'enfant à naître est relative à la vie intra-utérine, sans que l'on dispose d'information sur l'évolution de la grossesse après le diagnostic prénatal.

- *DPN avec déséquilibre antérieur* signifiant la même "histoire naturelle" que précédemment, mais avec notion de l'existence d'un enfant déséquilibré à terme (dans notre base de données) pour la même équation de translocation (selon le même mode de ségrégation non alterne).

- *Déséquilibre à terme*, traduisant la naissance d'un enfant viable porteur d'un syndrome malformatif et d'un retard mental sévère.

Tableau 2. La variable à expliquer

	Effectif		%		Total
Ségrégation méiotique observée					
alterne	6118		*82,9*		
					7383
non alterne	1265		*17,1*		
adjacent 1		901		*71,2*	
adjacent 2		58		*4,6*	
3:1 tertiaire		269		*21,3*	
3:1 échange		37		*2,9*	
Histoire naturelle					
ITG	36		*2,8*		
FCS	17		*1,3*		
mort-né (ou MFIU)	12		*1,0*		1265
DPN	9		*0,7*		
DPN + déséquilibre	28		*2,2*		
déséquilibre à terme	1163		*91,9*		
Déséquilibre à terme					
oui	1191		*16,1*		
					7383
non	6192		*83,9*		

3) les numéros des chromosomes impliqués

Une translocation est caractérisée par une paire de chromosomes, le premier et le deuxième de l'équation de la translocation. Ils sont forcément différents l'un de l'autre. Il existe 231 combinaisons possibles de ces paires de chromosomes (22 modalités * (22-1) modalités / 2), et la fréquence de ces paires ne suit pas une distribution au hasard.

4) les bras des points de cassure

Trois situations peuvent se rencontrer, elles caractérisent chaque équation de translocation. Les deux points de cassure sont situés :

- soit sur les bras courts des 2 chromosomes impliqués "pp",
- soit sur les bras longs des 2 chromosomes impliqués "qq",
- soit l'un sur un bras long et l'autre sur un bras court des 2 chromosomes impliqués "pq" ou "qp".

5) le sexe du parent porteur ou origine

L'origine de la translocation réciproque est soit maternelle, soit paternelle. Dans certains cas, elle est inconnue, lorsque le caryotype des parents n'a pas pu être obtenu. On rappelle que les individus dont l'origine est "de novo" ne sont pas retenus dans l'analyse statistique, puisque leur généalogie n'est pas informative.

6) l'âge des parents

Il peut s'agir aussi bien de l'âge du parent porteur de la translocation, que de l'âge de l'autre parent (non porteur de la translocation). On peut souligner dès à présent le grand nombre d'informations manquantes pour ces deux variables : âge maternel et âge paternel. Ceci obligera à une analyse séparée sur deux fichiers différents, l'un réduit aux individus possédant l'information sur l'âge des parents et l'autre global, mais ne prenant pas en compte le facteur "âge des parents".

Tableau 3. Les variables explicatives

	Variable	Distribution	
Variables quantitatives		min max	moyenne
	CS1	12,6 à 149,9	m = 81,4
	CS2	12,6 à 142,2	m = 45,3
segments	TS1	1,3 à 85,1	m = 20,4
centriques	TS2	1,3 à 80,4	m = 15,8
et transloqués	CSbandeR1	0 à 71,6	m = 38,1
	CSbandeR2	0 à 69,3	m = 20,4
	TSbandeR1	0 à 40,6	m = 11,2
	TSbandeR2	0 à 37,0	m = 9,5
âge des	âge maternel	15 à 51	m = 28,0
parents	âge paternel	16 à 72	m = 29,4
viabilité	nombre de gamètes en déséquilibre viables	0 à 12	m = 5,6
Variables qualitatives		Effectif	(%)
bras des	bras "qq"	n = 3441	(47 %)
points de	bras "pp"	n = 1080	(14 %)
cassure	bras "pq"	n = 2862	(39 %)
sexe du parent	origine maternelle	n = 3539	(48 %)
porteur	origine paternelle	n = 2025	(27 %)
	origine inconnue	n = 1819	(25 %)
	Xmes 1 à 8, 10 à 12, 16 à 20	n = 3449	(48 %)
chromosomes impliqués	Xme 9	n = 961	(13 %)
dans la	Xmes 13,14 et 15	n = 1335	(18 %)
translocation	Xmes 21 et 22	n = 963	(13 %)
	t(11;22)	n = 675	(8 %)

La distribution graphique de chacune de ces variables est présentée en annexe 5.

I. 2.3 Exploitation des données

En vue de l'analyse statistique de ces données disponibles, et plus particulièrement en vue de la modélisation de la mesure quantitative du risque de survenue d'un enfant malformé, il a été nécessaire d'obtenir d'autres données à partir des premières (variables construites ou regroupées, etc...). Nous les présentons ici telles qu'elles seront le plus souvent utilisées dans ce travail, et en distinguant la variable que l'on cherche à expliquer des variables qui peuvent servir à l'expliquer qui seront appelées variables explicatives.

La variable à expliquer

Dans le chapitre II, la variable à expliquer sera le mode de ségrégation méiotique en utilisant sa forme sous 2 modalités seulement, *alterne* versus *non alterne*. Suite à cette première analyse descriptive, il a semblé important de pouvoir dissocier, parmi les individus avec mode de ségrégation non alterne, ceux qui étaient venus à terme avec un syndrome clinique compatible avec la vie de ceux qui n'avaient pas vu le jour. On retrouve, parmi ces derniers, les FCS, les mort-nés et les interruptions thérapeutiques de grossesse.

C'est pourquoi, dans les chapitres III et IV, la variable à expliquer sera une combinaison des deux variables disponibles "mode de ségrégation méiotique" et "histoire naturelle". Elle sera sous forme dichotomique prenant :

- la valeur 1, s'il s'agit d'un individu avec mode de ségrégation non alterne et histoire naturelle "DPN avec déséquilibre antérieur" ou "déséquilibre à terme",
- et la valeur 0, sinon. Dans la valeur 0, seront donc regroupés des individus avec mode de ségrégation alterne et des individus avec mode de ségrégation non alterne, pour ces derniers, il s'agit de foetus ou d'enfants porteurs d'un déséquilibre en matériel chromosomique non observé "viable" dans notre base de données (tableau 2).

Les variables explicatives

1) les segments chromosomiques

Il s'agit de mesures caractérisant chaque équation de translocation réciproque (les individus avec la même équation de translocation auront les mêmes valeurs pour ces variables). Pour chaque équation de translocation une procédure permet d'obtenir la longueur absolue en mm de chaque segment chromosomique (d'après la nomenclature ISCN 1981). On obtient ainsi 4 longueurs pour les 2 chromosomes impliqués dans la translocation :

- longueur du segment centrique du premier chromosome (CS1), et idem pour le deuxième chromosome (CS2),
- longueur du segment transloqué du premier chromosome (TS1), et idem pour le deuxième chromosome (TS2).

Chacune des longueurs de ces segments peut aussi être exprimée en fonction du matériel génétique qu'elle contient. Une différence importante existe en génétique dans le rôle respectif des bandes R, des bandes G et de l'hétérochromatine. Les premières sont reconnues comme plus riches en gènes que les deuxièmes, quant à l'hétérochromatine elle est classiquement reconnue comme matériel génétique inerte (Sumner, 1982). D'où l'idée de mesurer ces longueurs non plus en longueur absolue totale, mais en longueur cumulée des bandes R que chacun de ces segments contient. On obtient ainsi 4 nouvelles mesures :

- longueur en bandes R du segment centrique du premier chromosome (CSbandeR1), et idem pour le deuxième chromosome (CSbandeR2),
- longueur en bandes R du segment transloqué du premier chromosome (TSbandeR1), et idem pour le deuxième chromosome (TSbandeR2).

2) les numéros des chromosomes impliqués dans l'équation de translocation réciproque

Les combinaisons des paires de chromosomes impliqués dans la translocation peuvent être regroupées de multiples façons différentes. Certains seulement de ces regroupements ont été étudiés en se basant sur des contraintes à la fois d'effectifs et de caractères cytogénétiques des chromosomes.

Liste des contraintes

- un groupe doit avoir un effectif suffisant (un minimum de 500 observations nous a paru raisonnable).
- une priorité est donnée à la fréquence des translocations réciproques. Par exemple, les individus issus d'une translocation réciproque (11, 22) représentent 675 observations, soit 9,1 % du total, et constitueront un groupe à eux seuls.
- pour les chromosomes acrocentriques et le chromosome 9, trois groupes sont constitués : le chromosome 9, les chromosomes 13, 14 et 15, et les chromosomes 21 et 22. Pour les autres chromosomes, le regroupement s'effectue par la taille et par la position du centromère.
- la variable doit avoir un nombre réduit de modalités.

Nous avons donc construit une Nouvelle Variable Chromosomique (NVC) à 8 modalités qui sont les suivantes :

	n°	Chromosome	effectif
modalité	1	11 et 22	675
modalité	2	9	960
modalité	3	13,14 et 15	1337
modalité	4	21 et 22	965
modalité	5	4	745
modalité	6	1 à 8	835
modalité	7	16 à 20	1133
modalité	8	10 à 12	740

Le tableau de l'annexe 6 permet de visualiser les chevauchements entre ces groupes. En effet, pour une translocation (4;13) par exemple, il a été nécessaire de prendre une décision quant à son affectation à la modalité 3 ou bien à la modalité 5 de cette nouvelle variable.
Cette nouvelle variable sera tantôt utilisée sous cette forme à 8 modalités, mais plus souvent sous une forme à 5 modalités, qui regroupe sous une seule modalité les modalités n°5 à n°8.

3) Notion de viabilité

La survenue d'un enfant malformé à terme est ce que l'on cherche à prédire. Cet enfant peut être la conséquence de l'un ou l'autre des 4 modes de ségrégation non alterne possibles. Disposant de méthodes permettant de prédire selon lequel de ces 4 modes la translocation réciproque a le plus de chance de donner lieu à un déséquilibre (cf prédiction du mode de déséquilibre), il devient important de pouvoir préciser la viabilité "potentielle" de chacun de ces modes de ségrégation non alterne.

Des procédures ont permis de générer, à partir de la base de données et pour chaque équation de translocation, la liste des 12 déséquilibres potentiels possibles en fonction du mode de ségrégation méiotique qui sera observé dans l'hypothèse d'une ségrégation non alterne. Pour chacun de ces déséquilibres potentiels, la longueur des segments en déséquilibre (en trisomie et/ou en monosomie) est rapportée à la longueur totale du Génome Haploide Autosomal (GHA) donnant ainsi un pourcentage de trisomie et de monosomie spécifique à chacun des 12 déséquilibres potentiels. Daniel en 1979 avait défini des seuils de viabilité pour ces déséquilibres chromosomiques (triangle défini par la ligne joignant les 4 % de trisomie et les 2 % de monosomie) (Daniel, 1979).

Ayant constaté à partir des 1150 déséquilibres observés de notre base qu'un pourcentage non négligeable de cas dépassait ces seuils de viabilité (5 à 8 %), de nouveaux critères de viabilité basés sur des seuils plus élevés en trisomie et en monosomie ont été définis (annexe 7) (Jalbert, 1992; Cohen, 1994).

Grâce à ces nouveaux critères, nous avons créé un indice de viabilité qui représente le nombre de déséquilibres potentiellement viables pour chaque équation (ce nombre peut aller de 0 à 12). L'intérêt de pouvoir disposer de cette variable est le suivant : moins il y a de déséquilibres potentiels viables et moins grand sera le risque d'observer la survenue d'un enfant déséquilibré à terme (en cas de ségrégation non alterne, l'issue de la grossesse sera plus volontiers une FCS ou une MFIU ou un mort-né). A l'opposé, plus il y a de déséquilibres potentiels viables, plus grand sera le risque de survenue d'un enfant malformé vivant et viable.

Dans notre base de données, on note 10 équations seulement qui ne présentent aucun déséquilibre potentiel viable parmi les 12 déséquilibres possibles. Parmi les 30 individus issus de parents porteurs d'une de ces 10 équations, aucun mode de ségrégation non alterne n'a été observé. Il se peut qu'il y ait eu des FCS très précoces, mais celles-ci ont été inapparentes.

Il nous a semblé que cette variable viabilité résumait de façon un peu "grossière" l'information contenue dans le graphique de l'aire de viabilité (annexe 7). Nous avons donc construit 6 autres variables représentant la viabilité des déséquilibres potentiels de manière plus détaillée, selon les principaux modes de ségrégation non alterne. De la même façon que la variable précédente, chacune de ces 6 variables caractérise une équation de translocation (tableau 4).

• Nombre de déséquilibres viables pour les déséquilibres potentiels à 47 chromosomes (dus à un mode de ségrégation 3:1) : variable quantitative discontinue prenant pour valeur 0,1,2,3 ou 4.

• Longueur minimum en trisomie/monosomie pour les 4 déséquilibres potentiels dus à des modes de ségrégation adjacent (1 ou 2) : variable quantitative continue (annexe 6).

• Valeur maximum du sinus de l'angle trisomie/monosomie pour les 4 déséquilibres potentiels dus à des modes de ségrégation adjacent (1 ou 2) : variable quantitative continue (annexe 6).

• Nombre de déséquilibres viables pour les 4 déséquilibres potentiels dus à un mode de ségrégation adjacent (1 ou 2) : variable quantitative discontinue prenant pour valeur 0,1,2,3 ou 4.

• Nombre de déséquilibres viables pour les déséquilibres potentiels à 45 chromosomes (dus à un mode de ségrégation 3:1) : variable quantitative discontinue prenant pour valeur 0,1,2,3 ou 4.

• Longueur minimum en trisomie pure ou en monosomie pure pour les 8 déséquilibres potentiels issus d'un mode de ségrégation 3:1 : variable quantitative continue.

Tableau 4. Les variables concernant la viabilité

Variables	Distribution			
	min		max	moyenne
Nbre gamètes viables à 47 chromosomes (modes 3:1)	0	à	4	m = 2,4
Longueur minimum pour les modes adjacent	0,17	à	5,36	m = 1,68
Maximum de sin(angle) pour les modes adjacent	0,73	à	0,99	m = 0,96
Nbre gamètes viables pour les modes adjacent	0	à	4	m = 2,2
Nbre gamètes viables à 45 chromosomes (modes 3:1)	0	à	4	m = 0,9
Longueur minimum pour les modes 3:1	0	à	8,46	m = 3,06

La distribution graphique de ces variables figure à l'annexe 5bis.

I. 3. Travaux existants sur la prédiction du mode de déséquilibre

Des travaux effectués dans le laboratoire de cytogénétique de Grenoble depuis 1980 ont permis d'élaborer une méthode de prédiction du mode de déséquilibre. Cette prédiction constitue la première étape dans la prédiction du risque de survenue d'un déséquilibre à terme. En effet, selon le mode de ségrégation non alterne qui aura lieu, on perçoit bien que le matériel chromosomique en déséquilibre sera différent, selon que les segments centriques ou transloqués seront impliqués en trisomie et/ou en monosomie. Ces variations en matériel chromosomique influencent peu la gravité des syndromes cliniques, par contre la viabilité des déséquilibres en dépend beaucoup, donc par voie de conséquence le risque de survenue d'un enfant déséquilibré à terme aussi.

Deux méthodes seront présentées : la méthode du diagramme du pachytène (DP) basée sur la configuration géométrique de ce diagramme, et une méthode basée sur l'analyse discriminante (AD). Puis une discussion sur les résultats de ces deux méthodes sera faite.

I. 3.1. Méthode du diagramme du pachytène* (DP)

La méthode dont nous allons rappeler les principes a été élaborée par Jalbert à partir de l'idée suivante : "la ligne de ségrégation passe entre les segments les plus longs du quadrivalent et détermine ainsi le mode de ségrégation" (pour plus de détails on peut se reporter à Jalbert, 1980). Un certain nombre de critères sont utilisés, de façon à prédire le mode de ségrégation non alterne le plus probable pour toute nouvelle translocation réciproque. Ces critères font intervenir la taille des segments centriques et transloqués ainsi que les numéros des deux chromosomes impliqués (annexe 8).

• **Résultats bruts**

Le fichier ayant servi à établir la méthode prédictive est le fichier 1, nous avons donc testé ces critères sur 1026 individus ayant présenté un mode de ségrégation non alterne et appartenant aux fichiers 2 et 3.

Globalement 674 individus, soit 65,7 %, ont été bien classés par cette méthode.
Selon les modes de ségrégation :

535 sur 724 avec ségrégation adjacent 1 ont été bien classés (73,9 %)
28 sur 51 avec ségrégation adjacent 2 ont été bien classés (54,9 %)
98 sur 227 avec ségrégation 3:1 tertiaire ont été bien classés (43,2 %)
13 sur 24 avec ségrégation 3:1 échange ont été bien classés (54,2 %)

mais 149 individus n'ont pu être classés, car ne satisfaisant pas aux doubles conditions de chaque groupe.

• **Amélioration de la méthode**

Afin de pouvoir affecter tous les individus sans exception, on a testé cette même méthode en supprimant le 2^{e} critère pour le groupe adjacent 1, puisque ce critère avait aussi été discuté par De Arce (De Arce, 1986). Ceci n'améliore pas vraiment la prédiction, car le groupe des adjacent 1 est ainsi artificiellement sur-représenté, et les autres groupes ne sont pas améliorés (on note un nombre non négligeable de 3:1 tertiaire affectés au groupe des adjacent 1).

Parmi les 149 "inclassables", on remarque que 100 individus appartiennent aux 3:1 tertiaire et 39 aux adjacent 1. Thomas avait lui aussi obtenu un certain pourcentage d'individus non classés: 16 sur 75, parmi lesquels 13 appartenant au groupe des 3:1 et 2 au groupe des adjacent 1 (Thomas, 1987).

---

* Le diagramme du pachytène représente l'appariement au cours de la prophase de la méiose des 2 paires de chromosomes impliqués dans la translocation. Ce phénomène observé sur les préparations est appelé stade pachytène en raison du caractère épaissi des chromosomes au moment de la prophase. Sur la figure 2, les chromosomes sont représentés sous la forme d'un diagramme du pachytène et constituent un quadrivalent.

On a donc poursuivi la classification sur ces 149 individus non classés, en acceptant qu'ils ne satisfassent plus qu'un seul des critères d'affectation à la 2e étape. Le critère retenu a été celui du numéro des chromosomes impliqués. On constate que, selon cette procédure en deux étapes, les résultats sont nettement améliorés (tableau 5).

Tableau 5. Prédiction du mode de ségrégation non alterne (méthode DP)

	sensibilité [IC]	(faux -)	spécificité [IC]	(faux +)
Méthode initiale				
adjacent 1	74 % [71-77]	(26 %)	94 % [92-97]	(6 %)
adjacent 2	55 % [41-69]	(45 %)	99 % [98-100]	(1 %)
3:1 tertiaire	43 % [37-50]	(57 %)	91 % [88-93]	(9 %)
3:1 échange	54 % [34-74]	(46 %)	90 % [88-92]	(10 %)
Méthode modifiée				
adjacent 1	**77 %** [74-80]	(23 %)	90 % [87-93]	(10 %)
adjacent 2	**55 %** [41-69]	(45 %)	99 % [98-100]	(1 %)
3:1 tertiaire	**82 %** [76-87]	(18 %)	87 % [85-89]	(13 %)
3:1 échange	**54 %** [34-74]	(46 %)	90 % [88-92]	(10 %)

• **Remarque**

Pour 115 équations sur 1381 (8 %), la méthode conduit à deux modes différents prédits (équations satisfaisants à la fois les critères du mode adjacent 1 et ceux des modes 3:1 par exemple). Et ceci indépendemment du fait qu'ils s'agissent d'équations nécessitant ou non une deuxième étape pour leur classification. On note, pour ces équations, la présence du chromosome 9 ou d'un acrocentrique pour au moins un des chromosomes de la translocation.

1e mode prédit	2e mode prédit
104 Adj 1	96 3:1 tert
11 Adj 2	19 3:1 éch
n = 115	n = 115

Pour ces équations, on a regardé le mode de ségrégation réellement observé (Tableau 6).

Tableau 6. Mode de ségrégation non alterne observé et prédit

Mode Observé	Nombre d'équations	1e mode prédit DP	2e mode prédit DP	mode prédit AD
aucun déséquilibre	27	1 Adj 2 26 Adj 1	3 3:1 éch 24 3:1 tert	1 Adj 2 26 Adj 1
Adjacent 1	59	2 Adj 2 57 Adj 1	9 3:1 éch 50 3:1 tert	2 Adj 2, 1 3:1 tert 56 Adj 1
3:1 tertiaire	23	3 Adj 2 20 Adj 1	**3 3:1 éch** **20 3:1 tert**	2 Adj 2, 1 3:1 éch 1 3:1 tert, 19 Adj 1
Adjacent 2	3	3 Adj 2	3 3:1 éch	1 3:1 tert, 2 Adj 2
2 modes différents	3	2 Adj 2 1 Adj 1	1 3:1 éch 2 3:1 tert	1 3:1 tert, 1 3:1éch 1 Adj 1

La case en caractère gras montre qu'avec le 2e mode, on pourrait reclasser correctement certaines observations en 3:1 tertiaire. Comme il n'est pas possible de distinguer les équations appartenant à cette case en caractère gras des autres, le choix du 2e mode conduirait alors à une perte en prédiction correcte pour un grand nombre d'équations adjacent 1 (59 équations sur 88, c'est à dire 67 %). Pour minimiser le taux d'erreur, il nous semble donc préférable de ne pas tenir compte de ce 2e mode "possible" d'après la méthode DP, et de ne se référer qu'à un seul mode de ségrégation le plus probable pour cette méthode.

I. 3.2. Méthode de l'analyse discriminante

C'est une méthode d'analyse de données qui permet d'affecter, après avoir défini au départ des sous-groupes d'individus, de nouveaux individus à ces sous groupes. Aux Etats-Unis, Thomas, en utilisant cette méthode pour la même problématique, a obtenu un taux d'erreur réél de 17 % seulement sur 75 individus testés à l'aide d'une analyse discriminante comportant 8 variables (Thomas, 1987). Au laboratoire de Cytogénétique de Grenoble, un premier travail réalisé en 1988 (Latche, 1988) montre un taux d'erreur réel de 16 % sur 975 individus testés par une analyse discriminante comportant 16 variables (critère d'affectation ne tenant pas compte des probabilités *a priori*). Cette dernière analyse discriminante a été ensuite améliorée, les améliorations portant sur les points suivants (Cans, 1990) :

- réduction du nombre de variables informatives à 4 seulement, en utilisant les valeurs du critère de Wilks et du pouvoir discriminant pour sélectionner les variables, et les valeurs du taux d'erreur réel pour le choix de la meilleure analyse discriminante.

- paramètres des équations estimés à partir d'un fichier plus grand (fichier 1 + fichier 3), donc meilleure précision dans l'estimation de ces paramètres.

- critère d'affectation tenant compte des probabilités a priori.

- discrimination entre les 4 principaux groupes non alterne et non pas entre 2 groupes seulement (adjacent versus 3:1).

Des précisions sur la méthode de l'analyse discriminante avec le logiciel Systat sont fournies à l'annexe 9.

Les 4 variables retenues ont été les longueurs en bandes R des 4 segments impliqués dans la translocation : longueur en bande R du segment centrique du 1^{e} chromosome, et même chose pour le 2^{e} chromosome, et longueur en bande R du segment transloqué du 1^{e} chromosome, et même chose pour le 2^{e} chromosome.

Les résultats de classification sur 797 observations non alterne du fichier 2 (taux d'erreur réel) sont donnés au tableau 7.

Tableau 7. Prédiction du mode de ségrégation non alterne par la méthode AD

	sensibilité [IC]	(faux -)	spécificité [IC]	(faux +)
adjacent 1	**94 %** [92-96]	(6 %)	66 % [60-71]	(34 %)
adjacent 2	51 % [37-65]	(49 %)	97 % [96-98]	(3 %)
3:1 tertiaire	54 % [47-61]	(46 %)	93 % [91-95]	(7 %)
3:1 échange	23 % [0-46]	(77 %)	99 % [98-99]	(1 %)

Tenant compte des probabilités a priori, cette méthode favorise nettement le groupe des adjacent 1 au détriment des autres groupes. Très peu d'observations adjacent 1 sont mal classées, par contre, dans les observations classées adjacent 1, un bon nombre sont des observations issues d'un autre mode de ségrégation non alterne.

I. 3.3. Combinaison des deux méthodes

Les résultats sont "satisfaisants" pour chacune des méthodes. La DP est peu sensible, mais plus spécifique, et c'est l'inverse pour l'AD (Cans 1993).

Nous avons donc testé le résultat de ces deux méthodes sur l'ensemble des équations de notre base de données. Dans ce cas, il n'est pas possible de vérifier l'exactitude de l'une ou de l'autre par rapport au mode de ségrégation observé, puisque, pour un certain nombre d'équations

(environ 50 %), il n'y a pas eu de mode de ségrégation non alterne observé. Sur 1381 équations, on note une discordance dans le mode de prédiction pour :

- 307 d'entre elles (22 %) si l'on tient compte de tous les modes prédits
- 278 d'entre elles (20 %) sans tenir compte du deuxième mode prédit par la DP
- 268 d'entre elles (19 %) en excluant les équations à 0 gamète viable (pour lesquelles le mode de ségrégation prédit n'est pas nécessaire, puisqu'aucun d'entre eux n'est viable)

La concordance entre les deux méthodes est montrée au tableau 8.

Tableau 8. Concordance entre les deux méthodes prédictives

		méthode DP				
		Adj 1	Adj 2	3:1 tert	3:1 ech	
méthode AD	Adj 1	**934**	0	80	21	1035 (74,9%)
	Adj 2	44	**20**	31	6	101 (7,3%)
	3:1 tert	43	5	**126**	2	176 (12,7%)
	3:1 ech	20	3	23	**23**	69 (5,0%)
		1041 (75,3%)	28 (2,0%)	260 (18,8%)	52 (3,8%)	**1381**

coefficient K = 0,50 concordance modérée
proportion observée de concordance : 79,9 %
proportion de concordance si l'on fait l'hypothèse d'un remplissage purement aléatoire des cellules de ce tableau : 59,2 %

La discordance entre les 2 méthodes est importante, et probablement trop pour pouvoir en déduire une stratégie unique dans la détermination du mode de ségrégation non alterne le plus probable. Plusieurs stratégies sont donc possibles, selon que l'on privilégie la sensibilité ou la spécificité de la méthode, ou bien la fiabilité par rapport à un groupe plutôt que par rapport à un autre. Le groupe adjacent 1 est différent des autres, dans la mesure où c'est le groupe pour lequel les individus en déséquilibre seront le plus souvent viables (quantité de déséquilibre plus faible). De plus, les erreurs ne sont pas de la même importance : en effet, il est plus grave de prédire un mode avec non disjonction (adj 2, 3:1 tert, 3:1 éch) à la place d'un mode adjacent 1 que de prédire un mode 3:1 tertiaire au lieu d'un 3:1 échange (en raison de la quantité globale de matériel chromosomique en déséquilibre). En privilégiant la sensibilité plutôt que la spécificité, on choisit aussi l'erreur la moins grave, car les observations qui ne seront pas des adjacent 1, mais qui seront prédites adjacent 1, présentent a priori un déséquilibre en quantité plus importante et seront donc plus sujettes (que les adjacent 1) à un déséquilibre non viable que viable (donc constituent une menace moins grande pour le généticien). L'erreur qui consisterait à prédire un mode avec non disjonction plutôt qu'un mode adjacent 1 serait plus grave, car elle rassurerait à tort en tablant sur une non-viabilité plus probable des gamètes en déséquilibre.

La stratégie que nous proposons est la suivante :

- si le mode prédit par l'AD est adjacent 1 ----> on garde ce mode comme résultat final. La sensibilité de l'AD pour le mode adjacent 1 étant très bonne, on privilégie en premier lieu cette méthode,

- si le mode prédit par l'AD est adjacent 2, 3:1 tertiaire, ou 3:1 échange ----> on regarde alors quel a été le mode prédit par la DP, puisque cette méthode présente une meilleure sensibilité/spécificité pour ces 3 modes. Ce mode peut être une des 4 modalités, et c'est celui prédit par la DP qui sera gardé comme résultat final.

Cette prédiction du mode de déséquilibre sera introduite comme variable explicative dans les chapitres III et IV.

Chapitre II

Le risque de survenue d'un enfant malformé : analyse descriptive

Dans ce chapitre nous allons voir les méthodes descriptives utilisées pour prendre connaissance des fichiers. Le but est d'obtenir une description ou un classement descriptif de la descendance d'un parent porteur d'une translocation réciproque. Certains auteurs ont exploré cet aspect par une analyse univariée simple. Par exemple, pour un individu présentant une translocation avec un point de cassure sur le bras court du chromosome 8, le risque de survenue d'un enfant malformé est de 9,1 ± 7,1 % (Stengel-Rutkowski, 1988). Ou bien, pour un individu repéré par un mode de détection "enfant déséquilibré à terme", le risque de survenue d'une ségrégation non alterne est de 22,1 ± 5,3 % (Daniel, 1986). Malgré tout leur intérêt, ces analyses possèdent des limites, puisque leurs résultats ne sont précis que pour des translocations fréquentes et qu'elles ne permettent pas la prise en compte de plus d'une variable à la fois. Nous avons donc essayé des méthodes descriptives multivariées, comme les méthodes d'analyse de données. Ces méthodes devraient permettre d'étudier les relations entre les variables explicatives et d'isoler les variables les plus influentes parmi celles disponibles.

II. 1. Méthodes d'analyse de données

Nous ferons d'abord un bref rappel des méthodes utilisées, puis nous présenterons les résultats obtenus pour chacune d'elles.

II. 1.1. Méthodologie

L'analyse en composantes principales (ACP) consiste à déterminer les combinaisons linéaires de variables explicatives qui procurent la plus grande dispersion des observations (Robert, 1989). L'idée est de repérer une quelconque structure du nuage notamment en relation avec les deux sous-groupes d'observations représentant respectivement les observations avec mode de ségrégation alterne et celles avec mode de ségrégation non alterne.

L'analyse factorielle discriminante (AFD) est une méthode qui essaye de construire des combinaisons linéaires de variables explicatives (scores discriminants), qui sépareraient au mieux les deux sous-groupes d'observations : mode de ségrégation alterne et mode de

ségrégation non alterne. Elle peut être considérée comme un résumé de l'information existante permettant de classer entre deux groupes (Romeder, 1973). Contrairement à l'ACP purement descriptive, c'est déjà une méthode de classification puisqu'elle permet, si le pouvoir discriminant de l'analyse est suffisamment bon, d'affecter de nouveaux individus à l'un ou l'autre sous-groupe.
Pour ces deux analyses, le logiciel SYSTAT a été utilisé. Les variables sont toutes "traitées" de façon quantitative (sous forme binaire 0/1 pour les variables qualitatives).

II. 1.2. Résultats

Ces deux analyses ont été réalisées sur 2993 observations appartenant aux fichiers 1 et 3 (les premières observations saisies dans la base de données). Parmi ces observations les 2 groupes à repérer sont le groupe mode de ségrégation "non alterne" (347 individus, soit 11,6 %) et le groupe mode de ségrégation "alterne" (2644 individus). Le fichier 2 issu de la littérature servira d'échantillon test pour le calcul du taux d'erreur réel.

• ACP

Une première ACP normale réalisée avec 13 variables* a montré 52 % de variance expliquée sur les trois premiers facteurs (fig. 3). Nous avons réduit le nombre des variables explicatives, en supprimant celles dont la somme des coefficients sur les principaux facteurs était la plus faible et celles qui étaient le moins corrélées entre elles.

Malgré cette réduction, l'ACP à 5 variables ne montre que 71 % de variance expliquée sur les trois premiers facteurs (fig. 3). Les 5 variables retenues étaient l'âge du parent porteur, l'appartenance au groupe des translocations comportant le chromosome 9 en deuxième chromosome, l'appartenance au groupe des translocations comportant les chromosomes 4 ou 18 en deuxième chromosome, l'existence d'un déséquilibre observé à terme pour la même équation de translocation, et le rapport de la somme des segments centriques sur la somme des segments transloqués. Lorsque l'on projette les individus sur les trois premiers facteurs principaux de cette ACP, on ne distingue pas de sous-groupe particulier et le nuage de points semble unique. On remarque juste qu'il semble y avoir très peu de mode de ségrégation non alterne dans la partie inférieure du graphe. Ce graphe a été obtenu à partir d'un échantillon randomisé de 122 observations issues des 2993 ayant servi à réaliser l'ACP (fig. 4a).

---

* CSbandeR1, CSbandeR2, TSbandeR1, TSbandeR2, Σ CS / Σ TS, 5 variables qualitatives (groupes de chromosomes impliqués dans la translocation), sexe et âge du parent porteur, existence d'un déséquilibre observé à terme.

Fig. 3. Valeurs propres des ACP

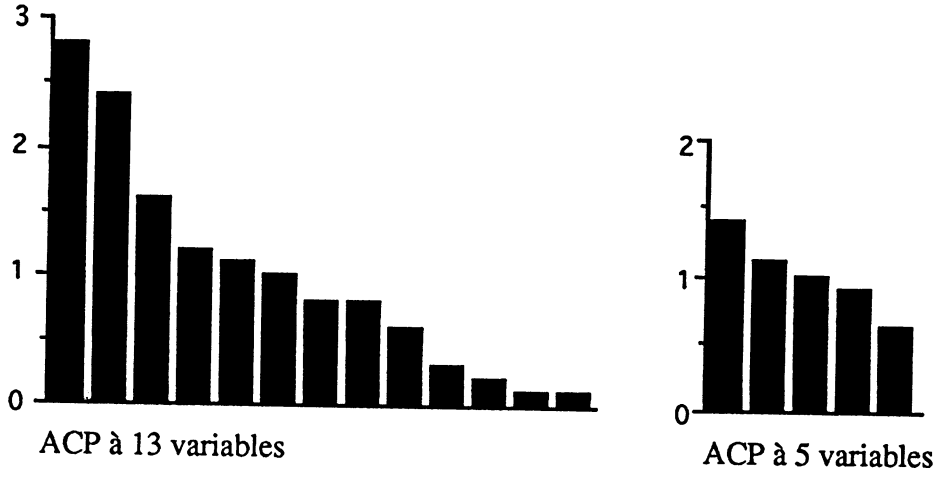
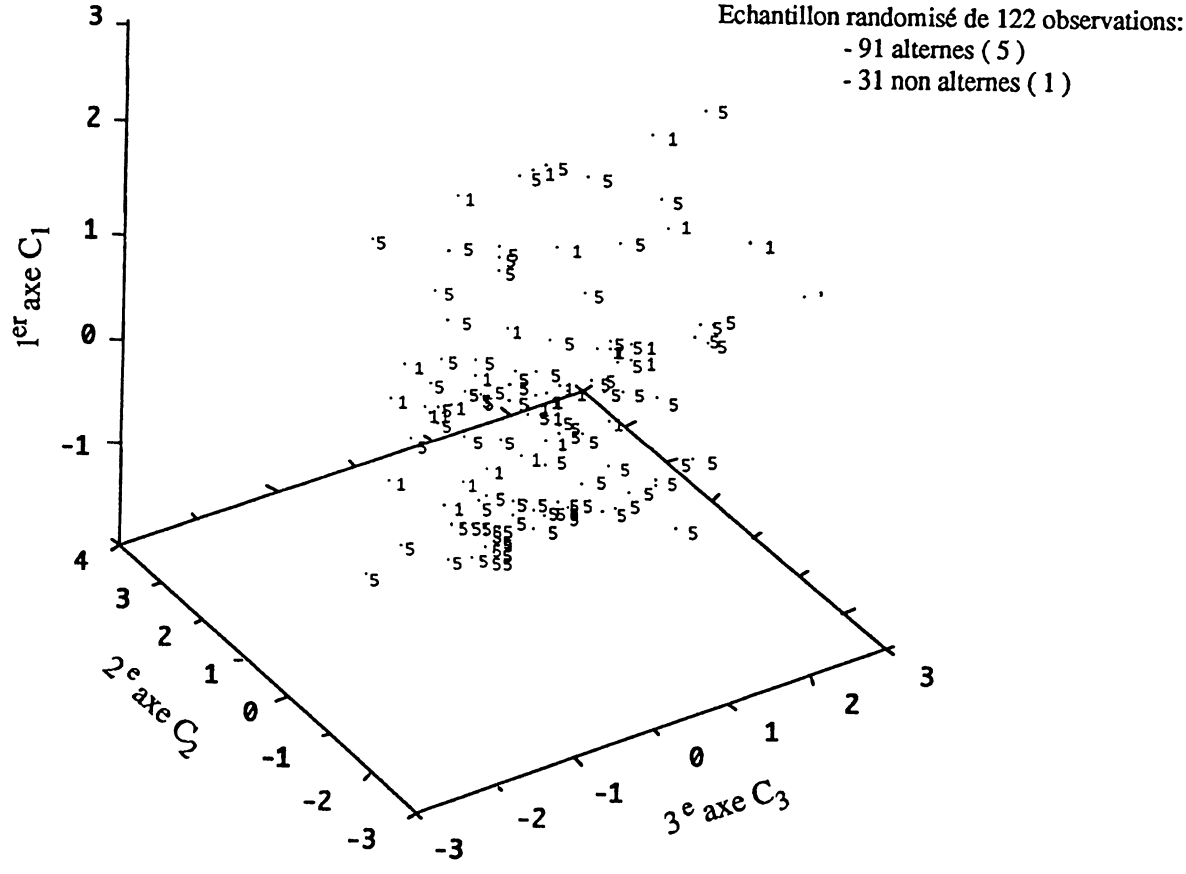


Fig. 4a. Représentation des individus sur les 3 premiers facteurs de l'ACP à 5 variables. Prédiction alterne/non alterne



1er axe C_1 représenté par l'existence d'un déséquilibre à terme et le rapport des segments centriques et transloqués
2e axe C_2 représenté par les groupes de translocation avec le chromosome 9 et avec les chromosomes 4 ou 18
3e axe C_3 représenté par l'âge du parent porteur

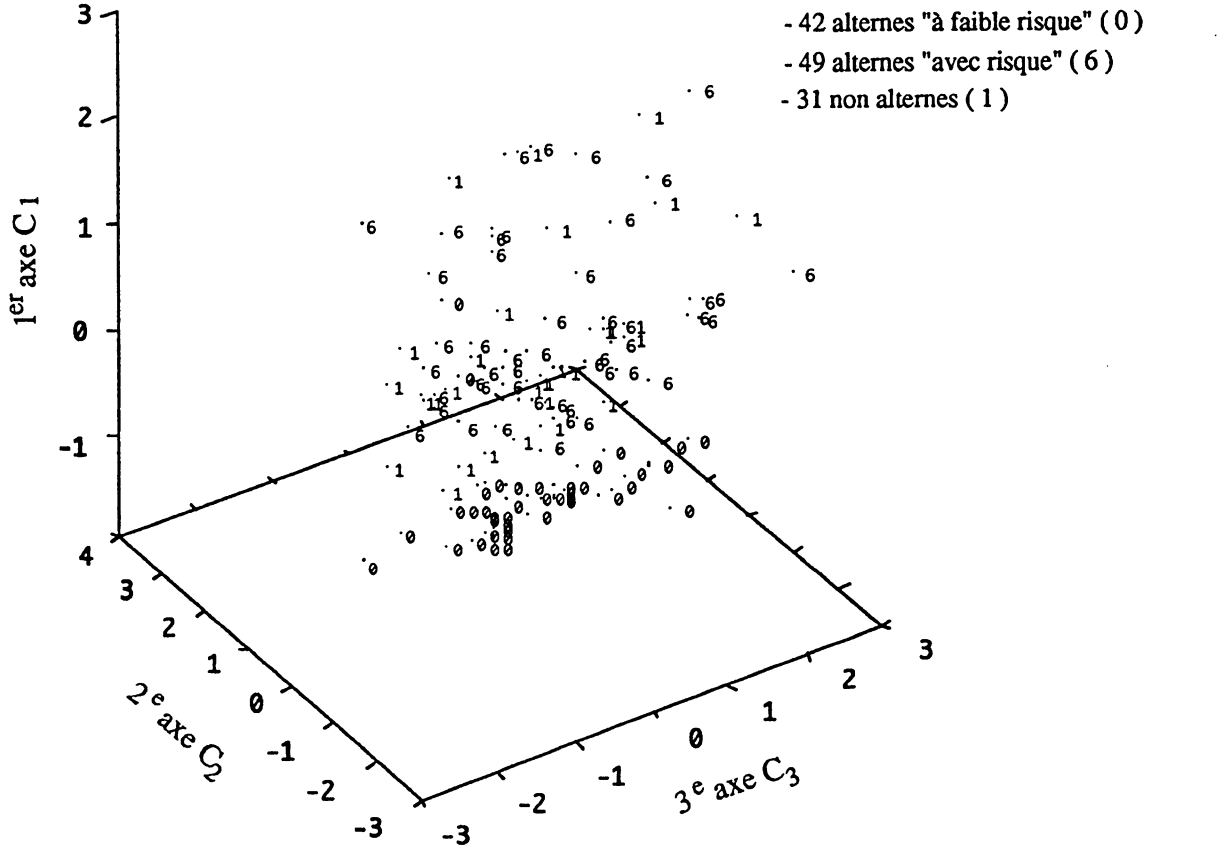
A partir de la remarque ci-dessus, on en est venu à se demander si on ne pourrait pas distinguer, parmi les translocations ayant conduit à une ségrégation alterne, celles qui ne donnaient lieu que très rarement à une ségrégation non alterne (pas de descendance déséquilibrée à terme, ni de fausses couches) de celles qui donnaient lieu aussi bien à une ségrégation alterne qu'à une ségrégation non alterne. Etant donné que, dans la base de données, seules ont été saisies les fausses couches caryotypées (une minorité des fausses couches), il nous a semblé intéressant de tenir compte de la variable "nombre de fausses couches dans la descendance d'un individu" de façon à séparer le groupe des alterne en deux sous-groupes :

- les "alterne à faible risque", représentés par 1063 individus n'ayant jamais eu ni fausses couches, ni enfant déséquilibré à terme dans leur descendance,

- les "alterne à risque", représentés par 1583 individus ayant présenté soit des fausses couches, soit des déséquilibres à terme dans leur descendance.

A partir du même échantillon randomisé de 122 observations, on procède à la représentation sur les trois facteurs principaux de la même ACP des individus regroupés de cette manière (fig. 4b).

Fig. 4b. Représentation des individus sur les 3 premiers facteurs de l'ACP à 5 variables. Prédiction alterne/non alterne

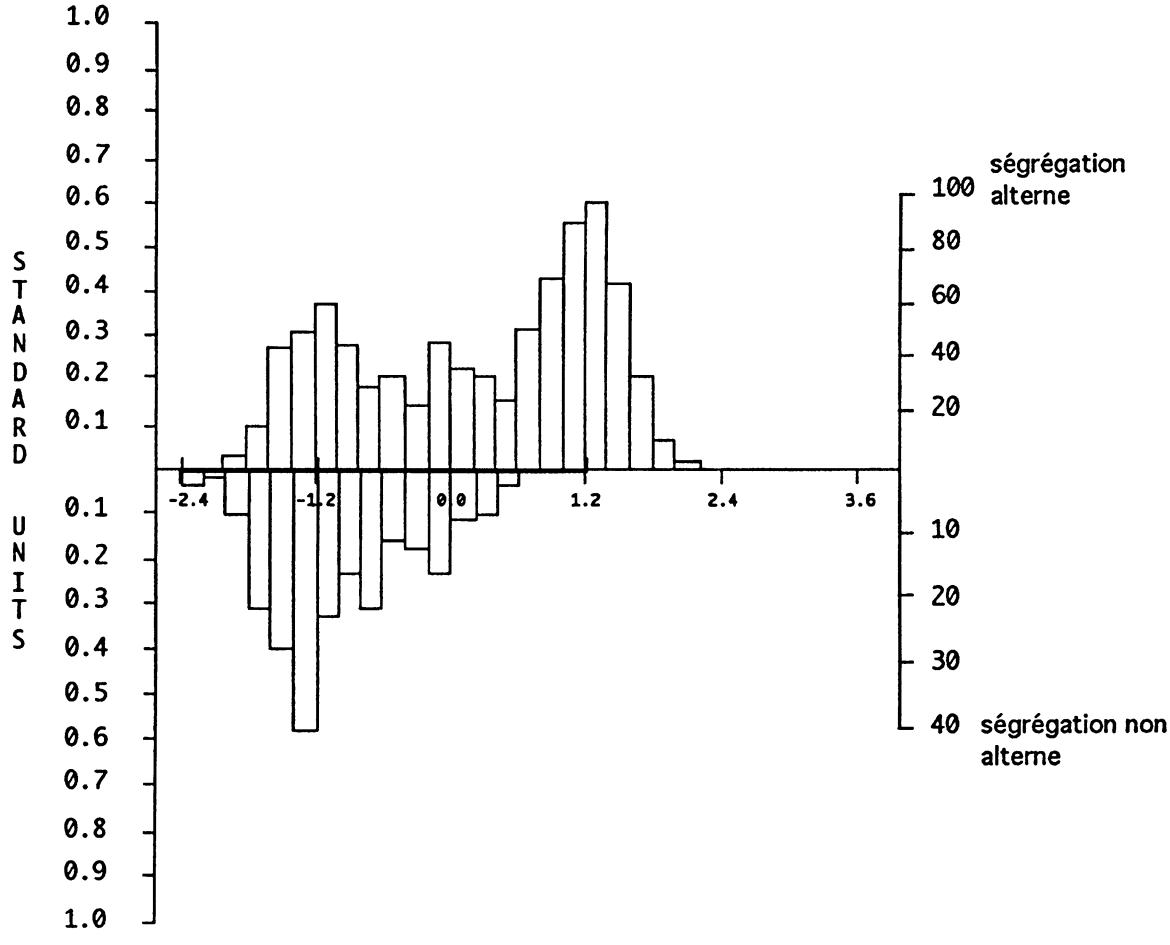


On remarque que le groupe des "alterne à faible risque" se distingue nettement du reste du nuage de points, suggérant que la distinction de sous-groupes parmi le groupe des alterne pourrait aider à discriminer les alterne des non alterne en procédant de manière séquentielle.

• **AFD**

Que ce soit l'AFD à 13 variables explicatives ou bien à 8 variables (après réduction selon les mêmes critères que ceux utilisés pour l'ACP), les résultats ne sont pas satisfaisants. Malgré 81,1 % d'individus bien classés (en taux d'erreur apparent), le pouvoir discriminant n'est que de 44 %, et, même si l'on note 96 % d'individus bien classés parmi les "alterne", on note seulement 22 % d'individus bien classés parmi les "non alterne". Le taux d'erreur réel n'a pas pu être calculé en raison du grand nombre d'informations manquantes pour la variable âge du parent porteur dans le fichier 2. L'histogramme du score discriminant montre l'important recoupement entre les deux groupes, sans valeur seuil possible de ce score permettant la distinction entre ces deux sous-groupes (fig. 5).

Fig. 5. Histogramme du score discriminant - AFD à 8 variables
Prédiction alterne / non alterne



II. 1.3. Interprétation

Au total, aucune de ces deux méthodes ne donne de résultats très probants. Bien qu'ayant utilisé des variables suggérées influentes dans la littérature, les variables explicatives semblent peu discriminantes. Les résultats que l'on obtiendrait si l'on tenait compte uniquement des probabilités *a priori* et non pas des valeurs des variables explicatives seraient les suivants :

68 % globalement d'individus bien classés
- 80 % pour les alterne
- 20 % pour les non alterne

Ils ne sont guère moins bons que ceux obtenus par l'AFD à 8 variables.

Certaines raisons peuvent être avancées pour tenter d'expliquer ces résultats décevants : 1) la plupart des variables explicatives sont des variables qualitatives qu'il a été nécessaire de traiter en variables binaires "0/1" ; 2) étant donné les effectifs des 2 sous-groupes très différents, des variables très discriminantes sont nécessaires pour dépasser l'influence des probabilités *a priori*. En effet, la prise en compte de celles-ci conduit à une affectation trop fréquente au groupe des alterne. Ces raisons nous ont incité à utiliser d'autres méthodes descriptives avant d'envisager l'utilisation de méthodes inférentielles (chapitres III et IV).

II. 2. Autres méthodes

Les deux méthodes présentées ci-dessous sont basées sur un apprentissage des caractéristiques de l'échantillon permettant une classification de nouveaux individus. Elles différent essentiellement par leurs méthodes d'apprentissage. Mais il faut souligner aussi que la première a pour but, en plus des résultats de classification, de tenter d'expliquer le phénomène ou du moins d'en faciliter la compréhension.

Pour ces deux analyses, les variables sont toutes traitées qualitativement (avec regroupement en classes pour les variables quantitatives).

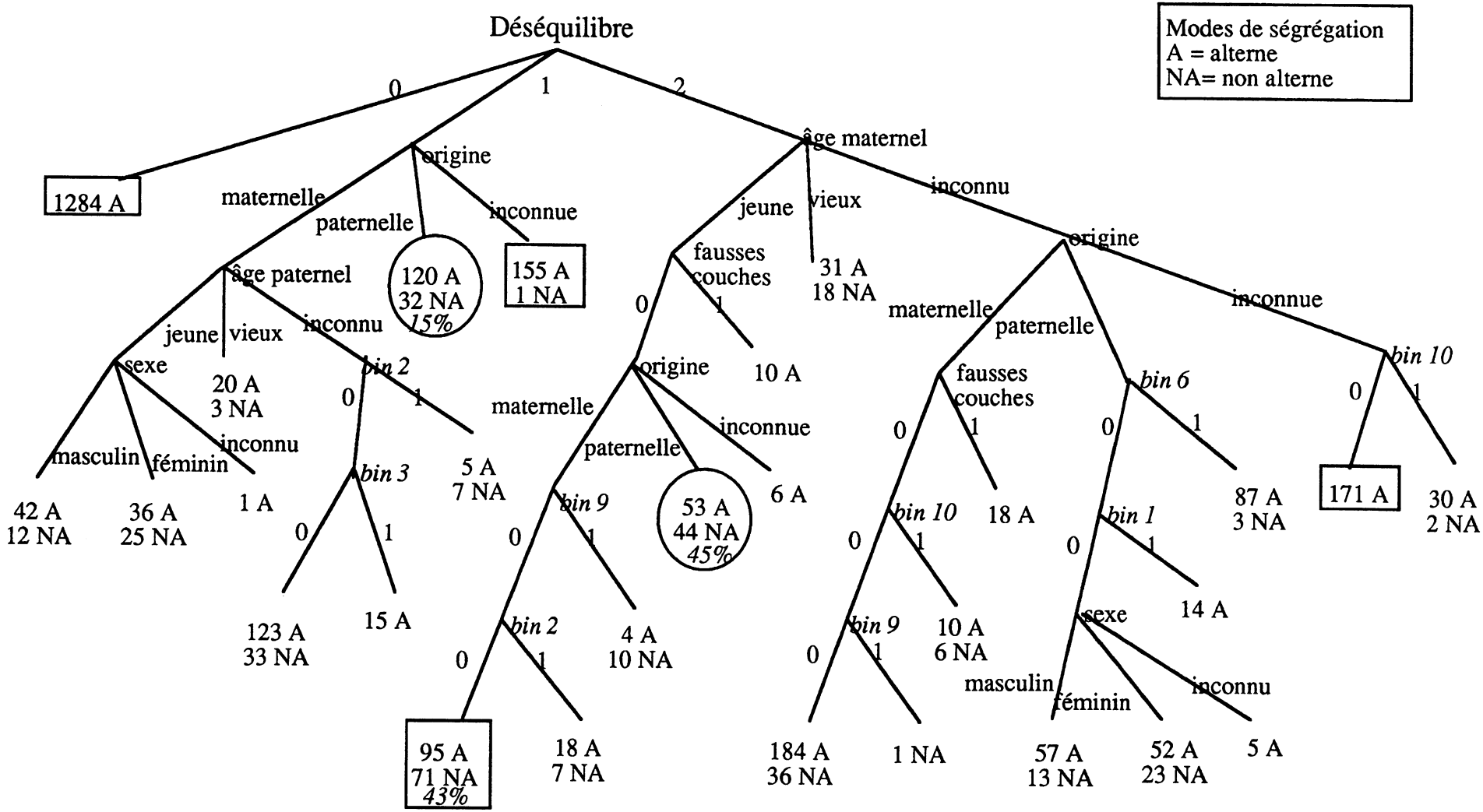
II. 2.1. Arbre d'induction ou régression qualitative

Cette méthode est la méthode anciennement nommée segmentation.

• Il s'agit d'algorithmes conduisant à la description de la base de données sous la forme d'un arbre. Sommairement, les branches de l'arbre correspondent aux valeurs des différentes variables explicatives et l'objectif est d'obtenir finalement des feuilles contenant un faible mélange des modalités de la variable à expliquer. L'arbre induit à partir d'un fichier d'exemples permet l'affectation des éléments d'une feuille à la modalité la plus fréquente de celle-ci. C'est alors un arbre de décision (Crémilleux, 1991). Un logiciel ARBRE mis au point au sein du laboratoire TIM3-IMAG de Grenoble a été utilisé pour cette application.
Cette méthode a été appliquée d'une part aux 2993 observations des fichiers 1 et 3 (premières observations saisies dans la base de données), et d'autre part à 3247 observations du fichier 2. Les objectifs étaient l'exploration de ces fichiers, mais aussi leur comparaison entre eux : si on obtient des arbres induits similaires, cela constituera un argument pour considérer ces fichiers homogènes et pour les analyser ensemble. La variable décision est le mode de ségrégation alterne ou non alterne, et les variables explicatives sont celles citées précédemment (chapitre I).
Le critère de sélection d'une variable est le gain d'information, et la construction d'un arbre est arrêtée lorsqu'un noeud comporte moins de 100 cas.

• Les arbres obtenus sont volumineux et présentent des architectures distinctes selon les fichiers de données et selon les variables explicatives utilisées. On note cependant que la première variable segmentant la racine est toujours soit la variable "origine", soit la variable "existence d'un déséquilibre à terme pour la même translocation". Ces deux variables ainsi que la variable "âge parental" sont en général présentes en 1^e, 2^e ou 3^e position dans chacun des différents arbres. L'indice de qualité globale, qui représente le taux d'information expliquée par les variables explicatives, est meilleur pour les arbres construits à partir des fichiers 1 et 3, que pour ceux construits à partir du fichier 2 (33 à 37 % contre 20 à 21 %). Cependant, tous les arbres permettent de dégager des feuilles intéressantes, c'est-à-dire clairement en faveur d'alterne ou de non alterne. Nous présentons à titre d'exemple le graphe d'un des arbres obtenu à partir des fichiers 1 et 3 (fig. 6).

Fig. 6. Arbre de régression qualitative



Légende :	
bin 1 à bin 10	10 variables chromosomiques décrivant les numéros des chormosomes impliqués dans la translocation
fausses couches	≥ 2 fausses couches dans la descendance d'un individu porteur d'une translocation réciproque
âge parental	il a été recodé en 3 catégories : moins de 35 ans, 35 ans ou plus et inconnu
déséquilibre	variable quantitative décrivant l'existence de 0, 1 ou ≥ 2 déséquilibres pour cette translocation

Sur ce graphe, on peut remarquer l'existence de quelques feuilles homogènes.

- Une première feuille déterminée par l'absence de déséquilibre pour ces translocations contient 1284 observations à ségrégation alterne. Ce groupe d'observations est semblable à celui décrit en p. 36 des "alterne à faible risque". Environ un tiers des fichiers 1 et 3 concerne des familles pour lesquelles il n'y a pas eu de conséquence pathologique dans leur descendance.

- Une deuxième feuille déterminée par l'origine inconnue de la translocation contient 155 observations à ségrégation alterne et 1 seule à ségrégation non alterne. Lorsque l'origine de la translocation est restée inconnue (absence d'enquête familiale ou des résultats de celle-ci) cela concerne le plus souvent des familles pour lesquelles il n'y a pas eu naissance d'un enfant malformé. Cela traduit le fait qu'en l'absence de pathologie exprimée la réalisation de l'enquête familiale est beaucoup plus difficile.

- Une troisième feuille, déterminée par l'implication des chromosomes 1 à 8 et 10 à 12 dans la translocation (variable chromosomique bin 10), contient 171 observations avec ségrégation alterne, alors qu'une autre feuille, déterminée par l'implication des chromosomes acrocentriques dans la translocation (variable chromosomique bin 2), contient plus de 40 % d'observations avec ségrégation non alterne. Cela est tout à fait concordant avec les suggestions déjà émises par Boué et Jalbert (p.9) concernant l'influence des numéros des chromosomes impliqués sur la survenue d'un déséquilibre à terme.

On peut souligner aussi le rôle différent de certaines variables selon le niveau où elles se situent dans l'arbre. C'est le cas de l'interaction entre la variable "origine" et d'autres variables explicatives. Lorsque l'origine est paternelle, on observe une feuille où 15 % des cas sont non alterne. Après prise en compte de l'âge maternel et du nombre de fausses couches, l'arbre conduit lorsque l'origine est paternelle à une feuille où 45 % des cas sont non alterne.

• Ce travail montre la complexité des données et la difficulté à discriminer. Les conclusions apportées sont de plusieurs ordres :

- les résultats suggèrent une certaine hiérarchie dans le pouvoir discriminant des variables : prédominance de "origine", "existence d'un déséquilibre" et "âge parental" parmi toutes les variables explicatives disponibles.

- il existe des interactions entre les variables et il semble nécessaire de les prendre en compte si l'on veut mieux préciser l'influence de chacune des variables sur la survenue d'un enfant malformé.

- même si le fichier 2 diffère des deux autres (en raison principalement de son mode de recrutement), il apparaît licite de vouloir effectuer une analyse statistique sur l'ensemble des trois fichiers compte tenu des configurations très proches des arbres induits (mêmes variables se retrouvant en premières positions à la racine de l'arbre).

II. 2.2. Méthode des réseaux de Kohonen

L'utilisation de cette méthode a pour but d'essayer une méthode de classement dans laquelle la disproportion entre les groupes n'intervient pas dans le critère d'affectation.

• Il s'agit d'une méthode de classement basée sur un modèle d'auto-organisation (Kohonen, 1988). Le processus d'auto-organisation consiste à apprendre à des cellules à réagir de mieux en mieux aux *stimuli* auxquelles elles sont les plus sensibles et à influencer dans le même sens leurs cellules voisines (rappelant le fonctionnement des neurones du cortex moteur).
Le stimulus est défini par le vecteur des variables explicatives pour un individu. Chaque cellule est caractérisée par un vecteur de poids "synaptiques" ou valeurs des variables explicatives correspondant à une réponse maximale de la cellule. Le critère de sélection de la cellule la plus sensible à un stimulus est une distance minimale entre le vecteur des variables explicatives et le vecteur des poids "synaptiques".

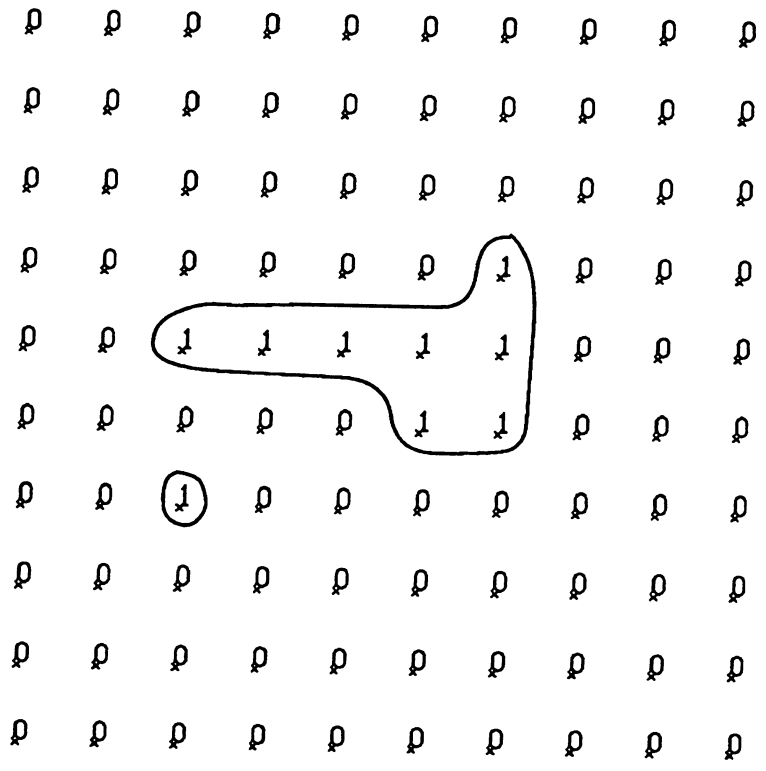
A l'issue du processus d'auto-organisation, les cellules d'un même voisinage doivent répondre toutes au même stimulus. Cette analyse des *stimuli* projetés sur les mêmes aires peut être susceptible de révéler des liaisons entre ces *stimuli*, que d'autres méthodes (ACP, AFD) n'auraient pas révélé.

• Cette méthode a été appliquée à un fichier randomisé de 1303 individus (dont 125 déséquilibres à terme soit 9,6 %). Le résultat après 1000 itérations est présenté sur la figure 7. Sur 100 cellules définissant l'aire de projection des 1303 *stimuli*, 9 d'entre elles ont reçu l'étiquette "déséquilibre à terme". En terme de classement (taux d'erreur apparent sur le même fichier) on note :

taux d'erreur apparent Kohonen	taux d'erreur apparent AFD à 8 variables
- 88 % ± 2 % d'individus bien classés globalement,	81 % ± 1 %
- 10 % ± 5 % d'individus bien classés parmi les 125 "malades" (déséquilibres à terme),	22 % ± 4 %
- et 96 % ± 1 % d'individus bien classés parmi les 1178 "non malades".	96 % ± 1 %

Fig. 7. Carte topologique des translocations réciproques

Gain initial	= 0,2	Nb. itérations	= 1000
Nb. neurones	= 100	Nb. *stimuli*	= 1303
Nb. réciprocités	= 38	Nb. affinités	= 98
Taux de réciprocité	= 38	Taux d'affinité	= 98
Nb. bien classés	= 1152	Nb. mal classés	= 151
Taux de bien classés	= 88 %		



Comme pour l'AFD, ces résultats ne sont pas meilleurs que ceux que l'on obtiendrait en tenant compte uniquement des probabilités *a priori*. Si l'on regarde les caractéristiques des vecteurs synaptiques pour les cellules étiquetées "déséquilibre à terme", on note une tendance à de grands segments transloqués pour le 1e chromosome de l'équation et de grands segments centriques pour le 2e chromosome. Mais il est important de souligner que ces cellules étiquetées "déséquilibre à terme" ne comprennent en fait que 10 % des individus effectivement malades, d'où la nécessité d'être très prudent dans l'interprétation de ces vecteurs synaptiques.

• Pour expliquer le manque de performance de cette méthode sur nos données nous pensons surtout qu'il existe un problème dans la prise en compte des variables qualitatives à plusieurs modalités. En effet, celles-ci sont considérées par le logiciel comme des qualitatives ordonnées ce qui n'est pas du tout le cas des variables suivantes : "bras des points de cassures", "numéro des chromosomes impliqués" et "mode de ségrégation non alterne prédit". Une

décomposition de ces variables en plusieurs variables binaires 0/1 (nombre de modalités moins une) aurait été nécessaire.

Au total :

1) Ces méthodes nous ont permis de prendre connaissance des données et de leur complexité, et aussi de mesurer l'importance de la prise en compte de la viabilité du déséquilibre chromosomique dans la variable à expliquer. Pendant très longtemps, nous avons considéré que le mode de ségrégation méiotique (alterne versus non alterne) était la variable à expliquer . Mais, compte tenu des biais de sélection inhérents au recrutement et des résultats de ces premières analyses il nous est apparu plus juste de considérer comme variable à expliquer la survenue d'un déséquilibre viable à terme versus son contraire. Le déséquilibre viable à terme traduit bien évidemment un mode de ségrégation non alterne; par contre, son contraire est le plus souvent le résultat d'un mode de ségrégation alterne, mais peut être aussi le résultat d'un mode de ségrégation non alterne "non viable" (ou sans confirmation de sa viabilité) comme c'est le cas pour les fausses couches.

2) Suite à ce travail, différentes options d'analyse étaient possibles. Une première possibilité consistait à procéder de façon séquentielle, en écartant pas à pas les observations pour lesquelles le mode de ségrégation a toujours été alterne (par exemple en écartant d'abord les 1284 observations de la 1^{e} feuille de l'arbre de régression ci-dessus). Ceci aurait permis alors d'effectuer des AFD sur une population plus restreinte avec un déséquilibre d'effectifs moins prononcé. Une deuxième solution consistait à expérimenter des méthodes de statistique inférentielle sur ces données, afin de construire un modèle prédictif permettant la quantification du risque de survenue d'un enfant malformé. Nous avons préféré la deuxième solution d'autant qu'il apparaissait plus raisonnable (car moins ambitieux) de vouloir quantifier un risque de survenue d'une maladie, plutôt que de vouloir affecter un individu à un groupe malade ou non malade. Existe-t-il des individus à fort risque et d'autres à très faible risque ? Il est très difficile de séparer les malades des non malades, mais il semble peut-être plus facile de séparer, parmi les non malades, un groupe pour lequel le risque de maladie est très faible. Ce travail de quantification du risque est décrit dans les chapitre III et IV.

Chapitre III

Le risque de survenue d'un enfant déséquilibré à terme : utilisation de la régression logistique

Dans le chapitre précédent, nous avons vu des méthodes statistiques descriptives, certaines classiques, d'autres moins souvent utilisées en médecine, ainsi que les limites de ces méthodes dans le problème qui nous concerne. Dans ce chapitre, nous aborderons la régression logistique, méthode qui devrait permettre d'une part de préciser par des tests le rôle des différentes variables explicatives, et d'autre part de proposer un modèle prédictif pour l'estimation de ce risque. La première partie sera consacrée aux aspects méthodologiques et la stratégie adoptée sera décrite. La deuxième partie exposera les résultats obtenus : description des modèles et résultat prédictif de la modélisation. La dernière partie est destinée à l'interprétation du modèle proposé, en précisant l'influence de certaines variables dans le risque de survenue d'un enfant malformé.

III. 1. Méthode

Au chapitre précédent, nous avons vu les limites des méthodes descriptives pour répondre à la question posée. Une modélisation avec utilisation de méthodes statistiques inférentielles est apparue nécessaire. Nous détaillerons successivement les raisons qui nous ont conduit à utiliser la régression logistique comme outil de modélisation, puis la théorie de la régression logistique et les stratégies utilisées (adéquation et choix du modèle).

III. 1.1. Pourquoi la régression logistique ?

Appartenant aux modèles linéaires généralisés, la régression logistique est un modèle robuste à la non normalité des variables (Hosmer, 1989), contrairement à l'analyse discriminante. Nous avons vu, lors de la présentation des données, que les variables quantitatives explicatives ne présentaient pas une distribution normale. Il en est évidemment de même pour les variables qualitatives impliquées (comme le numéro des chromosomes).

Cette méthode d'analyse multivariée permet de mesurer le rôle d'une variable ajustée sur toutes les autres, et autorise la prise en compte de variables aussi bien qualitatives que quantitatives, ainsi que des termes d'interaction. La régression qualitative avait suggéré cette hypothèse de l'existence d'interactions lors de l'analyse descriptive (ch. II. 2.1).

Cette méthode, très souvent utilisée en épidémiologie sur des données observées, présente l'avantage aussi d'une interprétation simple des paramètres estimés pour chacune des variables explicatives. Cette interprétation est basée sur la mesure du risque relatif, défini par le rapport entre la probabilité d'être malade parmi les sujets "exposés" et la probabilité d'être malade parmi les sujets "non exposés". Les sujets exposés seront par exemple les sujets présentant un chromosome acrocentrique impliqué dans la translocation, et les sujets non exposés ceux présentant un chromosome n'appartenant pas au groupe des acrocentriques impliqué dans la translocation. Pour estimer ce risque relatif on calculera l'*odds ratio* (OR) qui est proche du risque relatif lorsque la maladie est rare.

L'intérêt d'une modélisation par rapport à une analyse descriptive simple réside dans le fait de pouvoir prédire la variable "à expliquer" ou "variable réponse" à l'aide des valeurs des variables "explicatives" ("covariables", "régresseurs"). Connaissant la valeur des variables explicatives d'une nouvelle observation, il est alors possible de déduire, à partir du modèle, la valeur estimée de la variable à expliquer pour cette nouvelle observation.

III. 1.2. La régression logistique

Les hypothèses relatives aux modèles de régression linéaire sont trop restrictives, lorsque la variable réponse observée est binaire (variance non constante, variable réponse estimée comprise entre 0 et 1). Dès 1972, la théorie des modèles linéaires généralisés a été proposée, elle permet d'étendre les notions de la régression linéaire aux structures exponentielles (Nelder, 1972). La régression logistique est un cas particulier des modèles linéaires généralisés. Compte tenu des valeurs théoriquement possibles de l'espérance mathématique de la variable à expliquer, comprises entre 0 et 1 puisqu'il s'agit d'une probabilité, on considérera que chaque individu suit une loi binomiale avec dénominateur égal à 1 (~ B(1, p)). La fonction canonique appliquée aux modèles linéaires généralisés en cas de distribution binomiale est la fonction "logit", et le modèle ainsi défini un modèle de régression logistique.

Soit une variable à expliquer Y et des variables explicatives $X_1, X_2, X_3, ... X_p$

1) Dans le Modèle linéaire classique on cherche à expliquer une variable quantitative à l'aide de variables explicatives quantitatives ou qualitatives

$$E(Y/X_1, X_2, X_3, ...X_p) = \mu \qquad Y = \mu + \varepsilon$$

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = X\beta$$

$$\text{Hypothèses : } E(\varepsilon_i)=0 \ , \ Var(\varepsilon_i)=\sigma^2 \ , \ Cov(\varepsilon_i \varepsilon_j)=0$$

2) Dans le Modèle linéaire généralisé (GLM) la variable à expliquer est qualitative ou quantitative et on cherche à l'expliquer par des variables explicatives qualitatives ou quantitatives.
La loi de probabilité de la variable à expliquer appartient à la famille des lois exponentielles.

$$E(Y/X_1, X_2, X_3, \dots, Xp) = \mu \qquad Y = \mu + \varepsilon$$
$$\eta = g(\mu) = X\beta \qquad var(Y) = \phi\, V(\mu)$$
$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Dans le cas particulier de la régression logistique :

la fonction de lien est $g(\mu) = \log(\mu/(1-\mu))$
la fonction de la variance est $V(\mu) = \mu(1-\mu)$
$p = Pr(Y=1/X_1, X_2, X_3, \dots, Xp)$
et
$1-p = Pr(Y=0/X_1, X_2, X_3, \dots, Xp)$
Y suit une loi binomiale $B(1, p)$
$\text{logit}(p) = \log(p/(1-p)) = \eta \qquad p = e^\eta / 1 + e^\eta$

Le logit(p) est le logit de la probabilité d'être malade, c'est une combinaison linéaire des variables explicatives.
Si l'on considère 3 variables explicatives, dont une quantitative X et 2 qualitatives F à J modalités et G à K modalités, l'écriture détaillée d'un modèle s'effectue alors de la manière suivante.
Pour chaque individu i, $i = 1 \dots n_{jk}$

$$\eta_i^{jk} = \beta + \beta_j^F + \beta_k^G + \beta_{jk}^{FG} + \gamma X_i^{jk} + \gamma_j^F X_i^{jk} + \gamma_k^G X_i^{jk} + \gamma_{jk}^{FG} X_i^{jk}$$

et les contraintes linéaires sur les paramètres peuvent être les suivantes :

$$\beta^F \in R^J \quad \text{et} \quad \beta_1^F = 0$$
$$\beta^G \in R^K \quad \text{et} \quad \beta_1^G = 0$$
$$\gamma^F \in R^J \quad \text{et} \quad \gamma_1^F = 0$$
$$\gamma^G \in R^K \quad \text{et} \quad \gamma_1^G = 0$$
$$\beta^{FG} \in R^{JK} \quad \text{et} \quad \beta_{11}^{FG} = 0$$
$$\gamma^{FG} \in R^{JK} \quad \text{et} \quad \gamma_{11}^{FG} = 0$$

Le traitement d'une variable qualitative à plusieurs modalités nécessite de la coder différemment, ceci pouvant s'effectuer de différentes façons*.

Nous avons retenu la paramétrisation qui consiste à choisir la 1e modalité comme modalité de référence, en raison de sa plus grande facilité d'interprétation.

Exemple : Pour la variable "bras des points de cassure" qui est à 3 modalités ("qq", "pp", "pq"), la modalité "qq" sera considérée comme modalité de référence pour les deux autres. Les modalités bras "pp" et bras "pq" seront 2 variables binaires (oui=1, non=0) . Un paramètre sera estimé pour les modalités "pp" et "pq", mais pas pour la modalité "qq".

III. 1.3. Stratégie

Nous décrivons ici la démarche globale qui sera utilisée par la suite sur nos données. Cette démarche s'appuie sur les tests classiquement utilisés en régression logistique. Certaines contraintes sont induites par le sujet même de l'étude (nature des données) et d'autres sont proposées par les généticiens.

III. 1.3.1. Choix des variables du modèle

1) Pour chacune des variables explicatives, on procède d'abord à **une analyse univariée.** Le score Test mesure le rapport de vraisemblance entre deux modèles : différence des logvraisemblances entre le modèle "vide" (déviance maximum) et le modèle contenant la variable explicative. Il sert donc à mesurer l'influence "significative" ou non d'une variable dans le modèle. Le test de Wald mesure le degré de signification des différents paramètres d'une variable explicative. Dans le cas de variables qualitatives, il peut être utilisé pour choisir le meilleur regroupement de modalités (nombre de modalités et choix de ces modalités). Dans le cas où 2 variables sont très corrélées entre elles, on retient celle qui procure le meilleur gain en déviance.

2) On procède ensuite à **l'analyse multivariée.**

Lors de cette analyse, nous n'avons pas retenu de méthode pas à pas pour les raisons suivantes :

- cette méthode peut conduire à sélectionner des variables non pertinentes (Flack, 1987),

---

* Selon les logiciels utilisés, les méthodes proposées pour le traitement des variables qualitatives sont différentes :
  - **1** modalité de référence définie par la première modalité dans BMDP (partial method), EGRET, et GLIM (méthode de la cellule de reference)
  - **2** modalité de référence définie par la modalité la plus nombreuse dans SAS et SYSTAT
  - **3** modalité de référence définie par la moyenne dans SPLUS (déviation par rapport à la moyenne) et BMDP (méthode marginale)

Avec les logiciels GLIM et SPLUS que nous avons utilisés, nous avons toujours retenu la méthode de référence à la première modalité, méthode par défaut dans GLIM, et en option dans SPLUS.

- il n'existe pas *a priori* de supériorité du pas à pas par rapport à un choix orienté et avisé des variables (Greenland, 1989, Hosmer, 1989),

- disposant des résultats préliminaires de l'analyse descriptive des données concernant les variables chromosomiques les plus pertinentes, et compte tenu de l'existence de certaines variables contraintes à rester dans le modèle, il nous a semblé plus opportun de ne pas procéder à une sélection de variables "à l'aveugle".

La stratégie que nous avons suivie consiste à débuter avec le modèle "plein", pour en exclure petit à petit un certain nombre de variables. Par le score Test, on mesure l'effet de la non prise en compte d'une variable dans le modèle, et ceci pour chacune des variables explicatives successivement. S'il s'avère non significatif, alors la variable est retirée du modèle. Il existe cependant deux exceptions à cette règle :

- une variable pourra être maintenue de force dans le modèle si elle est comprise dans une interaction significative,

- certaines variables seront "forcées" à rester dans le modèle pour des raisons physio-pathologiques sur demande des généticiens.

Pour les variables contraintes à rester dans le modèle, même si elles ne sont pas significatives, on vérifie que leur présence dans le modèle ne provoque pas de changement notable dans l'estimation des paramètres des autres variables du modèle (et de l'erreur standard).

Les interactions d'ordre 2 sont systématiquement testées. Seules les interactions présentant un seuil de signification à 1‰ sont retenues. Les interactions d'ordre supérieur sont également testées afin de mesurer leur importance.

En respectant le principe de parcimonie, on espère, avec cette stratégie, pouvoir obtenir un modèle comportant les variables les plus influentes et le plus petit nombre de paramètres (ceci de manière à gagner en précision dans l'estimation des paramètres et en simplification dans l'interprétation du modèle). Le meilleur modèle "proposé" (dans le cadre de modèles présentant une hiérarchie entre eux) sera celui qui présente la plus petite déviance avec le moins de paramètres possibles (stratégie similaire à celle proposée par Greenland, 1989 et Mac Cullagh, 1989).

III. 1.3.2. Adéquation du modèle aux données

Avant d'accepter une modélisation, plusieurs vérifications sont nécessaires :

- n'existe-t-il pas des valeurs aberrantes de certaines observations dans les données?

- le modèle décrit-il correctement les données ?

Recherche de valeurs aberrantes

Les données étudiées sont des données observées qui peuvent être de mauvaise qualité. Dans la mesure où ces données peuvent avoir une influence erronée sur l'estimation des paramètres, il est important de les repérer. Les observations aberrantes sont détectées par l'analyse des résidus. Plusieurs méthodes de calcul des résidus sont disponibles : résidus de la déviance*, résidus de Pearson, résidus d'Anscombe (Mc Cullagh, 1989). Bien que les résidus de la déviance présentent une distribution plus symétrique que les deux autres, nous avons préféré travailler sur les résidus de Pearson étant donné la grande taille de notre échantillon de données. Tout résidu standardisé supérieur à 2 est considéré comme résidu d'une observation "suspecte" *a priori*. Toute observation "suspecte" est vérifiée avec retour à la source de l'information (article ou laboratoire d'origine). Des corrections sont alors effectuées, s'il s'agit d'erreurs de saisie ou d'interprétation des données. Pour les données vérifiées qui restent aberrantes, nous utilisons le test par délétion : y a-t-il changement dans la valeur des paramètres et de la déviance selon que l'observation à valeur suspecte est incluse dans les données ou non (Pregibon, 1981) ? S'il n'y a pas de modification dans la valeur des paramètres ou de la déviance, nous optons pour le choix de garder cette observation dans les données, dans la mesure où ces valeurs aberrantes peuvent être nécessaires à une estimation non biaisée des paramètres (Jennings, 1986).

Le modèle proposé est-il raisonnable pour décrire les données ?

Si ce n'est pas le cas, les données montrent une déviation systématique par rapport aux valeurs prédites, ou bien quelques données seulement ne collent pas du tout. En l'absence de biais d'échantillonnage, le choix du modèle doit être revu, si les résultats des tests sont en contradiction avec les données ou la connaissance actuelle (Bonneu, 1988).

Les tests d'adéquation d'un modèle logistique aux données ne sont pas très développés : ils sont peu nombreux, longs à effectuer et non disponibles en préprogramme dans les logiciels actuels, contrairement à ce qui se passe pour les tests d'adéquation des modèles de régression linéaire multiple.

Les résidus

En régression linéaire multiple, une configuration non uniforme des résidus résulte soit d'erreurs non distribuées normalement, soit d'observations aberrantes.

En régression logistique, les résidus n'ont pas une distribution bien connue, il s'agit donc de simuler des données de manière à obtenir une estimation de la distribution théorique de ces résidus. Selon Jennings, "chaque résidu aurait sa propre distribution, ce qui complique singulièrement les choses. Les composantes individuelles de la déviance, comme celles du chi2, ont une distribution à deux maximums. Dans tous les cas, les résidus n'ont aucune raison de suivre une distribution du chi2 ou une distribution normale" (Jennings, 1986).

---

* La déviance résiduelle est définie comme 2 fois la différence entre le logarithme du maximum de vraisemblance et le logarithme de la vraisemblance du modèle testé.

Pour vérifier s'il existe ou non une déviation systématique de ces résidus, on utilise le "graphe des probabilités empiriques". C'est un graphique qui représente en abscisse une distribution théorique de résidus (résidus simulés) et en ordonnée les résidus observés obtenus par régression logistique : il doit être linéaire, et permet de détecter des écarts des données par rapport aux hypothèses du modèle logistique (Landwehr, 1984). La distribution théorique est obtenue en simulant K nouvelles variables * à expliquer (binaires 0/1), suivant une loi binomiale ($\sim B(1, \hat{p})$), $\hat{p}$ étant le vecteur des valeurs prédites obtenu à partir des données observées sous le modèle considéré. Sur ces K nouvelles variables simulées, on applique le même modèle de façon à obtenir un vecteur de résidus pour chacune d'elles. K distributions de résidus sont alors obtenues, et on prend la médiane de ces résidus pour obtenir une approximation de la distribution théorique des résidus pour ce modèle. Les étapes et commandes nécessaires à ces simulations figurent à l'annexe 10.

Les résultats de classification

Une première approche consiste à regarder le taux d'erreur apparent du modèle, c'est-à-dire à calculer, sur les observations ayant servi à construire le modèle, le pourcentage d'observations bien classées lorsque l'on applique le risque prédit par le modèle. Ce taux d'erreur apparent est une surestimation de la réalité, puisqu'il teste le modèle sur les mêmes observations ayant servi à le construire.

Plusieurs méthodes sont possibles pour obtenir un taux d'erreur réel du modèle : soit tester le modèle sur de nouvelles observations, soit construire le modèle à partir d'une sélection du fichier global et le tester sur le reste du fichier, ce qui s'apparente à la méthode du jackknife ou cross-validation (Bunke, 1984).

Nous avons retenu la deuxième méthode pour obtenir ce taux d'erreur réel. Trente échantillons randomisés représentant chacun environ 30 % du fichier de données sont considérés comme échantillons de validation. Sur chacun de ces échantillons, on teste le modèle construit à partir des 70 % restants du fichier de données. Le modèle proposé, avec les paramètres estimés à partir de chacun des échantillons de construction, est appliqué à chacun des 30 échantillons de validation, avec calcul du pourcentage d'observations bien classées. On peut également vérifier la robustesse du modèle sur les 30 échantillons aléatoires de 70 % du fichier de données (appelés échantillons de construction), la robustesse étant définie par l'absence de changement dans la valeur des paramètres et de leurs écart-types pour les différents échantillons.

Le taux d'erreur réel n'est pas surestimé, et il permet de tester les capacités de prédiction du modèle. Dans la mesure où de nouvelles translocations seront rentrées progressivement dans la base de données, cela permettra de tester à nouveau le modèle proposé, mais cela nécessitera aussi une correction régulière de l'estimation des paramètres de ce modèle.

---

* Nous avons choisi K=45 de façon à obtenir une distribution théorique suffisamment fiable avec un nombre raisonnable de simulations.

III. 1.4. Odds ratio

Nous traiterons ici surtout de l'interprétation des coefficients "β" du prédicteur linéaire. L'OR , déjà défini ci-dessus, se calcule très simplement dans le cas d'un modèle sans interaction, c'est l'exponentielle du coefficient de la variable. Lorsque le modèle contient plusieurs variables, la mesure de ce risque est ajusté sur les autres variables (c'est à dire qu'il est mesuré en tenant compte de la valeur et du rôle propre des autres variables), et il n'a pas la même valeur qu'en analyse univariée.

Calcul de l'OR (en reprenant la notation de la p.47):

* pour la variable quantitative X, on dira que le risque d'être malade est :
  - multiplié par exp(γ) pour un changement de 1 unité de X ,
  - et qu'il est multiplié par exp(10γ) pour un changement de 10 unités de X . Dans le calcul de l'intervalle de confiance (IC) de l'OR il faudra aussi multiplier par 10 l'écart-type de γ (Hosmer Lemeshow 1989).

* pour la variable qualitative F, le risque d'être malade dû à la modalité j versus la modalité de référence j=1 est multiplié par l'expression $\exp(\beta_j^F)$. Il est important de noter que cette expression de l'OR "$\exp(\beta_j^F)$" dépend en fait des options de recodage des modalités des variables qualitatives. Si l'on avait retenu la "marginal" method de BMDP, l'OR aurait alors pris pour valeur "$\exp(2\beta_j^F)$" et non pas "$\exp(\beta_j^F)$". En fait, le résultat de la valeur de l'OR serait identique, simplement les valeurs des coefficients β_j^F ne seraient pas les mêmes.

Lorsqu'il y a des termes d'interaction, la mesure de ce risque se complique, puisque la définition même de l'interaction signifie que ce risque est variable selon les différentes modalités des variables de l'interaction. Dans ce cas, pour chaque modalité de l'une des deux variables de l'interaction, on calculera les valeurs de l'OR.

Exemples :

*Interaction entre une variable quantitative et une qualitative X * F : pour la modalité j=2 de la variable F, l'OR dû à X=20 versus X=1 sera de 2,5 par exemple; alors que pour la modalité j=3 de la variable F, l'OR dû à X=20 versus X=1 sera de 0,5.*

*Interaction entre deux variables qualitatives F * G : pour la modalité k=2 de la variable G, l'OR dû à la "modalité j" de F versus la "modalité 1" de F prendra une certaine valeur, qui peut être différente de la valeur du même OR pour la modalité k=3 de G.*

En exprimant les résultats sous la forme d'OR, on considérera qu'une variable explicative constitue un facteur de risque de maladie, si la borne minimale de l'intervalle de confiance de son OR est supérieur à la valeur 1 et ne contient pas cette valeur 1. Si l'intervalle de confiance de l'OR ne comprend pas la valeur 1 et lui est inférieur, on dira que la variable constitue un facteur protecteur pour la maladie.
L'intervalle de confiance à 95 % de l'OR est calculé selon la méthode du maximum de vraisemblance, en utilisant la variance du coefficient "β" notée $SE^2(\beta)$.

$$IC = [\exp(\beta - 1{,}96 * SE(\beta)) \ , \ \exp(\beta + 1{,}96 * SE(\beta))]$$

En cas d'interaction, le calcul de l'intervalle de confiance devient plus compliqué (Hosmer-Lemeshow, 1989). Si l'on reprend l'exemple de l'interaction F * G, une expression simplifiée du prédicteur linéaire pour une observation peut en être :

$$\eta_i^{jk} = \beta + \beta_j^F + \beta_k^G + \beta_{jk}^{FG}$$

Pour la modalité k de G, l' OR de F (modalité j versus modalité 1) s'écrira : $\exp(\beta_j^F + \beta_{jk}^{FG})$

La variance de log(OR) s'écrira : $\left(SE(\beta_j^F + \beta_{jk}^{FG})\right)^2 = var(\beta_j^F) + var(\beta_{jk}^{FG}) + 2cov(\beta_j^F, \beta_{jk}^{FG})$

et l'IC : $\left[\exp\left((\beta_j^F + \beta_{jk}^{FG}) - 1{,}96 * SE(\beta_j^F + \beta_{jk}^{FG})\right) , \exp\left((\beta_j^F + \beta_{jk}^{FG}) + 1{,}96 * SE(\beta_j^F + \beta_{jk}^{FG})\right)\right]$

III. 2. Les résultats

III. 2.1. Analyse préliminaire

La variable origine

La modalité "inconnue" de la variable origine n'est pas distribuée de façon identique selon les deux modes de ségrégation alterne et non alterne. L'origine est plus souvent connue lorsqu'il y a eu un mode de ségrégation non alterne. Alors que, globalement, l'information manque dans 25 % des cas, elle n'est absente que dans 1 % des cas si il y a eu ségrégation non alterne ($p < 0,001$). Cette liaison origine-mode de ségrégation est en fait due à un biais de recrutement important : lorsqu'il n'y a pas de conséquence clinique grave dans la descendance, l'identification du parent porteur de la translocation est moins souvent recherchée.

En régression logistique, nous avons observé que :

- les paramètres des différentes modalités de cette variable sont tous significatifs lorsque cette variable est testée isolément,
- la modalité origine paternelle devient non significative lorsque sont incluses d'autres variables dans le modèle,
- après exclusion des observations avec origine inconnue, on retrouve les deux constatations faites ci-dessus (tableau 9).

Tableau 9. Variable "origine" en régression logistique (analyse univariée et multivariée)

	Origine (prise comme seule variable)		Origine + autres variables	
	paramètre	test Wald	paramètre	test Wald
origine :				
maternelle	- 0,395	S	1,520	S
paternelle	- 0,248	S	- 0,148	NS
inconnue	- 2,395	S	- 2,353	S
origine :				
maternelle	- 0,385	S	1,506	S
paternelle	- 0,254	S	- 0,151	NS

Compte tenu du rôle important de la variable origine d'un point de vue cytogénétique, et malgré l'absence de significativité de l'origine paternelle, le choix est fait de contraindre cette variable à rester dans le modèle. Par contre, on ne retiendra pas les observations avec modalité inconnue pour cette variable, et ceci pour les deux raisons suivantes:

- la liaison avec le mode de ségrégation est purement due à un artefact (l'origine est plus souvent recherchée lorsqu'il existe un enfant déséquilibré à terme dans la famille),
- les paramètres restent inchangés selon que ces observations sont incluses ou non pour décrire le modèle.

------> La construction du modèle se fera alors sur un fichier de 5564 observations seulement et non pas de 7383 observations (exclusion des observations avec origine inconnue).

La variable âge du parent

L'information sur l'âge du parent n'est disponible que dans 25 % des cas (28 % pour l'âge maternel et 24 % pour l'âge paternel). Compte tenu de l'influence classiquement reconnue pour certaines anomalies chromosomiques de l'âge avancé de la mère (l'exemple le mieux documenté étant celui de la trisomie 21, Hook, 1981), c'est le rôle de l'âge de la mère que nous avons étudié plus particulièrement.

Aussi bien en analyse univariée qu'en analyse multivariée, on note une relation décroissante entre le risque de survenue d'un déséquilibre à terme et l'âge maternel. Par exemple, dans un modèle de régression logistique comportant l'âge maternel et cinq autres variables, pour une augmentation de 10 ans d'âge maternel, l'OR passe de 1,0 à 0,7. Pour une augmentation de 20 ans, l'OR passe de 1,0 à 0,5. Le risque serait donc, au vu de ce résultat, divisé par 2 pour une femme de 45 ans par rapport à une femme de 25 ans.

Ces résultats sont très contradictoires avec les connaissances physiologiques actuelles et il s'agit d'en comprendre la raison. L'hypothèse que nous émettons est que cette contradiction est due "aux données" et plus précisément au mode de recrutement des observations. Dans notre base de données, la population de mères peut être schématiquement divisée en deux groupes : les jeunes mères qui viennent précocément en consultation de conseil génétique, car elles ont connaissance d'un problème (translocation connue dans la famille, enfant précédent malformé, avortements répétés) et les mères plus âgées qui recourent au diagnostic prénatal de façon plus systématique (conduisant à des découvertes fortuites d'anomalies chromosomiques). Sans être vérifiée, cette hypothèse peut être confortée par l'analyse descriptive qui suit.

Si l'on regarde la distribution de l'âge maternel au sein de chacun des fichiers ayant servi à constituer la base de données, on constate que celle-ci diffère selon ces fichiers, traduisant bien les modes de recrutement différents (tableau 10).

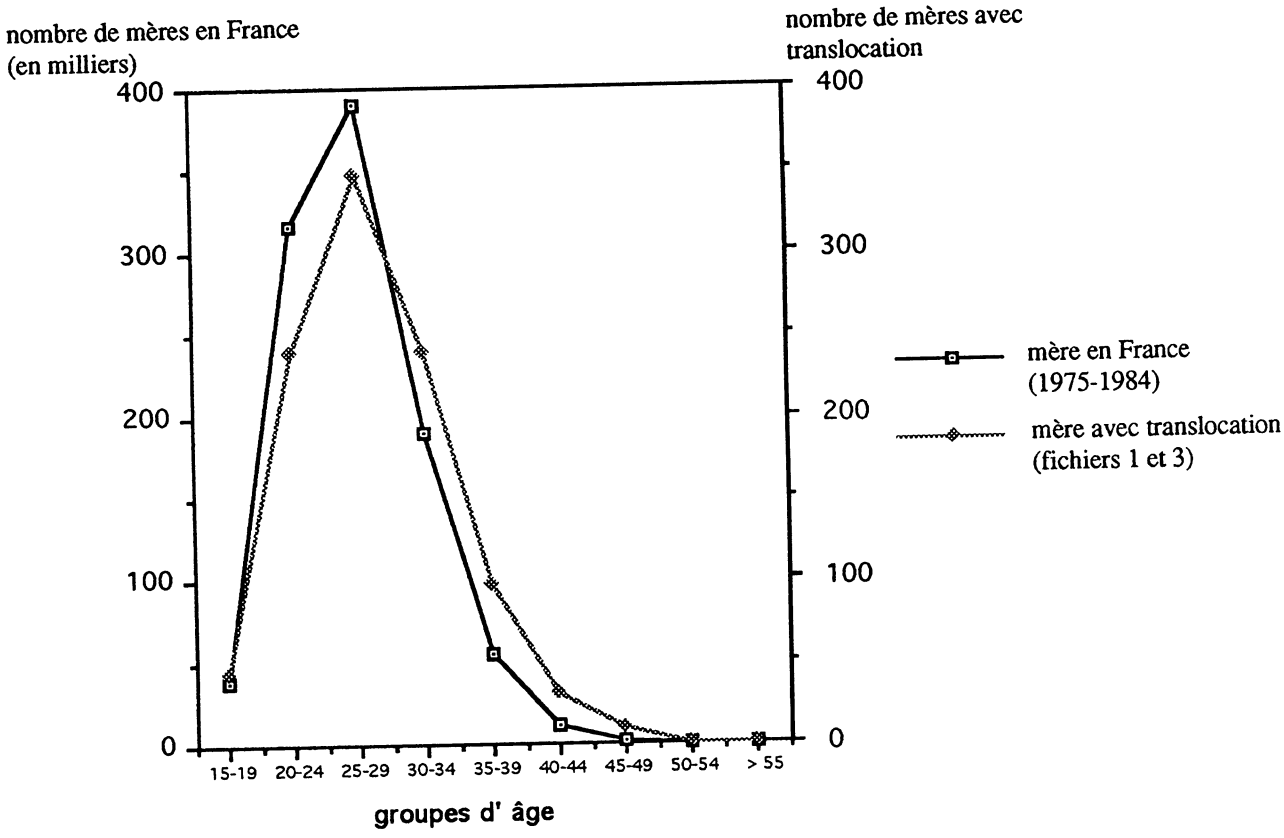
Tableau 10. Répartition de l'âge de la mère (3 groupes d'âge) selon les fichiers *(% en colonnes)*

	fichier 1	fichier 2	fichier 3	total
< 25 ans	139 *(25)*	323 *(38)*	134 *(20)*	596 *(28)*
25 - 37 ans	389 *(69)*	497 *(58)*	493 *(72)*	1379 *(66)*
≥ 38 ans	32 *(6)*	38 *(4)*	54 *(8)*	124 *(6)*
total	n = 560	n = 858	n = 681	n = 2099

L'âge moyen n'est pas le même pour chacun des fichiers : il est plus élevé pour les fichiers 1 et 3 (Analyse de variance : F significatif=21,5 p<0,001).

Si l'on compare cette distribution à celle de l'âge maternel en France (moyenne sur la décennie 1975 à 1984, source INSEE), on note une différence de distribution dans le sens d'une plus forte proportion de mères plus âgées chez les mères présentant une translocation ($p < 0,001$). Le fichier 2, issu de la littérature, présente une distribution très proche de l'âge des procréatrices françaises alors que les fichiers 1 et 3 s'en éloignent avec déviation vers la droite de la courbe pour ces deux fichiers (fig. 8)

Fig. 8. Distribution de l'âge des mères



Si l'on regarde maintenant la liaison "âge maternel - mode de ségrégation", on constate qu'elle existe dans les fichiers 1 et 3 ($p < 0,01$), alors qu'on ne la retrouve pas dans le fichier 2 où le risque de survenue d'un déséquilibre à terme est constant quel que soit l'âge de la mère (tableau 11).

Tableau 11. Proportion de ségrégation non alterne en fonction de l'âge de la mère et des différents fichiers *(en %)*

	fichier 1	fichier 2	fichier 3	total
< 25 ans	31/139 = *22*	206/323 = *64*	36/134 = *27*	*46*
25 - 37 ans	54/389 = *14*	316/497 = *64*	93/493 = *19*	*34*
≥ 38 ans	5/32 = *16*	25/38 = *66*	3/54 = *6*	*27*
total	90/560 = *16*	547/858 = *64*	132/681 = *19*	*37*

Conformément au recrutement des données, ces trois fichiers diffèrent aussi par la répartition de leurs modes de détection (tableau 12). Le mode de détection traduit le signe d'appel qui a permis la découverte de la translocation et il comporte 4 modalités.

Tableau 12. Répartition des modes de détection selon les fichiers *(% sur colonnes)*

	fichier 1	fichier 2	fichier 3	total
systématique	*71*	*40*	*78*	*61*
déséquilibre	*14*	*55*	*11*	*30*
échec reprod.	*10*	*1*	*1*	*4*
fortuit	*5*	*4*	*10*	*5*
effectif	n = 560	n = 858	n = 681	n = 2099

Le mode de détection présente une liaison significative avec l'âge de la mère allant dans le sens d'un plus grand nombre de mode "déséquilibre", plus la mère est jeune, et d'un plus grand nombre de mode "fortuit", plus la mère est âgée, et ceci pour chacun des fichiers. On note 41 %, 29 %, et 9 % de mode "déséquilibre" respectivement pour les 3 groupes d'âge du plus jeune au plus âgé ($p < 0,001$). On note l'inverse pour le mode de détection "fortuit" : 3 %, 5 %, et 12 % respectivement pour les 3 groupes d'âges du plus jeune au plus âgé. Ceci est particulièrement prononcé dans le fichier 3 avec 24 % de mode de détection fortuit chez les mères de plus de 38 ans. Le grand nombre d'amniocentèses pratiquées pour des femmes présentant uniquement un âge maternel avancé (> 38 ans pour la France) est spécifique à ce fichier 3 et peut expliquer cette plus grande proportion de modes de détection "fortuit".

Enfin, le tableau 13 montre la liaison "âge de la mère - mode de ségrégation" pour des modes de détection identiques, et on remarque, pour le mode de détection systématique (environ la moitié des observations), une tendance à une plus forte proportion de ségrégation non alterne lorsque la mère est plus âgée (tendance non significative $p > 0,10$). Cette relation est inversée pour le mode de détection "fortuit" ou bien même n'existe pas du tout (comme c'est le cas pour le mode de détection "déséquilibre").

Tableau 13. Proportion de ségrégation non alterne en fonction du mode de détection et de l'âge *(en %)*

	fortuit	déséquilibre à terme	échec reproduction	systématique	total
< 25 ans	3/16 = *19*	244/244 = *100*	1/19 = *5*	25/314 = *8*	273/593 = *46*
25 - 34 ans	3/50 = *6*	333/333 = *100*	6/43 = *14*	79/802 = *10*	421/1228 = *34*
≥ 35 ans	4/43 = *9*	50/50 = *100*	0/15 = *0*	21/158 = *13*	75/266 = *28*
total	10/109 = *9*	627/627 = *100*	7/77 = *9*	125/1274 = *10*	769/2087 = *37*

Note : les modes de détection "trisomie associée" ne figurent pas dans ce tableau (12 cas).

Au total, une liaison "âge de la mère - mode de ségrégation" existe, mais elle varie avec le mode de recrutement différent des fichiers. Les translocations pour lesquelles un mode de ségrégation non alterne est très rare n'ont le plus souvent aucune expression phénotypique, puisque conduisant à un individu porteur sain ou normal, et elles demeurent inapparentes parfois très longtemps en raison d'un dépistage tardif (où même ne sont jamais dépistées). Celles pour lesquelles un mode de ségrégation non alterne est plus fréquent, sont sujettes à un dépistage précoce lorsqu'elles présentent une forme clinique grave (enfant malformé vivant), et ceci chez des femmes parfois très jeunes dès leur première grossesse. Cet artéfact (présent au niveau des consultations de génétique) est suffisamment important pour masquer l'effet habituel de l'âge de la mère. Comme pour d'autres aberrations chromosomiques on s'attend à ce que les déséquilibres à terme surviennent plus fréquemment chez des mères âgées en raison de la diminution de l'effet de la sélection naturelle avec l'âge. On retrouve, pour les modes de détection systématique, une tendance à une plus forte probabilité de non alterne chez les femmes plus âgées (mais cette tendance n'est pas significative).

En pratique, en raison de ce biais d'échantillonnage, et du grand nombre d'informations manquantes pour l'âge du parent porteur (70 %), il apparaît impossible de vouloir étudier le rôle de la variable âge maternel sur la ségrégation. Cette variable âge de la mère n'est pas liée à beaucoup d'autres variables (on note seulement une liaison avec la variable TSbandeR2). Il est donc peu probable qu'elle puisse être un facteur de confusion important.

La solution qui consisterait à n'effectuer l'analyse avec l'âge que pour les observations avec mode de détection systématique peut difficilement être envisagée, en raison de la subjectivité entourant la variable mode de détection, et du petit nombre d'observations qui seraient alors disponibles (1280).

Les observations faites ici pour l'âge maternel sont aussi valables pour l'âge paternel (liaison "âge du père - mode de ségrégation" allant dans le même sens : plus forte proportion de non alterne chez les pères plus jeunes). Ceci est concordant avec le fait que ces 2 variables présentent un certain degré de corrélation entre elles ($r = 0,52$ pour les 1745 observations où l'âge des deux parents est disponible).

--------> Cette variable âge maternel ne fera donc pas partie des variables pouvant servir à construire le modèle.

III. 2.2. Le modèle proposé

On rappelle que la variable à expliquer est l'existence d'un déséquilibre à terme (1 si oui et 0 si non), variable obtenue par combinaison de deux autres : le mode de ségrégation non alterne et l'histoire naturelle (p.30). Les variables explicatives sont au nombre de 19. En fait, deux modèles (modèle 1 et modèle 2) seront proposés, car les variables explicatives caractérisant la viabilité des déséquilibres potentiels n'ont été disponibles que dans un deuxième temps.

Sur 5564 observations, la régression logistique pour chacune des variables explicatives montre les résultats suivants (tableaux 14-15).

Tableau 14. Variable "chromosome impliqué" en régression logistique (analyse univariée)

	ß	se(ß)	Score-test
NVC à 8 modalités			
t(11,22)	-1,380	0,11	**74,8**
Xme 9	0,248	0,14	7 ddl
Xmes 13,14,15	0,183	0,13	p < 0,001
Xmes 21,22	0,611	0,14	
Xme 4	0,106	0,15	
Xmes 1 à 8	-0,396	0,16	
Xmes 16 à 20	-0,101	0,14	
Xmes 10 à 12	0,032	0,15	
NVC à 6 modalités			
Xmes 1 à 8, 10 à 12, 16 à 20	-1,53	0,06	**66,8**
Xme 9	0,49	0,10	5 ddl
Xmes 13,14,15	0,33	0,09	p < 0,001
Xmes 21,22	0,76	0,10	
Xme 4	0,25	0,12	
t(11,22)	0,15	0,12	
NVC à 5 modalités			
Xmes 1 à 8, 10 à 12, 16 à 20	-1,56	0,05	**41,3**
Xme 9	0,48	0,10	3 ddl
Xmes 13,14,15	0,27	0,09	p < 0,001
Xmes 21,22	0,70	0,10	
t(11,22)	0,17	0,12	

Tableau 15. Autres variables explicatives en régression logistique (analyse univariée)

Variable	ß	se(ß)	Score-test
ORIGINE			**29,9**
paternelle	- 1,50	0,06	1 ddl
maternelle	0,37	0,07	p < 0,001
VIABILITE	0,08	0,01	**36,1**
			1 ddl
			p < 0,001
BRAS			
pp	- 1,32	0,05	**4,8**
qq	0,13	0,09	2 ddl
pq	0,15	0,07	NS
CSbandeR1			**3,7**
si ↗ de 1	- 0,004	0,002	1 ddl
			NS
CSbandeR2			**1,8**
si ↗ de 1	- 0,003	0,003	1 ddl
			NS
TSbandeR1			**68,9**
si ↗ de 1	- 0,036	0,004	1 ddl
			p < 0,001
TSbandeR2			**45,5**
si ↗ de 1	- 0,031	0,005	1 ddl
			p < 0,001

CSbandeR1, CSbandeR2 : longueur en bandes R des segments centriques (1er, 2e chromosome de la translocation)
TSbandeR1, TSbandeR2 : longueur en bandes R des segments transloqués (1er, 2e chromosome de la translocation)

Après cette analyse univariée, nous avons décidé de ne pas retenir les segments centriques car non significatifs et très corrélés aux segments transloqués. La variable chromosome (NVC) sera préférée dans sa forme à 5 modalités seulement, et la variable bras sera réduite à 2 modalités : "pp" versus les autres.

• **les interactions d'ordre 2**

Elles sont testées sur un modèle contenant les 6 variables suivantes : NVC, ORIGINE, VIABILITE, BRAS, TSbandeR1, TSbandeR2. On retrouve 8 interactions d'ordre 2 significatives, parmi lesquelles 5 seulement seront retenues. Le choix des 5 interactions retenues est fait à partir de la signification (Wald test) de chacun des paramètres de ces interactions, ainsi que de la contribution de chacune de ces interactions à la déviance (Score test). On utilise un seuil de significativité à 1‰ pour ces interactions (tableau 16).

Tableau 16. Interactions d'ordre 2 (analyse multivariée, Score test)

	NVC	TSbandeR1	TSbandeR2	ORIGINE	VIABILITE	BRAS
NVC	-	$p < 0,001$	$p < 0,001$	$p < 0,001$	$p = 0,05$	$p = 0,01$
TSbandeR1		-	X	$p = 0,20$	X	$p = 0,02$
TSbandeR2			-	$p < 0,001$	X	$p = 0,07$
ORIGINE				-	$p < 0,01$	$p = 0,15$
VIABILITE					-	$p = 0,07$
BRAS						-

La représentation graphique de ces interactions figure à l'annexe 11.

• **les interactions d'ordre supérieur**

La capacité mémoire du logiciel ne permet pas de tester des interactions à plus de 4 termes. Si l'on teste les interactions d'ordre supérieur à 2, une minorité d'entre elles sont significatives (4 sur 16 pour les interactions d'ordre 3, et 1 sur 5 pour les interactions d'ordre 4, avec un seuil de significativité à 1 ‰). Nous décidons de ne pas les retenir, en raison de la complexité de leur interprétation.

Les résultats de l'analyse multivariée sont présentés dans le tableau 17.

Tableau 17. Variables explicatives en régression logistique (analyse multivariée)

	analyse multivariée (sans interaction)	*Retenues*
TSbandeR1	S	oui
TSbandeR2	S	oui
BRAS		
"pp"	NS	oui
"pq" ou "qq"	NS	*(forcée)*
ORIGINE		
maternelle	S	oui
paternelle	NS	*(forcée)*
NVC		
Xmes 1 à 8, 10 à 12, 16 à 20	S	
Xme 9	S	
Xmes 13,14,15	S	oui
Xmes 21,22	S	
t(11,22)	S	
VIABILITE	S	non

Au total :

- la modalité ORIGINE paternelle n'est plus significative, mais la variable est contrainte à rester dans le modèle,

- la variable VIABILITE n'est plus significative lorsqu'on inclut les termes d'interaction, elle ne sera pas retenue,

- la variable BRAS est contrainte à rester, en raison de sa présence dans une interaction d'ordre 2 significative.

Le modèle proposé est le suivant :

Modèle 1

Il comporte 5 variables et 5 termes d'interaction d'ordre 2.

NVC + TSbandeR1 + TSbandeR2 + BRAS + ORIG + NVC.TSbandeR1 + NVC.TSbandeR2 + NVC.BRAS + NVC.ORIG + ORIG.TSbandeR2

	Value	Std Error	t value	
(Intercept)	0.45	0.17	2.6	
NVC(2)	-1.27	0.31	-4.1	Xme 9
NVC(3)	-1.10	0.28	-3.9	Xmes 13,14,15
NVC(4)	-1.22	0.31	-4.0	Xmes 21,22
NVC(5)	-3.40	0.62	-5.5	t(11,22)
TSbandeR1	-0.10	0.01	-10.8	TSbandeR1
TSbandeR2	-0.12	0.01	-8.9	TSbandeR2
BRAS	-0.43	0.14	-3.2	bras des points de cassure
ORIG	-0.26	0.14	-1.9	origine maternelle (•)
NVC(2)*TSbandeR1	0.08	0.01	5.4	
NVC(3)*TSbandeR1	0.08	0.01	5.6	
NVC(4)*TSbandeR1	0.09	0.02	5.7	
NVC(5)*TSbandeR1	0.05	0.06	0.8	(•)
NVC(2)*TSbandeR2	0.07	0.02	4.2	
NVC(3)*TSbandeR2	0.07	0.01	4.8	
NVC(4)*TSbandeR2	0.05	0.03	1.9	(•)
NVC(5)*TSbandeR2	0.12	0.05	2.5	
NVC(2)*BRAS	0.77	0.29	2.7	
NVC(3)*BRAS	0.07	0.31	0.2	(•)
NVC(4)*BRAS	0.67	0.32	2.1	
NVC(2)*ORIG	0.55	0.23	2.4	
NVC(3)*ORIG	0.12	0.21	0.6	(•)
NVC(4)*ORIG	0.86	0.21	4.0	
NVC(5)*ORIG	1.99	0.51	3.9	
ORIG*TSbandeR2	0.04	0.01	3.2	

Déviance nulle : 5718 avec 5563 degrés de liberté

Déviance résiduelle : 5337 avec 5539 degrés de liberté

(•) paramètre non significatif

Note : Il n'y a pas de paramètre estimé pour NVC(5)*BRAS, car il s'agit de l'équation t(11,22)(q,q), pour laquelle la variable BRAS (modalité "pp") vaut 0.

Ce modèle comporte 24 paramètres, en plus de la constante, dont 19 sont significatifs. Les valeurs estimées du risque de survenue d'un déséquilibre à terme sont comprises entre 0,001 et 0,496. Il existe 169 observations où cette valeur est ≥ 0,40.

Dans un deuxième temps, d'autres variables ont été introduites dans le modèle car devenues disponibles de façon automatique : il s'agit du mode de ségrégation non alterne le plus probable et des variables décrivant la viabilité des gamètes (décrites en p.25). En procédant de la même manière que pour le modèle précédent (dans la sélection des variables et des interactions à retenir), on a obtenu un deuxième modèle qui comporte 10 variables et 3 interactions d'ordre 2.

Modèle 2

BRAS + NVC + ORIG + VIA47 + LAMIN + SINUSADJ + CSbandeR1 + TSbandeR1 + TSbandeR2+ MOSEG + NVC.TSbandeR1 + NVC.ORIG + ORIG.TSbandeR2

	Value	Std Error	t value	
(Intercept)	-2.11	0.87	-2.42	
BRAS	-0.16	0.10	-1.61	bras des points de cassure (•)
NVC(2)	-0.24	0.24	-1.00	Xme 9 (•)
NVC(3)	-0.19	0.22	-0.86	Xmes 13,14,15 (•)
NVC(4)	-0.67	0.25	-2.73	Xmes 21,22
NVC(5)	-2.33	0.56	-4.18	t(11,22)
ORIG	-0.23	0.14	-1.65	origine maternelle (•)
VIA47	-0.17	0.05	-3.421	viabilité à 45 Xmes
LAMIN	-0.23	0.11	-1.99	longueur min adj
SINUSADJ	3.36	0.89	3.79	sinus adj
CSbandeR1	-0.01	0.004	-2.39	CSbandeR1
TSbandeR1	-0.09	0.01	-6.46	TSbandeR1
TSbandeR2	-0.09	0.02	-5.66	TSbandeR2
MOSEG(2)	1.79	0.30	6.05	mode de ségrégation adj 2
MOSEG(3)	0.51	0.15	3.44	mode de ségrégation 3:1 tert
MOSEG(4)	0.02	0.35	0.05	mode de ségrégation 3:1 éch (•)
NVC(2)*TSbandeR1	0.03	0.02	1.76	(•)
NVC(3)*TSbandeR1	0.05	0.01	3.71	
NVC(4)*TSbandeR1	0.07	0.02	3.63	
NVC(5)*TSbandeR1	0.05	0.04	1.07	(•)
NVC(2)*ORIG	0.58	0.23	2.53	
NVC(3)*ORIG	0.09	0.20	0.44	(•)
NVC(4)*ORIG	0.80	0.21	3.72	
NVC(5)*ORIG	2.00	0.51	3.96	
ORIG*TSbandeR2	0.04	0.01	3.07	

Déviance nulle : 5718 avec 5563 degrés de liberté

Déviance résiduelle : 5303 avec 5539 degrés de liberté

(•) paramètre non significatif

Avec le même nombre de paramètres à estimer, on a gagné 35 points en déviance et aussi de la simplicité en interprétation (moins de termes d'interaction). Parmi les 24 paramètres à estimer en plus de la constante, 16 sont significatifs. Les valeurs estimées du risque de survenue d'un déséquilibre à terme sont comprises entre 0,000 et 0,751. Il existe 183 observations où cette valeur est ≥ 0,40 et 29 observations où elle est ≥ 0,50.

III. 2.3. Adéquation du modèle proposé

Recherche des valeurs aberrantes (sur le modèle 1).

Sur le fichier global de 5564 observations, 94 résidus standardisés (environ 1,7 %) ont une valeur > 2,50 (ce sont tous des résidus positifs d'enfants nés avec un déséquilibre à terme). Parmi eux, 16 ont une valeur > 4. Leurs caractéristiques sont présentées au tableau 18.

Tableau 18. Résidus standardisés de Pearson > 4 (Modèle 1)

Numéro de famille	CS bande R1	CS bandeR2	TS bandeR1	TS bandeR2	origine	Chromosomes impliqués	Valeur du résidu
205	35.6	2.75	8.6	14.0	pat	11 - 22	5.20
333	69.25	25.9	2.25	25.13	pat	2 - 4	4.03
434	42.5	5.5	1.75	11.25	pat	11 - 22	4.30
440	42.5	11.1	1.75	5.6	pat	11 - 22	4.32
499 (2)	30.5	20.0	20.5	23.0	mat	5 - 10	6.42
566	33.0	12.25	18.0	26.75	pat	4 - 12	9.78
714	39.25	12.25	3.5	26.75	pat	8 - 12	4.73
793	10.6	3.5	23.6	26.25	pat	9 - 13	4.02
798	47.0	35.25	26.25	3.75	mat	1 - 12	4.05
805	35.6	2.75	8.6	14.0	pat	11 - 22	5.20
821	50.5	32.75	21.0	10.25	pat	2 - 10	4.31
963	10.6	0.88	23.6	27.9	pat	9 - 15	4.20
965 (2)	10.6	0.88	23.6	27.9	pat	9 - 15	4.20
1149	40.13	17.5	10.88	25.5	mat	5 - 10	4.64
	m = 31.0	m = 12.8	m = 14.9	m = 18.3			m = 5.00
moyenne sur le fichier global	M = 38.1	M = 20.4	M = 11.2	M = 9.5			

Tous ces cas proviennent de la littérature, excepté le cas 205 recensé lors d'une consultation génétique et le cas 1149 recensé lors d'une consultation de diagnostic prénatal.
(2) signifie qu'il y a 2 observations dans cette famille avec résidu > 4.

L'observation 566 présente la plus forte valeur de résidu (9,78), soit environ 10 fois supérieure à celle des autres résidus. Ces observations à fort résidu sont des translocations pour lesquelles il y a eu naissance d'un enfant déséquilibré à terme, alors que le modèle prédit un risque de déséquilibre très faible (entre 0,013 et 0,058). Dans le modèle 2, il existe en plus un résidu élevé mais négatif (< -4) pour la famille 219, il s'agit d'une translocation pour laquelle on n'a pas

observé de déséquilibre à terme, alors que le modèle prédit un risque très élevé de déséquilibre (0,741).

Une vérification des données de la base pour ces forts résidus montre que :

1) **pour la famille 566 :** l'information de l'article est en fait trop succinte pour pouvoir conclure et permettre de la rentrer dans la base de données. C'est le cas d'une fille de 15 ans présentant une malformation congénitale sans retard mental associé, et pour laquelle la vérification du caryotype est impossible : la présence de malformation est-elle due à la translocation de son père ? En concertation avec les généticiens, il est décidé de supprimer ce cas de la base de données.

2) **pour la famille 205 :** une erreur de saisie s'est produite, il s'agit en fait d'une translocation d'origine maternelle et non pas paternelle. Une correction est effectuée, qui entraîne la disparition de la valeur élevée d'un résidu pour cette famille.

3) **pour les familles 107, 499 et 875 :** il s'agit de familles pour lesquelles une vérification du caryotype n'a pas été possible. Ces déséquilibres à terme se situent en dehors des critères de viabilité cités par Daniel, ce qui a conduit les généticiens à vouloir en vérifier le caryotype. Ces observations sont donc maintenues dans la base, mais avec la caractéristique suivante : caryotype non vérifié.

4) **pour la famille 798 :** le déséquilibre à terme observé est un enfant décédé à 5 heures de vie, donc à la limite de la viabilité et probablement à rapprocher des enfants mort-nés considérés comme non déséquilibrés à terme dans la variable à expliquer.

Pour les observations à fort résidus, on procède au test par délétion.

- Si on supprime une à une chaque observation avec résidu > 4 , la différence de déviance obtenue est le plus souvent de l'ordre de 6 ou 7 pour 1 ddl. Le pourcentage de changement de déviance est faible, de l'ordre de 1‰, et le nombre de paramètres non significatifs reste inchangé.

- Si on supprime globalement toutes les observations avec résidus > 4, la différence de déviance est de l'ordre de 50 pour 1 ddl (après avoir soustrait la différence due à la diminution de la déviance du modèle nul). Le pourcentage de changement de déviance est de l'ordre de 1 %. Les paramètres significatifs et non significatifs restent les mêmes, ainsi que leur signe, sauf pour 1 paramètre sur 25 (signe inversé).

-------> En raison de ces résultats, on décide de garder ces observations à fort résidu pour construire le modèle.

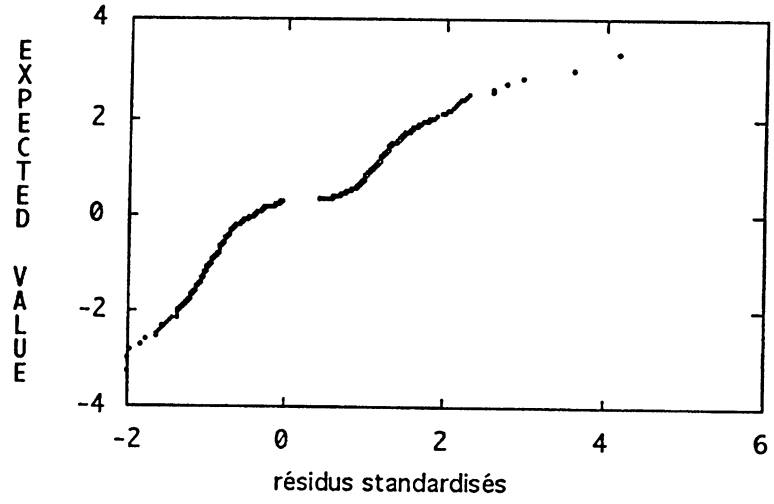
Adéquation du modèle (sur le modèle 1).

Lorsqu'on représente la fonction de répartition des résidus observés (fig. 9), celle-ci apparait bimodale (comme si deux distributions normales étaient accolées). Le graphe de la fonction de répartition des résidus standardisés montre bien ces deux distributions des résidus correspondant à

1) Yobs - Pr(y=1/x) pour les résidus positifs.

2) Yobs - Pr(y=0/x) pour les résidus négatifs.

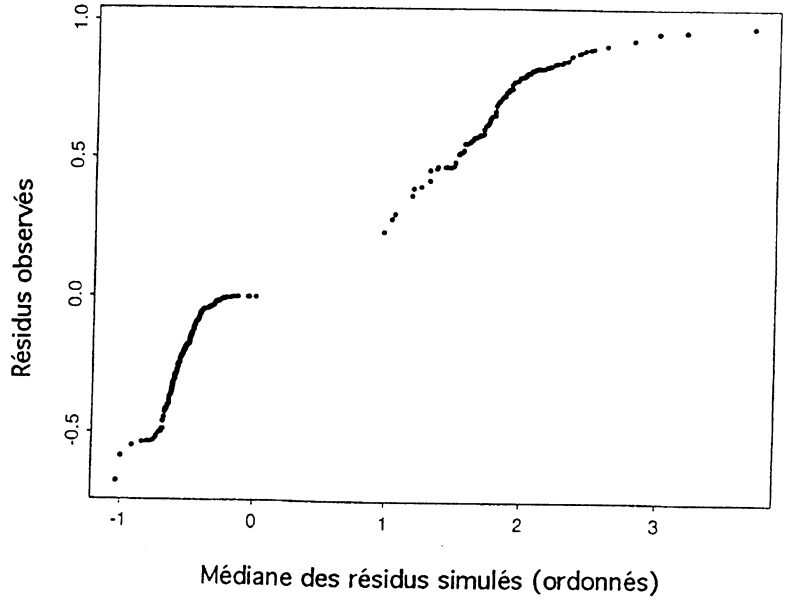
Fig. 9. Fonction de répartition des résidus standardisés de Pearson (modèle 1)



On note une absence ou presque de résidus de valeur comprise entre 0 et 0,5 ce qui traduit l'absence d'observations pour lesquelles la probabilité de déséquilibre est très forte (> 0,5).

Selon la méthode de Landwehr (Landwehr, 1984), on a simulé 45 variables à expliquer à partir du modèle 1 (cf méthodologie p.50 et annexe 10). Le graphe des résidus observés contre le vecteur "médiane" des résidus simulés montre la distribution théorique des résidus sous le modèle 1 (fig. 10).

Fig. 10. Graphe des résidus obtenu après simulations



Sur ce graphe on ne note pas de déviation systématique des résidus.

La robustesse du modèle a été testée sur le modèle 2.

Trente échantillons de 3900 observations environ (soit 70 % des données) ont été construits aléatoirement (générateur de nombres au hasard de Splus, branché sur l'horloge de l'ordinateur). On s'interroge sur les 2 points suivants :

1- est-ce que le modèle choisi est bien robuste ?

Chacun des 10 modèles (même prédicteur linéaire sur 10 échantillons différents) comprend 25 paramètres, en plus de la constante. On compare la valeur de chaque paramètre pour le modèle 2 (sur 5564 observations) et pour les modèles obtenus sur ces 10 échantillons (tableau 19).

Tableau 19. Robustesse du modèle proposé (Modèle 2)

Paramètre	Modèle 2 : valeur du coefficent		10 modèles - "échantillons" : [valeurs]
Constante	S	- 2,11	2 fois NS [-3,10 à -0,40]
BRAS	NS	- 0,16	2 fois S [-0,26 à 0,08]
NVC (2)	NS	- 0,24	1 fois S [-0,59 à -0,10]
NVC (3)	NS	- 0,19	toujours S [-0,37 à -0,05]
NVC (4)	S	- 0,67	3 fois NS [-0,69 à -0,39]
NVC (5)	S	- 2,33	toujours S [-3,00 à -2,00]
ORIGINE	NS	- 0,23	1 fois NS [-0,40 à -0,14]
VIA47	S	- 0,17	toujours S [-0,22 à -0,14]
LAMIN	S	- 0,23	6 fois NS [-0,39 à -0,11]
SINUSADJ	S	3,35	1 fois NS [3,10 à 4,51]
CSbandeR1	S	- 0,01	6 fois NS [-0,015 à -0,006]
TSbandeR1	S	- 0,09	toujours S [-0,10 à -0,08]
TSbandeR2	S	- 0,09	toujours S [-0,10 à -0,07]
MOSEG (2)	S	1,79	toujours S [1,57 à 2,05]
MOSEG (3)	S	0,51	toujours S [0,42 à 0,61]
MOSEG (4)	NS	0,02	toujours NS [-0,25 à -0,42]
NVC(2)*TSR1	NS	0,03	2 fois S [0,01 à 0,05]
NVC(3)*TSR1	S	0,05	toujours S [0,04 à 0,06]
NVC(4)*TSR1	S	0,07	toujours S [0,04 à 0,08]
NVC(5)*TSR1	NS	0,05	toujours NS [-0,02 à 0,08]
NVC(2)*ORIG	S	0,58	4 fois NS [0,42 à 0,73]
NVC(3)*ORIG	NS	0,09	toujours NS [-0,12 à 0,33]
NVC(4)*ORIG	S	0,80	toujours S [0,57 à 0,98]
NVC(5)*ORIG	S	2,00	toujours S [1,81 à 2,46]
ORIG*TSR2	S	0,04	1 fois NS [0,02 à 0,05]

Les différences sont principalement

- au niveau de NVC, qu'il importe de conserver étant donné sa présence dans deux interactions du modèle,

- et de CSbandeR1 et LAMIN, qui ne semblent pas apporter beaucoup au modèle (faible contribution à la déviance, que ce soit dans le modèle 1 ou dans le modèle 2).

De façon globale, on peut conclure que le modèle est suffisamment robuste, puisque la majorité des paramètres garde une valeur proche avec un test de Wald équivalent.

2- est-ce que le modèle choisi est bien le meilleur ?

Il est impossible de répondre à la question **"le modèle est-il le meilleur ?"**. Cependant, sur un des échantillons à 3929 observations, nous avons testé si d'autres modèles ne seraient pas meilleurs que le modèle proposé ou tout du moins équivalents. En reprenant la même démarche que précédemment (décrite pour le modèle 1 p.48-49), on obtient trois modèles pratiquement équivalents en déviance, dont le modèle 2. La différence entre les deux autres modèles et le modèle 2 consiste dans la prise en compte d'une interaction en plus ou en moins (ORIGINE*MOSEG) et dans la prise en compte ou non de CSbandeR1. La différence en déviance n'étant pas significative entre ces modèles et le modèle 2, ce dernier sera préféré en raison de sa simplicité et de sa plus grande proportion de paramètres significatifs.

III. 2.4. Résultats de classification

Peut-on tester la valeur prédictive des 2 modèles proposés ? Autrement dit, quelles sont les observations bien classées en prenant une "valeur - seuil" de discrimination pour la valeur prédite ?

Exemple : en prenant pour seuil f = 0,10, si f ≥ 0,10, on classera l'observation en déséquilibre à terme et si f < 0,10 on la classera en "non malade".

- **Le fichier comporte 5564 individus, dont 21 % sont "malades".**

Les résultats du taux d'erreur apparent sont assez semblables pour les deux modèles 1 et 2 (tableau 20).

Tableau 20. Taux d'erreur apparent (modèles 1 et 2)

risque prédit	Modèle 1		Modèle 2	
Proportion d'observations "non malades" bien classées				
< 0,01	52/53	*98,1 %*	35/35	*100 %*
0,01 ≤ - < 0,05	370/379	*97,6 %*	402/411	*97,8 %*
0,05 ≤ - < 0,10	431/463	*93,1 %*	514/550	*93,5 %*
	895		996	
0,10 ≤ - < 0,15	515/597	*86,3 %*	583/672	*86,8 %*
0,15 ≤ - < 0,30	2407/3110	*77,4 %*	2155/2809	*76,7 %*
Proportion d'observations "malades" bien classées				
0,30 ≤ - < 0,50	341/962	*35,4 %*	362/1058	*34,2 %*
≥ 0,50	0		18/29	*62,1 %*

On peut remarquer la supériorité du modèle 2, non pas en terme de taux d'erreur plus faible, mais en terme de nombre d'observations à faible risque bien classées. Le choix des valeur seuils s'est fait en concertation avec les généticiens et dépend étroitement de l'utilisation que ces derniers veulent faire de ces chiffres. On reviendra sur ce point au chapitre V.

Il existe une autre manière de présenter ces résultats de classification, en déterminant à quel rang se situe la 1^{e} observation mal classée. En partant des valeurs prédites les plus faibles et en les rangeant par ordre croissant, on obtient pour le modèle 2 les observations mal classées suivantes :

- la 1^{e} (fam 566) à la 56^{e} observation,
- la 2^{e} (fam 499) à la 150^{e} observation,
- la 3^{e} (fam 805) à la 294^{e} observation.

On peut remarquer que la 1^{e} observation mal classée doit en fait être exclue des données (p.66), et que la 2^{e} fait partie des observations particulières à caryotype non vérifié. On peut aussi exprimer ces résultats en disant que 293 observations à valeur prédite très faible sont en fait bien classées, soit 5,3 % des observations.

• **Le taux d'erreur réel** est calculé à partir de 30 échantillons composés de 30 % des observations du fichier de données, ces observations ne devant pas servir à construire le modèle. Les paramètres du modèle sont estimés à partir d'un fichier contenant 70 % des observations, et ce, pour chacun des 30 échantillons (méthodologie p.51) (tableau 21).

Tableau 21. Taux d'erreur réel et son IC (méthode jacknife sur le modèle 1)

risque prédit	Taux d'erreur réél (moyenne sur les 30 échantillons)	IC à 0,05
Proportion d'observations "non malades" bien classées		
< 0,01	98,8 %	[93,8 - 100,0]
0,01 ≤ - < 0,05	96,7 %	[94,5 - 98,2]
0,05 ≤ - < 0,10	93,1 %	[88,6 - 96,5]
Proportion d'observations "malades" bien classées		
≥ 0,30	34,1 %	[30,6 - 39,1]

Compte tenu des effectifs importants des échantillons, les taux d'erreur apparent et réel sont très proches avec une légère surestimation des capacités du modèle pour le taux d'erreur apparent.

• **Si l'on regarde maintenant la précision de la prédiction** (IC de la valeur du risque prédit), les résultats montrent une précision correcte pour les risques faibles, tandis que, pour les risques élevés, l'intervalle de confiance est plus large, et ceci indépendamment du nombre d'individus dans la famille ou pour la même translocation. Nous verrons dans la discussion comment ces résultats peuvent être exploités compte tenu de leur précision respective.

Exemples :

1) translocation t(6;9)(q27;p22) de la famille 1359 comportant 11 individus, dont 4 sont des déséquilibres à terme et 2 sont d'origine inconnue.

valeur prédite du risque : 49,7 % IC : [42,2-57,3] si l'origine est maternelle
38,9 % IC : [29,5-48,4] si l'origine est paternelle

2) translocation t(1;8)(q41;q23) comportant 42 individus appartenant à 4 familles différentes, avec 8 individus pour lesquels l'origine est inconnue. Il n'existe aucun déséquilibre à terme.

valeur prédite du risque : 14,6 % IC : [12,3-17,0] si l'origine est maternelle
13,4 % IC : [10,9-15,9] si l'origine est paternelle

3) translocation t(1;16)(p11;q11.2) de la famille 1062 comportant 6 individus, dont 1 d'origine inconnue et les autres d'origine maternelle. Il n'existe aucun déséquilibre à terme.

valeur prédite du risque : 0,09 % IC : [0,0 - 0,2] si l'origine est maternelle

III. 3. Interprétation du modèle

L'intérêt du modèle logistique est de pouvoir préciser l'influence spécifique de chacune des variables sur la survenue d'un déséquilibre à terme. C'est ce que nous allons voir dans ce paragraphe en utilisant la notion de l'odds ratio définie précédemment (p.51-52).

III. 3.1. L'origine de la translocation

Cette variable intervient de façon isolée dans le modèle, mais aussi en interaction, d'une part avec les groupes de chromosome et d'autre part avec la longueur des TSbandeR2.
Globalement, on observe que l'origine maternelle de la translocation constitue un facteur de risque pour la survenue d'un déséquilibre à terme par rapport à une origine paternelle de la translocation (risque plus grand de déséquilibre, si l'origine est maternelle). Ce facteur de risque peut être plus ou moins important en fonction des chromosomes impliqués dans la translocation et en fonction de la longueur de TSbandeR2.

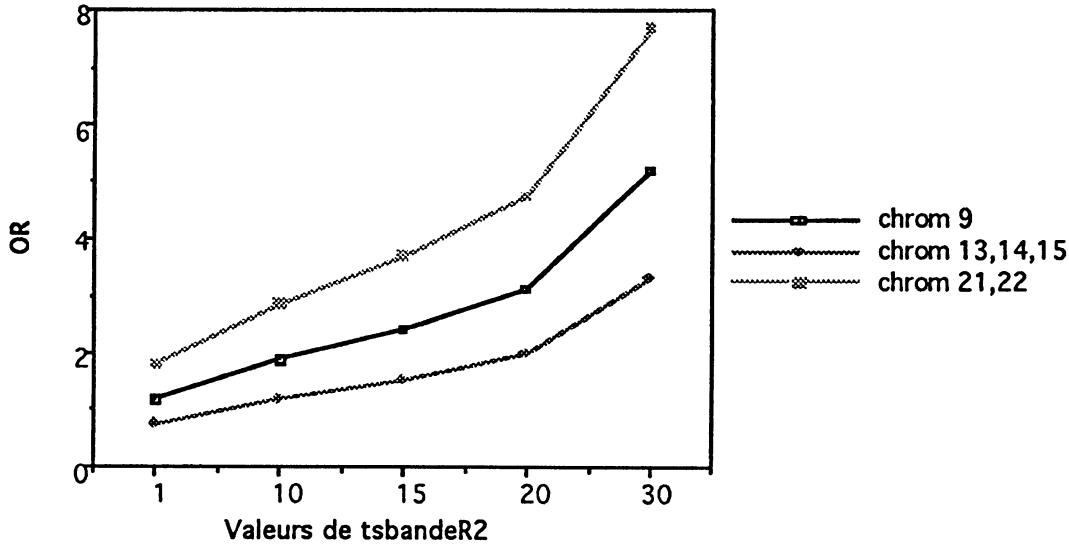
• **En cas d'origine maternelle,** le risque de déséquilibre à terme pour une t(11;22) est 10 fois plus grand que en cas d'origine paternelle. Par contre, pour une translocation impliquant un chromosome 13,14, ou 15, le risque de déséquilibre à terme est seulement 1 fois et demi plus grand en cas d'origine maternelle qu'en cas d'origine paternelle.

Ce facteur de risque est le plus important par ordre décroissant pour :

- les t(11;22) (risque multiplié par 10)
- les translocations impliquant les chromosomes 21 et 22 (risque multiplié par 4)
- les translocations impliquant le chromosome 9 (risque multiplié par 2)
- les translocations impliquant les chromosomes 13,14,15 (risque multiplié par 1,5).

• **Ce facteur de risque** est d'autant plus important que la longueur de TSbandeR2 augmente : lorsque ce segment passe de la taille 10 à 30, le risque est 2 fois plus grand. Cette influence de la longueur de TSbandeR2 existe quels que soient les chromosomes impliqués dans la translocation (fig. 11).

Fig. 11. Risque de déséquilibre à terme pour l'origine maternelle vs l'origine paternelle



L'existence d'un risque plus élevé en cas d'origine maternelle est une notion bien connue. Mais l'information apportée ici est plus précise, puisqu'elle montre que l'influence de cette origine maternelle n'est pas la même selon la translocation en cause (il existe des translocations où cette influence est très importante, et des translocations où elle est de moindre importance).

III. 3.2. Les chromosomes impliqués dans la translocation

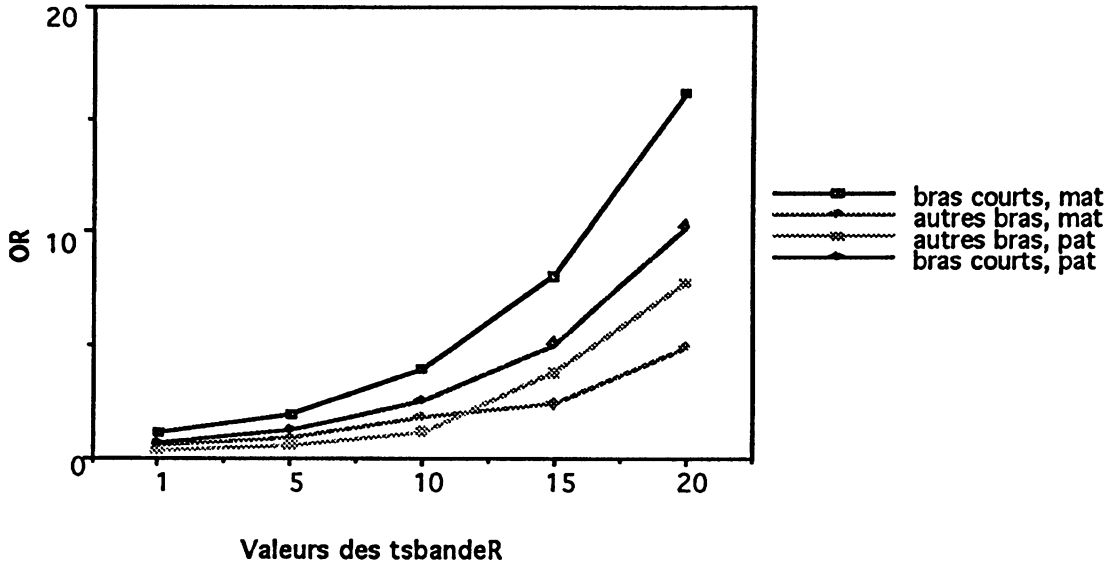
En ayant utilisé la variable NVC dans sa forme à 5 modalités, on avait déjà voulu souligner l'opposition qui existe entre des groupes de chromosomes "neutres" (les 1 à 8, 10 à 12, et 16 à 20) et des groupes de chromosomes "influents" (le 9, les 13,14,15, les 21,22 et la t(11;22)).

Le risque de déséquilibre à terme dû aux chromosomes impliqués sera donc toujours exprimé par rapport aux chromosomes 'neutres' (modalité de référence). De plus, comme pour la variable origine, la variable NVC existe aussi dans des interactions du modèle : ORIG*NVC et TSbandeR1*NVC. Il faut donc tenir compte de l'existence de ces interactions dans l'interprétation de son influence.

- **Risque dû au chromosome 9**

Pour une valeur moyenne de TSbandeR1, le risque de survenue d'un déséquilibre à terme est 2 à 5 fois plus grand si la translocation comporte un chromosome 9, que s'il s'agit d'une translocation impliquant des chromosomes "neutres" (fig. 12).

Fig. 12. Risque de déséquilibre lorsque la translocation implique le chromosome 9 vs les autres chromosomes

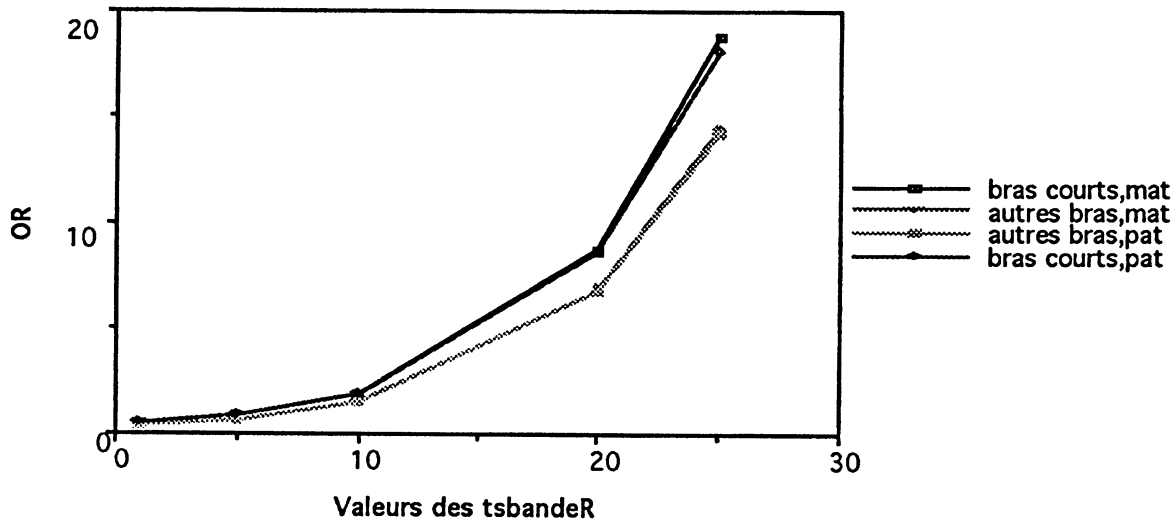


Ce risque augmente avec la valeur de TSbandeR1 : pour un segment passant de la valeur 10 à la valeur 20, il est multiplié par 3. Autrement dit, si ce risque vaut 3 ou 5 pour une valeur de TSbandeR1 de 10, il vaudra 9 ou 15 pour une valeur de TSbandeR1 de 20.

• **Risque dû aux chromosomes 13,14,15**

Plus les segments transloqués sont grands et plus l'influence de ce groupe de chromosomes est importante. Pour des segments transloqués de longueur égale à 11 (valeur moyenne sur l'échantillon), le risque de survenue d'un déséquilibre dû à la présence d'un de ces chromosomes dans la translocation est multiplié par 2, alors que, si les segments transloqués atteignent la valeur de 20, ce risque est multiplié par 8 ou 9 (fig. 13).

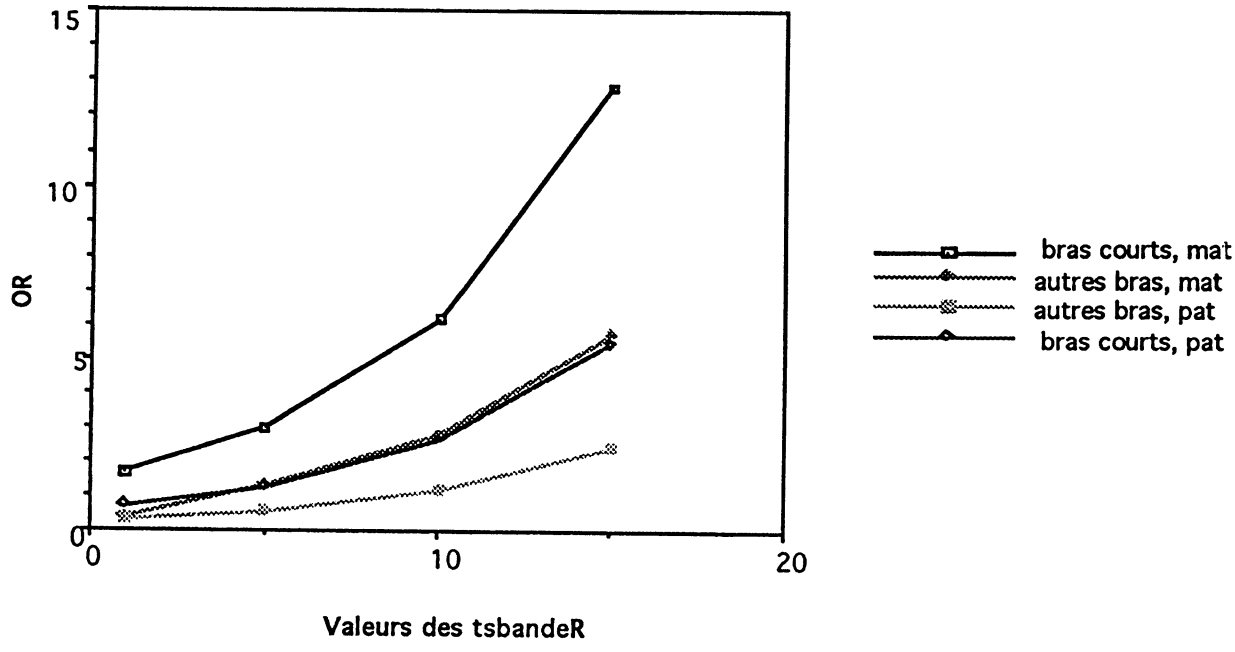
Fig. 13. Risque de déséquilibre lorsque la translocation implique les chromosomes 13, 14 ou 15 vs les autres chromosomes



• **Risque dû aux chromosomes 21 et 22**

La même influence est constatée, c'est-à-dire une augmentation du risque avec l'augmentation de la longueur des segments transloqués. Mais cette augmentation est encore plus sensible pour une origine maternelle que pour une origine paternelle (fig. 14).

Fig. 14. Risque de déséquilibre lorsque la translocation implique les chromosomes 21 ou 22 vs les autres chromosomes



• **Risque particulier dû à la t(11;22).**

Pour cette translocation, le risque de déséquilibre est le même que celui observé pour les chromosomes "neutres" (OR compris entre 0,1 et 2,0). C'est-à-dire qu'il n'y a pas plus de risque d'avoir un déséquilibre, si on a une t(11; 22) plutôt qu'une autre translocation. Par contre, le fait remarquable pour cette translocation, déjà cité plus haut, est le risque très augmenté de déséquilibre à terme lorsque l'origine est maternelle. Pour une descendance d'un père porteur d'une t(11,22), il n'y a pas plus de risque de déséquilibre à terme que pour des translocations impliquant les chromosomes 1 à 8, 10 à 12 ou 16 à 20.

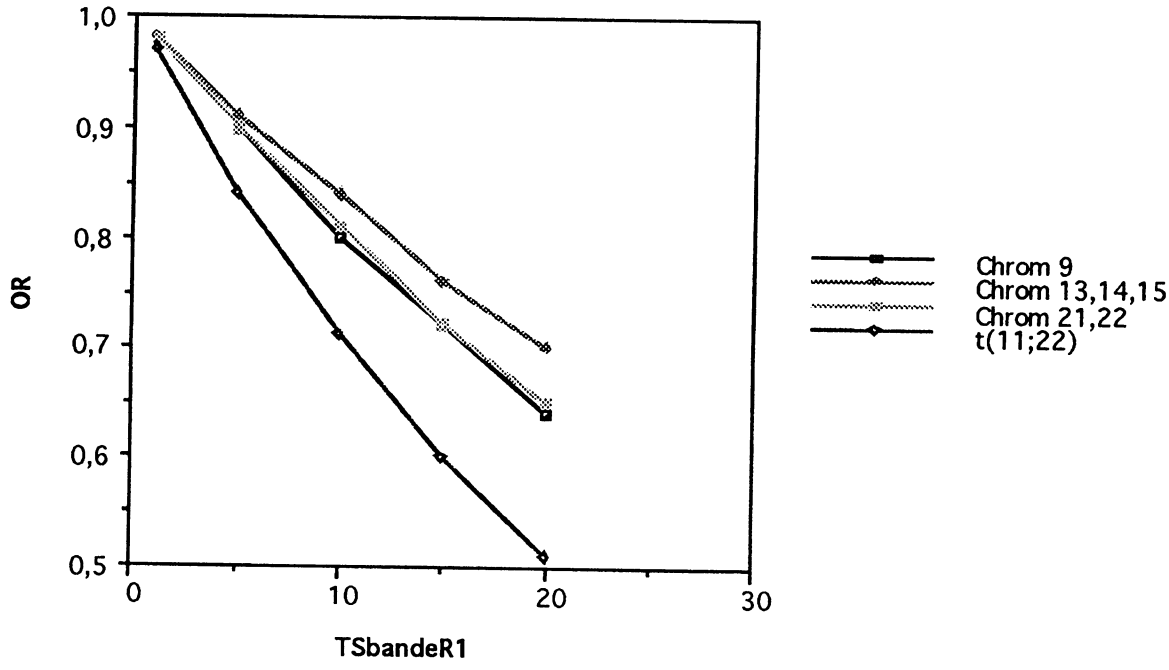
III. 3.3. Les segments transloqués

Comme les variables précédentes, ces segments sont aussi impliqués dans des interactions : avec les groupes de chromosomes, et avec l'origine de la translocation.

• On observe que des petits segments transloqués sont en faveur d'un déséquilibre à terme, alors que des grands segments transloqués auront plutôt un effet inverse favorisant un enfant normal ou un déséquilibre non viable. Cette constatation est tout à fait en accord avec les données de viabilité: plus les segments sont petits et plus le déséquilibre aura de chance d'être viable, plus les segments sont grands et moins le déséquilibre aura de chance d'être viable.

• Cet effet des grands segments transloqués se retrouve de façon identique pour tous les chromosomes impliqués. Lorsque l'on regarde le rôle d'un chromosome par rapport au risque de déséquilibre à terme, on observe que le risque dû à ce chromosome augmente avec la longueur des segments transloqués (traduisant l'influence de la longueur de ces segments sur le mode de ségrégation non alterne et donc sur l'apparition d'un déséquilibre). Par contre, si l'on regarde le risque dû à la longueur d'un segment transloqué, quel que soit le chromosome impliqué, on observe que l'existence de grands segments transloqués ne favorise pas la survenue d'un déséquilibre : le déséquilibre étant important, il sera probablement non viable et entraînera plus volontiers un échec de la reproduction (précoce ou tardif) (fig. 15).

Fig. 15. Risque de déséquilibre lorsque TSbandeR1 augmente



• On observe qu'en cas d'origine paternelle, le risque de déséquilibre à terme diminue régulièrement avec l'augmentation de la longueur des segments transloqués, tandis qu'en cas d'origine maternelle, le risque diminue également pour des petits segments transloqués (jusqu'à une valeur de 10), par contre il ne diminue plus pour des segments transloqués de valeur supérieure à 10. Cette influence particulière des grands segments en cas d'origine maternelle peut s'expliquer par la remarque suivante : en cas de grands segments transloqués, le mode de ségrégation sera plus volontiers une non disjonction (exemple adjacent 2), mais en cas de non disjonction avec de grands segments transloqués, le déséquilibre n'est pas forcément plus important que s'il y avait eu un mode adjacent 1. En effet, dans le cas particulier du mode adjacent 2, le déséquilibre porte sur les segments centriques, qui, dans le cas des chromosomes acrocentriques, sont relativement petits. Comme il a déjà été montré que l'origine maternelle favorisait la survenue d'une non disjonction (Jalbert, 1988; Betend, 1989), on peut supposer que l'origine maternelle avec de grands segments transloqués favorise un mode adjacent 2 qui présentera parfois un déséquilibre moins important, donc plus souvent viable.

En résumé, l'interprétation du modèle retrouve des facteurs de risque déjà connus comme l'origine de la translocation et les chromosomes impliqués. Elle permet de préciser les interactions existantes entre les variables explicatives, en montrant de quelle manière ces différentes variables exercent entre elles une influence qui, soit potentialise leur effet, soit au contraire l'annule.

Chapitre IV

Apport des modèles additifs généralisés dans la modélisation

Avec l'introduction des variables concernant la viabilité, le modèle 2 apparaît suffisamment adéquat pour envisager de l'améliorer. Dans la première partie de ce chapitre, nous exposons les modèles additifs généralisés, l'intérêt de leur utilisation, leur théorie, et la stratégie proposée. Dans une deuxième partie nous présentons les résultats obtenus après utilisation de cette méthode.

IV. 1. Méthode

IV. 1.1. Les modèles additifs généralisés (GAM)

Les modèles linéaires généralisés, et la régression logistique en particulier, font l'hypothèse d'une relation linéaire entre le prédicteur linéaire et les variables explicatives. Si cette hypothèse s'avère abusive (la relation peut être de type exponentiel ou polynomial), l'ajustement est de moins bonne qualité, puisque sa justesse dépend entièrement de l'adéquation des hypothèses "linéarité" et "pente constante". L'intérêt de l'introduction de la méthode des modèles additifs généralisés est de regarder la forme de la fonction non paramétrique de régression afin de :

- suggérer une influence non linéaire entre cette covariable et la variable à expliquer,
- proposer des transformations de variables susceptibles d'améliorer la non-linéarité.

Comme en régression logistique, dans les modèles GAM, on dispose d'une variable à expliquer qualitative ou quantitative et de variables explicatives quantitatives et qualitatives. Si l'on reprend les mêmes notations qu'au chapitre III (p.47), on peut écrire :

$$E(Y/X_1, X_2, X_3, ...Xp) = \mu \qquad Y = \mu + \varepsilon$$
$$\eta = S(X_1, X_2, X_3, ...Xp) = g(\mu) \qquad \eta = S(X)$$

S étant une fonction lisse à estimer (dans le logiciel Splus, ce lissage de la courbe est effectué par la méthode des splines)

Devant les problèmes de variance qui surgissent dans l'estimation d'une fonction de p variables, Hastie propose un modèle moins général représenté par la somme de p fonctions d'une seule variable (Hastie 1986).

$$\eta = S(X) \approx \Sigma S_j X_j = \beta_0 + s_1(X_1) + s_2(X_2) + ...+ s_p(X_p)$$

Pour visualiser la forme non linéaire de la fonction $s_1(X_1)$, on réalise un graphique avec, en abscisse, les valeurs de la variable (X_1), et en ordonnées, les valeurs de la fonction $s_1(X_1)$. La forme de cette fonction permet alors de suggérer une transformation de X_1, celle-ci pouvant être une transformation polynomiale, inverse, logarithmique ou autre.

En réintroduisant dans le modèle logistique les variables transformées, on peut espérer un meilleur ajustement du modèle aux données et des meilleurs performances de celui-ci en terme de résultats de classification.

Note : La recherche d'une transformation non linéaire ne peut s'appliquer qu'à des variables quantitatives et non pas qualitatives même codées numériquement (sauf celles ordonnées à plus de 3 modalités, qui sont alors considérées comme des variables quantitatives).

VI. 1.2. Stratégie proposée

Il s'agit, pour chaque variable explicative, de déterminer l'existence éventuelle d'une non linéarité, de proposer une transformation si la non linéarité existe, et de vérifier, après introduction de la variable transformée que l'on a amélioré les performances du modèle. La théorie des modèles additifs généralisés étant assez récente, il n'existe pas à notre connaissance dans la littérature d'étude de cas bien approfondie. Notre démarche a été totalement empirique. Schématiquement, nous distinguerons six étapes dans la stratégie que nous proposons.

1. Existe-t-il une non-linéarité ?

Une non linéarité sera retenue si la comparaison du modèle linéaire généralisé (GLM) et du modèle additif généralisé (GAM) montre un changement de déviance significatif. Soit la variable explicative V_j et la variable à expliquer Y. Un modèle GLM et un modèle GAM sont successivement testés :

glm (Y ~ V_j, binomial)
gam (Y ~ s(V_j), binomial)

Ces deux modèles étant emboîtés, on compare leurs déviances. Si le gain en déviance est significatif à 1 ‰, on concluera à l'existence d'une non linéarité; sinon, la variable V_j restera sous forme linéaire simple.

2. Quelle est la forme de la fonction $s(V_j)$?

Sous le modèle GAM, on regarde le graphe de cette fonction (avec les valeurs de la variable V_j et avec les valeurs des résidus). La forme de cette fonction permet peut-être de suggérer une transformation non linéaire qui peut, dans certains cas, être polynomiale, logarithmique, inverse, ou autre. S'il existe une non-linéarité, plusieurs transformations seront essayées.

3. Quelle est la meilleure transformation ?

Les différentes transformations suggérées sont introduites dans un modèle GLM (t_1, t_2 par exemple).

$$\text{glm (Y} \sim t_1(V_j), \text{binomial)}$$
$$\text{glm (Y} \sim t_2(V_j), \text{binomial)}$$

Ces modèles n'étant pas emboîtés, il n'est pas possible ici d'utiliser un test pour le changement en déviance. On retient la transformation qui procure le gain en déviance maximum pour la plus petite perte en ddl. La transformation polynomiale est systématiquement explorée, et, si elle est retenue, on procède à une analyse en composantes principales sur ce polynôme afin de ne pas retenir trop de paramètres (1^e et 2^e composantes principales d'un polynôme d'ordre 6 par exemple). En utilisant les composantes principales, on évite le problème d'une forte corrélation entre les puissances de la variable, grâce à un résumé de l'information.

4. Quelle est la forme de la fonction s() après transformation ?

Avec la "meilleure" transformation retenue, on vérifie, en utilisant un modèle GAM, que l'on a bien "linéarisé" la relation entre variable explicative et variable à expliquer. Pour ce faire, si la transformation retenue comme la meilleure est t_k , on réintroduit cette variable transformée dans un modèle GAM et l'on regarde ensuite la forme de la fonction $s(t_k(V_j))$ sous ce modèle.

$$\text{gam (Y} \sim s(t_k(V_j)), \text{binomial)}$$

Sur le graphique de la fonction $s(t_k(V_j))$, on s'assure que l'on peut admettre une linéarité pour ces nouveaux résidus. Et l'on s'assure également de la non significativité du gain en déviance de ce modèle GAM par rapport au modèle GLM du "3.".

5. On introduit toutes les variables transformées dans le modèle.

On introduit à nouveau les variables dans un modèle de régression logistique mais avec les variables transformées pour celles où une non linéarité a été détectée. C'est la forme transformée de la variable qui est utilisée aussi dans les termes d'interactions.

$$\text{glm (Y} \sim X_1 + X_2 + t_k(X_3) + t_{k'}(X_4) + \ldots + X_1 * t_k(X_3) + X_p)$$

Le modèle ainsi obtenu peut être considéré comme meilleur et mieux ajusté. On s'en assure en testant un modèle GAM sur le même prédicteur, et en comparant les déviances entre les deux modèles.

$$\text{gam}\ (\ Y \sim s(X_1) + s(X_2) + s(t_k(X_3)) + s(t_{k'}(X_4)) + ...+ X_1{*}t_k(X_3) + s(X_p))$$

Le gain en déviance obtenu sous GAM doit être non significatif à 1‰.

6. On vérifie que l'on a amélioré l'ajustement du modèle.

Pour vérifier si l'utilisation du modèle GAM a amélioré les performances du modèle, on regarde l'adéquation du modèle aux données et les résultats de classification. Une étude des résidus permet de vérifier leur nombre et leur distribution. Par les résultats de classification (taux d'erreur réel et taux d'erreur apparent), on regarde l'amélioration du modèle pour les différents groupes de risque prédit.

IV. 2. Résultats

Nous présentons d'abord les variables et les interactions du modèle 2 pour lesquelles une transformation apparaît intéressante. Puis nous proposons le modèle 3, très proche du modèle 2 mais avec certaines variables sous une forme transformée. Enfin, nous soulignons les améliorations apportées ainsi à la modélisation proposée.

IV. 2.1. Les variables transformées

Seules sont présentées ici les transformations retenues pour les variables présentes dans le modèle 2, ainsi que les résultats de la prise en compte des transformations dans les interactions. Pour les autres variables quantitatives VIA47 et SINUSADJ présentes dans le modèle 2, dès la première étape de la stratégie proposée, on ne retient pas de non linéarité.

Variable LAMIN (longueur minimum de trisomie/monosomie pour les adjacents) (tableau 22)

Plusieurs choix de transformation sont possibles, comme le montre le graphe de la fonction s (LAMIN) (fig. 16 p.88).

- une transformation en 2 droites : t_1(LAMIN) prend pour valeur 2,5 si LAMIN $\leq$ 2,5 et LAMIN sinon ; t_2(LAMIN) prend pour valeur 2,5 si LAMIN $\geq$ 2,5 et LAMIN sinon.
- un polynôme (jusqu'à l'ordre 5),
- une fonction logarithmique t_3(LAMIN)= log(-LAMIN+ 6)

Tableau 22. Transformation pour la variable LAMIN

modèle	ddl	deviance	test
GLM	5562	5567	
GAM	5559	5549	18 pour 3 ddl
GLM t_1(LAMIN)+ t_2(LAMIN)	5561	5554	13 pour 1 ddl
GLM polynôme d'ordre 2	5561	5560	7 pour 1 ddl
GLM polynôme d'ordre 3	5560	5557	10 pour 2 ddl
GLM polynôme d'ordre 4	5559	5556	11 pour 3 ddl
GLM polynôme d'ordre 5	5558	5549	18 pour 4 ddl
GLM C1(polynôme d'ordre 5)	5562	5559	8 pour 0 ddl
GLM C2(polynôme d'ordre 5)	5561	5558	9 pour 1 ddl
GLM t_3(LAMIN)	5562	5559	8 pour 0 ddl

*C1 signifie la première composante principale
C2 signifie les deux premières composantes principales

On peut retenir la première composante principale d'un polynôme d'ordre 5 ou la transformation logarithmique. En raison de sa simplicité d'utilisation, on préfèrera utiliser la transformation logarithmique.

Variable CSbandeR1 (tableau 23)

Plusieurs choix de transformation sont possibles (fig. 18 p.89) comme le montre le graphe de la fonction S (CSbandeR1).

- la fonction inverse t_1(CSbandeR1) = 1/CSbandeR1 ,
- un polynôme (jusqu'à l'ordre 4)

Tableau 23.Transformation pour la variable CSbandeR1

modèle	ddl	deviance	test
GLM	5562	5715	
GAM	5559	5688	27 pour 3 ddl
GLM polynôme d'ordre 2	5561	5713	2 pour 1 ddl
GLM polynôme d'ordre 3	5560	5701	14 pour 2 ddl
GLM polynôme d'ordre 4	5559	5682	33 pour 3 ddl
GLM C1(polynôme d'ordre 4)	5562	5717	aggravé
GLM C2(polynôme d'ordre 4)	5561	5715	0 pour 1 ddl
GLM C3(polynôme d'ordre 4)	5560	5706	9 pour 2 ddl
GLM t_1(CSbandeR1)	5562	5700	15 pour 0 ddl

*C1 signifie la première composante principale
C2 signifie les deux premières composantes principales
C3 signifie les trois premières composantes principales

On retiendra la transformation inverse comme la meilleure transformation. Avec la transformation polynomiale, il faudrait plus de 3 composantes principales pour obtenir un résultat aussi performant qu'avec la fonction inverse.

Variable TSbandeR1 (tableau 24)

On peut proposer les choix suivants dans la recherche d'une transformation (fig. 20 p.90).

- une fonction logarithmique t_1(TSbandeR1) = log[-TSbandeR1+50]),
- un polynôme (jusqu'à ordre 3)

Tableau 24. Transformation pour la variable TSbandeR1

Modèle	ddl	deviance	test
GLM	5562	5647	
GAM	5559	5637	10 pour 3 ddl
GLM polynôme d'ordre 2	5561	5641	6 pour 1 ddl
GLM polynôme d'ordre 3	5560	5638	9 pour 2 ddl
GLM C1(polynôme d'ordre 3)	5562	5640	7 pour 0 ddl
GLM C2(polynôme d'ordre 3)	5561	5640	7 pour 1 ddl
GLM C3(polynôme d'ordre 3)	5560	5638	9 pour 2 ddl
GLM t_1(TSbandeR1)	5562	5641	6 pour 0 ddl

*C1 signifie la première composante principale
C2 signifie les deux premières composantes principales
C3 signifie les trois premières composantes principales

On retiendra la transformation logarithmique. On aurait pu retenir la première composante principale du polynôme d'ordre 3 qui procure un gain en déviance légèrement meilleur. Mais la fonction logarithmique nous semble plus facile à utiliser, en vue de l'application du modèle à la prédiction du risque pour une nouvelle observation.

Variable TSbandeR2 (tableau 25)

Deux choix de transformation sont possibles d'après le graphe de la fonction S (TSbandeR2) (fig. 22 p.91).

- la fonction inverse t_1(TSbandeR2) = 1/TSbandeR2 ,
- un polynôme (jusqu'à ordre 6)

Tableau 25. Transformation pour la variable TSbandeR2

modèle	ddl	deviance	test
GLM	5562	5671	
GAM	5559	5648	23 pour 3 ddl
GLM polynôme d'ordre 2	5561	5657	14 pour 1 ddl
GLM polynôme d'ordre 3	5560	5652	19 pour 2 ddl
GLM polynôme d'ordre 4	5559	5652	19 pour 3 ddl
GLM polynôme d'ordre 5	5558	5646	25 pour 4 ddl
GLM polynôme d'ordre 6	5557	5638	33 pour 5 ddl
GLM C1(polynôme d'ordre 2)	5562	5680	aggravé
GLM C3(polynôme d'ordre 6)	5560	5653	18 pour 2 ddl
GLM t_1(TSbandeR2)	5562	5685	aggravé

*C1 signifie la première composante principale
C3 signifie les trois premières composantes principales

On retiendra le polynôme d'ordre 2 qui semble être la "meilleure" transformation.

Interaction origine*TSbandeR2 (tableau 26)

Pour vérifier s'il existe une non linéarité entre cette interaction et le logit de la survenue d'un enfant malformé, il est nécessaire de décomposer cette interaction en 2 variables :

- I1 qui prend la valeur de TSbandeR2 lorsque l'origine est paternelle, et 0 sinon
- I2 qui prend la valeur de TSbandeR2 lorsque l'origine est maternelle, et 0 sinon

Tableau 26. Existence d'une non linéarité pour l'interaction origine*TSbandeR2

modèle	ddl	deviance	test
GLM I1	5562	5645	
GAM I1	5559	5638	7 pour 3 ddl
GLM I2	5562	5718	
GAM I2	5559	5702	16 pour 3 ddl

Pour I1, la différence n'est pas significative entre GAM et GLM, donc on ne retient pas de non-linéarité pour cette partie de l'interaction. Pour I2, la différence de déviance est significative (p = 0,001), on peut retenir une non-linéarité. Les transformations proposées sont des fonctions polynomiales (jusqu'à l'ordre 6) (tableau 27).

Tableau 27. Transformation pour l'interaction I2

modèle	ddl	deviance	test
GLM	5562	5718	
GAM	5559	5702	16 pour 3 ddl
GLM polynôme d'ordre 2	5561	5714	4 pour 1 ddl
GLM polynôme d'ordre 3	5560	5711	7 pour 2 ddl
GLM polynôme d'ordre 4	5559	5706	12 pour 3 ddl
GLM polynôme d'ordre 5	5558	5705	13 pour 4 ddl
GLM polynôme d'ordre 6	5557	5684	34 pour 5 ddl
GLM C1(polynôme d'ordre 6)	5562	5718	inchangé
GLM C2(polynôme d'ordre 6)	5561	5717	1 pour 1 ddl
GLM C3(polynôme d'ordre 6)	5560	5715	3 pour 2 ddl

*C1 signifie la première composante principale
C2 signifie les deux premières composantes principales
C3 signifie les trois premières composantes principales

On ne trouve pas de transformation satisfaisante, si on ne veut pas trop perdre en ddl, donc on gardera cette interaction sous sa forme simple initiale.

Interaction NVC*TSbandeR1 (tableau 28)

De la même façon que pour la précédente, on décompose cette interaction en cinq variables : inter1 qui prend pour valeur TSbandeR1 si NVC = 1, et 0 sinon, inter2 qui prend pour valeur TSbandeR1 si NVC = 2, et 0 sinon, et même chose pour inter2, inter3, inter4, inter5.

Tableau 28. Existence d'une non linéarité pour l'interaction NVC*TSbandeR1

modèle	ddl	deviance	test
GLM inter1	5562	5588	
GAM inter1	5559	5551	37 pour 3 ddl
GLM inter2	5562	5718	
GAM inter2	5559	5700	18 pour 3 ddl
GLM inter3	5562	5718	
GAM inter3	5559	5714	3 pour 3 ddl
GLM inter4	5562	5701	
GAM inter4	5559	5681	20 pour 3 ddl
GLM inter5	5562	5717	
GAM inter5	5559	5713	4 pour 3 ddl

Parmi ces 5 variables, pour 2 d'entre elles : inter3 et inter5, il n'est pas nécessaire de retenir une non-linéarité.

Pour inter1, on essaye une transformation logarithmique, une transformation polynomiale et une transformation en 2 droites (d1 et d2) (tableau 29).

- d1 prend pour valeur 12 si inter1 $\leq$ 12, et inter1 sinon
- d2 prend pour valeur 12 si inter1 $\geq$ 12, et inter1 sinon

Tableau 29. Transformation pour l'interaction inter1

modèle	ddl	deviance	test
GLM	5562	5588	
GAM	5559	5551	37 pour 3 ddl
GLM polynôme d'ordre 2	5561	5557	31 pour 1 ddl
GLM polynôme d'ordre 3	5560	5551	37 pour 2 ddl
GLM polynôme d'ordre 4	5559	5550	38 pour 3 ddl
GLM t_1(log(50-inter1))	5562	5575	13 pour 0 ddl
GLM t_2(2 droites d1 et d2)	5561	5554	34 pour 1 ddl

C'est la transformation en 2 droites qui semble la "meilleure" transformation.

Pour inter2, on essaye des transformations polynomiales et une transformation en deux droites (d1 et d2) (tableau 30).

- d1 prend pour valeur 23 si inter2 ≤ 23, et inter2 sinon.
- d2 prend pour valeur 23 si inter2 ≥ 23, et inter2 sinon.

Tableau 30. Transformation pour l'interaction inter2

modèle	ddl	deviance	test
GLM	5562	5718	
GAM	5559	5700	18 pour 3 ddl
GLM polynôme d'ordre 2	5561	5708	10 pour 1 ddl
GLM polynôme d'ordre 3	5560	5708	10 pour 2 ddl
GLM polynôme d'ordre 4	5559	5695	23 pour 3 ddl
GLM polynôme d'ordre 5	5558	5695	23 pour 4 ddl
GLM polynôme d'ordre 6	5557	5692	26 pour 5 ddl
GLM C1(polynôme d'ordre 4)	5562	5718	inchangé
GLM C2(polynôme d'ordre 4)	5561	5707	11 pour 1 ddl
GLM C3(polynôme d'ordre 4)	5560	5707	11 pour 2 ddl
GLM t_1(2 droites d1 et d2)	5562	5704	14 pour 0 ddl

C'est la transformation en 2 droites qui procure également le meilleur gain en déviance.

Pour inter4, c'est une transformation polynomiale (ordre 2) qui donne les meilleurs résultats (tableau 31).

Tableau 31. Transformation pour l'interaction inter4

modèle	ddl	deviance	test
GLM	5562	5701	
GAM	5559	5681	20 pour 3 ddl
GLM polynôme d'ordre 2	5561	5685	16 pour 1 ddl
GLM polynôme d'ordre 3	5560	5684	17 pour 2 ddl
GLM polynôme d'ordre 4	5559	5681	20 pour 3 ddl

La 3^e interaction du modèle 2, NVC*ORIGINE, est une interaction entre variables qualitatives; cela n'a donc pas de sens de rechercher une éventuelle non linéarité pour cette variable.

Sur les 4 variables pour lesquelles on retient une transformation, on s'assure graphiquement que l'on a linéarisé les résidus en faisant tourner un modèle GAM sur chacune de ces variables transformées. Les figures 17, 19, 21 et 23 montrent les résidus après transformation.

Fig. 16. Fonction s (LAMIN)

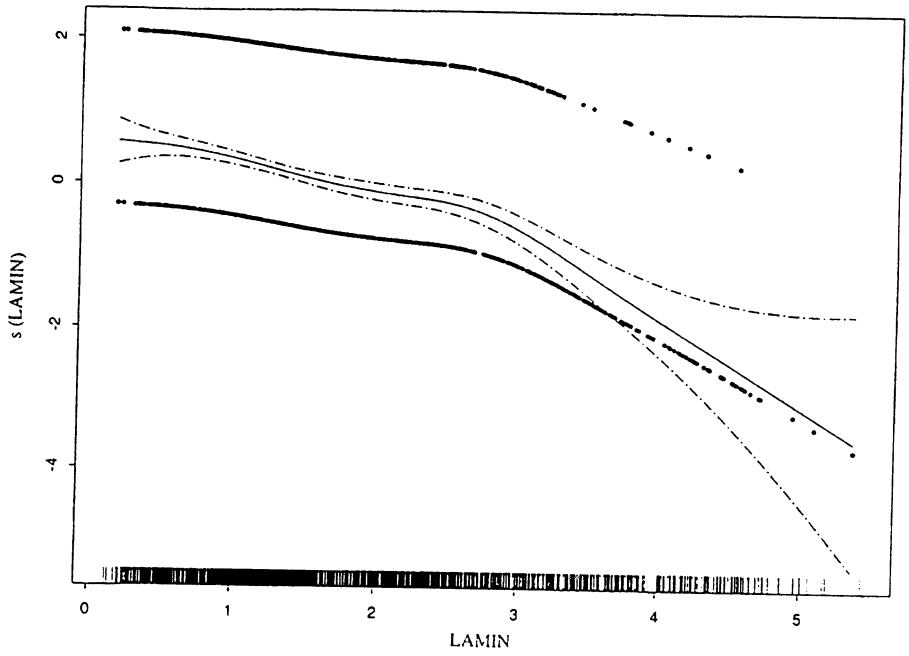
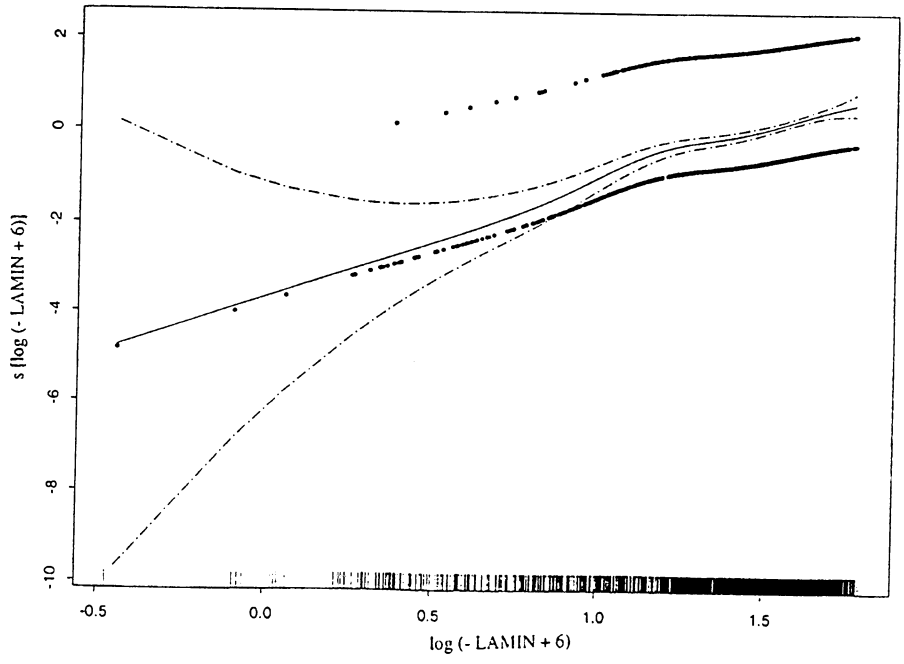


Fig. 17. Fonction s [log (- LAMIN + 6)]



Les traits verticaux sur l'axe des abscisses représentent les différentes observations pour chacune des valeurs de la variable LAMIN.

Fig. 18. Fonction s (CSbandeR1)

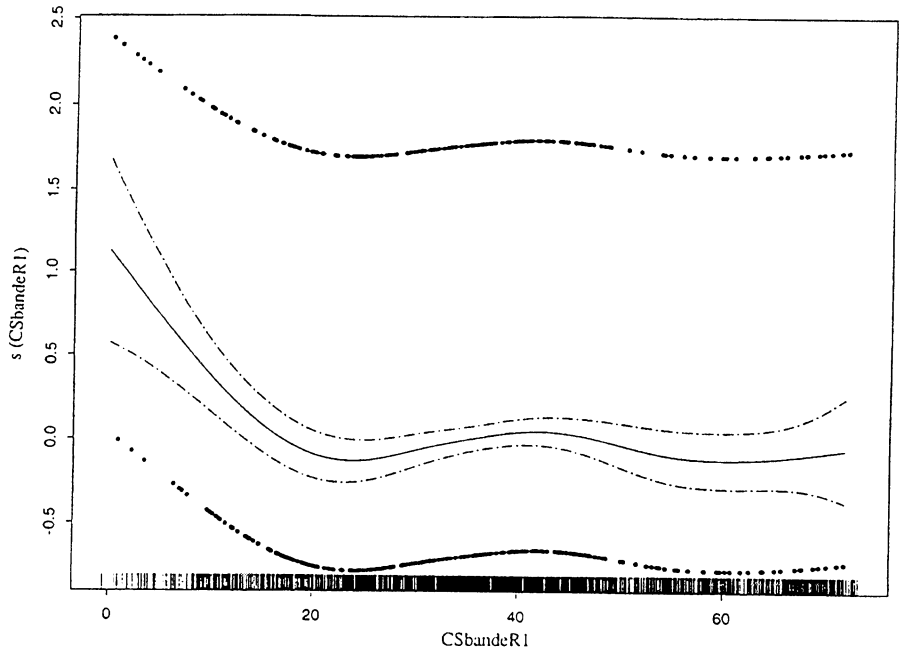
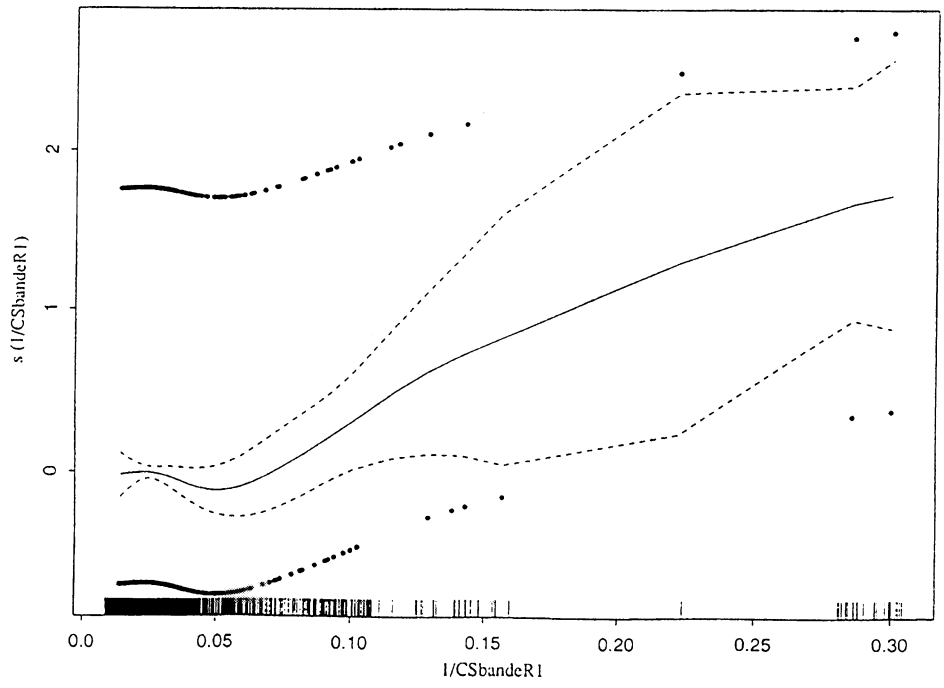


Fig. 19. Fonction s (1/CSbandeR1)



Les traits verticaux sur l'axe des abscisses représentent les différentes observations pour chacune des valeurs de la variable CSbandeR1.

Fig. 20. Fonction s (TSbandeR1)

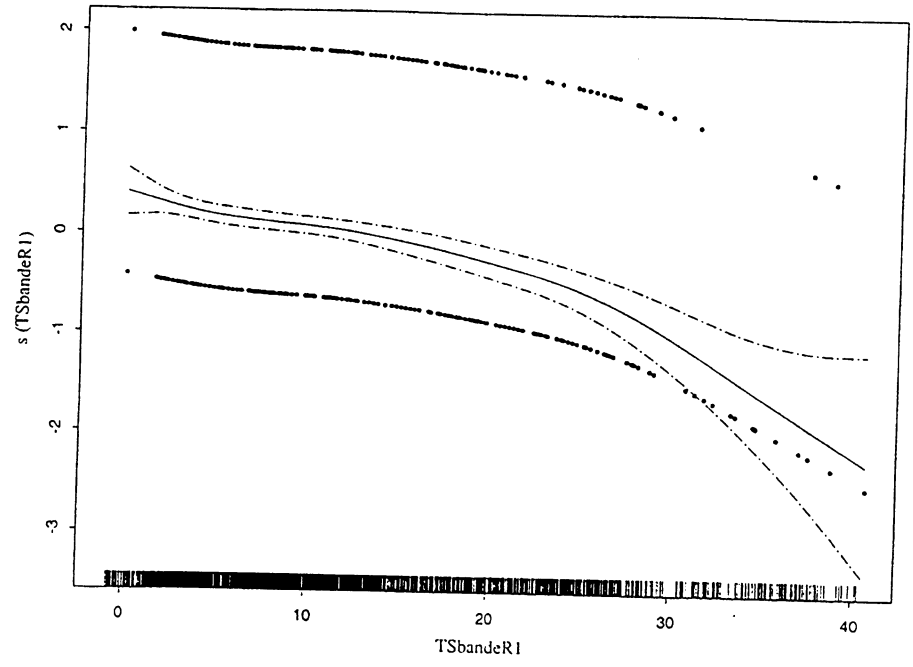
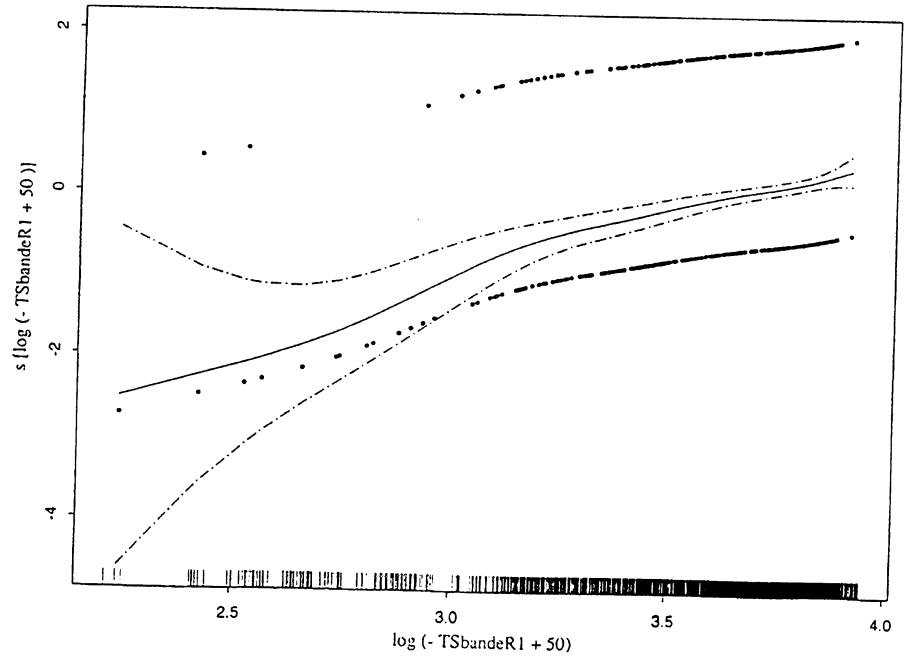


Fig. 21. Fonction s [log (- TSbandeR1 + 50)]



Les traits verticaux sur l'axe des abscisses représentent les différentes observations pour chacune des valeurs de la variable TSbandeR1.

Fig. 22. Fonction s (TSbandeR2)

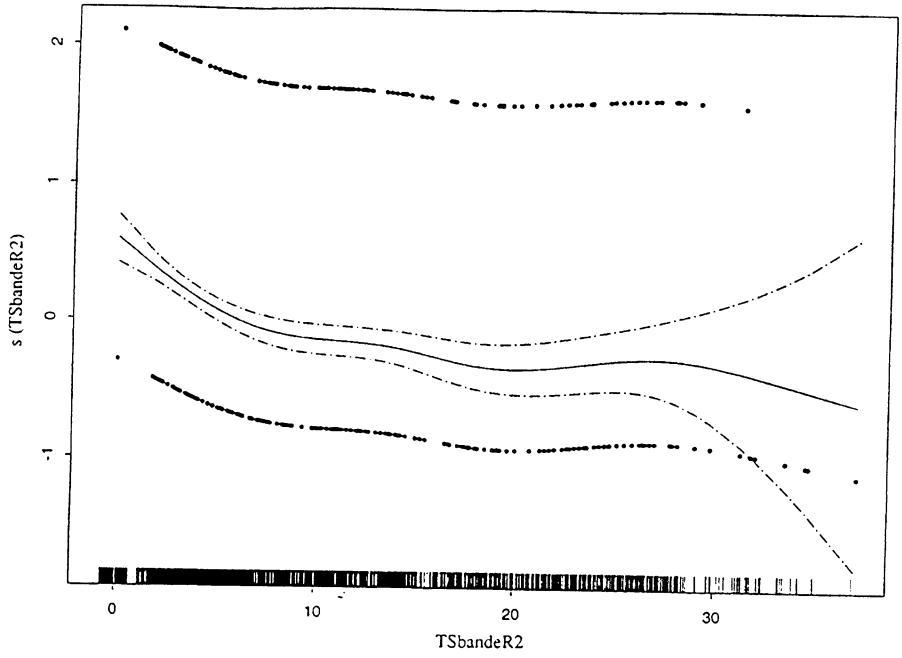
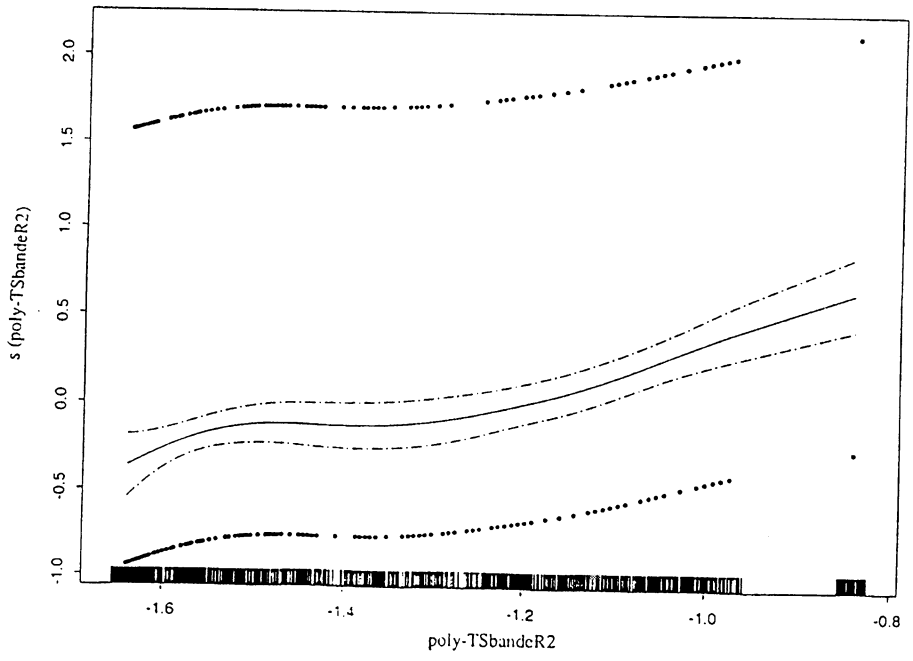


Fig. 23. Fonction s (poly-TSbandeR2)



Les traits verticaux sur l'axe des abscisses représentent les différentes observations pour chacune des valeurs de la variable TSbandeR2.

IV. 2.2. Le nouveau modèle proposé

Modèle 3

Il comporte 11 variables (dont 4 transformées) et 3 termes d'interaction.

BRAS + NVC + ORIG +VIA47 + t_3 de(LAMIN) + SINUSADJ + t_1 de (CSbandeR1) + t_1 de (TSbandeR1) + TSbandeR2 + (TSbandeR2)^2 + MOSEG + NVC*t_1 de (TSbandeR1)) + NVC*ORIG + ORIG*TSbandeR2

	Value	Std Error	t value	
(Intercept)	-17.11	1.74	-9.80	
BRAS	-0.28	0.10	-2.67	bras points de cassure
NVC(2)	4.65	2.20	2.12	Xme 9
NVC(3)	6.86	2.06	3.33	Xmes 13,14,15
NVC(4)	9.13	2.47	3.69	Xmes 21,22
NVC(5)	6.86	7.03	0.98	t(11,22) (•)
ORIG	-0.24	0.14	-1.73	origine maternelle (•)
VIA47	-0.13	0.04	-3.32	viabilité à 45 Xmes
t_3(LAMIN)	1.37	0.43	3.18	longueur min adj
SINUSADJ	3.11	0.87	3.58	sinus adj
t_1(CSbandeR1)	3.78	1.59	2.37	t_1 de CSbandeR1
t_1(TSbandeR1)	3.17	0.52	6.11	t_1 de TSbandeR1
TSbandeR2	-0.17	0.02	-8.72	TSbandeR2
(TSbandeR2)^2	0.004	0.0007	5.70	(TSbandeR2)^2
MOSEG(2)	1.40	0.31	4.48	mode de ségrégation adj 2
MOSEG(3)	0.45	0.15	3.01	mode de ségrég 3:1 tert
MOSEG(4)	0.26	0.36	0.72	mode de ségrég 3:1 ech (•)
NVC(2)*t_1 de TSbandeR1	-1.23	0.59	-2.089	
NVC(3)*t_1 de TSbandeR1	-1.80	0.56	-3.230	
NVC(4)*t_1 de TSbandeR1	-2.49	0.66	-3.750	
NVC(5)*t_1 de TSbandeR1	-2.34	1.85	-1.261	(•)
NVC(2)*ORIG	0.53	0.23	2.28	
NVC(3)*ORIG	0.09	0.21	0.45	(•)
NVC(4)*ORIG	0.82	0.22	3.82	
NVC(5)*ORIG	2.06	0.53	3.86	
ORIG*TSbandeR2	0.04	0.02	2.99	

Déviance nulle : 5718 avec 5563 degrés de liberté

Déviance résiduelle: 5261 avec 5538 degrés de liberté

(•) paramètre non significatif

Les interactions ont été laissées sous leur forme simple. Pour l'interaction NVC*TSbandeR1, si l'on inclut dans le modèle proposé (modèle 3) l'interaction sous la forme non linéaire des 3 décompositions inter1 (2 droites), inter2 (2 droites), et inter4 (polynôme d'ordre 2), on n'améliore pas significativement la déviance. C'est probablement parce qu'en ayant mis l'interaction sous forme NVC*t_1 de (TSbandeR1), on a déjà absorbé une bonne partie de l'effet non linéaire.

Comparativement au modèle 2, le modèle 3 comporte 1 seul paramètre de plus à estimer et on a gagné 42 points en déviance. Seuls 5 paramètres ne sont pas significatifs. Les valeurs estimées du risque de survenue d'un déséquilibre à terme sont comprises entre 0,000 et 0,994. La valeur prédite est supérieure à 0,40 pour 207 observations, et supérieure à 0,50 pour 31 observations.

En ce qui concerne l'étude des résidus de Pearson pour ce modèle 3, il n'y a pratiquement pas de changement par rapport aux modèles précédents, puisque 1,7 ‰ (n=93) des observations ont une valeur de résidu > 2,5 et seules 12 d'entre elles présentent une valeur > 4. Parmi les observations avec résidu > 4, on retrouve 9 observations déjà citées comme présentant un fort résidu sous le modèle 1 (III. 2. 3) et 7 d'entre elles ont fait l'objet de remarques sur la pertinence des informations les concernant (p.66)

Il nous reste à vérifier (cf étape 5) qu'un modèle GAM sur les variables transformées ne procure pas de changement significatif en déviance. Avant la transformation des variables, les résultats montrent l'intérêt de l'utilisation de GAM sur le modèle 2.

	déviance	ddl
modèle 2 GLM	5303	5539
modèle 2 GAM	5224	5521
	gain en déviance significatif à 1 ‰	

Le même tableau concernant le modèle 3 montre que le gain en déviance n'est plus significatif, témoignant de la prise en compte de la non linéarité.

	déviance	ddl
modèle 3 GLM	5261	5538
modèle 3 GAM	5221	5518
	gain en déviance non significatif à 1 ‰	

IV. 2.3. Classification

Selon la même méthodologie que celle utilisée dans le chapitre précédent, nous avons calculé les taux d'erreur apparent et réel de ce modèle 3. Les résultats sont montrés dans le tableau 31.

Tableau 31. Taux d'erreur sur le modèle 3

risque prédit	Taux d'erreur apparent		Taux d'erreur réel	IC de ce taux * d'erreur réel
Proportion d'observations "non malades" bien classées				
< 0,01	133/133	*100 %*	*99,9 %*	98,0-100
0,01 ≤ - < 0,05	348/357	*97,5 %*	*97,0 %*	95,5-98,3
0,05 ≤ - < 0,10	486/519	*93,6 %*	*93,1 %*	87,9-95,7
< 0,10	967/1009	*95,8 %*	-	
Proportion d'observations "malades" bien classées				
0,30 ≤ - < 0,50	355/1002	*35,2 %*	*33,0 %*	28,8-39,6
≥ 0,50	20/31	*64,5 %*	*57,4 %*	37,5-75,0

* L'intervalle de confiance a été calculé en supprimant les 2 valeurs extrêmes obtenues sur ces 30 échantillons

Si l'on classe les probabilités prédites par ordre croissant, les premières observations mal classées sont les suivantes :

- la 1e (fam 566) à la 157e observation
- la 2e (fam 499) à la 199e observation
- la 3e (fam 805) à la 332e observation.

Ce sont les mêmes que les premières observations mal classées avec le modèle 2 (cf p.70). Les résultats en taux d'erreur avec le modèle 3 sont donc surtout meilleurs pour les observations à valeur prédite très faible (< 0,01). Le nombre d'observations appartenant à ce groupe a augmenté, ainsi que la proportion d'observations bien classées, qui est très proche des 100 %. Comme nous le verrons au chapitre V, c'est essentiellement pour ce groupe que les résultats du modèle pourraient induire d'éventuels changements de pratique.

IV. 2.4. Utilisation de GAM en "boîte noire"

Si on pousse à l'extrême l'utilisation de GAM, on peut concevoir un modèle GAM et non plus GLM pour estimer la valeur du risque de déséquilibre à terme. En effet, le logiciel Splus donne la possiblilité d'obtenir une valeur prédite sous un modèle GAM. C'est ce que nous avons fait en construisant un modèle additif comportant les 10 variables du modèle 2 et les 3 interactions. En fait, les résultats du taux d'erreur de ce modèle ne montrent pas d'amélioration par rapport au modèle 3, ils figurent donc seulement en annexe 12, pour le lecteur qui voudrait s'y rapporter.

Ceci confirme les résultats de la stratégie que nous avons employée pour obtenir le modèle 3, et en particulier le fait que l'on a absorbé quasiment toute la non-linéarité avec les transformations proposées. Par contre, reste le problème de la fonction s(X) choisie. Celle choisie par défaut dans Splus est la méthode des splines, qui ne convient peut-être pas parfaitement à nos données puisque celles-ci ne sont pas distribuées normalement. Un choix de cette fonction par la méthode des noyaux pourrait peut être donner de meilleurs résultats.

En fait, la solution d'utiliser un modèle GAM "en boîte noire" n'est de toute façon pas satisfaisante, lorsque l'on attend d'un modèle non seulement un outil de prédiction, mais aussi une aide dans l'interprétation du phénomène à expliquer.

IV. 2.5. Risques dus à chacune des variables explicatives

Le retour au modèle logistique permet d'obtenir une estimation, propre à chaque variable, du risque de survenue d'un enfant malformé, ce que ne permet pas le modèle additif généralisé. Ces résultats d'odds ratio estimés à partir du modèle 3 sont présentés dans le tableau 32.

On retrouve les influences déjà décrites à la fin du chapitre III. L'apport de cette information dans la compréhension des mécanismes de ségrégation méiotique ou de viabilité d'un déséquilibre est discuté dans le chapitre V.

Compte tenu des données disponibles, il ne nous semble pas licite de continuer à chercher d'autres améliorations du modèle. Si l'on exprime les performances du modèle en pourcentage de déviance expliquée, les variables explicatives disponibles expliquent seulement 9 % de la déviance (il s'agit d'une méthode de calcul, très semblable à l'interprétation du R^2 de la régression multiple, proposée par Hastie). La seule manière pour augmenter ce pourcentage de déviance expliquée consiste à introduire de nouvelles variables pertinentes dans le modèle et à améliorer la qualité des données recueillies actuellement (ce qui sera proposé au chapitre V. 3.).

Tableau 32. OR des variables explicatives et leurs IC (modèle 3)

			OR	IC
BRAS	pq ou qq	réf		
	pp		0,76	[0,62-0,93]
NVC pour TSbandeR1 ≈ 10	Xmes 1 à 8, 10 à 12, 16 à 20	réf		
et ORIG maternelle	Xme 9		1,99	[1,38-2,87]
	Xmes 13,14,15		1,51	[1,08-2,10]
	Xmes 21,22		2,43	[1,17-5,06]
	t(11,22)		1,51	[0,18-12,69]
NVC pour TSbandeR1 ≈ 10	Xmes 1 à 8, 10 à 12, 16 à 20	réf		
et ORIG paternelle	Xme 9		1,17	[0,73-1,87]
	Xmes 13,14,15		1,37	[0,89-2,12]
	Xmes 21,22		1,07	[0,87-1,30]
	t(11,22)		0,19	[0,06-0,63]
ORIG pour TSbandeR2 ≈ 10	paternelle	réf		
et Xmes 1 à 8, 10 à 12, 16 à 20	maternelle		1,12	[0,89-1,40]
et Xme 9	maternelle		1,90	[1,28-2,82]
et Xmes 13,14,15	maternelle		1,23	[0,88-1,71]
et Xmes 21,22	maternelle		2,54	[1,81-3,57]
et t(11,22)	maternelle		8,78	[3,16-24,34]
VIA 47	pour une ↑ de 1		0,88	[0,82-0,95]
	pour une ↑ de 2		0,78	[0,67-0,90]
LAMIN	pour une ↑ de 1		1,32	[1,11-1,56]
	pour une ↑ de 2		1,98	[1,30-3,03]
SINUSADJ	pour une ↑ de 0,01		1,03	[1,01-1,05]
	pour une ↑ de 0,05		1,17	[1,07-1,27]
	pour une ↑ de 0,1		1,36	[1,15-1,62]
CSbandeR1	pour une ↑ de 1		1,01	[1,00-1,02]
	pour une ↑ de 10		1,10	[1,02-1,19]
	pour une ↑ de 20		1,21	[1,03-1,41]
TSbandeR1 pour				
Xmes 1 à 8, 10 à 12, 16 à 20	et une ↑ de 5		1,56	[1,35-1,80]
Xme 9	et une ↑ de 5		1,31	[1,09-1,58]
Xmes 13,14,15	et une ↑ de 5		1,21	[0,99-1,48]
Xmes 21,22	et une ↑ de 5		1,10	[0,90-1,34]
t(11,22)	et une ↑ de 5		1,12	[0,81-1,57]
TSbandeR2 pour				
ORIG maternelle	et une ↑ de 5		0,56	[0,43-0,71]
ORIG paternelle	et une ↑ de 5		0,47	[0,39-0,55]
MOSEG	adjacent 1	réf		
	adjacent 2		4,07	[2,20-7,51]
	3:1 tertiaire		1,58	[1,17-2,12]
	3:1 échange		1,29	[0,64-1,20]

Chapitre V

Applications et perspectives

Dans ce chapitre, seront montrées les applications directes de cette analyse et du modèle proposé, ainsi que les réflexions suggérées par l'interprétation des interactions. Puis nous discuterons les données disponibles actuellement, en précisant les améliorations nécessaires au recueil de ces données. Enfin, les perspectives de prise en compte de nouvelles variables seront exposées, ainsi que le lien avec le niveau de connaissance moléculaire dans le cas de la t(11;22).

V. 1. Application des résultats

Si on reprend les résultats commentés de la fin du chapitre III et les résultats d'OR du modèle proposé au chapitre IV, ces résultats sont très proches et conduisent aux réflexions suivantes.

V. 1.1. Apport à la compréhension du phénomène

• Comment expliquer l'influence des différentes variables et de leurs interactions, et en particulier l'effet dû à la longueur des segments transloqués ?

La longueur des segments transloqués et la longueur des segments en déséquilibre (% trisomie + % monosomie) varient en fonction des différents modes de ségrégation.Ces segments ont les valeurs les plus faibles pour les modes adjacent 1, puis ces valeurs augmentent pour les modes 3:1 et elles sont maximales pour les modes adjacent 2 (tableau 32).

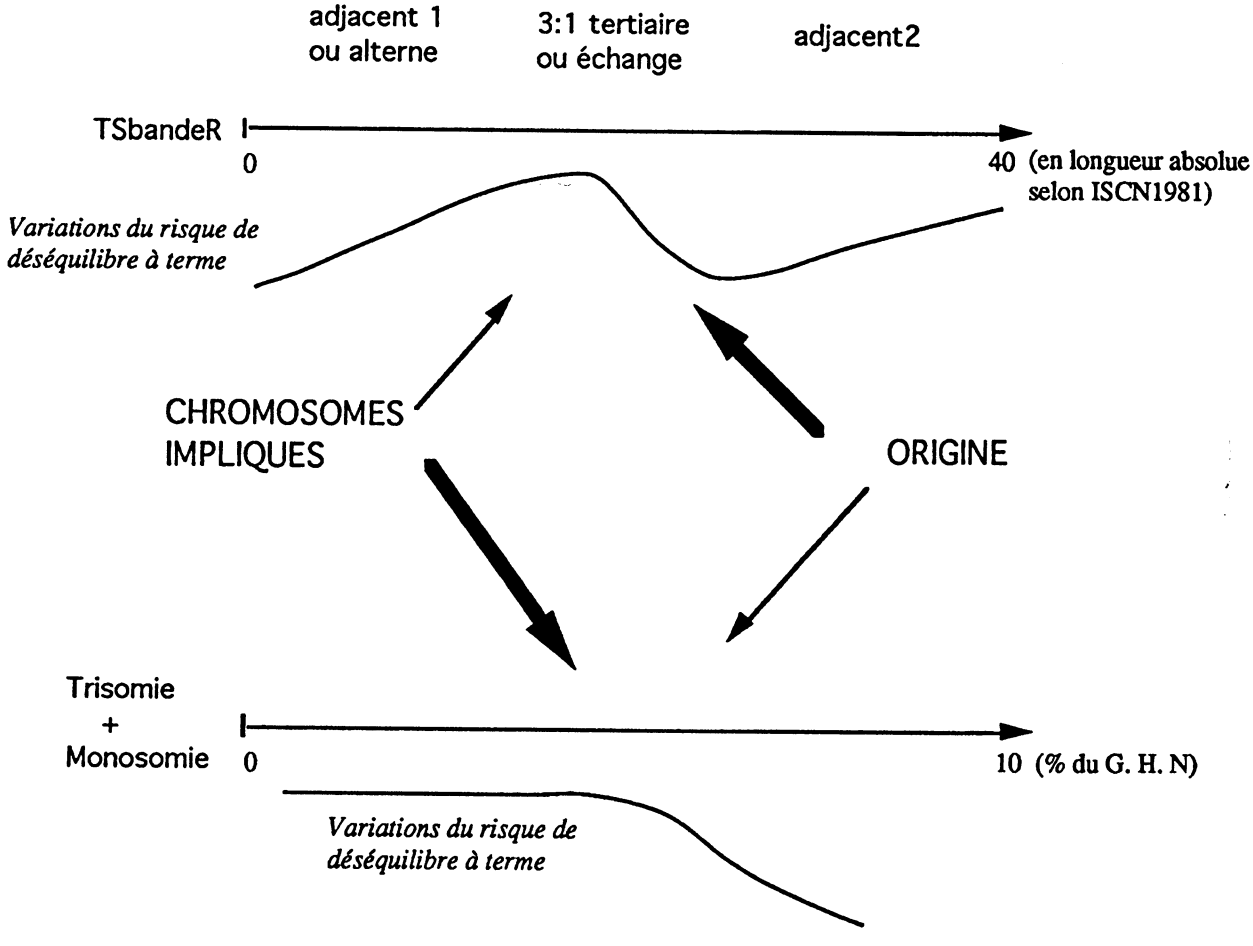
Tableau 32. Moyenne des longueurs des segments transloqués et des segments en déséquilibre

	TSbandeR1	TSbandeR2	TRISOMIE *	MONOSOMIE *
adjacent 1	8,8	6,1	1,06	0,45
3:1 tertiaire	8,6	13,9	1,74	0,16
3:1 échange	15,2	7,6	2,00	0,00
adjacent 2	19,7	15,2	1,85	1,11

* en % du Génome Haploïde Autosomal

A partir d'une certaine longueur des segments transloqués, le mode de non disjonction est préférentiellement un mode 3:1, puis adjacent 2. Ce dernier mode est souvent caractérisé par des pourcentages de trisomie et monosomie élevés (lorsque la translocation comporte des grands chromosomes) : il est donc plus souvent non viable. Mais il peut conduire parfois au contraire à une taille des segments en déséquilibre plus faible pour le déséquilibre observé (car portant sur des segments centriques de petits chromosomes) et donc plus souvent viable. D'où la proposition de visualiser l'influence combinée des segments transloqués, de l'origine de la translocation et des numéros des chromosomes impliqués de la manière suivante.

Fig. 23. Interactions entre l'origine de la translocation, la longueur des segments transloqués et le numéro des chromosomes



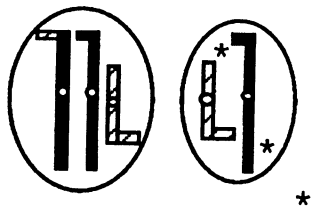
• Y a-t-il une explication physio-pathologique à l'influence de l'origine maternelle ?

Nos résultats confirment le fait que l'origine maternelle de la translocation est associée à la survenue d'un enfant malformé à terme. Est-ce la conséquence d'une influence de l'origine maternelle sur la non disjonction méiotique ? Et quelle est l'influence de l'origine paternelle sur cette ségrégation méiotique ?

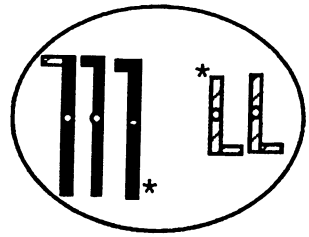
L'influence de l'origine maternelle sur la méiose se rencontre également dans d'autres anomalies chromosomiques (avec non disjonctions plus fréquentes, comme dans la trisomie 21 par exemple). Qu'en est-il de l'influence de l'origine paternelle ? Dans le cas de la stérilité masculine chez des pères porteurs d'une translocation, il a été montré que l'origine paternelle pouvait avoir une influence sur la gamétogénèse. Les auteurs qui ont étudié ce phénomène (Guichaoua, 1991) suggèrent l'existence d'une 'compétition' entre le quadrivalent de la translocation et la vésicule sexuelle favorisant un non appariement d'une portion du quadrivalent, et une éventuelle non disjonction. Si celle-ci aboutit à une stérilité masculine, c'est probablement en raison d'une extension de l'inactivation du X aux autres autosomes (avec développement anormal des gamètes). Ce phénomène d'interférence, entre la vésicule sexuelle et le quadrivalent, existerait principalement lorsque les chromosomes impliqués dans la translocation sont des chromosomes acrocentriques. Chez la femme, bien que l'on note également un risque augmenté lorsque les chromosomes de la translocation sont des acrocentriques, il est peu probable que des phénomènes similaires existent (en raison de l'appariement complet XX). En fait, la méiose féminine est très difficile à étudier, mais il n'existe pas à notre connaissance d'études ayant montré une plus grande fréquence des translocations chez les femmes stériles par défaut de l'ovogénèse.

Chez un père porteur d'une translocation, la sélection interviendrait donc à deux niveaux, d'une part au niveau gamétique (spermatogénèse défectueuse) et d'autre part au niveau zygotique, alors que chez une mère porteuse la descendance serait soumise uniquement à la sélection naturelle zygotique. Pour cette dernière, on peut supposer aussi l'existence d'un phénomène d'empreinte parentale : un même segment en déséquilibre conduirait à un déséquilibre plus volontiers viable s'il provient de la mère que s'il provient du père. Beaucoup d'arguments existent en faveur du phénomène d'empreinte parentale. Dans le cas des translocations robertsoniennes la question posée est de savoir si la fréquence des syndromes de Down en cas d'origine maternelle de la translocation est due au fait que le chromosome surnuméraire provient de la mère. La même question peut être posée dans le cas des translocations réciproques concernant la trisomie partielle de certains segments chromosomiques. Est-ce que ce matériel chromosomique en plus a des effets cliniques moindres lorsqu'il vient de la mère que lorsqu'il vient du père (ou bien l'inverse) ? (Hall, 1990). Selon cet auteur, ce phénomène d'empreinte ressemblerait à celui de l'inactivation de l'X dans les cellules germinales mais ne toucherait que certaines parties des autosomes.

t(2;22)(q14.1;q11.1)



paternel maternel*



trisomie du segment centrique du 2
monosomie du segment transloqué du 22

Dans l'exemple ci-dessus de translocation paternelle, la trisomie du segment centrique du chromosome 2 est composée de deux segments d'origine paternelle, et d'un segment d'origine maternelle. En référence à ce que l'on rencontre dans les syndromes de Prader-Willi et d'Angelman (Kaplan, 1989), cette situation pourrait correspondre à une absence d'expression du segment centrique d'origine maternelle et être non viable, alors que la même situation due à une translocation maternelle, correspondant à une absence d'expression d'un segment centrique d'origine paternelle, conduirait à un déséquilibre viable. Cette hypothèse d'empreinte parentale, confirmée pour les disomies (Hall, 1990), nécessiterait d'être testée dans le cas des translocations réciproques.

Dans le cas des translocations réciproques, l'effet de l'origine maternelle serait donc plus dû à la viabilité des déséquilibres entraînés par la non disjonction qu'à la non disjonction elle même.

La prédominance d'observations d'origine maternelle dans nos données peut être expliquée en partie par les conséquences de la stérilité masculine (en raison d'un certain nombre de ségrégations non alternes avec non disjonctions inapparentes). Mais il n'est pas possible d'apprécier la part respective de ce facteur et celle d'autres facteurs explicatifs comme les biais de recrutement ou les différences de viabilité des déséquilibres chromosomiques en fonction de l'origine parentale.

• Les seuils quantitatifs de viabilité ne suffisent probablement pas pour expliquer la survenue d'un enfant malformé à terme.

Dans la figure 23, seule la longueur des segments en déséquilibre est prise en compte : or il semble licite d'imaginer qu'il existe des segments chromosomiques toujours viables (quelle que soit la quantité du déséquilibre) et d'autres jamais. Un fait bien connu est que certaines trisomies complètes sont viables à terme (21,18,13) et d'autres non (indépendamment de la longueur du chromosome en cause). L'existence d'une interaction entre le numéro des chromosomes et les segments transloqués est aussi en faveur de cette hypothèse. Des arguments dans ce sens sont donnés également par l'étude cartographique des segments chromosomiques en déséquilibre.

- Pour le chromosome 5, on note une forte proportion (81/128 soit 63 %) de déséquilibres observés en monosomie, alors que, pour les autres chromosomes, les déséquilibres sont toujours plus nombreux en trisomie. Pour ce chromosome 5, la monosomie partielle du bras court est probablement plus volontiers viable que pour d'autres chromosomes (quelle que soit l'origine de la translocation).

- Il existe des segments très rarement observés en déséquilibre : les translocations impliquant le chromosome 19 sont peu fréquentes (1,6 % de nos observations) alors que ce chromosome est un petit chromosome. De plus ce chromosome comporte un grand nombre de sites particuliers : 7 Zn Finger protein (Lichter, 1992) et 2 sites fragiles. Il est probable que l'explication de sa très faible fréquence dans les anomalies chromosomiques constitutionnelles en soit l'existence de gènes "vitaux" sur ce chromosome.

Au lieu d'un schéma (simpliste comme celui proposé ci-dessus), il s'agirait de pouvoir proposer plusieurs schémas, en définissant des seuils pour chacun des 44 bras chromosomiques des 22 paires d'autosomes du génome. Il pourrait exister :

- un 1^{e} seuil au delà duquel il y aurait forte probabilité de ségrégation non alterne en raison de la configuration géométrique du quadrivalent (ce seuil est probablement le même pour des chromosomes de même taille sans hétérochromatine, comme les chromosomes 5 et 6 par exemple)

- un 2^{e} ou plusieurs autres seuils définissant des portions "vitales" du génome qui, si elles font partie d'un segment en déséquilibre, rendent celui-ci très probablement non viable quelle que soit sa longueur (ces seuils seraient spécifiques à chaque bras chromosomique).

Les valeurs de ces seuils pourraient être définies quel que soit l'autre chromosome de la translocation, et en fonction de l'origine de la translocation et de l'âge des parents. Même si on entrevoit rapidement la difficulté qu'il y aura, pour ces critères qualitatifs, à prendre en compte l'existence simultanée d'un déséquilibre sur 2 chromosomes différents, il apparaît intéressant de poursuivre la recherche dans ce sens, afin de pouvoir un jour définir des critères qualitatifs de viabilité.

-------> Au total l'interprétation de cette analyse pose plus de questions qu'elle n'en résoud. En plus des commentaires exprimés ci-dessus certaines interrogations sont soulevées et peuvent être adressées aux généticiens.

1. Lorsque les points de cassures sont situés sur les bras courts des 2 chromosomes le risque de survenue d'un enfant malformé est moins grand. Pourtant, des points de cassure sur les bras courts signifient plutôt des segments en déséquilibre petits, donc plus volontiers viables. Est-ce que l'effet de cette localisation des points de cassure traduirait une influence au niveau du diagramme du pachytène en faveur d'un mode de ségrégation alterne ?
2. Lorsque les chromosomes13, 14, ou 15 sont impliqués dans la translocation, le risque de survenue d'un enfant malformé est le même quelle que soit l'origine de la translocation, alors que lorsque les chromosomes 21, ou 22 (y compris les t(11;22)) sont impliqués, ce même risque est très différent selon l'origine de la translocation (plus élévé en cas d'origine maternelle). Il s'agit à chaque fois de chromosomes acrocentriques impliqués dans la translocation. Comment expliquer que l'influence de l'origine ne soit pas semblable pour des translocations impliquant des chromosomes ayant les mêmes caractéristiques ?

3. Lorsque le premier segment transloqué augmente en taille, quel que soit le chromosome impliqué et la taille de l'autre segment transloqué, le risque de survenue d'un enfant malformé augmente. A l'opposé lorsque le deuxième segment transloqué augmente en taille, quelle que soit l'origine de la translocation et la taille de l'autre segment transloqué, le risque de survenue d'un enfant malformé diminue. Comment expliquer cette influence antagoniste entre les 2 segments transloqués? Etant donné que TSbandeR1 est souvent plus grand que TSbandeR2 et que, d'après les règles conventionnelles d'écriture d'une translocation, le premier chromosome est toujours le plus grand, peut-on envisager un effet favorisant une non disjonction pour le premier segment transloqué, et un effet favorisant une non viabilité pour le deuxième segment transloqué ?
4. Le risque de survenue d'un enfant malformé est 4 fois plus important lorsque le mode de ségrégation non alterne prédit est un mode adjacent 2 plutôt qu'un mode adjacent 1, alors qu'il n'est que 2 fois plus important lorsque le mode prédit est un mode 3:1. Comment expliquer cette influence particulière de la prédiction d'un mode adjacent 2 ?

V. 1.2. Utilisation du modèle prédictif

• **Depuis 1989, une aide informatisée au conseil génétique pour les translocations réciproques** a été mise en place au laboratoire de Grenoble. Cette aide (Réci-Conseil) comprend les fonctionalités suivantes (Cohen, 1992).

- le dessin automatique du diagramme du pachytène
- la prédiction du mode de déséquilibre le plus probable (par les deux méthodes décrites au chapitre I. 3)
- la liste des déséquilibres potentiels avec mesure quantitative des segments en déséquilibre et mention de leur viabilité
- la projection de ces déséquilibres sur l'aire de viabilité (annexe 7)
- l' information qualitative sur l'existence ou non d'un déséquilibre observé identique pour chaque segment chromosomique
- les références bibliographiques pour cette translocation et pour les translocations voisines.

Toutes ces informations sont élaborées à partir de la base de données. L'avantage de ce système est de pouvoir renseigner très rapidement le généticien, même si celui-ci consulte pour une nouvelle translocation n'existant pas dans la base de données (actuellement la réponse se fait par fax, et sera progressivement remplacée par un accès télématique direct).

Deux cent vingt cinq demandes ont été effectuées auprès de Réci-Conseil, émanant pour les trois quarts de consultations de génétique en France (n = 162) et pour un quart de services de génétique européens (Allemagne, Belgique, Espagne, Norvège, Portugal, Suisse, n = 63). Les demandes ne sont réellement importantes que depuis 1992 en raison du temps nécessaire pour diffuser l'information sur l'existence de Réci-Conseil auprès des différents services.

Dans le cadre de cette aide aux généticiens cliniciens, ces informations sont précieuses mais encore insuffisantes dans la mesure où le souhait le plus souvent formulé, que ce soit par le couple porteur de la translocation ou par le clinicien, est la quantification du risque de survenue d'un enfant malformé à terme. Notre propos est donc de montrer maintenant quelques demandes effectuées durant l'année 1993, et pour lesquelles nous avons fourni, à titre exceptionnel, une estimation de ce risque. Aucune des translocations citées dans ces exemples n'existait dans la base de données.

• **Exemples réels**

1. t(8,21)(q24.2;q22).

Il s'agit d'un couple chez qui la femme est porteuse de la translocation survenue de novo (parents avec caryotype normal). Ce couple a déjà eu deux enfants déséquilibrés à terme (mode de déséquilibre adjacent 1, der(21)) et 2 interruptions thérapeutiques de grossesse en raison de la découverte, lors d'une choriocentèse, de la translocation à l'état déséquilibré (adjacent 1 der(21) également). Devant ce couple, et à côté des informations classiquement fournies par Réci-Conseil, la demande du généticien était un avis concernant un éventuel don d'ovocytes dans le cadre d'un conseil de parentalité.

Réci-Conseil montre un mode de déséquilibre adjacent 1 le plus probable, et 7 déséquilibres potentiels viables parmi les 12 (dont les 2 modes adjacent 1). "Il semble donc bien s'agir d'une translocation à très fort risque malformatif qui justifie un diagnostic prénatal, et une choriocentèse peut être proposée."

Si l'on ne dispose pas de la quantification du risque, l'aide de Réci-Conseil s'arrête là, et, pour le don d'ovocytes, l'avis est plutôt réservé : "On ne peut pas s'opposer à cette demande qui semble licite compte-tenu des antécédents. Cependant, il importe de rappeler à ce couple que la probabilité qu'il ait une grossesse normale reste très probablement supérieure à 50 % à chaque conception".

Un mérite de notre analyse a été de montrer l'existence de translocations à très fort risque (même > 50%). Dans le cas de cette famille, nous avons voulu connaître le risque estimé par le modèle proposé. Il s'élève à 70 % ± 17 %, ce qui est tout à fait en accord avec l'histoire clinique et plaiderait en faveur du recours au don d'ovocytes ou à l'adoption pour ce couple.

2. t(2,15)(p25;q21).

Il s'agit d'un couple jeune, chez qui le mari est porteur sain de la translocation. Ils ont eu 4 avortements spontanés (dont 1 caryotypé, avec translocation à l'état déséquilibré) plus un cinquième récemment n'ayant pu être caryotypé. Le père du mari est également porteur de la même translocation, il n'a eu que cet enfant unique sans notion d'avortements spontanés. La question posée par le généticien est : faut-il risquer une autre grossesse "naturelle" pour ce couple déjà éprouvé par ces 5 avortements ? ou faut-il envisager un autre mode de procréation ?

La première réponse avec les fonctionnalités actuelles de Réci-Conseil a été la suivante. Le mode déséquilibre le plus probable est le mode 3:1 tertiaire, et 5 déséquilibres potentiels sont viables parmi les 12. "Ces déséquilibres sont caractérisés par un faible pourcentage de trisomie et / ou de monosomie. Il s'agit donc d'une translocation à risque malformatif élevé, pour laquelle une choriocentèse peut être proposée au titre d'un diagnostic prénatal."

En utilisant la prédiction du modèle, la réponse est plus précise puisqu'elle évalue le risque de déséquilibre à terme à 19 % ± 8 %. A la question posée, on peut répondre au généticien que ce couple a environ 4 chances sur 5 d'avoir une grossesse normale si celle-ci arrive à terme.

3. t(3;8)(p21;p11.2).

Il s'agit d'un couple chez qui la femme est porteuse de la translocation. Pour ce couple ayant déjà eu 4 fausses couches précoces, la question d'un don éventuel d'ovocytes a été posée à la commission génétique du CECOS (Centre d'Etude et de Conservation des Oeufs et du Sperme humain).

La réponse de Réci-Conseil montre que seuls 4 déséquilibres sur 12 sont potentiellement viables et que le mode de déséquilibre le plus probable est un mode adjacent 1. "Le risque de cette translocation semble donc faible."

Le modèle proposé évalue le risque malformatif à 4,9 % ± 1,7 % ce qui est concordant avec les résultats de Réci-conseil. "Le don d'ovocytes demandé par cette patiente ne semble pas la meilleure solution, puisque de nouvelles tentatives devraient pouvoir se concrétiser par une descendance tout à fait normale."

• **Suite à ces exemples se pose la question de l'utilisation de la connaissance du risque de survenue d'un enfant malformé.** Cette information brute est certainement importante pour le couple porteur de la translocation. Pour le généticien, elle fait partie d'un ensemble d'informations nécessaires à connaître, mais pas forcément suffisantes, pour guider le conseil génétique, en particulier :

- choix de la méthode de diagnostic prénatal. Face aux deux méthodes possibles, la choriocentèse sera préférée, lorsque le risque malformatif sera fort, et l'amniocentèse sera préférée, lorsque ce risque sera faible. Il semble important d'insister sur ce choix entre les deux méthodes, puisqu'il semble à l'heure actuelle que ce choix soit surtout fonction de l'équipe médicale et non pas d'une appréciation du risque malformatif (même s'il est clair que l'on a plus souvent recours à l'amniocentèse qu'à la choriocentèse). Par ailleurs, reste le problème de la définition du seuil pour ces risques : 5, 10, ou 20 % ? Il semble que 10 % soit considéré comme un risque déjà élevé.

- abstention d'un diagnostic prénatal. Cet aspect est plus difficile à aborder dans la mesure où cela peut paraître choquant de ne pas recourir au diagnostic prénatal, lorsque celui-ci est possible. Même si les symptômes cliniques d'un enfant déséquilibré pour une translocation sont plus sévères en moyenne que les symptômes du syndrome de Down, cela rejoint le problème de la

trisomie 21 débattu récemment : faut-il proposer une amniocentèse systématique pour les femmes entre 35 et 38 ans, et à partir de quelle valeur de risque cela semble-t-il licite ? En terme de stratégie de diagnostic prénatal systématique on admet classiquement le seuil de 1 %, pour définir un risque élevé pouvant justifier un diagnostic prénatal systématique. Si l'on regarde les translocations pour lesquelles le risque est inférieur à 1%, même si elles sont rares (2,3 % de la base de données, avec, pour plus de la moitié d'entre elles, une borne supérieure de l'IC de la valeur prédite < 1%), on est en droit de discuter l'opportunité d'un diagnostic prénatal par amniocentèse pour ces couples. En effet le risque n'est pas plus élevé que celui de donner naissance à un enfant trisomique, par contre le risque de fausses couches après amniocentèse, même s'il est faible, existe. De plus si l'on imagine l'application de cette stratégie à un ensemble de personnes avec cette fréquence de pathologie (entre 1 % et 1 ‰), la valeur prédictive positive de l'amniocentèse est de 91 %, pour une fréquence de la maladie de 1 %, mais elle n'est plus que de 50 % pour une fréquence de la maladie de 1‰ (avec, dans les deux cas, une excellente valeur prédictive négative). Ceci constitue donc un argument de réflexion supplémentaire pour les couples porteurs et les généticiens face à la réalisation d'un diagnostic prénatal, lorsqu'il s'agit d'une translocation à très faible risque de déséquilibre (< 1 %). Compte tenu des biais de recrutement des données (détaillés en V.2.1), l'estimation de ce risque est surestimée, et le fait que les risques réels soient a priori moins élevés que ceux prédits par le modèle proposé constitue un élément rassurant.

- pour les risques élevés, notre travail a permis de faire prendre conscience aux cliniciens que certaines translocations ont un risque supérieur à 50 % de survenue d'un enfant malformé à terme, autrement dit un risque supérieur au risque de transmission de maladie autosomique dominante. Cette information est très importante, dans la mesure où elle peut conduire les couples à très fort risque à choisir éventuellement un autre moyen de procréation : soit FIV (Fécondation In Vitro) avec ou non diagnostic pré-implantatoire, soit don de gamètes, soit adoption. Les généticiens ont l'expérience clinique de certains couples gravement touchés par des avortements successifs (spontanés ou thérapeutiques) et pour lesquels il serait plus logique de déconseiller une grossesse naturelle. Dans ce cas, le biais de surestimation de la valeur de ce risque n'est par contre pas à négliger.

---> Il est clair que ces aspects de choix de diagnostic prénatal ou de stratégie de parentalité ne peuvent être qu'abordés seulement ici. Comme cela a été souligné dans le chapitre I (p.14), un travail approfondi, avec des arguments économiques complémentaires, nécessite d'être entrepris. Ce travail devra discuter les meilleures stratégies, d'une part au niveau individuel, et d'autre part au niveau collectif en terme de Santé Publique (en sachant que l'intérêt individuel diffère très souvent de l'intérêt collectif).

Il est important de souligner enfin que le modèle proposé doit être "réactualisé" régulièrement avec l'augmentation du nombre d'observations nouvelles dans la base de données. Ceci, dans le but d'obtenir une meilleure précision des paramètres, et de vérifier que ces nouvelles observations ne changent pas l'estimation de ces paramètres. Ce travail devrait être effectué tous les ans environ.

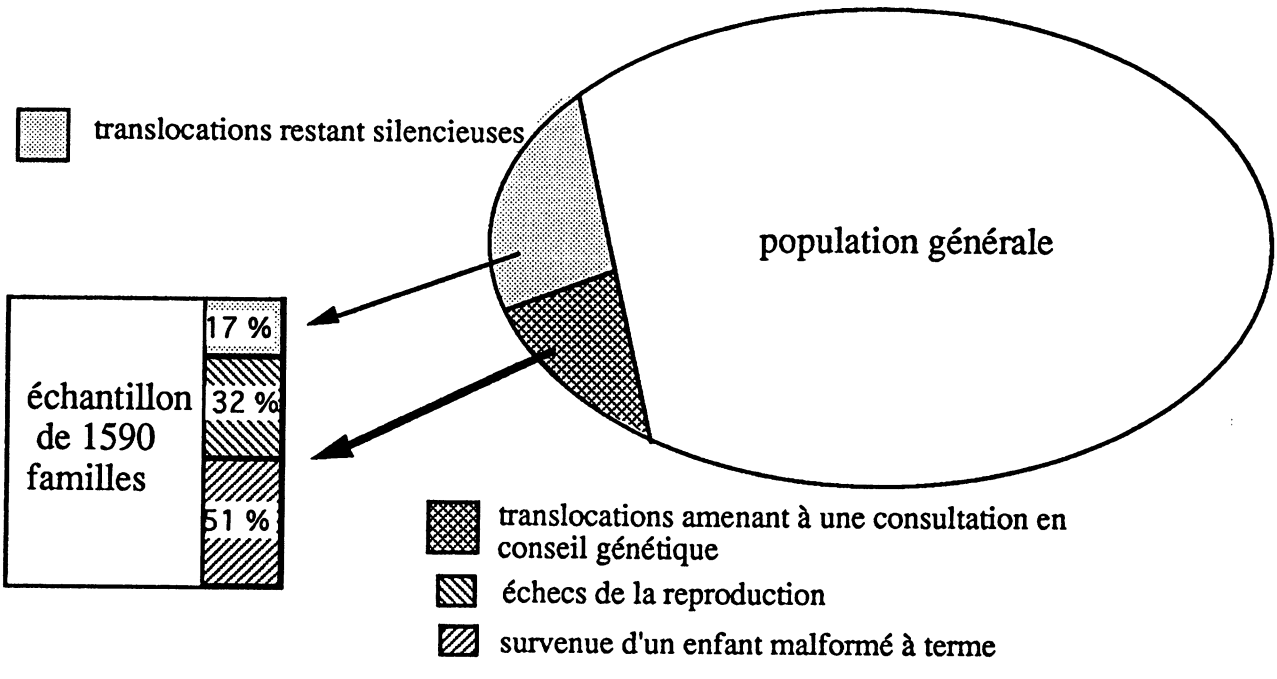
V. 2. Discussion sur les données

Nous discutons d'abord la qualité des données recueillies, ainsi que les différents biais qui les accompagnent, puis les choix de sélection effectués pour l'analyse.

V. 2.1. Le recrutement des données

• Représentativité de l'échantillon

Le mode de recrutement de nos données est représenté de façon schématique ci-dessous.



La composition des fichiers sur lesquels nous avons travaillé est influencée par le recrutement des cas. Dans notre échantillon, les familles porteuses d'une translocation (n=1590) ont été découvertes dans 51 % des cas en raison de la survenue d'un enfant malformé, dans 32 % des cas en raison d'échecs de la reproduction et dans 17 % des cas de manière fortuite (risque particulier, comme l'âge maternel avancé, conduisant à un diagnostic prénatal systématique).

Ce mode de recensement des cas ne fausse pas l'interprétation des résultats, dans la mesure où la prédiction issue de ces données ne pourra s'appliquer qu'à de nouveaux individus recrutés de la même façon. En effet, il est illusoire pour l'instant d'imaginer pouvoir réaliser des études ou des programmes de dépistage sur les translocations réciproques en population générale, car les méthodes de caryotype sont longues et coûteuses et les laboratoires de cytogénétique surchargés. Donc, l'outil prédictif ne peut s'appliquer qu'aux individus pour lesquels un conseil génétique ou prénatal est demandé, le plus souvent en raison de l'existence d'une pathologie ou d'un facteur de risque. Peut-être existe-t-il en population générale une fraction d'individus porteurs "inconnus" d'une translocation réciproque, mais on est en droit de penser que, si la translocation est restée silencieuse jusqu'alors (pas de fausses couches patentes répétées, ni d'enfants malformés), c'est que le risque de survenue d'un enfant malformé ou d'avortements à répétition pour ces individus est faible. Parmi les translocations découvertes de manière fortuite (translocations "silencieuses", 262 familles parmi 1590), 49 ont été découvertes en raison d'un âge maternel avancé (38 ans et plus). Dans la descendance de ces individus détectés fortuitement, on note seulement 11 modes de ségrégation non alterne (6 enfants malformés à terme et 5 diagnostics prénataux ou ITG), ce qui tend à conforter l'hypothèse ci-dessus.

• Il existe un biais de sélection

Le fait que le fichier de données issu de la littérature comporte une plus grande proportion d'enfants malformés est lié au biais de publication. Ce biais est un peu différent du précédent. Les conséquences en sont principalement une surestimation du risque de survenue d'un enfant malformé, cette erreur pouvant être un avantage ou un inconvénient selon les applications pratiques des résultats (cf ci-dessus).

V. 2.2. La sélection des données pour l'analyse

• Exclusion des individus avec origine inconnue.

Cette décision présentée au chapitre III (p.55) peut être critiquable. Elle nous a cependant semblé nécessaire dans la mesure où elle permettait d'éliminer un biais de sélection (origine moins souvent connue lorsqu'il n'y a pas eu d'enfant malformé) et rendait ainsi interprétable l'influence de l'origine parentale sur la survenue d'un enfant déséquilibré à terme. Si l'on essaye d'analyser les conséquences de cette exclusion, on sait que la proportion de déséquilibres à terme est plus importante après avoir éliminé les individus avec origine inconnue puisqu'elle passe de 16 % à 21 %. Ceci conduit donc à une surestimation du risque de déséquilibre à terme.

• Choix de garder le proposant.

Etant donné le mode de recrutement de nos données (cf ci-dessus), il ne s'agit pas de vouloir calculer un risque vrai de survenue d'un enfant malformé pour toute la population des porteurs de translocation réciproque, ce qui nécessiterait l'élaboration d'un plan d'expérience particulier. L'objectif est de déterminer, de façon empirique, le risque de survenue d'un enfant malformé pour

les familles porteuses de translocation et consultant les services de génétique. Dans la plupart des études descriptives sur les translocations réciproques (Boué, 1986, Stengel Rutkowski, 1988), les auteurs ont exclu, pour le calcul du risque, l'individu ayant permis la détection de la translocation (ou "proposant"). Cette exclusion du proposant ne nous paraît pas judicieuse dans notre étude pour les raisons suivantes :

- Il ne s'agit pas d'estimer, pour les translocations, le risque de récurrence d'une malformation. En prenant l'exemple de la fente palatine, il existe un risque de 1,7 ‰ en population générale de survenue d'un enfant avec cette malformation (Goodman, 1983). Le calcul du risque de récurrence, après exclusion du proposant, consiste à estimer l'accroissement du risque de base pour les familles présentant déjà un de leurs membres porteurs d'une fente palatine (risque de l'ordre de 2 à 5 %). Pour les translocations réciproques la problématique est différente. Environ 1,5 ‰ individus sont porteurs de translocations réciproques en population générale (voir p.12). Il existe des translocations pour lesquelles il n'y a jamais d'enfant malformé observé à terme (risque de base nul). L'hypothèse d'indépendance dans le cas des translocations est adoptée tacitement par la majorité des généticiens. Elle suppose, qu'au sein d'une même famille, le risque de survenue d'un enfant malformé est le même lors de deux grossesses successives, et que ce risque reste identique pour différentes familles avec la même translocation (Stengel-Rutkowski, 1988). **On ne se place donc pas dans les conditions d'un modèle de ségrégation mendélien (monolocus ou polylocus).**

- L'estimation du risque pour chaque translocation doit être faite indépendamment du mode de recrutement de celle-ci. En effet, ce mode de recrutement est souvent mal défini : 1) il peut être multiple comme un déséquilibre à terme et un échec de la reproduction pour une même famille, 2) il peut nécessiter de choisir un proposant parmi 2 individus d'une même famille, alors que le mode de recrutement de ces individus n'est pas indépendant, et 3) il peut être variable selon les habitudes médicales de la région concernée.

- Le fait de garder le proposant peut conduire tout au plus à une sur-estimation du risque de survenue d'un enfant malformé. Ce choix est préférable à celui d'exclure le proposant qui pourrait conduire à une sous-estimation dangereuse de ce risque.

A posteriori ce choix apparaît conforté par le fait que le modèle proposé reste "stable" (robuste dans l'estimation des coefficients, mêmes variables explicatives pour le meilleur gain en déviance) lorsqu'il est appliqué séparément sur les 3 fichiers de données, alors que ceux-ci diffèrent quant à leur mode de recrutement (dans le fichier issu de la littérature la survenue d'un enfant malformé est le mode de détection principal).

V. 3. Perspectives

V. 3.1. Amélioration du recueil des données

Si les problèmes de recrutement des fichiers nous semblent incontournables et donc difficilement améliorables, il n'en est pas de même de la nature des données recueillies.

• **En premier lieu, concernant l'âge des parents au moment de la grossesse.**

Cette information n'est disponible que pour un quart des individus seulement et plus souvent lorsqu'il y a eu naissance d'un enfant malformé. Nous avons vu au chapitre III (p.57) les conséquences dues au fait que cet âge est plus souvent connu, soit pour des parents jeunes avec enfant malformé, soit pour des parents âgés lors d'un diagnostic prénatal. L'analyse de ce facteur est donc rendu quasiment impossible, en raison du grand nombre d'informations manquantes. Nous avons déjà souligné que ce facteur pourrait être étudié si l'on disposait de données suffisantes concernant des translocations découvertes de façon systématique. En excluant l'individu ayant servi à repérer la translocation (qui appartient à des tranches d'âge particulières), on peut espérer ainsi supprimer l'effet du biais de recrutement sur l'âge du parent. Etant donné l'influence bien connue de l'âge parental sur un grand nombre d'anomalies chromosomiques, il est probable qu'il intervienne aussi pour les translocations réciproques, soit au niveau de la ségrégation méiotique, soit au niveau des mécanismes de sélection naturelle. Il est donc impératif qu'un effort important soit entrepris auprès des sources de données, afin de pouvoir disposer de cette information pour toutes les translocations. Sachant qu'il ne nous a pas été possible de récupérer cette information pour les translocations déjà saisies dans la base de données (que ce soit des données de la littérature ou celles issues des laboratoires), la seule solution pour remédier à cet état de fait serait l'obtention de cette information pour les nouvelles translocations signalées par les laboratoires depuis la mise en service de Réci-Conseil. Cette information sur l'âge doit être obtenue de manière précise (par exemple la date de naissance des parents n'est pas suffisante si la date des grossesses n'est pas mentionnée). Cela impliquera probablement un contact téléphonique avec les correspondants "sources de données", en attendant que l'habitude soit prise. Pour les articles qui sont aussi rentrés dans la base au fur et à mesure de leur parution, un contact avec l'auteur serait nécessaire afin d'obtenir cette information.

• **Concernant la procréation des individus porteurs d'une translocation.**

Autant la survenue d'un enfant en déséquilibre est toujours identifiée ou presque (certains cas peuvent ne pas l'avoir été ou l'avoir été à tort, mais ceci ne devrait plus arriver), autant les événements issus d'une ségrégation méiotique non alterne mais n'arrivant pas à terme sont moins bien connus. Par exemple :

* Les *ITG* pour lesquelles on ne sait pas ce qu'aurait donné l'évolution naturelle de la grossesse (décès in utéro ou enfant malformé à terme). A moins d'avoir la connaissance du devenir naturel

d'un déséquilibre chromosomique identique (même mode de ségrégation pour une même translocation), il ne sera jamais possible de prévoir ce qu'il serait advenu de cette grossesse en l'absence d'ITG. Il s'agit d'individus avec données censurées.

* Le *diagnostic prénatal* confirmant une ségrégation non alterne, mais pour lequel on n'a pas l'information sur l'issue de la grossesse. Par contact auprès de la source de données, l'information sur le suivi de cette grossesse doit à tout prix être recherchée (ITG, fausse couche tardive, mort-né, déséquilibre à terme ?), afin de limiter le nombre d'individus avec données manquantes.

* L'information sur les *fausses couches non caryotypées* est inexploitable, dans la mesure où l'on ne sera jamais sûr de la cause de la fausse couche. Pour les FC précoces, il n'y a pas de solution possible à ce problème, puisqu'un grand nombre d'entre elles sont inapparentes (Alberman, 1973). Par contre, pour les fausses couches après le troisième mois de grossesse, un caryotype devrait être systématiquement demandé et exécuté chaque fois que possible.

* Dans le même ordre d'idée l'existence d'une *stérilité masculine* n'est pas mentionnée actuellement dans la base de données. La raison en est principalement l'absence de signalement par les sources de données (littérature surtout) de ces pères stériles porteurs de translocation réciproque, puisque les problèmes de descendance sont très particuliers. Sa prise en compte pourrait aider à préciser les mécanismes de l'influence de l'origine parentale sur la survenue d'un déséquilibre à terme.

En améliorant ainsi le recueil des données concernant la procréation des individus porteurs d'une translocation, l'avantage serait de pouvoir étudier de façon plus précise le groupe des individus présentant un mode de ségrégation non alterne, lors de la méiose, mais ne donnant pas lieu à un enfant malformé à terme. En effet, jusqu'à présent, ce groupe est joint au groupe des individus avec ségrégation alterne, mais l'intérêt de l'étudier séparément serait double. D'une part, cela devrait permettre de pouvoir mieux préciser les critères de viabilité d'un déséquilibre (en disposant de plus d'information pour tous les déséquilibres non viables). D'autre part, même s'il est vrai que la survenue d'un enfant malformé à terme est probablement l'événement le plus grave, il semble aussi intéressant de vouloir essayer de quantifier le risque d'avoir des fausses couches à répétition, d'avoir un enfant mort-né ou de présenter une stérilité masculine. Si l'on considère que le but est de fournir une information aux individus porteurs d'une translocation, ces aspects ne sont pas à négliger. Ce serait à l'équipe de Réci-Conseil de sensibiliser les cliniciens sur ces points et de tenir à jour une liste pour chaque correspondant, avec les dernières informations à récupérer lors du prochain contact. D'un point de vue statistique, un exemple de méthode adaptée serait le modèle log-linéaire, puisque la variable à expliquer ne serait plus dichotomique, mais qualitative à plusieurs modalités (survenue d'un enfant malformé à terme, avortements répétés, mort-né, stérilité, individu normal ou porteur sain).

V. 3.2. Les nouvelles variables

En premier lieu seront décrites les variables facilement accessibles et pouvant être prises en compte dans un avenir proche. Ensuite, seront présentés des axes de travail dans le cadre de la recherche de nouvelles variables pertinentes explicatives.

• Il existe des variables qui pourraient être intéressantes et disponibles prochainement pour le recueil

La viabilité qualitative.

Concernant la viabilité des segments en déséquilibre, nous n'avons pu utiliser jusqu'à présent que des variables mesurant la quantité du déséquilibre chromosomique et non pas la qualité de celui-ci (à l'exception de la prise en compte de la nature du contenu chromosomique pour les segments issus de la translocation : bandes R des segments centriques et transloqués). Des arguments cytogénétiques (trisomies non viables) et des arguments issus de notre analyse (interaction entre le numéro des chromosomes et la taille des segments chromosomiques) laissent supposer que la viabilité du déséquilibre chromosomique dépend aussi de son contenu et non pas uniquement de sa taille. L'exploitation des données cartographiques concernant les translocations de la base de données, rendue possible grâce à son informatisation, pourrait être utilisée en extrayant les données suivantes pour chaque translocation.

1) Existe-t-il dans la base de données un déséquilibre observé à terme correspondant au mode de déséquilibre le plus probable* ? Cette information pourrait être recueillie sous la forme d'une variable dichotomique (oui/non). Mais, dans la mesure où la prédiction du mode de déséquilibre ne précise pas le chromosome dérivé, mais seulement s'il s'agit d'un adjacent 1, adjacent 2, 3:1 tertiaire ou 3:1 échange il serait nécessaire de compléter cette information. Pour chaque mode de déséquilibre le plus probable, il existe 2 possibilités pour les modes adjacent et 4 possibilités pour les modes 3:1 (se rapporter à la figure 2 du chap. I).

2) Dans le cas où un déséquilibre observé à terme correspondant au mode de déséquilibre le plus probable aurait déjà été observé viable à terme, on pourrait également prendre en compte les précisions suivantes :

- pour les modes adjacent, a-t-on observé à terme une seulement ou les deux possibilités de déséquilibre, et, de façon similaire, pour les modes 3:1, a-t-on observé à terme une seulement, deux, trois ou les quatre possibilités de déséquilibre ?

- combien de fois le mode de déséquilibre le plus probable a-t-il été observé viable à terme ? Cette information serait à recueillir soit sous la forme d'une variable catégorielle (1 fois, entre 1 et 5 fois, plus de 5 fois) ou sous forme quantitative discrète.

---

* Il est bien évident que, si un déséquilibre observé englobe (en qualité et quantité) un déséquilibre encore jamais observé, ce dernier sera considéré aussi comme déjà observé. Pour les segments caractérisés par une très grande taille, on s'efforcera de prendre en compte le fait qu'ils aient été validés ou non (c'est-à-dire que leur observation ait été confirmée ou non par l'étude de l'image caryotypique du déséquilibre).

3) Dans le cas où aucun déséquilibre observé à terme ne correspondrait au mode de déséquilibre le plus probable, il serait intéressant alors de tenir compte des différents segments impliqués dans le déséquilibre. La translocation peut n'avoir jamais conduit à un déséquilibre viable à terme ou avoir conduit à un mode de déséquilibre ne correspondant pas au mode le plus probable. Le déséquilibre est le plus souvent composite sur deux chromosomes différents, sauf pour le mode 3:1 échange; il est donc nécessaire de considérer la viabilité pour chacun des segments chromosomiques :

- Aucun des deux segments du déséquilibre n'a déjà été observé viable à terme.
- Un des segments du déséquilibre a déjà été observé viable à terme, mais l'autre non.
- Les 2 segments du déséquilibre ont déjà été observés à terme, mais pas de manière conjointe. Pour cette éventualité, qui risque d'être fréquente, il est impossible de se prononcer sur la viabilité du déséquilibre chromosomique global (résultant de la combinaison des deux segments). En effet, on peut envisager la situation suivante : un segment en déséquilibre du chromosome 5 associé à un segment en déséquilibre du chromosome 13 serait non viable, alors que le même segment du chromosome 5 associé à un autre chromosome serait viable et que le même segment chromosomique du 13 associé à un autre chromosome serait également viable.

------> Au total, afin de tenir compte de toutes les situations décrites en 1), 2) et 3), et en accord avec les généticiens, on pourrait proposer la construction d'une variable quantitative discrète prenant les valeurs suivantes :

0	correspondant à aucun déséquilibre composite, ni aucun segment observé à terme
0,5	correspondant à un segment observé à terme
0,9	correspondant à deux segments observés à terme, mais pas de manière conjointe
1	correspondant à un seul déséquilibre composite identique observé à terme
2 à 5	correspondant à plus d'un déséquilibre composite identique observé à terme
6 et plus	correspondant à plus de cinq déséquilibres composites identiques observés à terme

Chacune des possibilités du mode de déséquilibre le plus probable pourrait ainsi être prise en compte de façon additive simple. Avant d'utiliser réellement la variable ainsi décrite, une amélioration de la prédiction du mode de déséquilibre le plus probable serait souhaitable (qui tiendrait compte aussi des variables décrivant la viabilité quantitative).

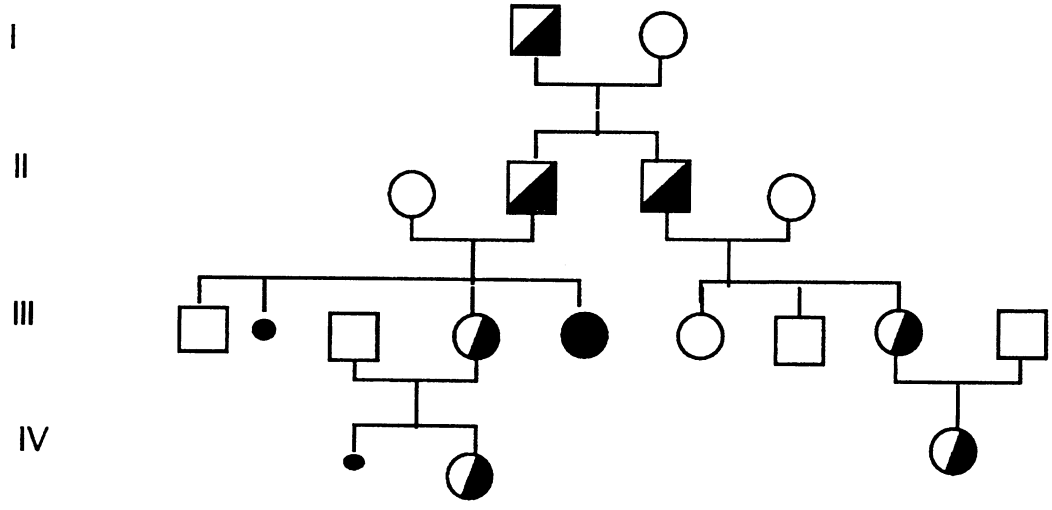
Contenu chromosomique des segments impliqués dans la translocation.

A côté de la prise en compte des bandes R, dont nous avons déjà montré l'intérêt (Cans 1993), des éléments tirés de la meilleure connaissance du contenu génique pour chaque chromosome pourraient être utilisés de façon équivalente. Récemment, il a été proposé une subdivision des bandes R en 4 sous-groupes différents, en fonction de leur richesse en séquence CG et en séquence ALU (Holmquist, 1992). Ce travail souligne l'importance des régions riches en séquences CG, correspondant aux bandes T (Dutrillaux, 1977), qui contiennent environ 65 % de l'ensemble des gènes, alors qu'elles ne constituent que 15 % du génome. Ces 4 types de bandes R ont été intégrées récemment dans la base de données (bandes R Alu+ CG+, bandes R Alu+ CG-, bandes R Alu- CG-, bandes R Alu- CG+). Il sera intéressant de tester si la prise en compte de la

taille des segments centriques et transloqués en longueur absolue de chacune de ces 4 "sous-bande R" pourrait améliorer les performances du modèle.

• **A côté des informations ci-dessus,** qui peuvent laisser envisager une amélioration du modèle dans un avenir proche, il existe aussi d'autres variables intéressantes, mais impossibles à analyser pour l'instant.

Utilisation d'individus apparentés dans l'analyse du contrôle génétique de la survenue d'un enfant malformé.



Le risque de survenue d'un enfant malformé à terme est-il le même pour III 3 et III 7, et pour IV 2 et IV 3 ?

La sélection naturelle ou la survenue d'une non disjonction pourraient-elles être sous contrôle de gènes intervenant dans la viabilité zygotique ou sur le déroulement de la méiose ? Ou bien encore, comme cela est suggéré dans la neurofibromatose, existerait-t-il des gènes modificateurs du phénotype (Easton, 1993) ? Pour étayer cette hypothèse, nous avons voulu vérifier si le risque de survenue d'un déséquilibre à terme était le même, au sein d'une même famille, quelle que soit la distance (degré de parenté) avec le proposant déséquilibré à terme. Malheureusement, le nombre d'observations par famille est restreint (5 en moyenne) et, même parmi les t(11,22) pour lesquelles on dispose souvent de généalogies plus complètes, il ne nous a pas été possible d'étudier cet aspect (une seule famille de 28 individus avec 2 déséquilibres observés à terme permettait d'étudier des degrés de parenté différents). L'augmentation régulière de la base de données ne suffira pas, si on veut continuer à travailler sur cette hypothèse. Il serait donc nécessaire de procéder à un enregistrement spécifique, auprès de quelques familles, pour obtenir suffisamment d'observations pour un même pedigree.

Les signes cliniques.

De la même façon que nous avons déjà suggéré (p.110) l'intérêt de disposer de renseignements plus précis quant au devenir (en particulier concernant les échecs de reproduction),

il serait également intéressant de préciser la gravité des syndromes cliniques des enfants porteurs d'une translocation à l'état déséquilibré. S'il est vrai que, dans la très grande majorité des cas, l'enfant est porteur d'un retard mental sévère associé à des malformations multiples, on connaît cependant la variation de ces pathologies, en particulier concernant leur durée de survie. Cette variabilité s'explique aisément par le contenu génétique différent pour chacun de ces déséquilibres. Même si cela nous semble d'un moins grand intérêt pratique immédiat, la gravité des symptômes cliniques des enfants malformés reste à explorer, et cet aspect pourrait aussi intéresser les biologistes moléculaires.

• **Enfin, toujours dans l'idée** que ce qui manque le plus dans cette problématique de prédire le risque de survenue d'un enfant malformé à terme, ce sont de nouvelles variables explicatives, un autre axe de recherche pourrait être exploré. Il s'agit de celui de l'étude des mécanismes méiotiques en biologie animale (nombreuses expériences réalisées sur la drosophile). Un contact avec les biologistes ou une revue exhaustive de la littérature sur la méiose dans le monde animal, pourrait peut-être suggérer de nouvelles variables caractéristiques du parent porteur ou de la fécondation et pouvant avoir une influence sur les phénomènes de ségrégation méiotique et / ou de sélection naturelle. En s'assurant de la faisabilité du recueil de ces variables dans le cas des translocations humaines (données pas toujours faciles à connaître chez l'homme comme le délai de fécondation par exemple), leur influence sur la survenue d'un enfant malformé à terme pourrait être testée pour les translocations de la base de données.

V. 3.3. Le cas de la t(11;22)(q;q)

Cette translocation est la plus fréquente des anomalies chromosomiques constitutionnelles (128 familles et 675 individus, soit 8 % de notre base de données), devant les t(1;4), t(4;18) et t(9;22). Elle sera étudiée particulièrement, et à titre d'exemple, pour montrer le type d'informations qu'il pourrait être utile de rassembler dans la perspective de développer l'utilisation cartographique de la base de données.

• Le point de vue clinique

Du point de vue clinique, les parents porteurs sains de cette translocation sont phénotypiquement sains, et ne semblent pas présenter un risque accru de cancer (Budarf, 1989), bien que cette translocation soit fréquemment impliquée en pathologie tumorale (sarcome d'Ewing, neuroépithéliome). La majorité d'entre elles sont détectées par la survenue d'un enfant malformé à terme (72 % des cas), le déséquilibre comportant une trisomie partielle composite : 11q23-->11 ter, et 22q11.2-->22qter, issue d'un mode de ségrégation 3:1 tertiaire à 47 chromosomes. Ce mode de ségrégation est d'autant plus fréquent que la translocation est portée par la mère (c'est pour cette translocation que l'influence de l'origine maternelle est la plus importante).

Le syndrome clinique résultant de la trisomie composite 11q23-->11 ter, et 22q11.2-->22qter a été étudié par de nombreux auteurs (Fraccaro, 1980), et sa description figure dans l'atlas de De Grouchy (1982). Il associe plusieurs malformations (crânio-faciales, urogénitales et squelettiques) et un retard mental sévère. Son évolution est caractérisée par une mortalité précoce. La localisation exacte des points de cassure par les techniques cytogénétiques classiques est difficile : certaines translocations t(11;22) ont été rapportées en 11q24, 11q25, 22q12, ou 22q13 mais le plus fréquemment il s'agit d'une t(11;22)(q23;q11.2). Seule une étude en biologie moléculaire peut déterminer précisément ces points de cassure. Après avoir réalisé cette étude sur des translocations constitutionnelles et tumorales, Budarf suggère que les points de cassure sont différents en fait pour ces deux anomalies, constitutionnelle et tumorale, apparemment identiques (Budarf, 1989).

• Le point de vue moléculaire

Les schémas ci-dessous montrent les données actuellement disponibles concernant les gènes, les sites fragiles et les séquences DNA connues pour les deux points de cassure impliqués dans cette translocation.

Bande 11q23 du chromosome 11

Cette bande comporte en fait 3 sous-bandes, 2 sous-bandes R et 1 sous-bande G, et le point de cassure des translocations constitutionnelles a été localisé sur la sous-bande G (11q23.2). On note la présence d' un site fragile sensible au folate (Hecht, 1984) et de deux Zn Finger protéines.
Le schéma ci-après (schéma 4) est emprunté à Junien (1991). Récemment, 112 microsatellites marqueurs ont pu être positionnés sur ce chromosome 11, soulignant l'évolution constante et rapide de cette carte physique (Couillin, 1994). En 1991, 155 gènes étaient assignés à ce chromosome, et le nombre de sondes connues actuellement s'élève à plus de 1400.

Bande 22q11.2 du chromosome 22

Il s'agit d'une bandeR, riche en séquences Alu, avec un site fragile et quatre Zn Finger protéines.
La représentation cartographique du chromosome 22 (schéma 5) est empruntée à Emanuel (1991). Pour ce chromosome, le nombre de gène assignés est de 68, et il existe plus de 450 sondes spécifiques.

Schéma 4. Chromosome 11

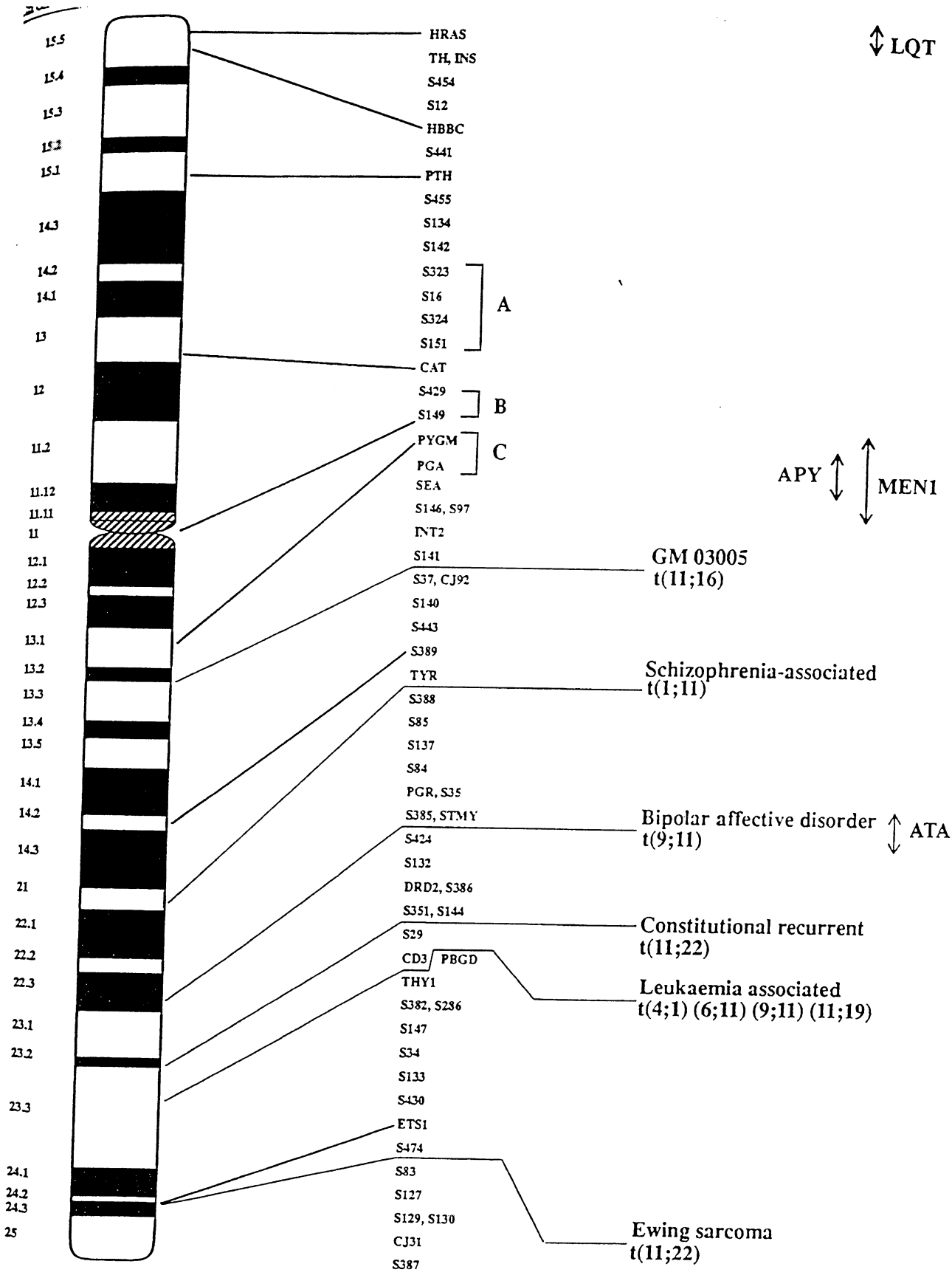


Figure: 5. Preliminary CEPH linkage map. Ordering of loci by genetic linkage in a preliminary (August 1991) CEPH consortium map. Well-defined 11q breakpoints have been superimposed. Refined disease linkage regions are indicated by arrows. Bracketed regions A, B, C show inconsistent order with firm physical assignments.

Schéma 5. Chromosome 22

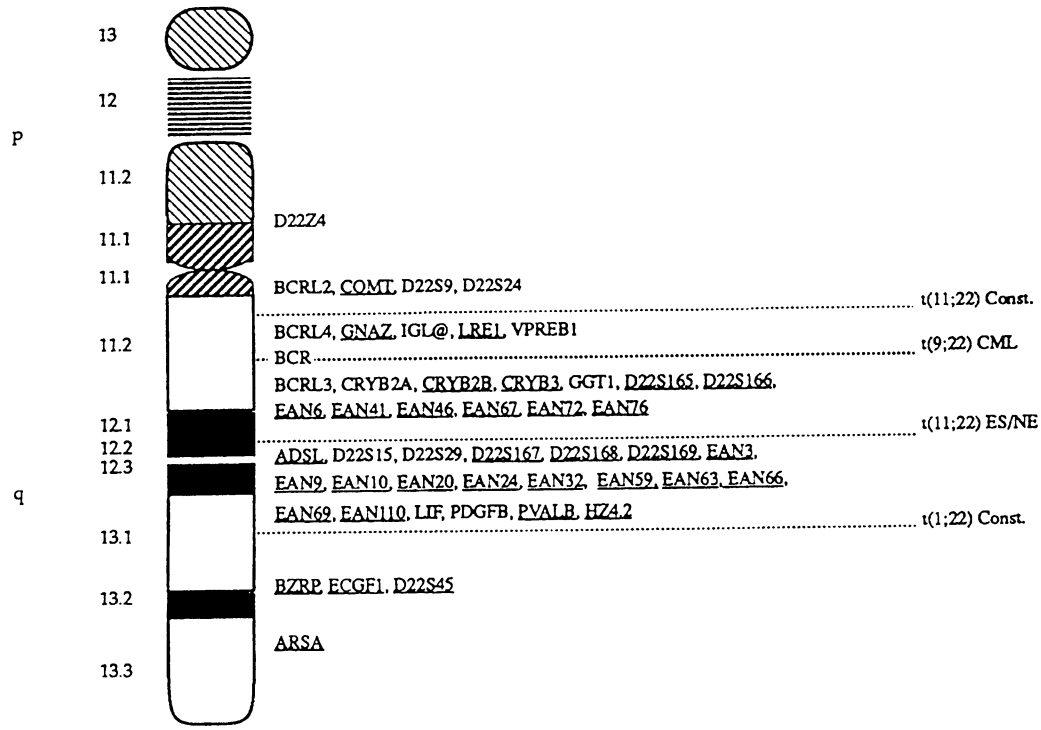


Figure: 3. Physical Location of some of the Chromosome 22 markers which have been mapped with respect to the reference hybrid panel. Underlined probes are sublocalisations since HGM10.5 was published. EAN probes are anonymous DNA markers reported at HGM11 (Fiedler et al.,1991). Markers are grouped, not ordered, within the breakpoint intervals.

• **Utilisation de ces informations :**

Les exploitations possibles de ces informations sont valables pour toutes les translocations réciproques et pas seulement pour la t(11;22)(q;q).

Pour certains individus porteurs d'une translocation à l'état apparemment équilibré, l'expression phénotypique montre un syndrome malformatif particulier (Mujica, 1989, Wilson, 1991, Imaizumi, 1991). Pour ces individus, la détermination exacte du point de cassure, en biologie moléculaire, peut se révéler particulièrement informative. Le syndrome clinique, dans ces cas, peut être dû, soit à une perte chromosomique infra-visible, soit au fait que la cassure s'est produite au niveau d'un gène lui-même. D'où l'intérêt de pouvoir analyser les prélèvements de ces individus pour les équipes de biologie moléculaire travaillant sur une région particulière du génome. Dans certains cas, une application pourrait même être de confirmer, dans les situations difficiles de conseil génétique, qu'il n'y a pas eu de perte chromosomique infra-visible pour une translocation apparemment équilibrée (Puissant, 1988).

L'étude phénotypique détaillée d'enfants malformés, issus d'une ségrégation non alterne, pourrait être complétée par une détermination exacte des points de cassure, permettant de mieux contrôler les séquences et les gènes en trisomie. Cette démarche pourrait suggérer des hypothèses quant à l'expression de caractères cliniques complexes, et surtout quant à la détermination de critères de viabilité pour les différents syndromes malformatifs présentés. Ce type d'étude serait d'autant plus fructueux qu'il existerait un nombre suffisamment important de prélèvements d' enfants malformés pour des translocations voisines ou apparemment identiques.

D'un point de vue plus général, le repérage des séquences connues d'ADN au niveau des points de cassure, suivi de l'étude de la corrélation entre ces séquences et des séquences télomériques ou centromériques connues, devrait pouvoir améliorer la connaissance quant au mécanisme de cassure dans les remaniements chromosomiques. Warburton (1991) souligne que l'apparente relation entre les sites fragiles et les points de cassure doit être regardée avec beaucoup de précaution quand la biologie moléculaire n'est pas utilisée pour préciser ces points de cassure. Et, dans le cas de la t(11;22), on ne peut écarter l'hypothèse que la fréquence de cette translocation soit due à une proximité physique de ces 2 chromosomes à l'intérieur du noyau, associée à une similitude génétique entre le 11q et le 22q (Budarf, 1989, Sèle, 1977), similitude retrouvée également dans l'expression clinique de la trisomie de chacun de ces segments. Une autre hypothèse serait l'existence de séquences ressemblant à des séquences télomériques, retrouvées en position intermédiaire sur certains chromosomes, rendant ces chromosomes plus susceptibles de se casser ou de se recombiner (Gillois, 1991). Seule une étude complète et détaillée, au niveau moléculaire, des points de cassure des translocations pourra éclairer ce mécanisme de cassure chromosomique, mais il faut garder à l'esprit que la plus grande fréquence "apparente" de certains points de cassure peut être due aussi bien à la viabilité des déséquilibres engendrés qu'à un mécanisme biologique particulier restant à identifier.

Conclusion

Si on reprend l'objectif de départ représenté par l'aide au conseil génétique en cas de translocation réciproque, ce travail constitue une étape importante, puisqu'il permet d'envisager de fournir au généticien une estimation quantitative du risque de survenue d'un enfant malformé à terme. Cette information est probablement celle qui est la plus attendue par les généticiens (et par le couple porteur), et on perçoit bien l'avantage représenté par le fait de savoir que le risque pour cette translocation est de "5 % ± 2 %" au lieu de "entre 0 et 40 %".
Concernant cette approche statistique, deux points nous semblent importants à souligner :

- la seule manière d'améliorer les performances de la modélisation sera l'obtention de nouvelles variables explicatives pertinentes,
- une grande vigilance s'impose, lorsque l'on fournit ainsi des résultats quantitatifs à des cliniciens ou à des patients, il est nécessaire d'insister sur le fait qu'il s'agit d'une estimation et non pas d'une réalité.

Il faut garder cependant à l'esprit qu'il existe aussi d'autres types d'approches (économique, santé publique, biologie de la reproduction, biologie moléculaire) qui seront nécessaires à développer dans un proche avenir. Même sans envisager une banque nationale de prélévements (il en existe déjà plusieurs en France), l'information sur le fait qu'il existe des prélèvements pour une translocation particulière dans tel centre est une information précieuse pour les différents chercheurs. Cependant, ceux-ci devront impérativement effectuer un retour de l'information concernant les familles qu'ils auront étudiées, afin de permettre l'intégration de ces données et l'amélioration des critères de viabilité qualitative pour chaque région du génome. Concernant les données cliniques, compte tenu de l'existence de nombreuses bases de données (GENDIAG, OMIM, POSSUM), il apparaît plus opportun d'envisager une connexion à ces bases plutôt que de concevoir une nouvelle base de données cliniques sur les aberrations chromosomiques.

Pour terminer, ce travail se situe dans le contexte d'un projet global d'étude de toutes les anomalies chromosomiques de structure au laboratoire de Cytogénétique de Grenoble, avec en particulier l'étude des inversions péricentriques, pour lesquelles la problématique est très similaire, bien que le nombre de cas différents beaucoup plus faible. Ce projet présente l'avantage de pouvoir mettre en commun les données cartographiques concernant toutes ces anomalies.

Références bibliographiques

Alberman E (1973) Epidemiology of spontaneous abortions and their chromosome constitution. Dans : Chromosomal errors in relation to reproductive failure. *INSERM (ed), Paris, 305-316.*

Becker RA, Chambers JM, Wilks AR.(1988). The new S language. Wadsworth and Brooks / Cole Advanced Books and Software, California.

Betend C (1989) Analyse d'un millier de translocations réciproques autosomiques humaines. *Thèse Médecine. Grenoble.*

Bonneu M (1988) Model choice for prediction in generalized linear models. *Statistics 19: 369-382.*

Borgaonkar DS, Reisor N.(1990). Repository of human chromosomal variants and anomalies. An international registry of abnormal karyotypes. 13th listing. Medical Center of Delaware, Delaware.

Boué J, Boué A, Lazar P (1975) Retrospective and prospective epidemiological studies of 1500 karyotyped spontaneous human abortions. *Teratology 12: 11-26.*

Boué A, Gallano P (1984) A collaborative study of the segregation of inherited chromosome structural rearrangements in 1356 prenatal diagnoses. *Prenat Diagn 4: 45-67.*

Boué A (1989) Médecine Prénatale. Médecine-Sciences Flammarion.

Bunke O, Droge B (1984) Bootstrap and cross validation estimates of the prediction error for linear regression models. *Annals of Statistics 13: 1400-1424.*

Budarf M, Sellinger B, Griffin C, Emanuel BS (1989) Comparative mapping of the constitutional and tumor-associated 11;22 translocations. *Am J Hum Genet 45: 128-139.*

Cans C (1990) Analyse statistique des translocations réciproques. *Mémoire DEA de Statistique et Santé. UFR Médicale de Kremlin-Bicêtre.*

Cans C, Cohen O, Mermet MA, Demongeot J, Jalbert P (1993) Human reciprocal translocations : is the unbalanced mode at nirth predictable ? *Hum Genet 91: 228-232.*

Chambers JM, Hastie TJ.(1992). Statistical models in S. Wadsworth and Brooks / Cole Advanced Books and Software, California.

Cohen O, Simonet M, Cans C, Mermet MA, Demongeot J, Amblard F, Jalbert P (1992) Human reciprocal translocations : a new computer system for genetic counseling. *Ann Génét 35: 193-201.*

Cohen O, Cans C, Mermet MA, Demongeot J, Jalbert P (1994) Viability thresholds for partial trisomies and monosomies. A study of 1159 viable unbalanced translocations. *Hum Genet 93: 188-194.*

Couillin P, Le Guern E, Vignal A, Fizames C, Ravisé N, Delportes D, Reguigne I, Rosier MF, Junien C, Van Heyningen V, Weissenbach J (1994) Assigment of 112 microsatellite markers on 23 chromosome 11 subregions delineated by somatic hybrids. Comparison with the genetic map. *Communication personnelle.*

Crandall BF, Lebherz TB, Rubinstein L, Robertson RD, Sample WF, Sarti D, Howard J (1980) Chromosome findings in 2500 second trimester amniocenteses. *Am J Med Genet 5: 345-356.*

Cremillieux B (1991) Induction automatique : aspects théoriques, le système arbre, applications en médecine. *Thèse informatique. Université J Fourier Grenoble.*

Daniel A (1979) Structural differences in reciprocal tranlocations. Potential for a model of risk in reciprocal translocations. *Hum Genet 51: 171-182.*

Daniel A, Boué A, Gallano P (1986) Prospective risk in reciprocal translocation heterozygotes at amniocentesis as determined by potential chromosome imbalance sizes. Data of the European collaborative prenatal diagnosis centres. *Prenat Diagn 6: 315-350.*

De Arce MA, Grace PM, Mc Manus S (1986) A computer model for the study of segregation in reciprocal translocation carriers : application to 20 new cases. *Am J Hum Genet 24: 519-525.*

Diday E, Lemaire J, Pouget J, Testu F.(1982). Eléments d'analyse de données. Dunod, Paris.

Dutrillaux B (1977) New chromosomes techniques. In: Yunis JJ(ed) Molecular structure of human chromosomes. *Academic Press, New York, 233-265.*

Easton DF, Ponder MA, Huson SM, Ponder BAJ (1993) An analysis of variation in expression of Neurofibromatosis (NF) type I (NF1): evidence for modifying genes. *The Am J Hum Genet 53: 305-313.*

Flack M, Chang S (1987) Frequency of selecting noise variable in subset regression analysis : a similar study. *Am Statistician 41: 84-86.*

Fortuny A, Carrio A, Soler A, Cararach J, Fuster J, Salami C (1988) Detection of balanced rearrangements in 445 couples with repeated abortion and cytogenetic prenatal testing in carriers.*Fertil Steril 49: 774-779.*

Fraccaro M, Lindstein J, Ford CE, Iselius L (1980) The 11q;22q translocation: a European Collaborative Analysis of 43 cases. *Hum Genet 56: 21-51.*

Fryns JP, Kleczkowska A, Kubien E, Van den Berghe H (1986) Excess of mental retardation and/or congenital malformations in reciprocal tranlocations in man. *Hum Genet 72: 1-8.*

Gardner RM, Sutherland GR (1989) Chromosome Abnormalities and Genetic counseling. *Oxford University Press, New York.*

Gillois M (1969) La pseudo-sexualité des cellules somatiques en culture : modèle de l'induction des "pseudo-méioses". *Ann Genet 12: 5-14.*

Gillois M (1991) Gene mapping today: applications to farm animals. *Genet Sel Evol 23:19s-48s.*

Glim (1987) System release 3.77 Manual. *Ed CD Payne, Oxford.*

Goodman RM, Gorlin RJ (1983) The malformed infant and child. *Oxford University Press, New York.*

Goujard J, Maillard F, Ancelin C, du Mazaubrun C, André F (1983) Enregistrement des malformations congénitales à Paris. *J Gyn Obst Biol Repr 12: 805-817.*

Greenland S (1989) Modeling and variable selection in epidemiologic analysis. *Am J Public Health 79: 340-349.*

de Grouchy J, Turleau C (1982) Atlas des maladies chromosomiques. *Expansion Scientifique Française, Paris.*

Guichaoua MR, de Lanversin A, Cataldo C, Delafontaine D, Alasia C, Fraterno M, Terriou P, Stahl A, Luciani JM (1991) Three dimensional reconstruction of human pachytene spermatocyte nuclei of a 17;21 reciprocal translocation carrier : study of XY-autosome relationships. *Hum Genet 8: 709-715.*

Hall JG (1990) Genomic imprinting: review and relevance to human diseases. *Am J Hum Genet 46: 857-873.*

Hastie T, Tibshirani R (1986) Generalized additive models. *Statistical Science 1: 297-318.*

Hecht F, Hecht BK (1984) Fragile sites and chromosome breakpoints in constituional rearrangements II. Spontaneous abortions, stillbirths and newborns. *Clinical Genetics 26: 174-177.*

Holmquist GP (1992) Review article: Chromosome bands, their chromatin flavors, and their functional features. *Am. J Hum Genet 51: 17-37.*

Hook EB, Cross PK, Jackson L, Pergament E, Brambati B (1988) Maternal age-specific rates of 47,+21 and other cytogenetic abnormalities diagnosed in the first trimester of pregnancy in chorionic villus biopsy specimens: comparison with rates expected from observations at amniocentesis. *Am J Hum Genet 42: 797-807.*

Hosmer DW, Lemeshow S (1989) Applied logistic regression. *John Wiley and sons, New York.*

Imaizumi K, Kuroki Y (1991) Rubinstein-Taybi syndrome with de novo reciprocal translocation t(2;16)(p13,3;p13,3). *Am J Medical Genetics 38: 636-639.*

ISCN (1981) An international system for human cytogenetic nomenclature. High resolution banding. *Cytogenet Cell Genet 11: 88-94.*

Jacobs PA, Aitken J, Frackiewicz A, Law P, Newton MS, Smith PG (1970) The inheritance of translocations in man : data from families ascertained through a balanced heterozygote. *Ann Hum Genet, Lond 34: 119-131.*

Jacobs PA (1977) Epidemiology of chromosome abnormalities in man. *Am J Epidemiol 3: 180-191.*

Jalbert P, Sèle B, Jalbert H (1980) Reciprocal translocations : a way to predict the mode of imbalanced segregation by pachytene diagram drawing. A study of 151 human translocations. *Hum Genet 55: 209-222.*

Jalbert P, Cohen O, Cans C, Mermet MA, Demongeot J (1992) Unbalances of reciprocal translocations (rcp): maximum thresholds of viability. *Am J Hum Genet 51 : A1240.*

Jennings DE (1986) Outliers and residual distributions in logistic regression. *J Am Statist Assoc 81: 987-990.*

Kaplan JC, Delpech M (1989) Biologie moléculaire et médecine. *Médecine-Sciences Flammarion, Paris.*

Kohonen T (1982) Self organized formation of topologically correct feature maps. *Biological Cybernetics 43: 59-69.*

Landwehr J, Pregibon D, Shoemaker AC (1984) Graphical methods for assessing logistic regression models. *J Am Statist Assoc 79: 61-71.*

Latche ML, Le Cohu I (1988) Les translocations réciproques de chromosomes. Etude prédictive de la ségrégation gamétique. *Grenoble Rapport Ensimag.*

Lichter P, Bray P, Ried T, Dawid IB, Wards DC (1992) Clustering of C_2-H2 zinc finger motif sequences within telomeric and fragile site regions of humac chromosomes. *Genomics 13: 999-1007.*

Lippman A, Tomkins DJ, Shime J, Hamerton JL (1992) Canadian multicentre randomized clinical trial of chorion villus sampling and amniocentesis. *Prenat Diagn 12: 385-476.*

Luciani JM, Guichaoua MR, Delafontaine D, North MO, Gabriel-Robez O, Rumpler Y (1987) Pachytene analysis in a 17;21 reciprocal translocation carrier : role of the acrocentric chromosomes in male fertility. *Hum Genet 77: 246-250.*

MC Cullagh P, Nelder JA (1989) Generalized linear models. *Chapman Hall, London.*

Mujica P, Morali A, Vidailhet M, Pierson M, Gilgenkrantz S (1989) A case of Alagille's syndrome with translocation (4;14)(q21;q21). *Ann Genet 32:117-119.*

Nelder JA, Wedderburn RWM (1972) Generalized linear models. *J R Statist Soc A 135: 370-384.*

Pregibon D (1981) Logistic Regression Diagnostics. *Ann Statistics 9: 705-724.*

Puissant H, Azoulay M, Serre JL, Larget Piet L, Junien C (1988) Molecular analysis of a reciprocal translocation t(5;11)(q11;p13) in a WAGR patient. *Hum Genet 79: 280-282.*

Romeder JM (1973) Méthodes et programmes d'analyse discriminante. *Dunod, Paris.*

Robert C (1989) Analyse descriptive multivariée. *Flammarion, Paris.*

Sachs ES (1985) Chromosome studies of 500 couples with 2 or more abortions. *Obstet Gynecol 65: 375-378.*

Sèle B, Jalbert P, van Cutsem B, Lucas M, Mouriquand C, Bouchez R (1977) Distribution of human chromosomes an the metaphase plate using banding techniques. *Hum Genet 39: 39-61.*

Simoni G, Fraccaro M, Gimelli G, Maggi F, Bricarelli DF (1987) False-positive and false-negative findings on chorionic villus sampling. *Prenat Diagn 7: 671-672.*

Stengel-Rutkowski S, Stene J, Gallano P (1988) Risk estimates in balanced reciprocal translocations. *Expansion Scientifique Française (ed) Monographie des Annales de Génétique, Paris.*

Sumner AT (1982)The nature and mechanisms of chromosome banding. *Cancer Genet Cytogenet 6: 59-87.*

Thomas IT, Carter RL, Gaitanzis YA, Frias JL (1987) Reciprocal translocations and segregation type : a predictive equation. *Department of Pediatrics and Hattie B Murroe Center for Human Genetics. University of Nebraska Medical Center (personal communication).*

Warburton D (1991) De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am J Hum Genet 49: 995-1013.*

Wilkinson L (1987) Systat: The system for statistics. *Evanston, IL: Systat, Inc.*

Wilson GN, Stout JP, Schneider R, Zneimer SM, Gilstrap LC (1991) Balanced translocation 12/13 and situs abnormalities: homology of early pattern formation in man and lower organisms ? *Am J Medical Genetics 38:601-607.*

Annexes

Annexe 1. Rappels de cytogénétique et quelques définitions

Chaque cellule du corps humain possède 23 paires de chromosomes, qui constituent le génome.

Il existe classiquement 7 groupes de chromosomes.

Groupe	Chromosomes
A	1, 2, 3
B	4, 5
C	6, 7, 8, 9, 10, 11, 12, X
D	13, 14, 15
E	16, 17, 18
F	19, 20
G	21, 22, Y

Chacun de ces chromosomes est caractérisé par sa taille, la position de son centromère et son contenu. Ce dernier peut être étudié à différents niveaux : niveau microscopique (différents types de bandes), niveau génique (séquence comportant de 20 à 30 000 nucléotides), et niveau moléculaire (nucléotides). Seuls sont abordés ici les aspects concernant le premier niveau. Les chromosomes d'une même paire sont appelés chromosomes homologues. Le nombre de chromosomes est réduit de moitié lors de la formation des gamètes. En fonction du nombre de chromosomes les cellules somatiques sont dites diploïdes tandis que les cellules sexuelles sont dites haploïdes.

Constitution des chromosomes (au niveau microscopique)

Les techniques de coloration ont remplacé maintenant la technique d'autoradiographie dans l'identification précise de chacun des chromosomes. Elles utilisent soit le colorant Giemsa (méthode G) soit la Quinacrine (méthode Q) et permettent de distinguer différentes bandes dont la répartition est spécifique à chaque chromosome.

- les bandes C correspondent à des complexes de séquences d'ADN hautement répétitifs (ADN satellite). Tous les centromères, les constrictions secondaires de certains chromosomes (1, 9 et 16, et bras long du Y) et les bras courts des chromsomes acrocentriques (13,14,15,21,22) sont constitués par ce matériel qui est un matériel génétiquement inerte ou hétérochromatine. Il n'y a pas de transcription de l'ADN, ni de crossing over dans ces régions conduisant donc à un moins bon appariement).
- les bandes G correspondent à de la chromatine rendue compacte par des protéines non histoniques, leur réplication est tardive, elles sont pauvres en gènes. Elles apparaissent colorées en foncé avec la méthode G.

- les bandes R correspondent aux régions du génome les plus riches en gènes, elles ont une réplication précoce et sont riches en guanine et cytosine. Elles apparaissent colorées en clair avec la méthode G, et colorées en foncé avec la dénaturation inverse de la méthode G.

La méiose

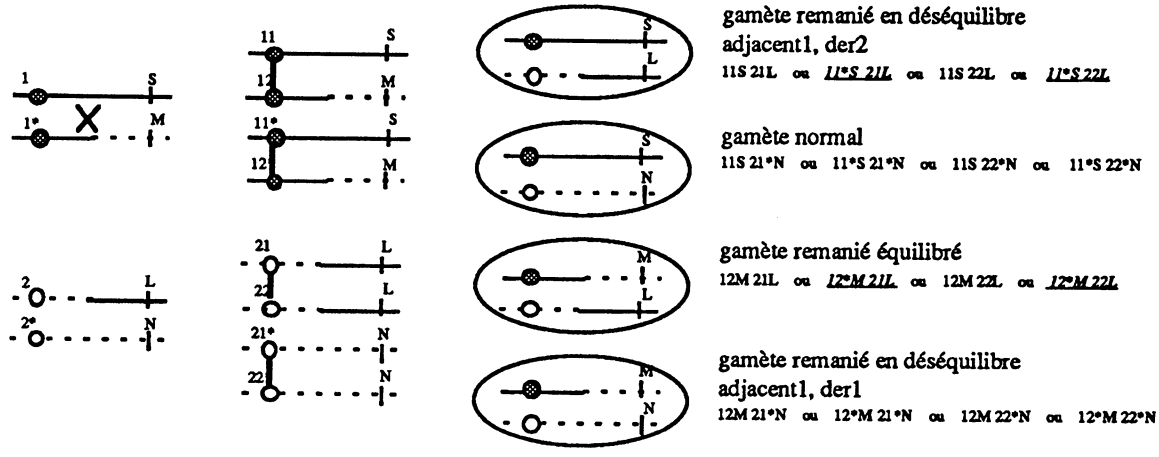
C'est une division nucléaire conduisant à la formation de 4 cellules haploïdes pour une cellule diploïde qui subit la méiose. Ses objectifs sont d'une part la réduction du nombre de chromosomes de 46 à 23 pour la reproduction, et d'autre part la transmission des caractères héréditaires avec brassage de l'information génétique. Au cours de la prophase de la méiose il y a appariement des chromosomes, ce phénomène peut être observé sur les préparations au stade pachytène. A ce stade, l'appariement des chromosomes, lorsqu'ils sont remaniés, constitue le diagramme du pachytène. Au cours de l'anaphase qui suit (et qui réalise le passage de l'état diploïde à l'état haploïde) ces chromosomes peuvent subir diverses ségrégations. Des phénomènes de non disjonction peuvent survenir suite à un défaut d'appariement de 2 chromosomes homologues (chromosomes de la même paire) et conduire à la naissance d'un enfant malformé à terme.

Le caryotype

C'est l'étude de la consitution chromosomique d'une cellule. La cellule étudiée doit être en division cellulaire afin que les chromosomes puissent être individualisés. En pratique c'est au cours d'une mitose spontanée ou provoquée (après culture cellulaire in vitro) que le caryotype est effectué. Un blocage des cellules en mitose permet alors d'observer les chromosomes au microscope électronique après utilisation ou non des techniques de coloration (citées ci-dessus).
Le plus souvent c'est un prélèvement sanguin qui est réalisé pour l'étude du caryotype avec culture de lymphocytes sanguins in vitro. D'autres milieux peuvent être prélevés : les prélèvements de sperme, de liquide amniotique et de villosités choriales sont souvent rencontrés dans l'étude des translocations réciproques.

Annexe 2. Méiose avec crossing over

Deux paires de chromosomes homologues remaniés
(chromosomes n°1 et n°2, t(1;2))
avec crossing over sur les bras longs du chromsome n°1.



La combinaison (1*-2) entraîne, en l'absence de crossing over, la formation de gamètes remaniés en équilibre (schéma 3 p.5). En cas de crossing over tel qu'il est montré ci-dessus, la même combinaison (1*-2) donne lieu soit à des gamètes remaniés en équilibre (12*M21L, 12*M21L), soit à des gamètes remaniés en déséquilibre (11*S21L, 11*S22L).

Annexe 3. Liste des laboratoires "sources de données"

Liste des laboratoires français :

Besançon (Bresson)
Chambéry (Noël)
Clermont-Ferrand (Malet)
Dijon (Turc)
Grenoble (Jalbert)
Lille (Deminatti)
Marseille (2) (Mattei, Stahl)
Montpellier (Emberger)
Nice (Ayraud)
Paris (3) (Taillemite, Girard, Desangles)
Strasbourg (Ruch)
Toulouse (Bourrouillon)
+ Aarhus au Danemark

Liste du Fichier Européen :

Allemagne	Aix-La-Chapelle
	Bonn
	Erlangen
	Göttingen
	Hanovre
	Lübeck
Belgique	Bruxelles
Danemark	Aarhus
	Copenhague
Finlande	Helsinki
France	Angers
	Besançon
	Bordeaux
	Caen
	Dijon
	Grenoble
	Lille
	Lyon
	Marseille
	Montpellier
	Paris
	Rouen
	Strasbourg
	Toulouse
	Tours
Grande Bretagne	
Angleterre	Amston-U-Lyne Lancs
	Birmingham
	Bristol
	Cardiff
	Leeds
	Londres
	Mitcham Junction
	Newcastle
	Nottingham
	Radlett Herts
	Salisbury
	Sheffield

Ecosse	Aberdeen
	Edimbourg
	Glasgow
Ulster	Belfast
Israël	Tel-Hashomer
Italie	Bologne
	Gènes
	Milan
	Padoue
	Rome
Pays-bas	Amsterdam
	Groningue
	Rotterdam
Suède	Göteborg
	Lund
	Stockholm
	Uppsala
Suisse	Bâle
	Genève
	Lausanne
	Zürich

Annexe 4. Fiche de saisie

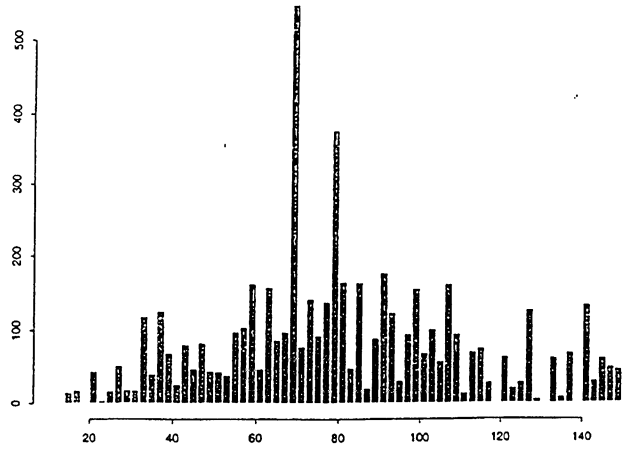
OBSERVATION D'UNE TRANSLOCATION RECIPROQUE

identificateur de famille	1047
identificateur d'individu	II1
équation de la translocation et origine	t(1 ; 6)(q32 ; q26) mat
sexe	masculin
mode de détection	déséquilibre
état	porteur déséquilibré
mode de ségrégation	adjacent 1

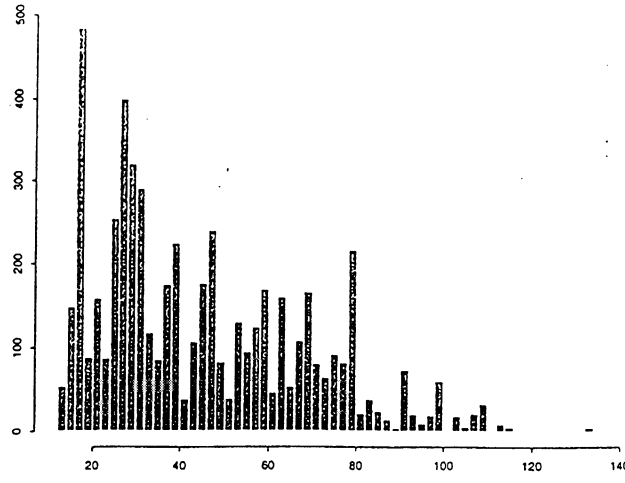
nombre de chromosomes	46	chromosome transloqué transmis	1
trisomie/monosomie	0	nombre de fausses couches	0
nombre de morts nés	0	nombre total de grossesse	0
age maternel	0	age paternel	0

identificateur du parent porteur	I1
Nbre de Déséquilibres caryotypés dûs à la RCP et issus d'un parent porteur	0
Histoire naturelle	
Vérification du caryotype	vnk

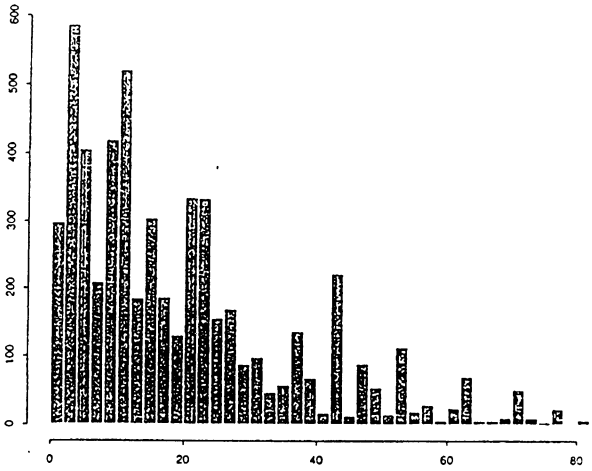
Annexe 5. Distribution des variables explicatives



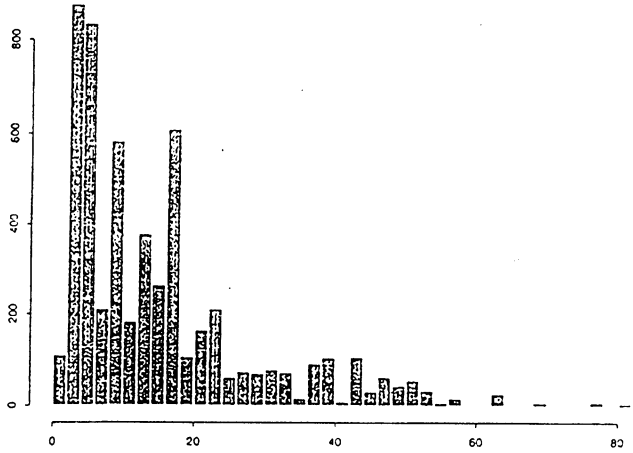
CS1



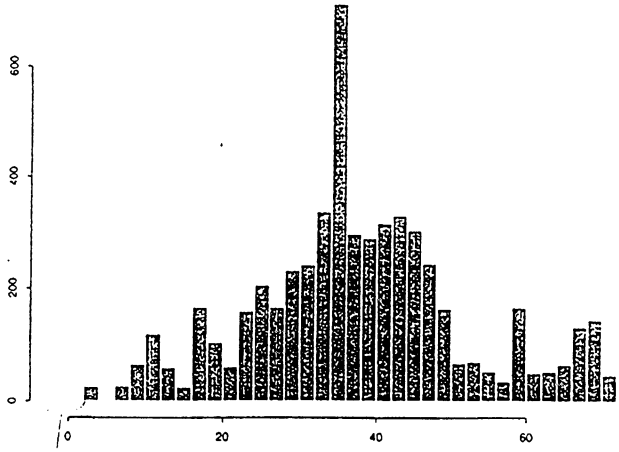
CS2



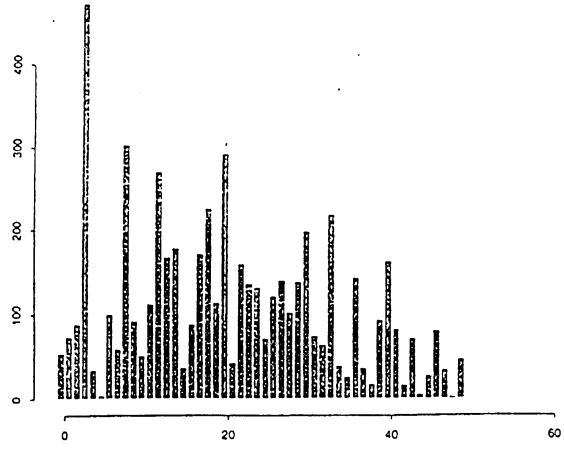
TS1



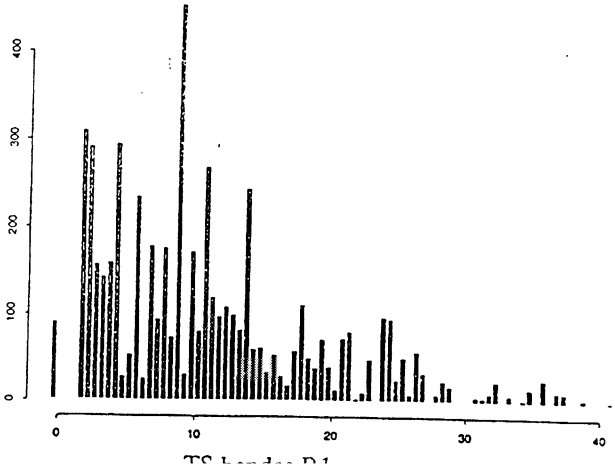
TS2



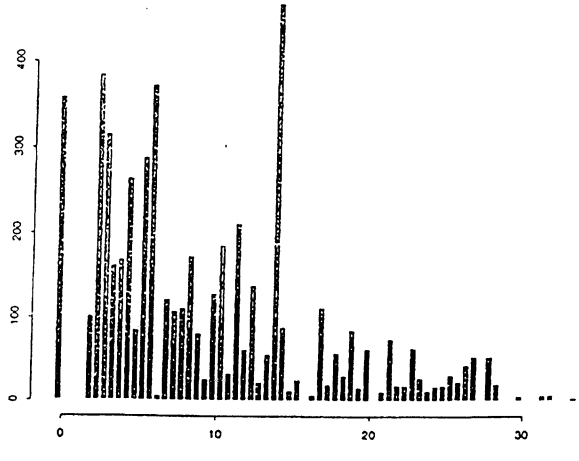
CS bandes R1



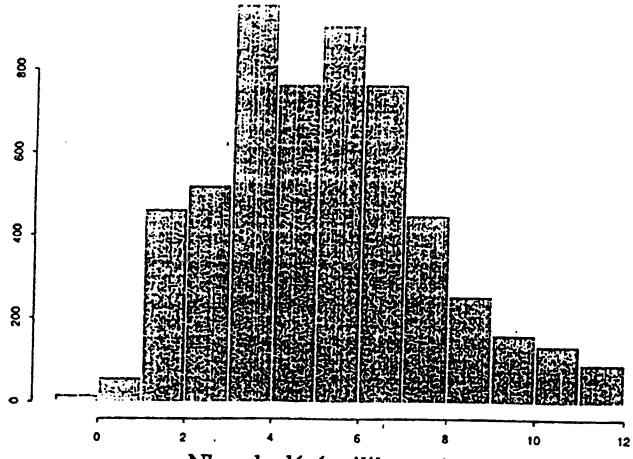
CS bandes R2



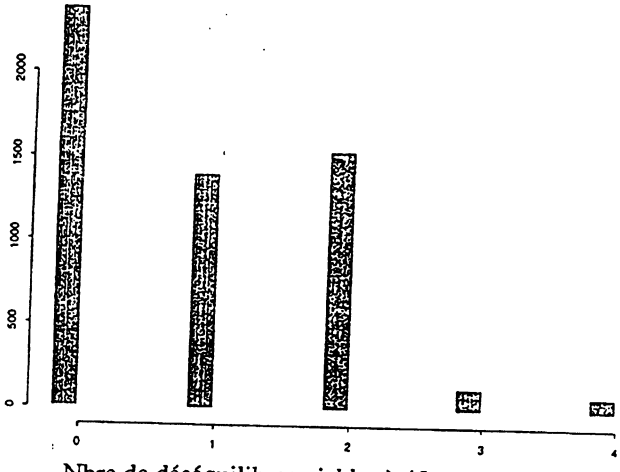
TS bandes R1



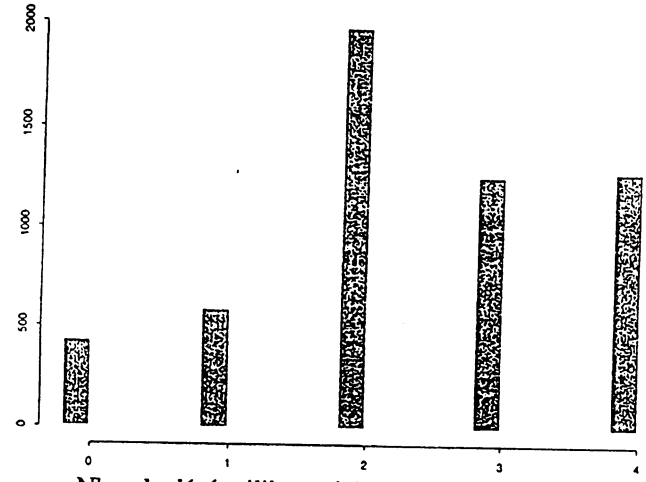
Annexe 5 bis. Distribution des variables concernant la viabilité



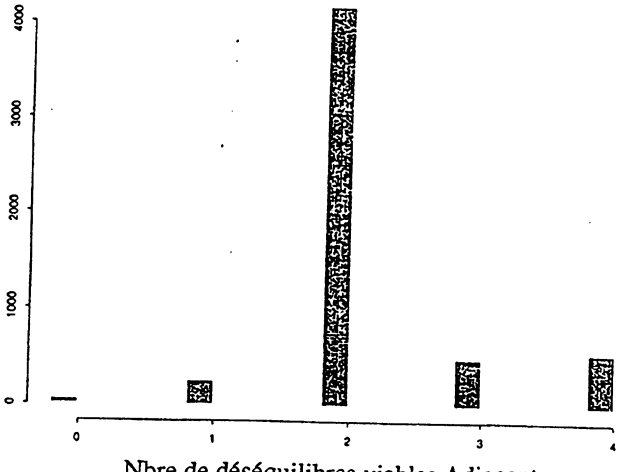
Nbre de déséquilibres viables



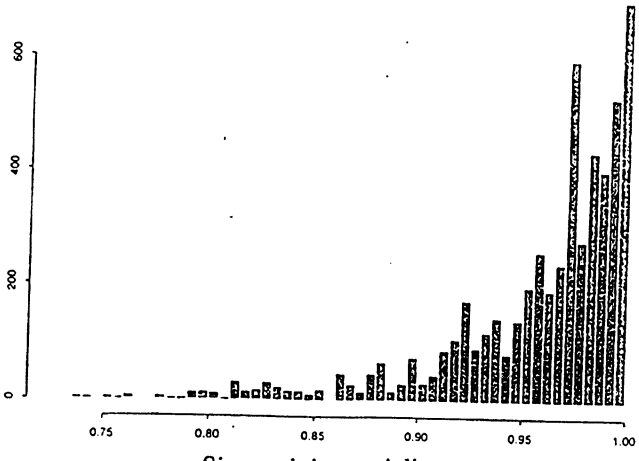
Nbre de déséquilibres viables à 45 chromosomes



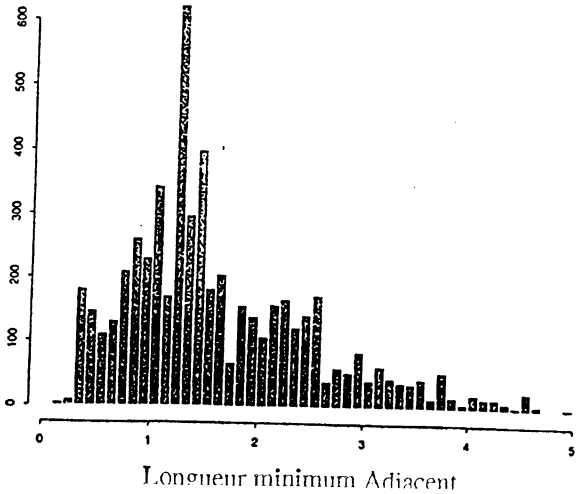
Nbre de déséquilibres viables à 47 chromosomes



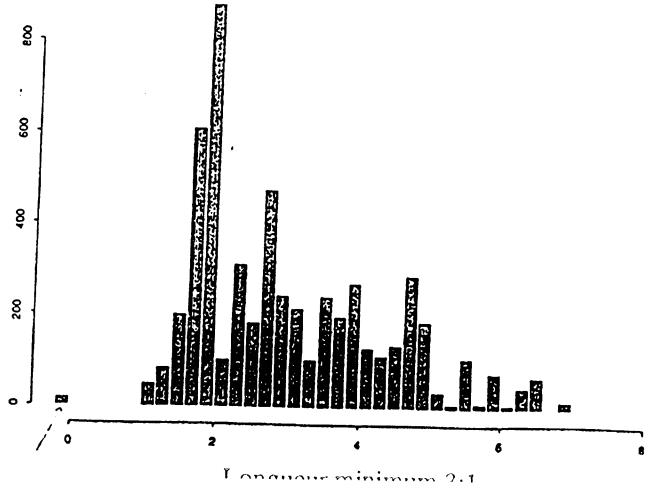
Nbre de déséquilibres viables Adjacent



Sinus minimum Adjacent



Longueur minimum Adjacent



Longueur minimum 2:1

Annexe 6. Numéros des chromosomes impliqués dans la translocation

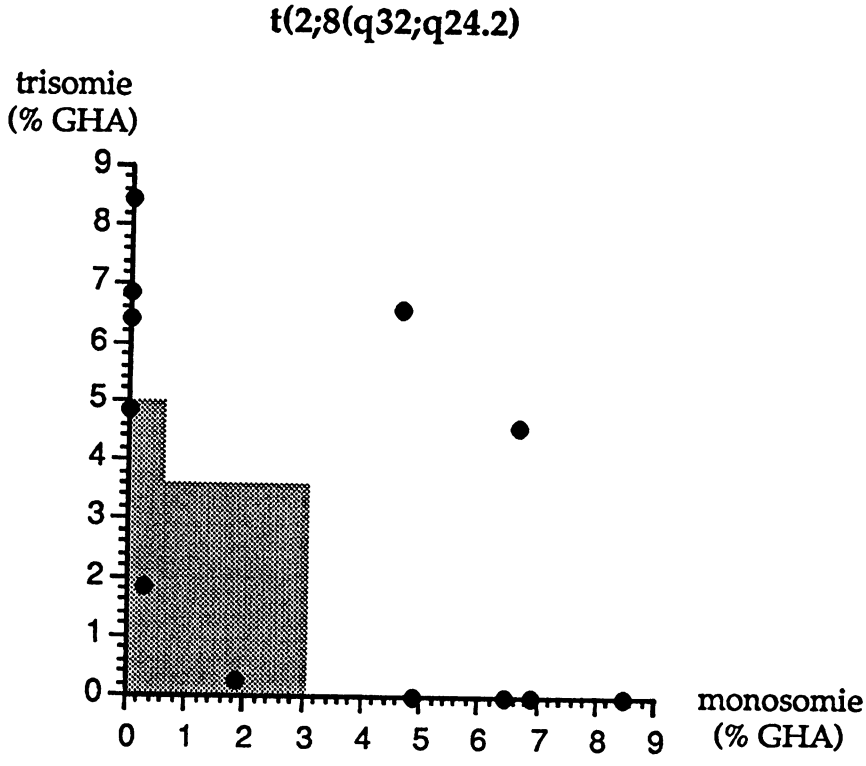
	Xme 2	Xme 3	Xme 4	Xme 5	Xme 6	Xme 7	Xme 8	Xme 9	Xme 10	Xme 11	Xme 12	Xme 13	Xme 14	Xme 15	Xme 16	Xme 17	Xme 18	Xme 19	Xme 20	Xme 21	Xme 22
Xme 1	30	39	146	52	26	50	48	**39**	48	19	11	*16*	*23*	*11*	54	49	49	1	62	17	52
Xme 2		17	18	67	52	55	71	**26**	53	4	56	*18*	*39*	*27*	22	29	41	3	16	10	9
Xme 3			41	16	42	38	51	**22**	10	16	22	*23*	*23*	*52*	26	14	24	8	7	18	23
Xme 4				36	30	35	98	**72**	56	36	46	*64*	*12*	*57*	51	15	108	0	29	46	50
Xme 5					20	47	25	**50**	88	74	22	*57*	*44*	*101*	26	29	60	13	3	41	19
Xme 6						15	25	**36**	18	37	12	*33*	*8*	*21*	3	13	24	5	54	16	27
Xme 7							51	**71**	29	21	49	*39*	*28*	*18*	5	3	29	5	17	41	28
Xme 8								**85**	24	21	32	*65*	*35*	*35*	10	9	24	0	4	0	32
Xme 9									**40**	**19**	**31**	**51**	**42**	**89**	**18**	**5**	**55**	**28**	**50**	**30**	**101**
Xme 10										21	36	*23*	*55*	*23*	8	51	86	0	15	49	42
Xme 11											13	*36*	*10*	*4*	38	12	14	0	6	36	***675***
Xme 12												*3*	*30*	*4*	12	6	23	0	8	59	6
Xme 13													*23*	*70*	*6*	*25*	*80*	*0*	*11*	*23*	*33*
Xme 14														*9*	*2*	*4*	*36*	*0*	*7*	*4*	*9*
Xme 15															*6*	*7*	*3*	*6*	*4*	*26*	*12*
Xme 16																14	21	7	7	92	8
Xme 17																	11	7	0	14	28
Xme 18																		2	24	23	8
Xme 19																			20	2	10
Xme 20																				21	26
Xme 21																					2

Variable NVC :

modalité n°	Chromosomes	Effectif n=	modalité n°	Chromosomes	Effectif n=
1	11 et 22	675	5	4	745
2	9	960	6	1 à 8	835
3	13, 14, 15	1337	7	16 à 20	1133
4	21,22	965	8	10 à 12	740

Annexe 7. Aire de viabilité

Exemple d'une translocation t(2;8). L'aire de viabilité est la partie grisée du graphique. Parmi les 12 déséquilibres potentiels de cette tranlocation, seuls 3 d'entre eux se projettent à l'intérieur de la surface de viabilité.



(GHA: Génome Haploïde Autosomal)

Note : L'aire de viabilité décrite par Daniel (1979) serait représentée par la surface du triangle délimité par la ligne joignant les 4 % de trisomie et les 2 % de monosomie.

Annexe 8. Critères de la méthode du Diagramme du pachytène (DP)

Les critères d'affectation pour chacun des 4 principaux modes sont les suivants.
(CS= segment centrique, TS= segment transloqué)

◊ Adjacent 2
- Σ CS < Σ TS,
- les segments centriques des 2 chromosomes impliqués comportent des zones hétérochromatiques (alors que les segments transloqués n'en comportent pas) : chromosome 9 + acrocentriques.

◊ Adjacent 1
- Σ CS > Σ TS,
- Le plus court des CS ≥ le plus long des TS.

◊ 3:1
- CS et TS sont de longueur inégale : les CS entre eux, les TS entre eux (dans un rapport > 2),
-un des chromosomes est un acrocentrique ou le chromosome 9 (avec partie hétérochromatique sur le segment centrique).

* 3:1 tertiaire
- TS court attaché au CS court,
- moyenne Σ CS et moyenne Σ TS : elles différent significativement et souvent moyenne Σ CS > moyenne Σ TS.

* 3:1 échange
- TS long attaché au CS court.

Note : La longueur des segments hétérochromatiques n'a pas été inclue pour le calcul de la longueur des segments centriques et transloqués (pour le chromosome 9 et les acrocentriques).

Annexe 9. Analyse discriminante

Le logiciel Systat que nous avons utilisé traite l'analyse discriminante comme un cas particulier de l'analyse de variance multivariée à un seul facteur. La variable dépendante à expliquer est considérée comme un facteur à k modalités, et une régression multiple sur les différentes modalités du facteur est effectuée pour chaque variable explicative. On teste l'hypothèse nulle représentée par le fait que les différents groupes de la variable dépendante sont équivalents. Le pouvoir discriminant est donné sous la forme d'un coefficient de corrélation canonique.
Comme critère d'affectation c'est la distance de Mahalanobis qui est utilisée, et en raison des effectifs inégaux des différentes classes, les probabilités apriori et les règles Bayésiennes sont introduites (option "prior" du logiciel).

Critère d'affectation sous SYSTAT avec option "prior"

Le logiciel calcule tout d'abord des distances (appelées distances de Mahalanobis) pour chaque individu à chacun des centres de gravité des classes (autant de distances que de modalités de la variable dépendante). Ces distances ont une expression de la forme :

$$d = (x - \mu_i)' \; V_{G_i}^{-1} (x - \mu_i)$$

V_{G_i} étant la matrice de variance covariance du groupe i (intragroupe),

μ_i étant les coordonnées du centre de gravité du groupe i,

et x étant les coordonnées de l'individu à classer.

Ensuite, il applique la règle Bayésienne suivante:

$$\text{Prob}\,(G_i / x) = \text{Prob}\,(G_i) * \text{Prob}\,(x / G_i) / \text{Prob}\,(x).$$

- Prob (G_i / x) est la probabilité aposteriori d'appartenir au groupe i (ce que l'on cherche).
- Prob (G_i) est la probabilité apriori d'appartenir au groupe i (fréquence observée sur l'échantillon de base).
- Prob (x / G_i) est la densité de probabilités dans le groupe i (ou distribution conditionnelle de x).
- Prob (x) est la densité de probabilités sur la population (ou distribution inconditionnelle de x).

La distribution conditionnelle de x se met sous la forme suivante :

$$f_i(x) = \frac{1}{2}\pi^{p/2} * \frac{1}{|V_{G_i}|^{1/2}} * \exp(-\frac{1}{2}(x - \mu_i)V_{G_i}^{-1}(x - \mu_i)) = \frac{1}{2}\pi^{p/2} * \frac{1}{|V_{G_i}|^{1/2}} * \exp(-\frac{1}{2}d)$$

Dans ce logiciel des hypothèses très fortes sur la normalité des distributions et sur l'égalité des matrices de variance-covariance sont faites lors du calcul du critère d'affectation. Mais le point intéressant est qu'il permet de tenir compte des probabilités apriori. On classe alors l'individu dans la classe ayant la plus forte probabilité aposteriori (et donc aussi la plus petite distance d).

Taux d'erreur réel

Une fois les individus de l'échantillon de base classés (taux d'erreur apparent d'individus bien classés), qu'en est-il des individus de l'échantillon-test à classer ? Pour ceux-ci, le logiciel propose une fonction, combinaison linéaire des variables $x_1,....x_p$, pour chaque groupe appelée "group classification function". Il y a autant de fonctions que de modalités de la variable dépendante. Un nouvel individu est affecté au groupe pour lequel la valeur de la fonction est la plus élevée.

$$d = (x-\mu_i)' V_{G_i}^{-1}(x-\mu_i) = x' V_{G_i}^{-1} x - 2x' V_{G_i}^{-1}\mu_i + \mu_i' V_{G_i}^{-1}\mu_i$$

Au lieu de minimiser une distance on maximise le terme négatif 'T' de cette distance d :

$$T = 2x' V_{G_i}^{-1}\mu_i - \mu_i' V_{G_i}^{-1}\mu_i$$

où $(\mu_i' V_{G_i}^{-1}\mu_i)$ est une constante, tout en tenant compte aussi de la probabilité apriori par l'intermédiaire de la valeur de la constante. En effet pour l'échantillon-test on suppose que les effectifs de chaque groupe sont sensiblement proportionnels aux probabilités apriori de l'échantillon de base (Diday 1982).

Annexe 10. Distribution simulée des résidus

Les étapes décrites par Landwehr sont les suivantes :	nom de la variable dans le fichier de commande Splus
1. A partir d'un modèle M obtenir les résidus observés simples. $\eta = \hat{\beta} X$ n valeurs prédites $\hat{p}$ $r = Y - \hat{p}$	RES
2. Ordonner ces résidus observés.	SRES
3. Créer des données simulées : à partir d'une matrice de 45 colonnes et de n lignes on fait en sorte que chaque colonne suive une loi binomiale $(1, \hat{p})$.	
4. On applique le même modèle M ($\eta = \hat{\beta} X$) à ces données simulées (sur chaque colonne de n observations). On obtient ainsi une matrice de 45 vecteurs de résidus simulés.	RESI
5. Ordonner ces résidus simulés.	SRESI
6. Prendre la médiande de ces 45 vecteurs de résidus simulés.	MAT4
7. Faire un plot de la médiane des résidus simulés contre les résidus observés. Il doit être approximativement en ligne droite avec une pente unique.	

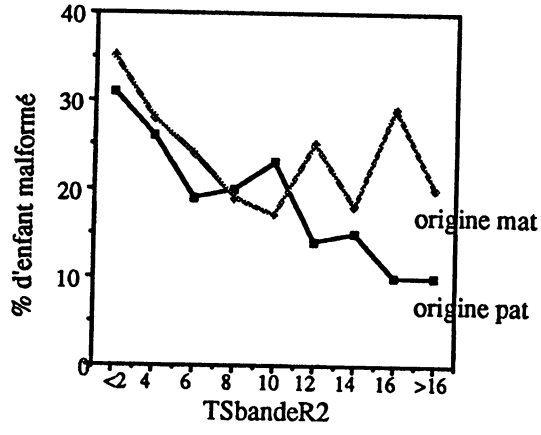
PROGRAMME Splus *pour obtenir la distribution des résidus simulés*

```
(u1 <- runif(5564,1,5564))
(u2 <- sort(u1))
glmres<- glm(V11~V3bf+V4bf+V21f+V15+V18+V4bf*V15+V4bf*V18
+V4bf*V3bf+V4bf*V21f+V21f*V18, binomial, f123dim, subset=u1<500)
pred <- fitted(glmres)
matr <- matrix((runif(23670)), 526, 45)
compa <- function(matr) ifelse(matr<pred, 1, 0)
mat2 <- apply(matr, 2, compa)

resid <- matrix(1, 526, 45)
for(i in 1:45){
resid[,i] <- fitted(glm(matess[,c(i)]~V3bf+V4bf+V21f
+V15+V18+V4bf*V15+V4bf*V18+V4bf*V3bf+V4bf*V21f
+V21f*V18, binomial, f123dim, subset=u2<500))
}
print(resid)

sresi <- matrix(1, 526, 45)
for(i in 1:45){
sresi[,i] <- sort(resid[,c(i)]) }
med <- function(sresi) median(sresi)
mat4 <- apply(sresi, 1, med)
res <- V11[u2<500] - pred
sres <- sort(res)
motif()
plot(mat4, sres)
```

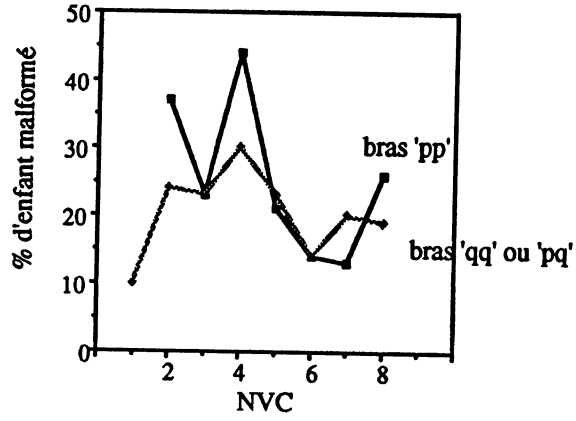
Note : il peut arriver qu'un des vecteurs simulés donne lieu à des problèmes de convergence lors de la régression logistique, il est alors exclu des données

Annexe 11. Représentation graphique des interactions

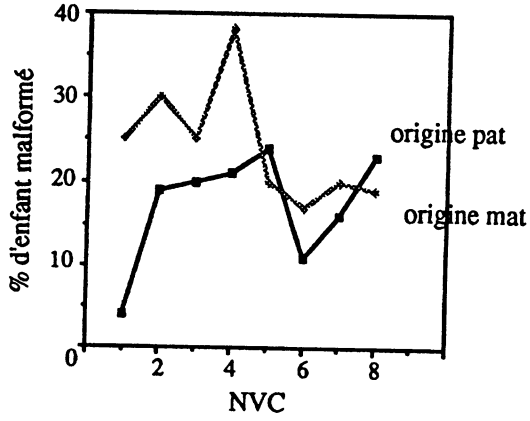
Interaction ORIG.TSbandeR2



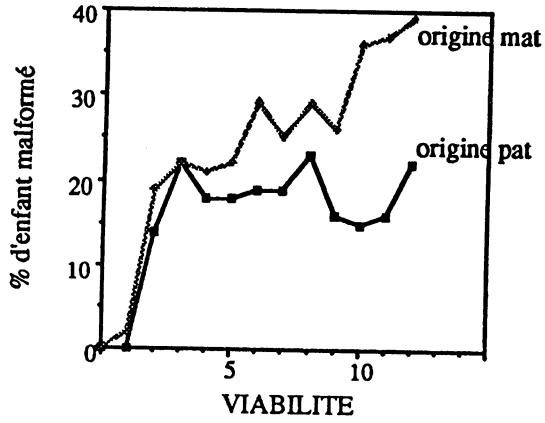
Interaction NVC.BRAS



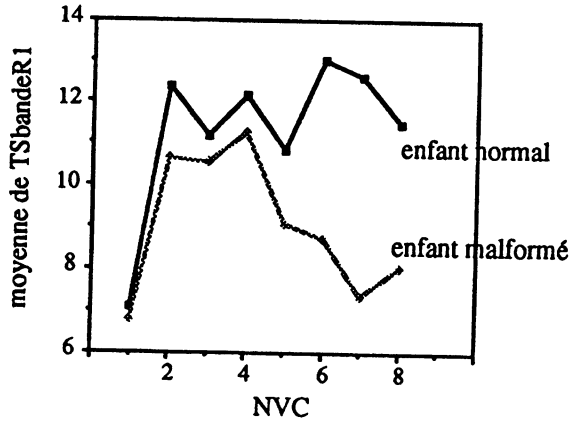
Interaction NVC.ORIG



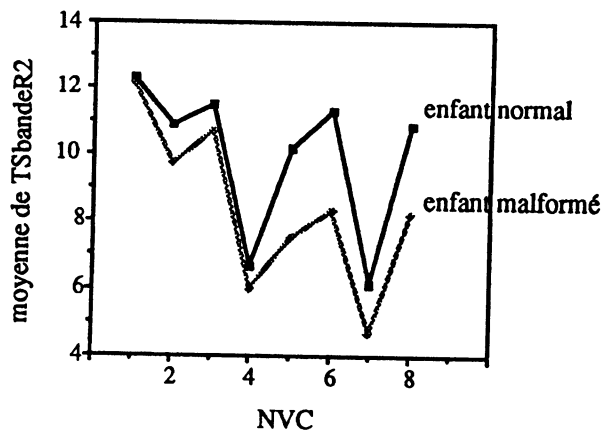
Interaction ORIG.VIAB



Interaction NVC.TSbandeR1



Interaction NVC.TSbandeR2



Annexe 12. Modèle GAM en "boîte noire"

Le prédicteur linéaire du modèle est le suivant :

BRAS + NVC + ORIG + s(VIA47) + s(LAMIN) + s(SINUSADJ) + s(CSBandeR1) + s(TSBandeR1) + s(TSBandeR2) + MOSEG + NVC*s(TSBandeR1) + NVC*ORIG + ORIG*s(TSBandeR2)

Déviance = 5524 pour 5521 ddl

Les valeurs prédites vont de 0.000 à 0.865. Le taux d'erreur apparent ne montre pas d'amélioration par rapport aux résultats du modèle 3.

risque prédit	Modèle 3 (GLM)		Modèle 4 (GAM)
Proportion d'observations 'non malades' bien classées			
<0,01	133/133	100 %	104/104 soit 100 %
≥0,01 et <0,05	348/357	97,5 %	392/401 soit 97,8 %
≥0,05 et <0,10	486/519	93,6 %	498/526 soit 94,7 %
< 0,10	967/1009	95,8 %	994/1031 soit 96,4 %
Proportion d'observations 'malades' bien classées			
≥0,30 et <0,50	355/1002	35,2 %	354/1014 soit 34,9 %
≥0,50	20/31	64,5 %	28/49 soit 57,1 %

De la même façon les premières observations mal classées (taux d'erreur apparent après avoir ordonné les valeurs prédites) sont les famille 566 en 56^{e} position et famille 499 en 150^{e} position ce qui est en dessous des performances du modèle 3.