

UNIVERSITE PAUL SABATIER - TOULOUSE III

U.F.R. SCIENCES DE LA VIE ET DE LA TERRE
ECOLE DOCTORALE BIOLOGIE SANTÉ BIOTECHNOLOGIE

THESE

Pour obtenir le grade de
Docteur de l'Université Toulouse III

Discipline: ECONOMIE DE LA SANTÉ

Mention: Immunogénétique, Epidémiologie, Santé publique

Présentée et soutenue le lundi 27 mars 2006

par

Frédérique Fève

TITRE:

Economie et Statistique du Don de Cellules Souches
Hématopoïétiques : Contributions à la Gestion
de Registres de Donneurs Volontaires

Sous la direction de

Jean-Pierre FLORENS Professeur de Mathématiques Université Toulouse I
Membre de l'Institut Universitaire de France
Anne CAMBON-THOMSEN Directrice de Recherche CNRS, Inserm U558 Toulouse

MEMBRES DU JURY

Jean-Pierre FLORENS Professeur de Mathématiques Université Toulouse I
Anne CAMBON-THOMSEN Directrice de Recherche CNRS, Inserm U558 Toulouse
Alvin ROTH George Gund Professor of Economics, Harvard University
Gérard DE POUVOURVILLE Directeur de recherche CNRS, INSERM, Le Kremlin Bicêtre
Jean-François ELIAOU Professeur d'Immunologie, Université de Montpellier
Florence TABOULET Professeur de Sciences Pharmaceutiques, Université Toulouse III

Lise ROCHAIX Professeur d'Economie Université Aix-Marseille, membre invité
Colette RAFFOUX Praticien Hospitalier, Paris, membre invité

— PRÉFACE —

Une pensée toute particulière va à Jean-Jacques Laffont, dont la disparition est toujours ressentie avec autant de tristesse. Si sa personnalité d'exception a profondément marqué l'esprit de la Recherche en Economie, elle m'aura également ouvert les portes du GREMAQ comme assistante de recherche puis plus tard celles de l'Institut d'Économie Industrielle. J'apprécie à quel point il est important d'être intégrée dans un laboratoire de recherche. Mais tout le mérite de ce travail de thèse revient à Jean-Pierre Florens, qui dirige à l'Institut les études d'économétrie appliquée auxquelles je participe depuis quatre ans. Sa curiosité intellectuelle et sa rigueur scientifique se sont avérées communicatives: et grâce à ses encouragements, après avoir travaillé comme actuaire, enseigné les sciences économiques et sociales au lycée, l'instruction civique et l'histoire de France au collège, je me suis inscrite en thèse à Toulouse à l'Université Paul Sabatier, à l'âge de trente-sept ans. Si mon activité professionnelle à l'Institut se prêtait davantage à des travaux de recherche en économie postale, le consortium européen MADDO¹, que dirigeait Anne Cambon-Thomsen et auquel j'ai participé m'a incité à faire de la recherche... en santé. La co-direction de cette thèse m'a paru naturelle.

Mes remerciements les plus sincères vont à Anne Cambon-Thomsen et à Jean-Pierre Florens pour leur aide précieuse et riche d'enseignements. Travailler sous leur direction fut pour moi une école de rigueur et de curiosité scientifique, de disponibilité et d'enthousiasme. Ma profonde admiration va à Jean-Pierre Florens pour m'avoir montré toute l'importance des contributions statistiques et économétriques dans le fonctionnement des mécanismes économiques qui nous régissent. Le mérite de ce que j'ai compris des aspects de génétique abordés dans cette thèse revient quant à lui plus naturellement à Anne Cambon-Thomsen: le concept d'hétérogénéité (immunogénétique dans notre cas particulier) si cher à un statisticien comme Jean-Pierre Florens rencontre ici le souci de "pluridisciplinarité" défendu par Anne Cambon-Thomsen.

¹pour MArrow DONors. Objectifs du projet: optimiser l'organisation des Registres de donneurs volontaires de Cellules Souches Hématopoïétiques

Je tiens à remercier H el ene Grandjean pour m'avoir accueillie   Toulouse au sein de l' quipe U558 de l'Inserm, dans le d partement  pid miologie de la facult  de m decine.

Un grand merci  galement aux professeurs de m decine qui m'ont accueillie au sein de leur laboratoire d'immunologie. Par ordre alphab tique (ce qui est la r gle chez les  conomistes), je citerai l' quipe du professeur Michel Abbal   l'h pital de Rangueil   Toulouse, celle du professeur Jean-Fran ois Eliaou (avec une mention particuli re pour Odile Avinens) du CHU Saint Eloi de Montpellier, le professeur Gottfried Fisher   Vienne et le laboratoire d'immunologie de Budapest dirig  par Katalin Rajczy et Gyozo Petranzy. Le travail d' quipe r alis  en particulier avec Jean-Fran ois Eliaou me fait entrevoir des extensions int ressantes   ce travail de th se et de nouvelles perspectives de recherche. A ce titre je n'oublierai pas de remercier  galement les participants du consortium MADDO, particuli rement le Docteur Colette Raffoux, qui a mis   notre disposition les donn es du Registre France Greffe de Moelle.

Je suis tr s reconnaissante aux professeurs Jean-Fran ois Eliaou, G rard de Pourville, Lise Rochaix, Alvin Roth et Florence Taboulet d' tre membres du Jury.

Cette th se doit  galement beaucoup   mes professeurs, co-auteurs, coll gues et amis, parmi lesquels figurent (toujours par ordre alphab tique) Etienne Billette de Villemeur, Marie-Pierre Boe, Erwan Bossis, Cathy Cazals, Fabrice Collard, Philippe De Donder, Martial Dupaigne, Marie-H el ene Dufour, Jacques Cr mer, Helmuth Cremer, Pierre Dubois, Bruno F ve, Patrick F ve, Alain Guay, Pascal Lavergne, Thierry Magnac, Fabien Moizeau, Michel Moreaux, Costas Meghir, Michel Mouchart, Franck Portier, Sophie Richard, Jean-Charles Rochet, Leopold Simar, Anne Vanhems, Alban Thomas, Yan Yohannes.

Je tiens aussi   remercier ma soeur St phanie, pour ses courriers  lectroniques si r guliers. Et enfin mais surtout... ce serait faire preuve d'ingratitude que d'oublier de remercier, pour leurs encouragements et leur soutien sans faille, ceux qui se r jouissent certainement le plus de l'ach vement de ce travail de th se : Patrick et nos trois gar ons, Lucas, S bastien et Jean-Jacques.

Table des matières

Introduction	13
1 Calcul économique et théorie de la décision	20
2 Evaluation économique de l'organisation d'un Registre	25
3 Contributions Statistiques à l'organisation d'un Registre de donneurs	28
1 Matching Models and Optimal Registry for Voluntary Organ Donation Registries	33
1.1 Introduction	33
1.2 Donors, Receivers and Registry	34
1.3 The theory of the Optimal Registry	38
1.4 Continuous types models	43
1.5 Optimal Registry: Some simulations	47
1.6 Filter and implementable improvement of a Registry	52
1.7 A Monte Carlo evaluation of an implementable improvement of a registry	56
1.8 Efficiency of Bone Marrow Donors' Registries: Models and Orders of Magnitude	65
1.8.1 The reference model	68
1.8.2 Numerical simulations	70
1.8.3 Back to the number of HLA types	71
1.8.4 Microsatelites filtering	74

1.8.5 Main conclusions and extensions.....	75
2 Evaluation Economique de l'Organisation d'un Registre	81
2.1 Un modèle de référence.....	81
2.1.1 Introduction.....	81
2.1.2 Le modèle structurel.....	82
2.1.3 La calibration du modèle.....	89
2.1.4 Résultats.....	92
2.2 Collaboration internationale entre Registres.....	98
2.2.1 Introduction.....	98
2.2.2 Le modèle.....	98
2.2.3 Les hypothèses de calibration.....	100
2.2.4 Etude sans sélection des donneurs.....	103
2.2.5 Etude avec sélection des donneurs.....	106
2.2.6 Un modèle de jeu entre plusieurs Registres.....	108
2.3 Choix de la précision du typage.....	111
2.3.1 Introduction.....	111
2.3.2 Choix d'un niveau de typage dans un seul pays sans sélection des donneurs	114
2.3.3 Comparaison des registres optimaux dans les cas 2 Digits et 4 Digits par la probabilité de mise à disposition.....	115
2.3.4 Comparaison des registres optimaux dans les cas 2 Digits et 4 Digits par le bien-être social.....	117
2.3.5 Choix entre un registre à 2 Digits et un registre à 4 Digits dans un contexte de concurrence internationale.....	119

2.4 ANNEXE: UN CADRE ANALYTIQUE DE RÉFÉRENCE	122
2.4.1 La collecte des données	125
2.4.2 Le choix des inputs de la grille d'analyse et la calibration du modèle	127
2.4.3 Les grilles d'analyse du coût du typage HLA	132
3 Contributions Statistiques à l'organisation d'un Registre de donneurs.	137
3.1 Introduction	137
3.2 A moment estimation of the haplotypes distribution using genotypes data....	139
3.2.1 The latent model and the observables.....	139
3.2.2 Likelihood.....	142
3.2.3 A moment estimation: an introductory case	144
3.2.4 Moment estimation : the general case.....	148
3.2.5 Asymptotic distribution	150
3.2.6 Correction for negative probabilities	152
3.2.7 A Monte Carlo simulation	153
3.2.8 Application to the relation between the microsatellite MOGc and gene HLA-A.....	155
3.3 Implementation of the estimation algorithm.....	160
3.3.1 A five Loci simulation	163
3.3.2 An application to HLA haplotype frequencies and comparison with EM algorithm	168
3.4 How many different HLA genotypes exist in a population?.....	170
3.4.1 The problem.....	170

3.4.2 A Polya urn scheme for types' generation	170
3.4.3 Exact and approximate number of types	171
3.4.4 Estimation of n_0 and prediction of the number of types	173
3.4.5 A repeated observations case and a test of the model.....	175
Conclusions et Extensions.....	179
Bibliographie	183
Résumé en anglais	189
Résumé en français	190

Introduction

Pour guérir certaines maladies graves et notamment des maladies hématologiques malignes, comme les leucémies par exemple, le médecin dispose de diverses modalités de traitement, notamment l'allogreffe de Cellules Souches Hématopoïétiques (CSH). Celle-ci ne constitue pas un recours systématique mais représente, dans certains cas, une possibilité thérapeutique. La greffe de CSH est un traitement reconnu, et parfois le seul traitement curatif, pour les hémopathies malignes (leucémies) et les défauts génétiques du système immunitaire. Une transplantation de CSH constitue un traitement efficace puisque le greffon permet la reconstruction immunitaire complète de l'hôte, ce qui est parfois la seule solution pour des individus dotés d'un système immunitaire défaillant. Cette transplantation provient d'un donneur qui doit posséder un système HLA (Human Leucocytes Antigens) identique à celui du patient pour minimiser les risques de rejet. On dit alors que le donneur est compatible HLA (histocompatible) avec le malade. On recherche d'abord des donneurs apparentés (car plus proches génétiquement), puis non apparentés. C'est parmi les frères et soeurs que l'on trouve le plus facilement des sujets ayant le même groupe tissulaire. Lorsqu'une greffe de CSH est envisagée (les immunologistes utilisent le terme "transplantation"), le médecin va d'abord chercher un donneur dans la fratrie. On parle alors de greffe intra-familiale ou d'allogreffe apparentée de CSH. Ce type de greffe représente 68.3% des allogreffes de moelle osseuse en 2003. Mais parfois, ce donneur familial n'existe pas, soit parce que le malade n'a ni frère ni soeur, soit parce que ces derniers ne sont pas compatibles avec lui. Or 70% des malades n'ont pas de soeur ou de frère compatible. Les médecins font alors appel au fichier mondial des volontaires de don de CSH: pour ces malades, il existe une possibilité de trouver un donneur HLA identique (Rendine 1999). Plutôt que de typer l'ensemble des gènes HLA pour chaque individu, l'usage est de ne typer que les antigènes HLA-A, HLA-B, HLA-DR, qui portent

l'essentiel de l'information importante sur le plan de la réaction immunitaire. Dans un souci de simplification, on appellera dans cette thèse, Haplotype HLA, l'ensemble de ces trois gènes, portés sur un même chromosome d'origine paternelle ou maternelle.²

Le fichier mondial compte près de 9 millions de donneurs, dont 130 000 sont issus de l'actuel fichier français³. Mais il reste parfois difficile, voire impossible, de trouver un donneur compatible car le système HLA est très complexe et très polymorphe. En 2003, en France, 31.7% des allogreffes de moelle osseuse ont été réalisées en faisant appel aux fichiers répertoriant les volontaires de don de moelle osseuse. Ces Registres nationaux ont été constitués depuis une vingtaine d'années afin de répertorier les donneurs volontaires de Cellules Souches Hématopoïétiques (CSH) et de mettre à la disposition des patients des possibilités thérapeutiques efficaces. La compatibilité HLA entre patient et donneur de CSH (prélevées dans la moelle osseuse ou à partir de sang périphérique) est un paramètre essentiel qui détermine le succès de la greffe: le risque de rejet de la greffe augmente avec le degré de disparité HLA.

Les greffes de CSH provenant de donneurs non apparentés représentaient 8.5% des greffes allogéniques en 1990 , 22% en 1996, 31% en 2000 (Gratwohl, Baldomero, Herisberger et al. 2000). Cette constante augmentation résulte de la mise en réseau par le Registre BMDW (Bone Marrow Donors Worldwide)⁴ de près de 9 millions de donneurs volontaires de groupes HLA connus, inscrits dans les différents Registres nationaux. L'inscription des donneurs sur les Registres nécessite plusieurs étapes qui sont réalisées au sein de centres agréés : information du donneur, acceptation d'un engagement volontaire, visite médicale, typage HLA. Le typage (ou phénotypage HLA) consiste à identifier des antigènes caractéristiques de l'individu. Cette analyse est réalisée avant toute greffe de cellules pour apprécier la compatibilité tissulaire. Chaque individu possède sa propre combinaison d'antigènes. Actuellement, en France, la quasi-totalité des donneurs sont typés en basse résolution pour HLA-A et HLA-B (techniques sérologiques) et -DR (techniques de biologie

²see Marsh et al. 2004 for Nomenclature factors of the HLA system

³Le nombre total de donneurs inscrits au 31 décembre 2004 est de 129 042. (source: Rapport d'activité FGM, <http://www.fgm.fr>)

⁴<http://www.bmdw.org>

moléculaire). Les techniques de biologie moléculaire ont progressivement remplacé, dans la pratique courante, la sérologie tout au moins, pour l'analyse du polymorphisme HLA de classe II (HLA-DR, -DQ, -DP). Deux niveaux de résolution peuvent être atteints par les techniques de typage HLA basées sur l'analyse de l'ADN. Classiquement, dans un premier temps, la recherche de donneurs potentiels dans les Registres est réalisée en comparant les typages HLA de classe I (HLA-A, B) et de classe II (HLA-DR et/ou -DQ) de basse ou de moyenne résolution des donneurs et des receveurs. En cas d'appariement à ce stade, le polymorphisme HLA de classe I (HLA-A, -B, -C) et de classe II (HLA-DR, -DQ,-DP) est analysé de façon exhaustive aboutissant au typage HLA de haute résolution indispensable à l'appariement entre les receveurs et les donneurs de CSH sélectionnés.

Le Registre national géré par France greffe de Moelle (FGM⁵) comprend un peu plus de 120 000 donneurs volontaires de CSH, soit 0.2% de la population française. Créé en 1986 par les professeurs Jean Bernard et Jean Dausset, ce Registre recense les volontaires pour le don de moelle osseuse pour les patients en attente de greffe de moelle et n'ayant pas de donneur familial compatible. Malgré le nombre apparemment très élevé de donneurs de CSH, occasionnant des "frais de recrutement", de typage HLA et de gestion des Registres importants, la situation n'est pas satisfaisante. L'amélioration de la qualité et de l'efficacité des Registres est une priorité afin de pouvoir satisfaire les besoins thérapeutiques des patients receveurs potentiels. L'origine des donneurs tend à être relativement homogène contrastant avec celle de l'origine des receveurs qui reflète l'extrême diversité de la population humaine. Ainsi l'amélioration des Registres, en particulier du Registre français, nécessiterait a priori d'augmenter la diversité et le nombre des donneurs ainsi que d'améliorer (au sens rendre "plus informatif") le typage HLA. De nouvelles techniques pourraient contribuer à définir différents niveaux de typages et ainsi de gagner, en temps et en coût, donc en efficacité, lors du recrutement des donneurs. Le Plan Greffe (2002), relatif au recrutement de donneurs volontaires de moelle osseuse, mis en place pour 3 ans par le gouvernement français, est reconduit pour 10 ans: il s'agit de sensibiliser de nouveaux donneurs grâce à une campagne d'information étalée dans le temps

⁵La loi de Santé Publique du 13 août 2004 a confié la tutelle du registre à la nouvelle Agence de Biomédecine

et de faciliter l'accueil des donneurs dans les centres de l'Établissement français du sang et les établissements de soins.

L'objectif que l'Agence de biomédecine s'engage ainsi à remplir est d'accroître le fichier français de 10 000 nouveaux donneurs de CSH par an, sur 10 ans. L'agence de biomédecine, a repris en mai 2005 les activités de prélèvement et de greffes d'organes, de tissus et de cellules confiées à l'établissement Français des Greffes depuis plus de 10 ans. Cet établissement était notamment chargé de l'enregistrement de l'inscription des patients, de la gestion de celle-ci et de l'attribution des greffons, qu'ils aient été prélevés en France ou en dehors du territoire national. Ce plan d'action de 10 000 nouveaux donneurs par an tient compte des recommandations d'un groupe d'experts et du souhait des associations de voir s'étendre le fichier français. Ces dispositions permettent également de participer à l'effort mondial de recrutement des nouveaux donneurs et d'être en mesure de proposer aux patients étrangers les ressources du fichier français, comme le fichier français fait appel, pour les patients français, aux ressources du fichier mondial. Ce recrutement devrait conduire à augmenter le taux de greffes non-apparentées réalisées à partir du fichier français de 25% à 50%. En 2003, en effet, 75% des allogreffes de moelle osseuse non apparentées ont été réalisées à partir de donneurs issus du fichier mondial. Si l'objectif est de pourvoir réaliser un nombre maximum d'appariements entre donneurs et receveurs, ce travail de thèse montre qu'il n'est pas optimal, contrairement à une idée reçue, d'augmenter l'hétérogénéité génétique des registres ou d'accroître de façon systématique le nombre de donneurs. L'objectif poursuivi est double : il s'agit d'une part de réduire l'inégalité des chances pour les patients de trouver un donneur non apparenté HLA identique, d'autre part de diminuer significativement les ressources affectées (inutilement ?) au recrutement des donneurs, porteurs de types HLA communs déjà présents dans les Registres. Le financement du système de santé en France évolue vers l'attribution d'enveloppes affectées aux différents secteurs de la santé: cette thèse utilise le calcul économique pour modéliser l'organisation de la filière du don de Cellules Souches Hématopoïétiques en vue de greffes dans une optique d'aide à la décision.

Dans un article pionnier, Kennet Arrow (1963) soulignait les spécificités du calcul économique

en santé dans le champ de la théorie micro-économique, sous le titre "Uncertainty and the Welfare Economics of Medical care". Lorsque le financement d'un système de santé est administré (public), des arbitrages collectifs se succèdent aux arbitrages individuels. On dit que la contrainte de ressources, qui d'ordinaire s'impose aux ménages sous la forme d'une contrainte budgétaire individuelle, est "externalisée" au niveau collectif. Ainsi un conflit d'éthique peut apparaître au niveau des choix thérapeutiques, si l'éthique individuelle du couple médecin-patient peut se résumer par "la santé n'a pas de prix" tandis que la définition d'une éthique collective part du constat selon lequel ce qui est alloué à certains n'est plus disponible pour les autres. Le financement du système de santé en France évolue vers l'attribution d'enveloppes affectées aux différents secteurs de la santé. Ce choix implique le nécessaire développement de la mesure du coût et de l'efficacité des soins dispensés. Si on souligne l'inefficacité d'un accroissement de taille des registres en termes de Bien-être-Social, une valeur raisonnable des paramètres du modèle que l'on propose dans le chapitre 2 justifie toutefois le plan d'augmentation de 100 000 nouveaux donneurs sur 10 ans.

Ce travail de thèse utilise le calcul économique pour modéliser l'organisation de la filière du don de cellules en vue de greffes. Si cette évaluation économique des Registres propose un mécanisme d'appariement, qui représenterait un gain pour le Registre lorsqu'il existe un donneur de même type que le receveur, d'une perte dans le cas contraire, elle devrait s'accompagner d'une étude de l'automatisation des techniques de typage ou suggérer une organisation différente de la filière favorisant les économies d'échelle : l'idée étant de réduire de façon significative les étapes consommatrices de ressources nécessaires à l'inscription des donneurs sur les Registres. La démarche de l'économiste est ainsi souvent "mal perçue" par les professionnels de la santé. L'Économie de la santé est parfois considérée comme "économies de santé", comme on parlerait d'"économies d'énergie", à savoir ne s'intéressant qu'à la maîtrise des dépenses, qu'aux économies à tout prix, y compris au détriment de l'intérêt du malade. Ceci n'est pas sans heurter l'éthique du médecin. En conséquence, à la demande des intéressés, ne seront mentionnés aucun des

noms des laboratoires d'immunologie concernés, sans lesquels les aspects empiriques de ce travail de thèse n'auraient pu être abordés.

Cette thèse se compose de trois chapitres, qui traitent de problèmes liés à la modélisation décisionnelle de l'organisation du système des fichiers de CSH en Europe. Les motivations et le contexte de ce travail de recherche étant exclusivement européens, cette thèse n'envisage pas les modalités ou les pratiques existantes aux Etats-Unis. En revanche l'étude du typage centralisé et davantage "ciblé" tel qu'il se pratique dans le cadre du *National Marrow Donor Program* serait une extension intéressante de ce travail de thèse. Le premier chapitre propose une mesure d'efficacité d'un Registre de donneurs volontaires de CSH. Il retient comme critère d'optimisation la probabilité pour un receveur quelconque de trouver un donneur. Ce chapitre met en évidence la valeur du Registre si les donneurs arrivent à la même fréquence que les receveurs (pas de sélection) ainsi que la valeur du Registre liée à la sélection optimale des donneurs. Quel que soit le critère de sélection retenu, ce chapitre souligne la faible efficacité de l'organisation d'un registre de CSH au niveau d'une entité géographique (pays, région, ethnie): la probabilité de ne pas trouver de donneur compatible avec un receveur potentiel demeure très élevée même pour des tailles de registres importants si l'on néglige la dimension internationale des registres. Le second chapitre traite de l'analyse économique de la gestion d'un registre. Etant donné le coût élevé de l'accueil, du typage des donneurs et de gestion du registre, est-il socialement et économiquement efficace d'avoir un registre de donneurs ? Les responsables du registre français soulignent en outre le fait que l'accroissement du registre est davantage limité par des considérations budgétaires que par manque de volontaires : l'objectif du chapitre est d'évaluer l'efficacité d'un registre de donneurs de CSH à la lumière du calcul économique et d'identifier les éléments-clé qui entrent en jeu. Certains peuvent être évalués à partir d'arguments statistiques (nombre et distribution des types HLA en France), d'autres peuvent évoluer en fonction de la politique de gestion du fichier et d'autres enfin relèvent d'une évaluation plus complexe liée au bénéfice attendu d'une greffe. L'intérêt de ce calcul est double : évaluer la valeur des paramètres qui rendent optimale la politique actuellement suivie ou à partir de valeurs de paramètres a priori

calculer une gestion optimisée du registre. On pourra aussi examiner la sensibilité de ces résultats aux paramètres "statistiques". La question posée est la suivante : étant donné le coût élevé de l'accueil, du typage des donneurs et de la gestion du registre, est-il socialement et économiquement efficace d'avoir un Registre de donneurs ? On modélise ici la situation actuelle où d'une part, le financement du registre est assuré en grande partie par des fonds publics et où d'autre part, le typage des donneurs est réalisé par des laboratoires hospitaliers ou appartenant à l'établissement français du sang. Utilisant les données collectées dans le cadre du projet européen MADDO, la seconde partie de ce chapitre tente de mesurer le coût du typage HLA, peu dépendant de la structure institutionnelle et des règles de tarification en vigueur. Il s'agit d'une étude économique descriptive du coût qui ne se limite pas aux seules dépenses exprimées en unités monétaires (tarifs des réactifs), mais considère également toutes les ressources, notamment celles dont la consommation n'est pas représentée par les prix de marché, comme par exemple le travail ou le capital. Il s'agit d'une étude empirique réalisée grâce aux données collectées et recueillies dans un certain nombre de laboratoires d'Immunologie et d'Histocompatibilité. Pour des raisons de confidentialité, aucune évaluation de coûts réalisée au sein des laboratoires concernés ne sera communiquée: on choisit d'explicitier le tarif "Plan greffe" 2002⁶. On construit ainsi une structure d'évaluation économique des coûts, présentée en annexe au chapitre sous la forme d'une grille d'analyse, représentative des actes de laboratoire de type typage HLA. Les chiffres obtenus recouvrent une forte disparité en termes de coût due à l'utilisation de techniques différentes dans les laboratoires.

Le troisième chapitre apporte une contribution statistique à la sélection des donneurs : il pose le problème de la distribution jointe de plusieurs variables discrètes (Haplotypes⁷): connaissant les phénotypes des individus et propose une méthode simple d'estimation. On

⁶183.13 euros par typage

⁷A l'exception des cellules sexuelles matures, toutes les cellules humaines contiennent des paires de chromosomes. L'un des chromosomes est hérité du père, l'autre de la mère. Mais ce ne sont pas des copies exactes des chromosomes qui passent d'une génération à l'autre. En effet, lors de la formation des ovules et des spermatozoïdes, les paires de chromosomes subissent un processus de recombinaison. Les deux chromosomes formant une paire s'unissent et échangent des séquences. Le résultat est un chromosome hybride contenant des segments des deux chromosomes de la paire, et c'est ce chromosome hybride qui sera transmis à la génération suivante

étudie dans ce chapitre les propriétés de cette méthode et l'on compare par simulations son efficacité à celle de la méthode utilisée par les statisticiens du corps médical. Cette question statistique est un élément du problème décisionnel relatif à l'optimisation du fichier de donneurs de CSH défini dans le chapitre premier: L'information contenue dans les haplotypes (variables latentes du modèle) est indispensable à l'analyse des données génétiques qui intéressent le biologiste or le typage HLA ne procure qu'une association de deux haplotypes, appelée phénotype.

1 Calcul économique et théorie de la décision

L'utilisation du calcul économique peut montrer son efficacité dans des domaines dans lesquels son évocation même suscite en général une réaction immédiate de rejet et l'exemple du don de CSH, le plus souvent prélevées dans la moelle osseuse, en vue de greffes pourrait être de cette nature. Malgré un nombre important de donneurs, plus du tiers des malades ne trouvent pas de donneur compatible dans un délai raisonnable. L'introduction de nouveaux donneurs est une opération coûteuse car elle nécessite des campagnes de sensibilisation, un accueil et le typage HLA des volontaires. Par ailleurs, le mécanisme d'arrivée spontanée des donneurs va entraîner la réplique de types HLA connus (que l'on a déjà dans le Registre au moins une fois) alors que les types rares demeurent absents.

On pourrait imaginer une présélection des donneurs potentiels sur la base d'un test génétique peu coûteux et permettant de prévoir avec une bonne fiabilité le type HLA. L'efficacité de cette procédure dépend des différentes composantes économiques du mécanisme (coût du typage, coût du test de sélection) qu'il convient d'évaluer. L'automatisation des techniques de typage, selon les techniques modernes de la génomique, l'organisation différente de la filière favorisant les économies d'échelle sont également des voies à évaluer économiquement pour leur potentialité à réduire de façon significative les étapes consommatrices de ressources nécessaires à l'inscription des donneurs sur les Registres. L'objectif de ce chapitre est d'estimer la valeur d'un Registre afin de modéliser ce que serait le mécanisme de construction d'un Registre de donneurs de CSH. Si le point de départ de ce

travail s'appuie sur le cas particulier du don de CSH, le modèle proposé est applicable à l'ensemble des Registres de dons d'organes. La greffe de moelle osseuse, pour sa réussite, est soumise à un respect rigoureux de la compatibilité HLA entre le donneur (vivant) et le receveur. Les combinaisons possibles des groupes HLA de chaque individu sont très nombreuses et la probabilité de trouver une "bonne" correspondance entre un receveur et un donneur est infime. Le polymorphisme des loci du système HLA est très important, mais il existe toutefois des associations préférentielles d'allèles. Un allèle est une version possible d'un gène, il est constitué d'un enchaînement de nucléotides (un fragment d'ADN). Chez un individu, chaque gène est représenté par deux allèles, situés sur le même locus (place du gène sur le chromosome), et que ces deux allèles sont soit identiques dans leur composition nucléotidique (individu homozygote pour ce gène), soit différents dans leur composition (individu hétérozygote). Un gène peut correspondre, dans une population donnée, à plusieurs allèles. On dit alors que ce gène est polyallélique; si un gène est représenté par plus de deux allèles et que ceux-ci se retrouvent dans une population avec une fréquence supérieure ou égale à 1%, il est polymorphe. Tous les allèles ne sont pas équiprobables au sein d'une population et il existe des allèles fréquents, d'autres plus rares ou inexistantes. De plus, les associations préférentielles d'allèles à divers locus sur le même haplotype (phénomène connu sous le nom de "déséquilibre de liaison") concourent au fait qu'il existe également des haplotypes fréquents et des haplotypes rares : la probabilité de trouver un individu strictement identique dans un fichier de donneurs non apparentés est généralement très faible. Ainsi un donneur identique n'est pas toujours trouvé et dans certains cas des greffes avec des donneurs partiellement compatibles sont réalisées. Ces greffes ont moins de chances de succès, mais l'acceptation d'une erreur augmente considérablement la probabilité de trouver un donneur dans un fichier.

Etant donné le fait que le nombre d'allèles possible pour chaque gène varie entre 10 et 40, cela représente une immensité (en théorie) de possibilités pour le phénotype HLA. Si le fichier mondial des volontaires de don de CSH compte près de 9 millions de donneurs (dont 120 000 sont issus de l'actuel fichier français), il reste parfois difficile, voire impossible, de trouver un donneur compatible car le système HLA est très complexe et très polymorphe.

Le modèle proposé dans ce chapitre montre la faible efficacité de l'organisation du fichier français : Si l'on utilise les données de France Greffe de Moelle, la probabilité pour un malade français de ne pas trouver un donneur dans le fichier FGM serait proche de 90%⁸. De plus, on montre que l'efficacité des Registres (mesurée en termes de proportion de patients qui trouvent un donneur), augmente très lentement avec la taille du Registre (see Oudhshoorn, Van der Zanden, Bakker, Van Rood (2005), Fève, Florens(2005)). Ce chapitre tente de formaliser le problème de l'optimisation de la composition des Registres. La conviction du médecin greffeur estime que la chance de trouver un donneur compatible dépend de la taille du Registre (Speiser, Tiercy, Rufer, Chapuis et al. 1994). Rendine, Borelli, Barbanti et al. (1999) proposent une stratégie de recrutement des donneurs afin de retracer l'hétérogénéité génétique de la fraction de la population italienne étudiée. Déjà en 1989, Sonnenber, Eckman et Pauker, défendaient l'idée d'une relation croissante entre la taille d'un Registre et le fait d'augmenter le nombre potentiel de "matchings". A l'appui du modèle mathématique proposé, ce chapitre souligne au contraire, la faible efficacité d'un Registre composé d'un nombre important de donneurs. La mesure d'efficacité définie par le modèle estime la valeur du Registre: il s'agit de la probabilité pour un receveur quelconque de trouver un donneur. Un accroissement important de la taille des Registres n'augmenterait pas leur efficacité de manière sensible. Et enfin, un troisième point souligne l'importance négligeable des mécanismes de sélection des donneurs. La quatrième partie de ce chapitre propose plusieurs simulations afin de tester le modèle économique que l'on propose.

Tous les modèles proposés utilisent les données FGM. La calibration de départ reprend les chiffres de 2003 : Le Registre comporte 937 donneurs français et la France compte 62 malades receveurs greffés (parmi 813 malades en attente de greffe qui ne trouvent pas de donneur disponible). La probabilité pour un malade français de ne pas trouver un donneur serait de 92%; On compare dans ce chapitre la valeur du Registre si les donneurs arrivent à la même fréquence que les receveurs (pas de sélection) à la valeur du Registre liée à la sélection optimale des donneurs ainsi qu'à la valeur du Registre associée à des

⁸estimation faite sur la base déléments chiffrés du rapport d'activité 2004

mécanismes de sélection réalisables. La dernière partie du chapitre souligne les principales conclusions issues des simulations réalisées : une augmentation significative de la taille des Registres (i.e. multiplier par 2 ou 3 le nombre de donneurs) a un impact relativement faible sur la probabilité de trouver un donneur: si on double la taille du Registre, on augmente cette probabilité de 10% à peine. L'impact d'un mécanisme de sélection est lui aussi très faible. Le premier modèle montre que l'efficacité du Registre entre arrivée aléatoire des donneurs (sans sélection) et sélection optimale n'est que de 2% si la taille du Registre est multipliée par 2. De plus, la règle de sélection que l'on suggère dans ce chapitre élimine un très petit nombre de "fréquents" et la majeure partie des types "rares" (i.e. présents une seule fois dans le Registre). La dernière section du chapitre propose un design pour un Registre optimal mais souligne le fait que ce Registre n'est pas implémentable. Le mécanisme de sélection retenu vise à utiliser un test (supposé moins coûteux qu'un typage HLA complet) comme pré-filtre avant inscription du donneur sur le Registre. En effet, le polymorphisme de la région HLA peut correspondre à diverses différences: variation du type des nucléotides constituant l'ADN, inversion, insertion ou répétition de certains motifs. Les plus intéressants - car les plus facilement lisibles - se traduisent par des différences de longueur d'allèles, c'est-à-dire qu'ils comptent un nombre de paires de base variable par individu. C'est le cas des microsatellites: le nombre d'unités de répétition à un locus donné est variable d'un individu à l'autre au sein d'une même espèce, ce qui en fait d'excellents marqueurs génétiques. Les microsatellites sont des loci (ou régions) où de courtes séquences d'ADN (constituées d'un petit nombre de nucléotides: A, G, C, T, Adénine, Guanine, Cytosine et Thymine) sont répétées consécutivement. Les SNP's (Single Nuclear Polymorphism) constituent également des marqueurs génétiques de choix du génome humain. Ils représentent une variation d'une séquence d'ADN, qui peut provenir de la simple modification d'un seul nucléotide (A, G, C ou T). 90% de toutes les variations génétiques humaines sont de type SNP's: ceci est très informatif sur la plupart des maladies de l'homme, des virus, des drogues ou des bactéries. Les SNP's sont en outre invariants dans le temps et sont donc très utilisés dans les études de génétique des populations.

On exploite ainsi deux procédures de test différentes les microsatellites et les SNP's. A chaque fois les questions posées sont les mêmes: comment utiliser de manière la plus efficace possible les informations fournies par le test ? N'est-il pas plus efficace de laisser arriver les donneurs de façon aléatoire ? De façon plus générale, en imaginant que de nombreux tests (utilisés comme pré-filtre) soient disponibles, quelle est la meilleure façon d'augmenter l'efficacité du Registre ? les simulations réalisées en terme de stratégie optimale (utilisant un test comme pré-filtre ou non) se heurtent à un problème de taille : celui de la dimension de l'espace paramétrique. Si l'on prend pour exemple la greffe de CSH, le nombre de phénotypes HLA est au moins de 500 000 types en France et la liste n'est pas exhaustive. Ce chapitre propose une méthodologie permettant d'évaluer le design d'un Registre de donneurs de CSH. Le critère d'évaluation retenu est la probabilité pour un receveur de trouver un donneur. On compare les mécanismes d'arrivée aléatoire des donneurs (le cas aujourd'hui) avec des registres dont on supposerait un fonctionnement optimal (mécanisme de tri et ou de sélection approprié). L'efficacité d'un registre est fonction de sa taille (le stock de donneurs), de la probabilité pour un donneur d'être disponible pour une greffe, et du nombre de phénotypes existant dans la population. La moyenne géométrique des fréquences des phénotypes présents dans le registre définit la distribution des fréquences des phénotypes. La calibration de notre modèle appliqué à plusieurs scénarios nous donne les résultats suivants : un registre de donneurs de CSH (possédant les caractéristiques du registre français en 2003) n'est pas très efficace (seuls 10% des patients trouvent un donneur). On montre par ailleurs que dans le meilleur des cas, multiplier la taille du registre par 2 ou par 3 (à condition qu'un mécanisme de sélection optimal des donneurs soit mis en place) n'augmenterait l'efficacité du registre que de 20%. Notre petit modèle examine l'influence de la taille d'un Registre sur son efficacité ainsi que celle d'un mécanisme de sélection des donneurs: les simulations réalisées sur la base de l'échantillon MADO soulignent la faible efficacité des registres de CSH. Mais si l'on se limite aux types "fréquents" (soit 0.52% d'individus sur les 107 925 présents dans le Registre FGM au moment de notre étude), notre modèle montre que 92% des malades "fréquents" trouvent effectivement un donneur.

Notre étude pourrait être transposée au niveau mondial en supposant que tous les Registres fonctionnent comme un Registre "intégré". On pourrait également imaginer différentes sources de CSH: les registres de donneurs volontaires et le sang de cordon. La modélisation de ces deux sources de CSH fait l'objet de travaux en cours.

2 Evaluation économique de l'organisation d'un Registre

L'inscription des donneurs sur les Registres nécessite plusieurs étapes réalisées au sein de centres agréés : information du donneur, acceptation d'un engagement volontaire, visite médicale, étude des sérologies anti-virales et typage HLA. Les stratégies de typage HLA varient dans les laboratoires en fonction de l'expérience locale et du type d'équipement disponible. Quelle que soit la technique de typage utilisée, le typage a un coût que ce second chapitre se propose de mesurer. La première partie de ce chapitre propose un petit modèle économique du fonctionnement et du coût de la gestion d'un registre de donneurs volontaires. Les registres de donneurs ont pour fonction de fournir des Cellules Souches Hématopoïétiques (CSH) issues de donneurs volontaires compatibles avec les malades pour lesquels aucun donneur familial n'est disponible. L'organisation de ces registres pour de très nombreux pays et au niveau mondial nécessite une organisation complexe et coûteuse. L'article se propose de modéliser le bien-être social apporté par cette organisation et de proposer une stratégie d'optimisation de l'évolution des registres. Le modèle calibré que nous proposons montre l'incidence relativement faible de l'accroissement des registres comparé à l'accroissement de la disponibilité des donneurs et à la diminution des coûts. On évalue le gain pour les patients associé à l'existence d'un registre (le surplus pour les patients) ainsi qu'une mesure de l'avantage pour les laboratoires associé au fait de devoir réaliser les typages HLA (le surplus pour les laboratoires). On modélise ainsi le Bien-être Social (défini comme la somme du surplus pour les patients et la somme du surplus pour les malades) apporté par le Registre français et on propose une stratégie d'optimisation de l'évolution des registres. Le modèle détermine la valeur implicite d'une greffe de CSH

($V = 55\,831\text{€}$). Ce résultat semble tout à fait acceptable compte tenu du coût général du traitement des leucémies. Un autre résultat découlant de ce modèle est la justification du tarif de 183€ qui permet de recalculer les coûts fixes des laboratoires. Trois conclusions se dégagent ainsi de notre étude : L'accroissement de la taille du Registre actuel n'aura que peu d'impact sur son efficacité, donc sur le bien-être social. Malgré ce peu d'impact une valeur raisonnable des paramètres justifie le plan d'augmentation de 100 000 nouveaux donneurs sur 10 ans. La politique de l'agence régulatrice du système devrait être d'accroître la probabilité qu'un donneur compatible soit disponible et d'accepter la greffe. De manière moins importante, toute réduction du coût marginal aurait une incidence positive sur le BES et la quantité Q optimale conduirait à une augmentation de l'efficacité du système. Notre étude souligne que l'organisation d'un Registre au niveau d'un pays a une faible efficacité. On a montré que la probabilité de ne pas trouver de donneur compatible avec un receveur potentiel demeure très élevée même pour des tailles de registre importantes. Au niveau d'un seul pays, on peut conclure que l'accroissement du fichier des donneurs ne se justifie pas au delà d'une certaine taille mais que d'autres mesures peuvent être plus utiles (augmentation de la disponibilité des donneurs, meilleure qualité de typage). Cette analyse néglige toutefois la dimension internationale des registres: grâce à leur interconnection, l'ensemble des registres peut être consulté et un donneur peut être choisi dans n'importe quel pays. Un argument en faveur de l'accroissement des registres nationaux est donc d'augmenter leur utilisation pour des receveurs étrangers. C'est l'objet du petit modèle proposé dans la seconde partie du chapitre. Les troisièmes et quatrièmes parties traitent du choix de la précision du typage. La technologie de typage HLA a longtemps consisté en une procédure en deux temps: typage grossier d'abord puis fin ensuite (les coûts de ces deux typages s'ajoutent dans ce cas). De manière expérimentale, un typage fin d'emblée est envisagé. Praticué à grande échelle par des moyens optimisés, il diminuerait considérablement l'écart de coût entre les deux techniques, et ferait prendre la décision pour un typage fin, en cas de nouveaux entrants dans le Registre. On peut ainsi envisager le choix entre les deux registres dans le cas d'un seul pays ou en concurrence internationale. La dernière partie du chapitre considère un modèle de jeu entre

les différents registres dans chaque pays, ce qui devrait permettre de mesurer l'incidence de la collaboration internationale sur les tailles optimales des registres. L'annexe I est une évaluation empirique du coût du typage HLA⁹, peu dépendant de la structure institutionnelle et des règles de tarification en vigueur. La méthodologie proposée consiste à découper chaque technique de typage en tâches élémentaires pour lesquelles le temps de travail et l'équipement requis sont évalués à l'aide de grilles de rémunération et de prix des éléments connus. La définition et la quantification des tâches ont été réalisées par interviews, sous forme d'entretiens directs. La grille d'analyse ainsi construite nous permet de définir un protocole de laboratoire "standardisé": L'évaluation "empirique" du coût du typage se réalise en deux temps : la collecte des données (celle-ci détermine les quantités consommées), la mesure des quantités consommées (les structures de coût sont-elles toutes identiques ? existe-il des économies d'échelle ? des économies de gamme. On obtient ainsi autant de coûts pour le typage HLA que de laboratoires d'immunologie interviewés. Les coûts de séquençage réalisés par exemple dans une structure centralisée ont été analysés et comparés avec les coûts des méthodes classiques de typage basse et haute résolution(et avec le coût des typages effectués de façon non centralisée). Dans cette optique, le calcul distingue le coût de l'extraction d'ADN du reste de la technique. Les différentes composantes du coût, tenant compte, en particulier des coûts de la centralisation (frais d'envoi des échantillons d'ADN par exemple), ainsi que la présence de rendements d'échelle et d'économie de gamme (incidence sur le coût du typage d'activités complémentaires du laboratoire,... typage d'autres régions génétiques) ont été étudiés. L'évaluation du coût du typage proposé intègre celle de l'utilisation d'un test (de deux tests : les microsatellites et les SNPs) comme étape du processus de constitution d'un Registre; est-il économiquement pertinent et techniquement envisageable d'utiliser un test comme filtre permettant d'optimiser le processus d'échantillonnage des donneurs au sein de la population ? L'introduction de nouveaux donneurs est une opération coûteuse car elle nécessite campagnes de sensibilisation, accueil et typage HLA des volontaires. Un registre optimal -au sens où il maximise la probabilité pour un malade de trouver un

⁹réalisée dans le cadre du Work Package 6 de MADO

donneur - ne doit pas nécessairement refléter la population ni offrir la plus grande diversité possible de donneurs. On peut également envisager de pré-sélectionner les donneurs potentiels sur la base d'un test génétique peu coûteux et permettant de prévoir avec une bonne fiabilité le type HLA. L'efficacité de cette procédure dépend des différentes composantes économiques du mécanisme (coût du typage, coût du test de sélection) qu'il convient d'évaluer. L'automatisation des techniques moléculaires de typage, selon les techniques modernes de la génomique, l'organisation différente de la filière favorisant les économies d'échelle sont également des voies à évaluer économiquement pour leur potentialité à réduire de façon significative les étapes consommatrices de ressources nécessaires à l'inscription des donneurs sur les registres.

3 Contributions Statistiques à l'organisation d'un Registre de donneurs

Le matériel génétique humain est organisé sous forme de paires de chromosomes, dits chromosomes homologues (23 paires de chromosomes homologues): tous les loci (donc tous les gènes sont représentés deux fois. Ainsi les humains possèdent deux allèles de chaque gène (un sur chaque chromosome), l'un hérité du père l'autre de la mère. Par exemple, on peut dire qu'un individu possède les allèles $A1$ et $A3$ du gène HLA-A. Un haplotype se définit comme étant la combinaison, sur un chromosome, d'un nombre donné d'allèles, à différents locus voisins. Sur chacun des deux chromosomes 6 existe un haplotype correspondant aux groupes des allèles de tous les gènes du Complexe Majeur d'Histocompatibilité (CMH). Chaque sujet hérite un haplotype de ses deux parents. L'un des haplotypes est hérité du père, l'autre de la mère. Mais ce ne sont pas des copies exactes qui passent d'une génération à l'autre. En effet, lors de la formation des ovules et des spermatozoïdes, les paires de chromosomes subissent un processus de recombinaison. Les deux chromosomes formant une paire s'unissent et échangent des séquences. Le résultat est un chromosome hybride contenant des segments des deux chromosomes de la paire, et c'est ce chromosome hybride qui sera transmis à la génération suivante. A mesure que les humains se

sont répandus dans le monde, la fréquence des haplotypes a varié d'une région à l'autre par le jeu du hasard et de la sélection naturelle. Ainsi aujourd'hui, la fréquence d'un haplotype peut varier d'une population à l'autre, en particulier entre des populations très éloignées l'une de l'autre et ayant peu de chances d'échanger de l'ADN par l'accouplement. Enfin, de nouvelles variations dans la séquence d'ADN (les mutations) ont contribué à la création d'haplotypes. La plupart de ces nouveaux haplotypes n'ont pas eu le temps de se propager au-delà de la population et de la région géographique où elles ont apparu.

L'information contenue dans les Haplotypes HLA est indispensable à l'analyse de données génétiques, permettant par exemple la localisation géographique de certaines maladies (Risch and Merikangas 1996) ou encore l'interprétation de fragments d'extraction d'ADN (Wang, Kidd, Zhao 2003). L'estimation de probabilités haplotypiques est une question importante à la fois pour les études de génétique des populations (Single, Merger et al. 2002) mais également pour la détermination de gènes responsables de certaines maladies (Niu, Qin et al. 2002). On constate par exemple que l'estimation de fréquences haplotypiques utilisée dans les études prenant en compte plusieurs locus permet de révéler certaines associations entre marqueurs et maladies, que l'analyse uni-locus ne peut identifier. Les méthodes classiques de typage ne donnent pas d'information sur la phase (chromosome). Celle-ci peut néanmoins être obtenue grâce au typage de membres d'une même famille (Duldbridge, Kolleman et al. 2000).¹⁰.

Cet article pose le problème de l'estimation de la distribution jointe des Haplotypes connaissant les Phénotypes (variables observables) et propose une nouvelle méthode simple d'estimation. Ces variables sont associées à des positions (loci) sur un chromosome et correspondent à des gènes. On s'intéresse en particulier à la dépendance entre ces variables (déséquilibre de liaison) et à la loi conditionnelle des gènes du système HLA qui définissent l'histocompatibilité étant donné un ensemble de microsatellites. La difficulté du

¹⁰Si l'on ne possède aucune information concernant les membres de la famille d'un malade, on a recours à une méthode statistique. A partir des observations des phénotypes, on estime la distribution des Haplotypes et la connaissance conjointe de cette distribution et du phénotype d'un individu permet de reconstituer la phase

problème tient à la dimension de l'espace paramétrique (on s'intéresse à une dizaine de variables pouvant prendre une vingtaine de modalités) et au mécanisme d'observation. Pour chaque individu, on observe deux réalisations de chaque variable relatives aux deux chromosomes mais l'observation sur la même phase (chromosome) n'est pas disponible. On appelle phénotype cette séquence de couple d'observations. Les deux méthodes les plus courantes sont d'une part la méthode du maximum de vraisemblance, mis en oeuvre via l'EM algorithm (Excoffier and Slatkin, 1995), et une méthode plus parcimonieuse proposée par Clark (1990). Une troisième méthode a été proposée par Stephens, Smith and Donnelly (2001). Leur méthode de reconstruction de la phase (de type Bayésien) utilise le Gibb's sampling, une forme d'algorithme MCMC. Nous proposons une nouvelle méthode statistique, basée sur l'équilibre d'Hardy-Weinberg, permettant de reconstruire la phase du phénotype d'un individu. Il ne s'agit ici ni d'un algorithme EM, ni d'une approche MCMC de type bayésien mais d'une méthode de moments, basée sur l'adéquation entre moments théoriques et moments empiriques, permettant l'estimation de la distribution des Haplotypes à l'aide des phénotypes. Même si l'estimateur de cette méthode de moments ne possède pas les propriétés asymptotiques du maximum de vraisemblance, il a pour principale caractéristique d'être très simple à calculer. Il ne dépend pas d'une règle d'arrêt et les simulations réalisées montrent que lorsque l'échantillon est petit les résultats sont meilleurs que ceux obtenus avec le maximum de vraisemblance. La présence de grandeurs non observables entraîne que la distribution des observables est un mélange de probabilités: la technique habituelle d'estimation repose sur l'EM algorithm et plus rarement sur les méthodes MCMC. Ces techniques sont très générales et n'exploitent pas toute la structure du modèle. Ce chapitre propose une méthode différente fondée sur des moments permettant d'obtenir une estimation qui ne soit pas la limite d'un algorithme. On étudie les propriétés de cette méthode et l'on compare par simulations son efficacité à celle des méthodes déjà utilisées. Cet estimateur n'est donc pas construit comme la limite d'un algorithme récursif (dépendant de conditions initiales et d'une règle d'arrêt) mais est immédiatement calculable. Cet estimateur est convergent et asymptotiquement normal et bien qu'il ne possède pas toutes les propriétés asymptotiques du maximum de

vraisemblance, on montre que les résultats obtenus peuvent être meilleurs notamment s'il s'agit d'un petit échantillon. Grâce ce à la rapidité de calcul de l'estimateur proposé, il est facile de faire une analyse en termes de Bootstrap de sa distribution. On a testé l'efficacité de notre méthode sur une analyse empirique du déséquilibre de liaison entre MOGc et le gène HLA-A. Deux extensions de cet article sont en cours: la première étant l'extension de cette méthode d'estimation à un nombre toujours plus important d'allèles et de locus (optimisation de la méthode de comptage), la seconde étant l'étude du cas où les probabilités jointes des allèles sont égales à 0 (strictement). On propose ensuite une présentation intuitive de l'algorithme d'estimation en s'appuyant sur une simulation réalisée avec 5 loci. La construction et la conception informatique de l'algorithme ont été réalisées avec STATA. Des travaux d'amélioration et d'optimisation des procédures de dénombrement et de calcul sont actuellement en cours. La dernière partie du chapitre suggère un petit modèle statistique qui montre le lien de proportionnalité existant entre le nombre de phénotypes HLA (différents) dans une population et le logarithme de la taille de cette population.

Chapitre 1

Matching Models and Optimal Registry for Voluntary Organ Donation Registries

1.1 Introduction

The purpose of this paper is to perform an evaluation of a mechanism for the constitution of organ donor registries. The aim is to increase the adequacy between the file of potential volunteer bone marrow donors and the needs of patients. Alive voluntary organ donors are needed in order to treat some diseases and to perform grafts. These organs are taken just in time before transplantation. It's impossible to preserve them. One example is typically the bone marrow transplantation (CSH) aimed to treat blood diseases especially leukemias and immune-deficiencies diseases. A graft from a donor to a patient is possible only a compatibility condition. This condition is, in theory, the identity of the HLA system where the HLA (Human Leucocytes Antigens) is characterized by a double sequence of alleles of a set of genes on the pair of the 6th chromosome; A simplified view of the HLA consists by considering only three genes, A, B and DR and the type of an individual is described for example by (1,2) (2,44) (3,4) which means that the pair of gene A is 1 and 2, of gene B 2 and 44 and of gene C is 3 and 4. Each pair is ordered because we cannot observe on which chromosome the alleles are. It should be underline that, contrarily to many problems in economics, individuals do not know their own type. Moreover the typing of an individual has a substantial cost. As the number of possible alleles for each gene varies between 10

and 40, the number of theoretical possible HLA is huge (several millions). The number of possible types and their inequal distribution implies that a large number of potential volunteer bone marrow donor is needed. The French Registry contains approximatively 120 000 donors and is interconnected with the worldwide file. The total number of potential donors in the world is more than 6 millions but less than half of patients find a compatible donor ¹. Moreover it has been empirically verified that the efficiency of the registries (in term of proportion of patients who find a donor) increases very slowly when the size of the registry increases (Oudshoorn, 1997). The aim of this paper is to formalize the problem of how to improve the organisation of this kind of registries. Basic definitions are given in section 2. A theoretical concept of optimal registry is conducted in section 3. Then we consider implementable improvement of the registry based on a filtering system where a low cost information associated to the type may be obtained. The optimal use of an information and the selection of optimal information are theoretically computable. The problem is practically untractable due to its dimensionality. We propose in section 4 a feasible strategy based on simulations.

1.2 Donors, Receivers and Registry

Each individual of the population is characterized by a type j belonging to a finite set $1, \dots, j$. For example the type is the HLA phenotype, i.e. the list of the alleles on the two chromosomes of loci A,B and DR registered with a given precision ("two digits" or "four digits" in the HLA case). The receivers are drawn in the population and the frequencies of types for receivers is described by a probability vector p_j ($p_j > 0 \sum_{j=1}^J p_j = 1$). By construction of the list of the types all the p_j are strictly positive. A registry is a list of donors recorded by their identity and their type. Two conditions are necessary in order to do a transplant to a given receiver :

- The existence in the registry of donors of individual of the same type. We assume that only perfect matching transplants are realized and we never introduce an idea

¹We eliminate patients for who a family donor may be found

of distance or almost compatibility between donors and receivers.

- At least one compatible donor should accept the transplant (or should be available for the transplant). Multiple causes exist for non availability (pregnancy, professional requirements, illness...). We summarize this complex phenomena by assuming that a compatible donor "accept" the transplant with a probability a . Acceptation decisions for different donors are independent events.

A registry design is defined by two components :

- an initial registry : this initial registry is characterized by its size N_0 and by the number N_{0j} of donors of type j . This number may be equal to 0 for many types.
- an increment process defined by the number N of new donors introduced in the registry and by a sampling mechanism of the types described by a vector q_j ($q_j \geq 0$ $\sum_{j=1}^J q_j = 1$) of frequencies. For example if donors and receivers are drawn randomly in the same population $p_j = q_j$ and are equal to the frequency of type j in the population.

We should underline that individuals and registry management ignore the types. Typing is a complex operation only realized where a new donor is introduced in the registry. Then, in practice, the registry management cannot choose q_j . However we first imagine situation where $(q_j)_j$ can be selected arbitrarily by the registry management ("first best" approach) and we consider secondly implementable choices of q_j ("second" best approach). In allow analysis N is given. In practice N may be constrained by the arrival process of new potential donors and by the budget of the registry organization because introducing a new donor is costly.

We consider in this paper only a single period model : starting for an initial registry we consider its improvement by a unique increment and not by several increments through multiple periods. The mathematical tools to model and to solve this dynamic version used more technicalities are needed to solve this dynamic programming problem.

As an illustration consider the french registry of bone marrow donors. The types are defined by HLA haplotypes A,B,DR recorded in two digits. The current registry contains approximately 100 000 donors and an increment of 10000 by year is scheduled. If we want to analyze the one year mechanism N is 10 000 but we may also consider a long term variation ($N = 100\ 000$ or more). The number J of possible types is a crucial element of the analysis and will be discussed later on. More than 60 000 types are present in the registry. At the world level the interconnection between the national registries gives a total registry of more than 6 millions which also increases by several hundred of thousands people each year. More than 400 000 types have been observed. ² A registry system is defined by J , the p'_j 's, the N_0j' 's, a , N and the q'_j 's. Such a system may be evaluated.

We propose to evaluate a registry by the expected probability to not find a donor for a receiver. To illustrate this concept consider just the simple case where $N_0 = 0$ (no stock) and $a = 1$ (all compatible donors accept the transplant). For any receiver the non-realization of a transplant (all the donors have a type different of j is a random event which has a probability of $(1 - q_j)^N$ if the type of the receiver is j . As the type of the future receivers are not given when the registry is designed we consider the expectation of this probabilities through the different types :

$$L = \sum_{j=1}^J p_j (1 - q_j)^N$$

This quantity may be viewed as the evaluation of the registry system and $1-L$ is equal to the probability of any receiver to find a donor.

Remark 1: An alternative criteria for the evaluation of the registry design would be based on the expected waiting time of a patient. Let us assume that N new donors are drawn with a probabilities $(q_j)_{j=1,\dots,J}$. For a patient where type is j , the waiting time is 0 with probability $1 - (1 - q_j)^N$, 1 with a probability of $(1 - q_j)^N(1 - (1 - q_j)^N)$. In general

²The list of possible alleles on each locus A,B, DR is probably known and then defines a huge list of potential types. However most of associations don't exist and the number of sequences A,B, DR is smaller than the product of the number of alleles on each locus.

the waiting time is t with a probability of $(1 - q_j)^{Nt}(1 - (1 - q_j)^N)$. The expected waiting time of this Pascal distribution is equal to $\frac{1}{1 - (1 - q_j)^N}$. We average this expected time with respect to the patient's type and we get the following evaluation criteria:

$$L_1 = \sum_{j=1}^J p_j \frac{1}{1 - (1 - q_j)^N}$$

Remark 2: In the previous definitions we consider the cases where a patient does not find a donor only. We may also take into account cases where a donor is found. For example if A is the cost of not find a donor and B the value where a donor is present in the registry, the evaluation of the registry design is :

$$L_2 = \sum_{j=1}^J p_j (A(1 - q_j)^N - B(1 - (1 - q_j)^N)) + B \sum_{j=1}^J p_j$$

It may be very easily shown that optimal designs of registries based on L_1 or L_2 instead of L will give same results. Then we keep K as an evaluation criterium but we extend its evaluation to case where $N_0 \neq 0$ and $a \neq 1$.

Proposition 1: If N is large and the q_j are small the evaluation of a registry system is equal to :

$$L = \sum_{j=1}^J p_j (1 - a)^{N_0j} e^{-aNq_j}$$

Proof : let fix j the type of a receiver. The number of donors of this type is $N_0j + m_j$ where m_j is drawn by a Binomial distribution :

$$Prob(m_j) = C_N^{m_j} q_j^{m_j} (1 - q_j)^{N - m_j}$$

given j and m_j the probability of to not find a donor is $(1 - a)^{N_0j+m_j}$. Then given j only, the probability to not find a donor is

$$\sum_{m_j=0}^N (1 - a)^{N_0j+m_j} C_N^{m_j} q_j^{m_j} (1 - q_j)^{N-m_j}.$$

Then

$$\begin{aligned} L &= \sum_{j=1}^J p_j \sum_{m_j=0}^N (1 - a)^{N_0j+m_j} C_N^{m_j} q_j^{m_j} (1 - q_j)^{N-m_j} \\ &= \sum_{j=1}^J p_j (1 - a)^{N_0j} E[(1 - a)^{m_j}] \end{aligned}$$

where m_j is drawn by the binomial distribution. For large N and small q_j it is well known that this Binomial distribution is approximatively a Poisson distribution parametrized by $\lambda_j = Nq_j$. Moreover an elementary computation shows that $X \sim \mathfrak{S}(\lambda)$ implies $E(b^X) = e^{\lambda(b-1)}$. Then :

$$L = \sum_{j=1}^J p_j (1 - a)^{N_0j} e^{-aNq_j}$$

1.3 The theory of the Optimal Registry

In this section we assume that the institution which has in charge the management of the registry optimizes the registry design in order to minimize the evaluation criterium under a budget constraint. This institution has a fixed total budget B and it is assumed that the cost of the registry is linear (any donor costs b). The budget constraint reduces in that case to the elementary relation :

$$B = bN$$

for which N is determined and equal to B/b . A more sophisticated cost function $C(N)$ may be introduced and the constraint becomes $B = C(N)$ but in all cases the number of new donors N follows from the budget constraint and this number N is not a random element. The main element of this assumption is that the cost function of the treatment of donors

(which contains essentially the typing cost) cannot be influenced by the institution which manages the registry. This assumption may be false in practice : registry management and typing laboratories are often both controlled by the public health administration which may be modified the cost function. In this section we only consider the case where N is determined by the budget constraint.

The problem then reduces to minimize the evaluation criterium with respect to the drawing design of the donors $(q_j)_{j=1,\dots,J}$. This analysis has only a theoretical objective because the result will require to be implemented the knowledge of the types. The result has however an interest as a reference theoretical optimal registry.

Equivalently the problem is to minimize

$$L = \sum_{j=1}^J p_j (1-a)^{N_0 j} e^{-aNq_j}$$

with respect to the q_j 's under the constraints:

$$\sum_{j=1}^J q_j = 1 \text{ and } q_j \geq 0 \forall j = 1, \dots, J.$$

Let us denote by \tilde{q}_j and \tilde{L} the solution of this problem.

This optimization problem has no solution in a closer form and should be performed numerically. Theoretically this optimisation is not difficult even if the objective is a non linear function of q_j 's. The constraint are linear : one an exact constraint and J are inequality constraints. However the problem is almost untractable in practice because the dimension J of the q vector is extremely large. We will show later on some example of this computation in models with "small" J (1 around 1000).

One can remark that the minimization of L under the equality constraint has an elegant solution which provides a bound of the efficiency the registry.

Proposition 2 : The minimum of L with respect to the q_j under $\sum_{j=1}^J q_j = 1$ is reached

for

$$q_j^0 = \frac{1}{J} + \frac{1}{aN} \left\{ \ln p_j - \frac{1}{J} \sum_{\ell=1}^J \ln p_\ell \right\} + \frac{\ln(1-a)}{aN} \left\{ N_{0j} - \frac{N_0}{J} \right\}$$

and the optimal value of L is equal to

$$L^0 = J\bar{p}(1-a)^{\frac{N_0}{J}} e^{-\frac{aN}{J}}$$

where $\ln \bar{p} = \frac{1}{J} \sum_{j=1}^J \ln p_j$

Proof: We replace q_j by $1 - \sum_{j=1}^{J-1} q_j = 1q_j$ and we compute the first order condition of the minimization:

$$\frac{\partial L}{\partial q_j} = p_j(1-a)^{N_{0j}} aN e^{-aNq_j} + p_J(1-a)^{N_{0J}} aN e^{-aNq_J} = 0 \quad \forall j = 1, \dots, J-1.$$

Then:

$$p_j(1-a)^{N_{0j}} e^{-aNq_j} = \text{constant and } \sum_{j=1}^J q_j = 1$$

$$\Rightarrow q_j^0 = \frac{1}{aN} \{ \ln p_j + N_{0j} \ln(1-a) \} + C$$

Using $\sum_{j=1}^J q_j = 1$ we get

$$C = \frac{1}{J} - \frac{1}{aN} \left\{ \frac{1}{J} \sum_{\ell=1}^J \ln p_\ell + \frac{N_0}{J} \ln(1-a) \right\}$$

from which the q_j^0 's are divided.

The solution of the first order conditions is a minimum because the function L is convex as a function defined on the q_j 's.

The value of L^0 is immediately obtained by replacing q_j by q_j^0 . ■

These results are easy to interpret and require several comments.

1. The value L^0 is obtained by relaxing some constraint satisfied by \tilde{L} . As a consequence :

$$L^0 \leq \tilde{L}$$

or

$$1 - L^0 \geq 1 - \tilde{L}$$

The value $1 - L^0$ then gives an upper bound to the probability to find a donor and can be view as an (optimistic) measurement of the maximal efficiency of a registry.

2. This upper bound $1 - L^0$ depends in a few number of characteristics of the registry system: It depends only:

- on the sizes of the initial and the incremental registries N_0 and N .
- on the probability a to of a donor to be available for a transplant
- on the number of types J
- on a characteristic of the dispersion of the distribution of the types in the receiver's population. This characteristic is the geometrical mean \bar{p} .

More precisely the key elements of the efficiency of the registry are $J\bar{p}$ (ratio the geometrical mean) \bar{p} and of the arithmetic mean $1/J$), relative size of the registries relatively to J ($\frac{N_0}{J}$ and $\frac{N}{J}$) and a .

3. The q_j^0 depends on three components:

- The uniform distribution $\frac{1}{J}$
- A measure of the importance of p_j with respect to \bar{p} : types j for which p_j is greater than \bar{p} should have q_j^0 greater than $\frac{1}{J}$.

- A measure of the importance of N_{0j} relative to the mean size of the registry $\frac{N_0}{J}$. This measure is weighted by $\ln(1-a)$ which is negative. Then over represented types in the initial registry have a q_j^0 inferior to $\frac{1}{J}$.
4. If q_j is taken at the optimal value q_j^0 the joint probability for any patient to have a type j and to find a donor, $p_j(1-a)^{N_0} e^{-a N q_j^0}$, is constant for any type and equal to $\bar{p} e^{-a \frac{N}{J}} (1-a)^{\frac{N_0}{J}}$. This optimal selection preserves the equity between individual with respect to joint event. However the probability to find a donor given the type is not constant and equal to $\frac{1}{p_j} (1-a)^{\frac{N_0}{J}} e^{-a \frac{N}{J}}$
5. The minimum of $\sum p_j e^{-a_j N q_j}$ under the constraint $\sum_{j=1}^J q_j = 1$ is reached if

$$q_j = \frac{1}{\sum_{l=1}^L \frac{1}{a_l}} + \frac{1}{a_j N} \left(\ln a_j p_j - \frac{\sum_{l=1}^J \frac{\ln a_l p_l}{a_l}}{\sum_{l=1}^J \frac{1}{a_l}} \right)$$

and the value at the maximum is equal to

$$11 - \tilde{J} \tilde{p} \exp^{-\frac{N}{\tilde{J}}}$$

where

$$\tilde{J} = \sum_{l=1}^J \frac{1}{a_l}$$

and

$$\ln \tilde{p} = \frac{\sum_{l=1}^J \frac{\ln a_l p_l}{a_l}}{\sum_{l=1}^J \frac{1}{a_l}}.$$

Unfortunately direct application of this formulae may lead to negative values if J is large and N relatively small.

For large N the optimal q_j^0 converges to $\frac{1}{J}$ (the uniform distribution) and are then positive. However we will see in the next section that the value of N for which $\frac{1}{J}$ may be accepted is extremely large if J is also large.

1.4 Continuous types models

What is the maximal gain of optimal selection of donors ?

In the previous section we have proof that the probability to find a donor for any receiver is equal to:

$$\pi = 1 - \sum_{j=1}^J p_j \exp(-a N p_j)$$

if donors arrive with the same frequencies as the receivers. Remember that the types are $1, \dots, J$ with frequencies p_j , N is the size of the registry (for simplicity we don't consider case with initial stock) and a is the probability for a compatible donor to be fully compatible (see chapter 2). In case of optimal selection of donors we have also proof that the upper bound of π is:

$$\pi_0 = 1 - J\bar{p} \exp\left(-\frac{aN}{J}\right)$$

where \bar{p} is the geometrical mean of the p_j 's. Our objective is to compute π_0 and π for realistic scenarios for a large population. The computation of π_0 is easy because it depends on few parameters only: the size of the registry, the number of types and a characteristic of the heterogeneity of the frequencies, $J\bar{p} \in [0, 1]$. This value is related to the Entropy divergence between the $(p_j)_j$ and the uniform distribution. The problem is to compute π which depends on all the p_j , typically unknown for a large population. The model we suggest consists to replace the p_j by a parametric family (typically a single parametric model, the parameter of which calibrated in order to get $J\bar{p}$ at a given level) and to compute π . This objective is difficult to reach in a discrete case but we will simplify this model by considering continuous approximations of the discrete distributions.

Continuous types models

We assume that the types are elements of an interval $[0, J]$ and that the distribution of the types in the population is a density probability $p(x)$. The donors are drawn from a density $q(x)$. The probability to find a donor is then:

$$1 - \int_0^J \exp(-aNq(x) p(x)) dx$$

If the donors arrive from the population without any selection this probability is:

$$\pi = 1 - \int_0^J \exp(-aNp(x)) p(x) dx$$

An elementary extension to continuous distribution of the argument presented in chapter 1 gives us the minimum of

$$\int_0^J \exp(-aNq(x)) p(x) dx$$

w.r.t. q and under the constraint

$$\int_0^J q(x) dx = 1$$

. Indeed, the functional first order conditions are:

$$-aN \int_0^J \tilde{q}(x) \exp(-aNq(x)) p(x) dx - \rho \int_0^J \tilde{q}(x) dx = 0$$

for any $\tilde{q}(x)$ where ρ is a Lagrange multiplier. Then:

$$\exp(-aNq(x)) p(x) - \rho = 0$$

or

$$q(x) = \frac{\ln p(x)}{aN} + C$$

and

$$\int_0^J q(x) dx = 1 = \frac{1}{aN} \int_0^J \ln p(x) dx + CJ.$$

Then

$$C = \frac{1}{J} - \frac{1}{aN} \left(\frac{1}{J} \int_0^J \ln p(x) dx \right).$$

We may conclude that:

$$\pi_0 = 1 - J \exp\left(\frac{1}{J} \int_0^J \ln p(x) dx\right) \exp\left(-\frac{aN}{J}\right)$$

This expression is analogous to the one obtained with discrete types.

Specification of a parametric family for $p(x)$

We assume that $p(x)$ is a truncated exponential distribution:

$$p(x) = \frac{\lambda \exp(-\lambda x)}{1 - \exp(-\lambda J)} \mathbb{I}(x \leq J).$$

In order to calibrate this distribution we compute

$$J \exp\left(\frac{1}{J} \int_0^J \ln p(x) dx\right).$$

Elementary computation shows that

$$J \exp\left(\frac{1}{J} \int_0^J \ln p(x) dx\right) = \frac{\lambda \exp(-\lambda x)}{1 - \exp(-\lambda J)} \exp\left(\frac{-J\lambda}{2}\right)$$

which only depends on $\alpha = \lambda J$. For example if $\alpha = 3$ this expression is equal to 0.7 and $\lambda = \frac{3}{500\,000}$ if $J = 500\,000$.

We may now compute $1 - \pi$ without selection of donors where an also elementary integral computation gives:

$$1 - \pi = \frac{1}{\alpha \frac{aN}{J}} \left[\exp\left(-\frac{aN}{J} \frac{\alpha \exp(-\alpha)}{1 - \exp(-\alpha)}\right) - \exp\left(-\frac{aN}{J} \frac{\alpha}{(1 - \exp(-\alpha))}\right) \right].$$

For example, if $a = \frac{1}{3}$, $N = 200\,000$, $J = 500\,000$, $\alpha = 3$ and $J \exp\left(\frac{1}{J} \int_0^J \ln p\right) = 0.7$, we get $\pi_0 = 0.39$ and $\pi = 0.19$. The maximal gain for optimal donor selection is then 0.2.

More numerical results

We have computed π and π_0 for different values of N, J and $J\bar{p}$ using this parametric specification (see Tables 1 and 2). The main result is that: the difference between π and π_0 decreases if $J\bar{p}$ increases if Q increases or if J increases. The relevance of selection of donor is then specifically important if the registry is small, the heterogeneity in the frequencies large and if the number of types is high.

Probability to find a compatible donor

without (π) or with donor selection function(π_0)

if the size of the registry (N) and if the heterogeneity of the frequencies ($J\bar{p}$)

Table 1: Number of types = 500 000

$J\bar{p}$ Q	0.9		0.8		0.7		0.6		0.5	
	π	π_0	π	π_0	π	π_0	π	π_0	π	π_0
100 000	0.08	0.16	0.09	0.25	0.10	0.35	0.12	0.44	0.14	0.53
200 000	0.15	0.21	0.17	0.30	0.19	0.39	0.22	0.47	0.25	0.56
300 000	0.21	0.26	0.24	0.35	0.27	0.43	0.31	0.51	0.34	0.59
500 000	0.32	0.36	0.36	0.43	0.40	0.50	0.44	0.57	0.48	0.64
1 000 000	0.53	0.54	0.57	0.59	0.61	0.64	0.65	0.69	0.69	0.74

Table 2: Number of types = 700 000

$J\bar{p}$ Q	0.9		0.8		0.7		0.6		0.5	
	π	π_0	π	π_0	π	π_0	π	π_0	π	π_0
100 000	0.06	0.14	0.07	0.24	0.08	0.33	0.09	0.43	0.10	0.52
200 000	0.11	0.18	0.13	0.27	0.14	0.36	0.16	0.45	0.19	0.55
300 000	0.16	0.22	0.18	0.31	0.20	0.39	0.23	0.48	0.26	0.57
500 000	0.24	0.29	0.28	0.37	0.31	0.45	0.35	0.53	0.39	0.61
1 000 000	0.42	0.44	0.47	0.50	0.51	0.57	0.55	0.63	0.59	0.69

1.5 Optimal Registry: Some simulations

In order to shed light on the previous mathematical results and to get more intuition on their contain we have done several simulations based on specific models of the Registry system. These simulations are done using the following principles: we consider a list of types and their frequencies, a value of a , a size N_0 of the initial registry and of values of N_{0j} . Finally different sizes N of the incremental registry are considered. Essentially the objective is to compare the value L if donors arrive at the frequencies p_j (no selection), "optimistic" optimal L^0 and in some cases real optimal value of the registry \tilde{L} .

The different models considered are all based on the French Registry (France Greffe de Moelle Registry) hereafter FGM). In order to calibrate our simulation we should underline the following result. In 2003, the registry had 120 937 donors³ and 62 French receivers under 813 found an available donor. The current value of our criteria L is then equal to 0.92.

We have exact three examples of a list of types from FGM and we then have three values of J : 1156 , 4621 and 62 220. The last one corresponds to the list of observed type in France provided with their frequencies. In our simulation we have assume that this list represents the whole list of types. This is false in reality but this defines the model. The sample of 4621 was the "MADO sample" used in other studies and the first one is 1/4 of the MADO sample. If $a = \frac{1}{3}$, conditions of the simulations are summarized in Table 0, where N represents the increment of the registry ($N = 10\%, 50\%, 100\%, 200\%$) and N_0 the initial stock. N_0 is the exact value observed in the first FGM database used ($N_0 = \sum_{j=1}^J N_{0j}$). The total number of individuals was exactly equal to 95 934. Models 1 and 2 are subsample of Model 3 and we respect the order of magnitude of the ratio $\frac{N_0}{J}$.

³Data are given on the web site of FGM, <http://www.fgm.fr>

Table 0

Model	J	N_0	N			
1	1 156	255	25	127	255	510
2	4 621	1 000	100	500	1 000	2 000
3	62 220	22 195	25 000			

For the first model the "small" size of the number of types allows a numerical computation of the \tilde{q}_j and of \tilde{L}^4

Results for model 1 are summarized in Table 1 and graphs of the \hat{q}_j are given in figure 1.1 and 1.2.

Table 1

Expected probability to not find a donor

N	Selection	L_0	L
25	0,72	0,63	0,71
127	0,69	0,61	0,6664
255	0,65	0,589	0,6294
510	0,59	0,547	?

⁴Computation has been done using Matlab...

For the model 2, see Table 2, (823 types), we have to compare L with no selection, L_0 and L with some particular non optimal q_j (for instance frequent types in the initial registry have been eliminated with three definitions of frequent : more than 3, 4 or 5 times in the initial registry. We have also eliminated frequent and rare types).

Table 2
Expected probability to not find a donor

Elimination based on the initial registry						
N	N_0 Selection	Optimal	$n_{0j} \geq 3$	$n_{0j} \geq 4$	$n_{0j} \geq 5$	$n_{0j} \geq 5$ et $N_{0j} = 0$
			0,1263	0,0826	0,041	0,553
0	0,75	x	x	x	x	x
100	0,74	0,641	0,737	0,736	0,735	0,731
500	0,7	0,623	0,706	0,703	0,697	0,687
1 000	0,666	0,601	0,67	0,6666	0,657	0,648
1 500	0,634	0,58	0,638	0,633	0,622	0,619
2 000	0,606	0,559	0,609	0,604	0,592	0,597

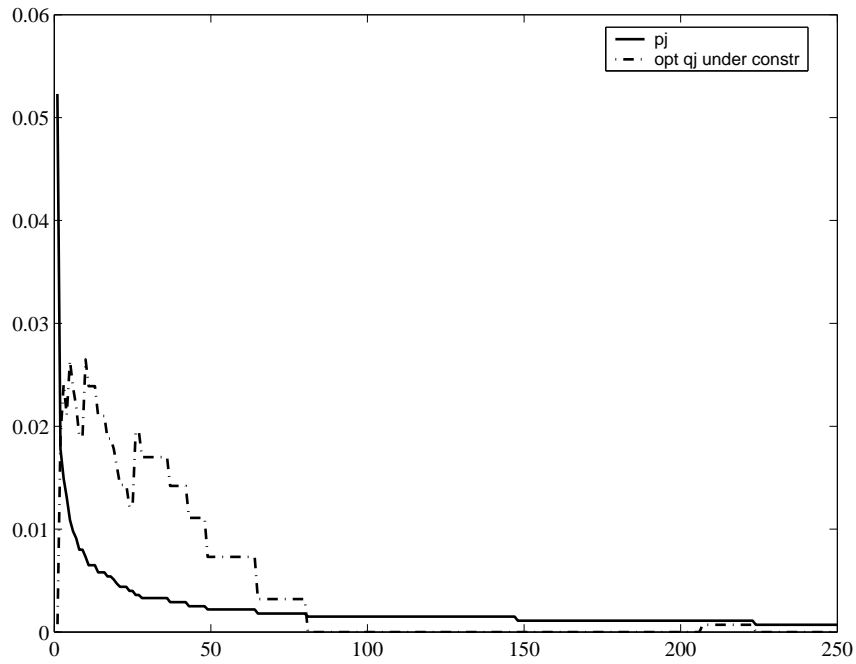


Figure 1.1: Increment of the Registry of 50%

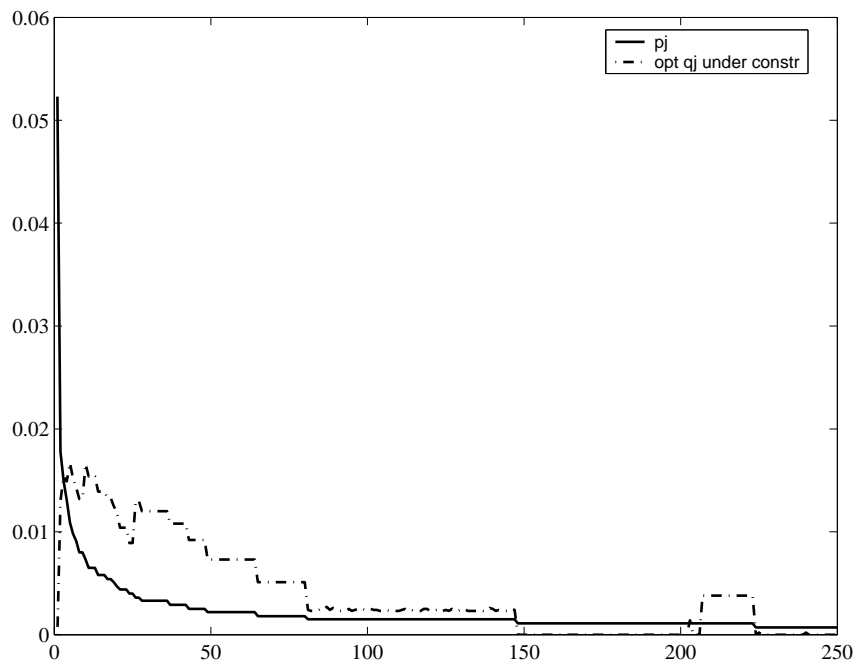


Figure 1.2: Double of the size of the Registry

Finally we have simulated in model 3 (62 210 types) an increment of 25 000 donors to an initial registry of 22 195.

The value of L for the initial registry was 0.71 and becomes 0.61 after an optimal (q_j^u) selection of donors.

The main conclusions of these simulations are the following:

1. An important increment of the donors files (multiplication by 2 or 3 the number of donors) has a relatively low impact on the probability to find a donor. Roughly speaking if the size is double this probability may increase of 10% approximatively.
2. The impact of the selection mechanism is also very low. In the model 1 the efficiency difference between N_0 selection and optimal (implementable) selection is only 2 % of the size of the registry is multiplied by 2.
3. The optimal selection rule eliminates a very few number of very frequents types present in the initial registry but essentially eliminates numerous rare types.

1.6 Filter and implementable improvement of a Registry

In the last section we have determined the optimal registry design but we have also shown that this optimal design is not implementable. We now consider implementable procedure based on presignaling or filtering which may be used in order to improve the registry. As in the previous section we start with an initial registry of N_0 donors and N_{oj} is the number of donors of type j . The total budget of the institution managing the registry is B and the cost of typing and of introducing a new donor in the registry is b .

A test is define by the following elements:

- it's cost c , typically smaller than b and the cost $b_1 (\leq b)$ of introducing in the registry somebody who has been tested.
- a list of possible results of the test $\{1, \dots, S\}$ and s is a possible result of the test.
- a joint distribution which represents the frequency in the population of potential donors of the test results and of the type j . This distribution is described by the probabilities

$$d(s, j) \quad s \in \{1, \dots, S\} \quad j \in \{1, \dots, J\}$$

$$d(s, j) \geq 0 \quad \sum_{s, j} d(s, j) = 1$$

This probability is assumed given. If the populations of donors and of patients are identical then the marginal distribution on j deduced from d , i.e. $d(s, j) = \sum_{s=1}^S d(s, j)$ should be equal to p_j .

In case of HLA typing we have in mind two examples of tests. The first one is the observation of microsatellites on the two chromosomes. Micro satellites are characterized by loci between the genes on the 6th chromosome, more easy to observed then the HLA genes but strongly correlated to the HLA types. The information contained by microsatellites

is "random" in the sense that the association between s and j is not perfect. In particular the conditional distribution $d(j|s)$ has in general a support not reduced to a singleton.

The other example we have in mind is the observation of SNP. Remember that a gene is characterised by a sequence of nucleotid (A,G,C or T) and it is possible to observe a subsequence only for each gene of the HLA system. The information contained by a given sequence of SNP is the exact knowledge of a partition of the list of types. In that case the conditional distribution of s given j has a support reduced to a singleton. However this is not true for the conditional distribution of j given s (which is the conditional distribution of the types given a subset). We will illustrate by examples below this two situations which appear to be particular cases of our general model.

Moreover let us remark that in this two cases, testing procedure requires the DNA extraction which is also the first step of typing. Then the testing step reduces the cost of typing and this explain the change of b into b_1 .

A test procedure rises two questions : First how to use in the more efficient way the informations provided by the test ? Second is it efficient to implement the test in comparison with the random arrival of donors? More generally if several tests are available (no test is a particular case of these different choices) what is the best test in order to improve the registry?

Consider first the optimal use of the test information. A strategy is defined by three elements :

- A number N_1 of individuals who are directly introduced in the registry (i.e. typed without test)
- A number of individuals M who are tested.
- A sequence of probability $\sigma(s) \in [0, 1]$ for each possible value of the test result. The number $\sigma(s)$ represents the proportion or the probability of an individual with test result equal to s to be introduced in the registry and fully typed.

As the test as a cost, it is not obvious that all the potential donors should be tested before typing. It is in general optimal (if $b - b_1 < c$) to directly type a group of people and to use the test strategy in order to correct the natural arrival process of the different types. An advantage of this presentation is that the no test case is contained in our pre-selection ($M = 0$).

This specification contains pure strategies ($\sigma(s) = 0$ or 1) for which a potential donor is typed or not depending on the result of the test. In other way, a pure strategy is equivalent to a partition of the set $\{1, \dots, S\}$ in two subsets \mathcal{S}_0 and \mathcal{S}_1 . In that case if $s \in \mathcal{S}_0$ the potential donor is fully typed and introduced in the registry and if $s \in \mathcal{S}_1$ the process stops. As usual in decision theory it is powerful to consider mixed or random strategies where $\sigma(s)$ may be any element in $[0, 1]$. If s is observed the registry manager drawn between "typing" or "stop" with probabilities $\sigma(s)$ and $1 - \sigma(s)$.

Let us consider an initial registry ($N_{0j=1, \dots, J}$) and a test strategy defined by N_1 , M and the $\sigma(s)$. This strategy defined:

- A random size of the increment of the registry:

$$N = N_1 + \left(\sum_{s=1}^S \sigma(s) d(s) \right) M$$

where $d(s) = \sum_{j=1}^J d(s, j)$ is the marginal probability of type s and $\sum_{s=1}^S \sigma(s) d(s)$ is the probability to be types after a test.

- the frequency of type j deduced from the test strategy is equal to:

$$q_j = p_j \frac{N_1}{N} + \frac{\sum_{s=1}^S \sigma(s) d(s, j)}{\sum_{i=1}^S \sigma(s) d(s)} \left(\frac{N - N_1}{N} \right)$$

The first element of this sum corresponds to individuals introduced in the registry without pretest and the second correspond to the tested people.

The expected cost of a test strategy is equal to

$$bN_1 + cM + b_1 \left(\sum_{s=1}^S \sigma(s)d(s) \right) M$$

An optimal use of the information contain of a test is then obtained by minimizing

$$\sum_{j=1}^J p_j (1-a)^{N_{0j}} e^{-a\{p_j N_1 + (\sum_{s=1}^S \sigma(s)d(s,j))M\}}$$

with respect to N_1 , M and the $\sigma(s)$'s under the constraints

$$b_1 n N_1 + cM + b_1 \left(\sum_{s=1}^S \sigma(s)d(s) \right) M = B$$

$$N_1 \geq 0 \quad M \geq 0$$

$$\forall s = 1, \dots, S \quad 0 \leq \sigma(s) \leq 1.$$

This optimization problem has no solution in closer form and should be done numerically. This numerical problem is however almost impossible to solve in practice due the dimension of S . An other important question is the knowledge of the joint distribution d . In the case of a "random" test like micro satellites informations, d should be derived from an estimation procedure based on a sample of individuals for whom s and j are observed. Here also the dimension of J and S is so large that no possible sample may carry a sufficient information about the joint distribution.

1.7 A Monte Carlo evaluation of an implementable improvement of a registry

As we have remarked in the previous section, the derivation of an optimal strategy based on a test in order to improve a registry faces to the curse of dimensionality. In the example of bone marrow transplant, the number of HLA types is extremely large (more than 60 000 types have been observed in France and don't constitute an exhaustive list) and the number of possible signals (observation of several microsatellites) is certainly greater than the number of inhabitants of France. The list of possible values of j and s are unknown and the joint probability $d(0, j)$ is very often calibrated using very small samples respectively to the number of possible values of the couple (s, j) .

A realistic situation (corresponding to actual improvement problem of the french bone marrow registry) may be described by the following arguments.

- we have a current registry, typically large from which a list of types and an evaluation of their probabilities may be derived. We assume here that donors and receivers are drawn from the same population.
- In this registry only type is available and not the signal s . However a sample of individuals is drawn (usually from the registry) for whom the signal is observable.

In the bone marrow application, the current registry contains more than 110 000 individuals and the sample has a size below 5 000.

In that case the use of the sample to estimate the joint distribution $d(s, j)$ is impossible. However partial statistical analysis may be done. For example the HLA type is defined by three loci and the signal by 15 microsatellites. Partial analysis of dependence between a gene on one loci and a small (one to three) number of microsatellite is possible. We suggest the following procedure:

Step 1. The sample is usually not randomly generated from the registry, in particular in order to obtain some "rare" types. However it is necessary to construct a system of

weights of individuals in the sample in order to have in the sample the same shape of the p_j as o, reality. If a type is represented ℓ_j type 0.

The expected cost of a strategy is given by

$$bN_1 + cM + b_1\left(\sum_{s=1}^S \sigma(s)d(s)\right)M$$

when $d(s) = \sum_{j=1}^J d(s, j)$ is the marginal distribution on the tests' results. Indeed $\sum_{s=1}^S \sigma(s)d(s)$ represents the probability for any individual who is tested to be introduced in the registry. A strategy respects the budget constraint if

$$bN_1 + cM + b_1\left(\sum_{s=1}^S \sigma(s)d(s)\right)M = B$$

Let us consider a registry strategy associated o a test. It generates :

- a (random) size of the registry :

$$N = N_1 + \left(\sum_{s=1}^S \sigma(s)d(s, .)\right)M$$

An individual in the registry has a probability $\frac{N_1}{N}$ to have been directly typed and $1 - \frac{N_1}{N}$ to have been typed after a test.

- probabilities

$$q_j = d(., j) \frac{N_1}{N} + \frac{\sum_{s=1}^S \sigma(s)d(s, j)}{\sum_{s=1}^S \sigma(s)d(s)} \left(1 - \frac{N_1}{N}\right).$$

The first term of the sum corresponds to donors directly typed (where

$$d(., j) = \sum_s d(s, j)$$

is the probability of type j in the donor's population) and the second term corresponds to donors preliminary tested before typed. Indeed:

$$\begin{aligned}
P(\text{donor type } j | \text{tested and typed}) &= \frac{P(j \text{ and typed} | \text{tested})}{P(\text{typed} | \text{tested})} \\
&= \frac{\sum_s P(s | \text{tested}) P(\text{typed} | s, \text{tested}) P(j | s, \text{typed}, \text{tested})}{\sum_{s,j} P(s | \text{tested}) P(\text{typed} | s, \text{tested}) P(j | s, \text{tested and typed})} \\
&= \frac{\sum_s d(s, \cdot) \sigma(s) d(j | s)}{\sum_{s,j} d(s, \cdot) \sigma(s) d(j | s)} \\
&= \frac{\sum_s d(s) \sigma(s) d(s, j)}{\sum_s \sigma(s) d(s)}
\end{aligned}$$

We have used the fact that the probability of j given s is not dependant of the decision to type or not. Finally note that the q_j constructed in that way are implementable. let us consider one of the evaluation criteria described in section 2 and 3. For example we may analyzed the approximated criterium L_3 depending on N and the q_j . Now, both N and the q_j 's depend on N_1 , M and $\sigma(s)$ and we should minimize this evaluation with respect to N_1 , M and $\sigma(s)$ under the budget constraint and $0 \leq \sigma(s) \leq 1$. In this presentation we have consider that the number of potential donors is unlimited. This assumption may be relaxed by imposing a supplementary inequality constraint ($N_1 + M$ should be smaller or equal to the number of volunteers). The number of volunteers is exogenous (to the decision problem) but may be modified by advertising campaign. This analysis may be generalized to the case where several tests may be used. Comparaison between tests may be done from the optimal value of evaluation criterium in each test. Unfortunately the optimisation program we have developped have no solution in closer form and should be solved numerically. It should be underline that both function to minimize and constraints are non linear and that the dimension of the problem is huge.

A test described by $d(s, j)$ and a strategy defined by α_s determines jointly an implementable probability q_j of drawing donors with type j . More precisely, an elementary computation shows that:

$$q_j = \sum_{s=1}^S \sigma(s) d(s, j)$$

Note that in that case $\sum q_j = \sum_{s=1}^S \sigma(s)d(s) < 1$. Given M , the number N_j of new donors of type j in the registry is generated by a binomial distribution parametrized by q_j and M . We can use the approximated evaluation criterium L'_3 :

$$L'_3 = \sum_{j=1}^J \alpha_j e^{-aM(\sum_{s=1}^S \sigma(s)d(s,j))}$$

and L'_3 should be minimized w.r.t. M and $\sigma(s)$ under the budget constraint and the inequalities

$$0 \leq \sigma(s) \leq 1 \quad \forall s = 1, \dots, S$$

The optimization of L'_3 w.r.t M requires that the number of potential donors is unlimited. A simplification of the problem is obtained by assuming M given. The organism in charge of the registry does not control the arrival of donors and any voluntary donors should be considered, at least at the test level.

In that case the problem of the selection of $\sigma(s)$ reduces to:

$$\min \sum_{j=1}^J \alpha_j e^{-aM(\sum_{s=1}^S \sigma(s)d(s,j))}$$

under $\sum_{s=1}^S \sigma(s)d(s) = \varepsilon$ and $0 \leq \sigma(s) \leq 1 \quad \forall s = 1, \dots, S$.

This problem has no analytical solution in general but usual operation research software may be used to solve this minimisation under constraints.

In order to illustrate the solution we consider two examples where we assume for simplicity that $N_0 = 0$ and $a = 1$. We also consider the first evaluation criteria L and the problem becomes:

$$\min \sum_{j=1}^J p_j \left(1 - \sum_{s=1}^S \sigma(s)d(s,j)\right)^N$$

under $\sum_{s=1}^S \sigma(s)d(s) = \varepsilon$ and $N = \varepsilon M$ and $a \leq \sigma(s) \leq 1$.

example 1: We consider an elementary case where there exists two types and two values of the test only. The joint probability d is given by $d(1, 1) = 0,08$, $d(1, 2) = 0,02$, $d(2, 1) = 0,09$ and $d(2, 2) = 0,81$. The patient are drawn in the same population ($p_1 = 0,17$ and $p_2 = 0,83$). Moreover $B = 1500$, $b = b_1 = 150$ and $c = 15$. Then without test, only 10 individuals may be types and L is equal to 0,026 in that case (which may be compare to the minimal value of L equal to 0,00071 obtained where $p_1 = 0,456$). An optimal use of the test consists to type all the potential donors for which $s = 1$ and $\frac{1}{3}$ of donors for which $s = 2$ ($\alpha_1 = 1$, $\alpha_2 = \frac{1}{3}$). In that case the value of L becomes 0,013.

example 2: Even if this example is not realistic (we only consider one chromosome, a single element of the HLA system (A) and a single microsatellite (MOGC)) it constitute a step in a direction of the application of this theory.

We consider a model with 24 possible types and 17 values for the signal and the joint distribution has been estimated and is given by table I. The original motivation of this research was to analyse the capacity of a set of microsatellites to predict groups of HLA types in the framework of optimizing HLA typing policies of Bone Marrow Donor Registries . In this paper we just present a preliminary step of this study concentrated on a single microsatellite MOGC and the A locus of the HLA system. We consider a sample (of size 2117)⁵ of phenotypes used for the estimation of the joint distribution of size 2117 of MOGc and A on a single chromosome.

The result of our estimation is given in table II where "0" denotes pairs of alleles of MOGc/A never observed. Probability values are rounded off.

The precision of this estimation result is analyzed by a non parametric bootstrap. From the original sample we have contracted 1 000 samples by random drawing with replacement. Each sample is used for a new estimation of the joint probability (Efron (1982),Hall (1999)).

We just illustrate the power of this analysis by two examples. We have constructed the

⁵This sample was randomly extracted from the France Greffe de Moelle Registry. In this data set missing data are reconstructed by answering homozygoty

bootstrap distribution of two measures of the linkage disequilibrium. The first one is the entropy measure defined by

$$I = \sum_{j,k} p(j, k) \ln \frac{p(j, k)}{p(j, \cdot) \cdot p(\cdot, k)}$$

where j is the index of possible alleles of MOGc, k is the index of possible alleles of A.

The estimated value of I (9.1) is 0,9701. The bootstrap mean is 0,9676 and a confidence interval of I at 95 % is [0,9215; 1,0173]. The distribution of I is given by the histogram in table V.

It is well know that entropy has some undesirable features and a better association measure is provided by Hellinger distance between the joint distribution and the product of its marginals, i.e.

$$H = \frac{1}{\sqrt{2}} \left[\sum_{j,k} (\sqrt{p_{jk}} - \sqrt{p_j \cdot p_k})^2 \right]^{\frac{1}{2}}$$

In particular, by construction, it is normalized in order to be between 0 and 1 where 0 is equivalent to independence. The actual estimated value of H is 0,4270. The bootstrap mean is 0,4271 and a confidence interval is [0,4033; 0,4520]. Histograms of bootstrap distribution of this, linkage disequilibrium measure is given in tables V and VI.

Table I

Joint Distribution of MOGC and HLA-A on a single chromosome

MOGCHLA-A	1	2	3	9	10	11	23	24	25	26	28	29	30	31	32	33	34	36	43	66	68	69	74	80	
121	0,076	0,003	0	0	0,001	0,003	0,015	0,016	0,035	0,001	0,052	0,002	0,004	0,002	0,010	0	0	0	0,002	0,014	0	0	0	0	0,24
123	0	0	0	0	0,001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
125	0	0	0	0	0	0	0	0	0	0	0	0	0,001	0	0,001	0	0	0	0	0	0	0	0	0	0,00
127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
129	0,005	0,016	0,117	0	0,002	0,022	0,022	0,022	0,005	0	0,005	0	0,001	0,014	0,003	0,001	0	0,001	0	0,001	0	0	0	0,001	0,19
131	0,005	0,169	0,003	0	0,001	0,004	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,013	0,001	0,001	0	0	0	0,012	0	0	0	0	0,21
133	0,002	0,002	0	0	0	0	0	0	0,001	0,011	0,016	0,001	0	0,016	0,001	0	0	0	0	0	0	0	0	0	0,03
135	0,002	0	0,002	0	0,051	0,020	0,001	0,001	0,001	0	0,001	0	0	0	0,001	0	0	0	0	0	0	0	0	0	0,08
137	0,003	0	0	0	0,002	0,001	0,024	0	0,001	0	0,016	0,004	0,001	0	0,016	0,004	0,001	0	0,010	0	0	0	0	0	0,06
139	0,001	0,001	0	0	0	0	0	0	0	0	0,002	0	0	0,002	0	0	0	0	0	0	0	0	0	0	0,00
141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
143	0,052	0,003	0	0	0,002	0,002	0,002	0,002	0	0,001	0,001	0	0,001	0	0,001	0	0	0	0	0	0	0	0	0	0,06
145	0,001	0,001	0	0	0,003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01
147	0,061	0,006	0,002	0	0,001	0	0,030	0	0	0	0	0	0	0,001	0	0	0	0	0,001	0	0	0	0	0	0,10
149	0	0	0	0	0,001	0,001	0,001	0,001	0	0	0	0	0	0	0	0	0	0	0	0,001	0,001	0	0	0	0,00
151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
153	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00

0,13 0,28 0,13 0,00 0,00 0,06 0,03 0,10 0,02 0,04 0,00 0,06 0,04 0,03 0,03 0,02 0,00 0,00 0,00 0,00 0,04 0,00 0,00 0,00 0,00

In section 2, we define L (our registry evaluation criteria) in order to evaluate a registry by the expected probability for a receiver to not find a donor. This quantity may be viewed as the evaluation of the registry system. To illustrate this concept, we evaluate different values of L in different cases.

Given the Agency Budget (15 000), the typing cost (150€), the cost of the test (15€), the table II aims at sum up results obtained if 200 new donors arrive. Knowing the sequence of probabilities $\sigma(s)$ for each possible value (denoted s , the μ_{sat}) of the test result, this example shows that it is possible to type 100 donors without test, against 80 with a test. The table II gives also the value of L without test which is equal to 0.0244. Without test, 100 individuals may be typed but when using a test based on a random strategie (see the strategie used for MAD0 when considering a single μ_{sat} MOGCc and the HLA-A). This value becomes then equal to 0.0228.

Table II

New donors	$\sigma(s)$	s
200	0.7022	121
Budget	0.5350	123
15000	0.5686	125
	0.5024	127
Typing Cost	0.4217	129
150	0.0000	131
	0.8204	133
Test Cost	0.6211	135
15	0.6615	137
	0.4908	139
Typings without TEST	0.0938	141
100	0.4327	143
	0.4652	145
Typings if test	0.0000	147
80	0.5047	149
	0.0945	151
Expected probability	0.0972	153
0.0244 (without test)		
0.0228 (with a test based on a random strategie)		

For three different values of the test (given in €), the Table III compares our registry evaluation criteria without and with a test. Without test, if budget is equal to 15 000€, 200 individuals may be typed and the value of L is equal to 0.0244. If budget is divided by two, we may type only 100 donors and the value of the registry would be worse (L=0.0692). If using a test, results depend on the cost of the test. For example, if the cost of the test is equal to 15€, it's more efficient in terms of efficiency of the registry (according to our criteria L) to type using a test. If the test is more expensive, our registry evaluation criteria is higher (remember we want to minimize it) and then it's not efficient to use the test.

Table III

	New donors 200 Budget 15000 Euros	New donors 100 Budget 7500 Euros
Without any test	0.0244	0.0692
With a test based on a random strategy		
Test cost 30 Euros	0.0341	0.0991
Test cost 15 Euros	0.0228	0.0643
Test cost 7.5 Euros	0.0191	0.0534

1.8 Efficiency of Bone Marrow Donors' Registries: Models and Orders of Magnitude

This section contains numerous simulations and computations derived from the results given in chapter 1. It is actually based on conferences given at the 6th (October 2004, Paris, France) and the 7th plenary MADO meetings (June 2005, Toulouse, France).

We provide a methodology for the evaluation of a donor's registry design based on the probability for a receiver to find a donor. We compare essentially registries based on random arrival of donors and registries where the selection of the donors is made by an optimal mechanism. Practicable implementations of such a mechanism by filtering processes are not discussed here.

The theoretical results exhibit the main elements of the maximal efficiency of a Registry:

Essentially the efficiency is determined by the size of the Registry, the probability for a donor to be available for a graft, the number of types in the population and the dispersion of the frequencies of types captured by the geometric mean of these frequencies.

The calibration of the model for different scenarii shows essentially the following result: a donor registry system of size corresponding to the actual ones in the main countries is not very efficient (approximately 10% of receivers find a donor) and this efficiency is difficult to increase. A huge increment of the registry (multiplication by 2 or 3) and an extremely efficient selection of the donors would lead in the best case to increase the efficiency to 20%.

An initial Registry generated without selection

We start with an initial registry generated without selection. This initial registry is "relatively large" (more than 120 000 donors). We compute two scenarios : "optimistic" (derived from MADO file), "pessimistic" (derived from FGM file). We define the scenario derived from MADO file as "optimistic" because this 5 000 individuals sample was constructed "ad-hoc".

Variation of the probability to find an (available) donor

Increment of the registry	No selection		Optimal selection	
	Pessimistic	Optimistic	Pessimistic	Optimistic
10 %	0.005	0.01	0.01	0.02
25 %	0.014	0.025	0.03	0.05
50 %	0.028	0.05	0.05	0.10
100 %	0.05	0.09	0.06	0.18

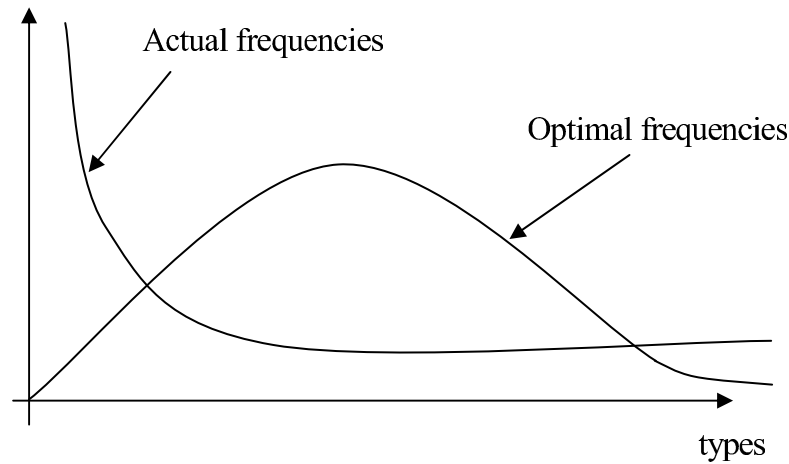
The numbers presented in the previous table are orders of magnitude valid for "any" size of the initial registry. If we had to interpret the results, let us say that, if with the initial registry the probability to find a donor is 0.25 and if the registry increases of 50 %, this probability becomes 0.30 without selection and 0.35 with an optimal selection (in the optimistic scenario). Our first conclusions are then:

1. the variation of the efficiency of the registry system changes slowly when the size of the registry increases
2. the gain of an optimal selection (perfect filter, no cost) is also low or almost negligible.

The optimal donor selection mechanism consists of:

- avoidance of a very few number of frequent phenotypes (0.04 %) (~ 30 phenotypes for 66 164 types)
- massive avoidance of rare phenotypes (more than 80 % of phenotypes representing 50 % of the population should be avoided).
- transformation of the frequencies of other types

The following figure shows the shape of "optimal frequencies" we would then design.



Impact of microsatellite filtering according to the predetermined rules

The designed rules that have been used for prefiltering for frequent HLA using μsat are worse than the "no selection" mechanism and then should not be implemented (whatever the cost is). This is due to the *selection rule* which does not eliminate the rare phenotypes, not to the absence of predictability power of μsat . We note that other rules of prediction may be constructed from the (small) available sample but are difficult to test for their robustness. Whereas in some cases, μsat filtering seems to be useful, an optimal use of μsat information is extremely hard to compute and difficult to implement. Due to the small difference between "actual" and "optimal" efficiency and to the fact that filtering leads to an efficiency not better than "optimal", *the use of prefiltering* is questionable in any case.

Role of the probability for a compatible donor to be actually available for a graft when contacted

It is probably more important to increase this probability than to increment the size and to improve the selection rule (the key element is the product of this probability by the size of the increment of the file).

1.8.1 The reference model

The Conditions for our model are the following ones: let us consider Phenotypes HLA A, B, DR, a type belonging to a finite set $\{1, \dots, J\}$ ($J \geq 2$), the probability for a recipient to have the type J , denoted p_1, \dots, p_j , an initial Registry with N_0 donors and N_{0j} donors of type j (may be equal to 0), the probability for a donor to be available for a graft $a \in [0, 1]$ (see chapter 2 for comments about this parameter "a").

If we increment the registry of N donors with frequency of donors types equal to q_1, \dots, q_j , two distributions for frequencies should be retained: a "First best" approach (any scenario available for the q_j) and a "second best" approach (filtering of donors (μsat , SNP,...) and restrictions to q_j realistic by filtering).

In order to simplify, we compute results for a one period model. An extension would be to set a dynamic model.

The measure of efficiency of the registry is the probability for any recipient to find a donor : $1 - \pi$. The aim of the model is to minimize π (probability of "not finding" a donor). The result is then

$$\pi = \sum_{j=1}^J p_j (1 - a)^{N_{0j}} e^{-aNq_j}$$

Given N , minimizing π with respect to the q_j under constraint $\sum q_j = 1$ is equal to:

$$q_j^0 = \frac{1}{J} + \frac{1}{aN} \left\{ \ln p_j - \frac{1}{J} \sum_{e=1}^J \ln p_e \right\} + \frac{1}{aN} \left\{ N_{0j} - \frac{N_0}{J} \right\} \ln(1 - a)$$

which implies :

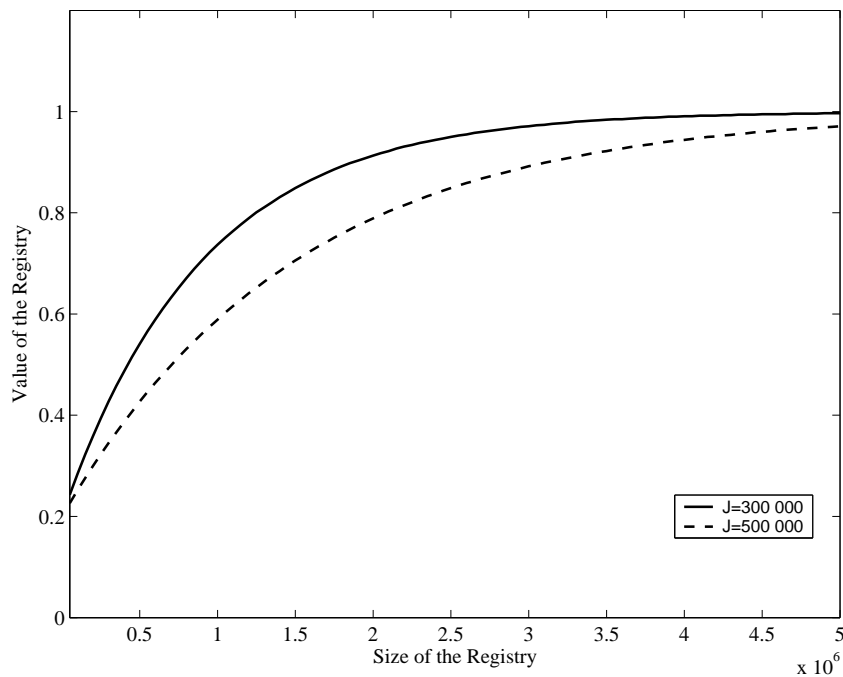
$$\pi^0 = J\bar{p}(1 - a)^{\frac{N_0}{J}} e^{-\frac{aN}{J}}$$

The efficiency system depends on a few parameters:

- J number of types
- \bar{p} geometric mean of frequencies
- a "probability to be available" for a graft
- $\left. \begin{array}{l} N_0 \\ N \end{array} \right\}$ sizes of the stock and of the increment
 $J\bar{p}$ measures the impact of the shape of the frequencies

We must notice that q_j^0 may be negative and then not realistic and $\pi^0 =$ is an "optimist" bound (minimum of π under positivity constraint is greater than π^0).

The following figure represents the value of the registry (measured as the probability for any patient to find a donor) as a function of its size. For the determination of the number of the types retained ($J=300\ 000$ or $J = 500\ 000$), see chapter 3 section 4.



1.8.2 Numerical simulations

The principle of our numerical simulations is the following one. Given a list of types and of their frequencies (p_j), a choice of a ($a = \frac{1}{3}$ ou $\frac{1}{6}$), an initial registry : N_0 is (in our example) proportional to p_j below a definite threshold and equal to 0 after, different sizes for N (the size of the increment), we compare

- the value of the registry if donors "are coming" with the same frequency as the recipients (no selection).
- the value of the registry related to the optimal selection of donors without or with any positivity constraints
- the value of the registry associated to "realistic" selection mechanisms

The base of our work is: a list of types issued from FGM registry : in fact 66 164 different observed phenotypes and their frequencies (see chapter 3 section 4 for statistical details), FGM data coming from *source* : *http://FGM website www.fgm.fr*

- Registre : 107 925
- French patients in 2003 : 813
- Grafted French patients : 62

The probability of "not finding" a donor is equal to 0.92. This implies that if we double the registry in an optimal way the expected number of new grafts by year is around 48 (the probability moves from .08 to .14). In the scenario based on FGM file, *an increment of the registry without selection will give around 40 new grafts by year only.*

If the actual registry is optimal :

$$J\bar{p}e^{-\frac{aN_0}{J}} = 0,92, J\bar{p} \leq 1. \text{ Then}$$

$$0,92 \leq e^{-\frac{aN_0}{J}}, \quad a = \frac{1}{3}, \quad N_0 = 121\,000$$

$$\Rightarrow J \geq 484\,000 \quad (J \geq 4N_0)$$

It seems to be reasonable to simulate situations in which the initial registry is 5 times smaller than the number of types

Then we test different models with an increment of the registry N , equal to $N = 0, 10\%, 25\%, 50\%, 100\%$ of N_0 , where J is the number of types and \bar{p} the geometric mean of frequencies. (see chapter 2 section 1 for the interpretation of $J\bar{p}$).

	J	N_0	$J\bar{p}$
Model 1	1 162	255	0.48
Model 2	4 648	1 000	0.49
Model 3	66 164	11 952	0.78

1 : A Mado quarter (allows most of numeric computations)

2: Mado

3: FGM

1.8.3 Back to the number of HLA types

Simulations without increment of registry and assuming that registry donors and recipients are drawn from the same population leads to a probability of "not finding" a donor between 0.63 and 0.78 (lower than the observed 0.92).

Model with 66 164 types

Size of registry	Efficiency
11 952	0,78
3 375	0,87
1 574	0,91

The observed efficiency is consistent with a number of types equal to 50 times the size of the registry (about 4 000 000 types in France, see chapter 3 section 4 for statistical computations).

But it is true for $a = \frac{1}{3}$. If $a = \frac{1}{6}$, the result is consistent with lower number of types.

Simulations are given for each of the models we tested.

Results of model 1 ($J = 1\,162$, $N_0 = 255$, $a = \frac{1}{3}$):

Expected probability of "not finding" a donor

N	N_0 selection	Optimal implementable	Optimal
10 % N_0	0.61	0.55	0.43
25 % N_0	0.59	0.54	0.43
50 % N_0	0.57	0.51	0.42
N_0	0.53	0.47	0.41
$2N_0$	0.47	0.42	0.38

Results of model 2 ($J = 4\,648$, $N_0 = 1\,000$, $a = \frac{1}{3}$) 823 types:

Expected probability of "not finding" a donor

"Elimination" based on the initial registry									
N	$a = 1/3$			$a = 1/6$			$a = 1/2$		
	N_0 selection	optimal	optimal implement- able	N_0 selection	optimal	Optimal implement- able	N_0 selection	optimal	optimal implement- able
0	0.637	x	x	0.7732	x	x	0.5388	x	x
100	0.6253	0.45	0.6175	0.7624	0.47	0.7619	0.5283	0.42	0.5139
250	0.6088	0.44	0.5874	0.747	0.47	0.7431	0.5137	0.41	0.4768
500	0.5841	0.44	0.5404	0.7286	0.46	0.7128	0.4917	0.40	0.4208
1000	0.5425	0.42	0.4575	0.6831	0.46	0.6558	0.4546	0.38	0.3277

Results of model 3 ($J = 66\ 164, N_0 = 11\ 952, a = \frac{1}{3}$):

Expected probability of "not finding" a donor

MADO sample

	N_0 selection	Optimal selection (non implementable)
0	0.777	0.726
10 %	0.771	0.721
25 %	0.763	0.715
50 %	0.749	0.704
100 %	0.723	0.683

Elimination of types and renormalization of the p_j

Expected probability of "not finding" a donor				
	$N_{0j} > 3$	$N_{0j} > 2$	$N_{0j} = 0$	$N_{0j} > 3$ & $N_{0j} = 0$
10 % N_0	0.771	0.771	0.769	0.767
25 % N_0	0.762	0.762	0.759	0.758
50 % N_0	0.748	0.745	0.743	0.736
100 % N_0	0.720	0.721	0.719	0.709

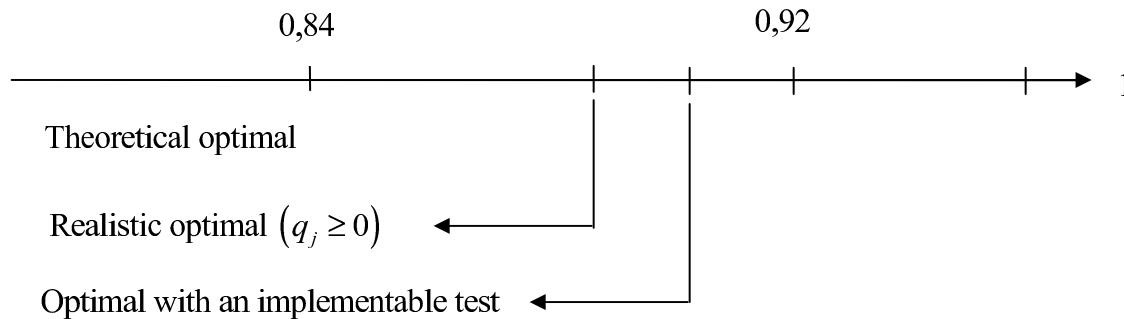
Simulations at the national level : we consider the FGM registry as the initial registry and we specify some assumption on the population of phenotypes

The efficiency observed is equal to 0.92

$$\begin{cases} J\bar{p} = 0.85 \\ a = \frac{1}{6} \\ \frac{N_0}{J} = \frac{1}{10} \Rightarrow 1\ 200\ 000 \end{cases}$$

The optimal efficiency of the registry (size 120 000) is equal to 0.84 in this case. The optimal efficiency would be equal to 0.82 with 240 000 donors (with 1 000 000, 0.74).

We represent the results on an horizontal axe:



1.8.4 Microsatelites filtering

Our simulations are based on the MADO file (4 648 types weighted from FGM registry which represent a population of 21 329 individuals) and we consider only 5 μ sat (D6S265, MIB, TNFa, DQCARI, G51152) and for any value of the μ sat phenotype a decision rule associates 1 (continue towards HLA typing) or 0 (stop typing). We evaluate this filtering mechanism by simulation. We try three different rules:

- Rule 1 : "simple decision rule MADO" (goal : elimination of frequents)
- Rule 2 (rule 1 modified) : less elimination of frequents
- Rule 3 : "another rule", defined as follow,
 1. We retain all the μ sat phenotypes of frequent and rare HLA (2.2 % population and 4 249 phenotypes representing 49 % population)
 2. any new donor with μ sat phenotype present in the previously defined set is then avoided

We conclude that rules 1 and 2 are worse than the no selection rule and should not be implemented. Rule 3 is better than no selection mechanism. If we evaluate the cost of rule 3, it is equivalent to increase the registry of 150 donors with no selection and to test 210 individuals, retain 100 using rule 3 and HLA type these individuals.

Then the test is useful if :

$$210 \text{ cost of test} + 100 \text{ cost of typing after test} \\ \leq 150 \text{ cost of typing}$$

e.g. if cost of typing after test = cost of typing

$$\text{Cost of test} \leq 0.24 \text{ cost of typing}$$

We finally compare the three models of microsatellite filtering in terms of our evaluation criteria.

Variation of the probability to find a donor for a recipient as function of the size of the registry and of the selection mechanisms

variation of the size	No selection	Optimal selection	μ -sat filter 1	μ sat filter 2	μ sat filter3
10 %	0.011	0.019	0.007	0.008	0.017
25 %	0.029	0.05	0.019	0.023	0.039
50 %	0.053	0.097	0.03	0.045	0.07
100 %	0.09	0.18	0.06	0.08	0.11

We must remark that the rule is an ad hoc rule based on the MADDO file. Is it robust to the extension to a larger population ? Moreover the MADDO file is optimistic in terms of the dispersion of the frequencies.

1.8.5 Main conclusions and extensions

We note the low efficiency of hematopoietic stem cells Registries and we have to underline the fact that an important increase of registries doesn't improve efficiency in a sensitive way. Moreover, there exists a low impact of selection mechanisms of donors.

We then try to improve the results of our model.

We decide first to limit the model to "frequent" types. We consider 14 815 types (the most frequent ones) which represent 0.5242 % of individuals. Assuming that we consider only this "sub-population" then:

$$J\bar{p} = 0,77$$

$$J = 14\,815$$

$$a = \frac{1}{3}$$

If $N = 100\,000$ (doubling of actual frequents registry) the optimum gives an efficiency for the registry equal to 0.08 (92 % of frequent patients will have a donor). The question asked is :” Should we limit the model to the frequents ?”.

The following table gives details of the model ”limited to frequents”. For three different sample size, we compute the value of the registry if N is equal to 50 000 or N is equal to 100 000.

Model ”limited to frequents”
Expected probability of ”not finding” a donor

66 164 types FGM: 107 925 indiv.	$N_j = 1$ in FGM dropped 14 815 types 0.5242 % of individuals		$N_j = 1 \& N_j = 2$ in FGM dropped 6 526 types 0.3707 % of individuals	
	$J\bar{p}$	Value of the registry	$J\bar{p}$	Value of the registry
$N = 100\,000$	0.7659	0.08 (92 % of patients find a donor)	0.7733	0.0047 (95 %)
$N = 50\,000$		0.25 (75 %)		0.06 (94 %)

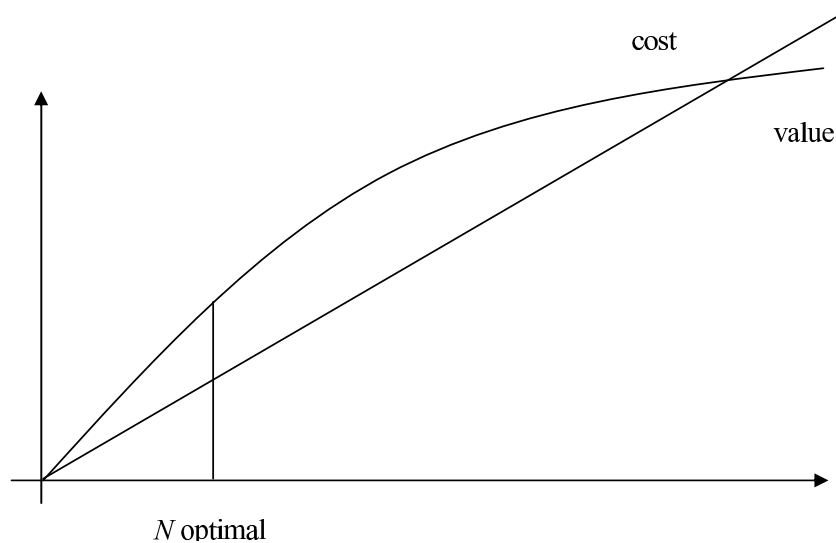
Economic value of the Registry (see chapter 2 for details):

Is it useful to have a bone marrow donor registry ?

We imagine a country, composed of, J types, \bar{p} (geometric) mean of frequencies, N recipients (during 10 years), V value of available matching, c cost of typing.

What would be in this case the value of a registry of size N ?

$$VM \underbrace{\left(1 - J\bar{p}e^{-\frac{aN}{J}}\right)}_{\substack{\text{probability to find} \\ \text{a donor in an} \\ \text{optimal world}}} - CN$$



$$N_{opt} = \frac{J}{a} \ln \frac{VM\bar{p}a}{c} = \frac{J}{a} \ln \frac{VMJ\bar{p}a}{Jc}$$

Is it useful to have a registry ? $N \geq 0$

$$\Leftrightarrow \frac{VM\bar{p}a}{c} \geq 0$$

$$\Leftrightarrow \frac{VM\bar{p}Ja}{c} \geq J$$

Given the following calibration, in this model $\frac{VM\bar{p}Ja}{c}$ represents the upper value for the number of types.

Calibration :

$$a = 1/3 \quad c = 200 \text{ euros} \quad M = 10\,000 \text{ patients} \quad J\bar{p} = 0.8$$

$V = ?$ Revealed value for France

2 millions euros / year

60 available matchings

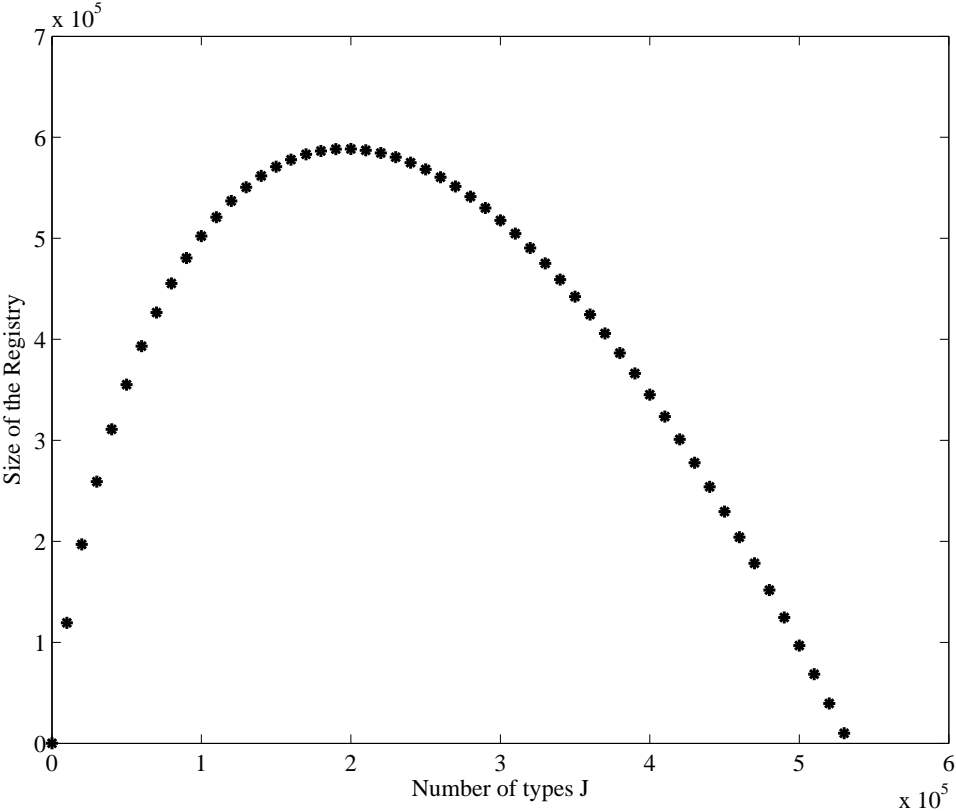
$$V = \frac{2\,000\,000}{60} = 40\,000$$

$$J \leq \frac{40\,000 \times 10\,000 \times 0.8 \times 1/3}{200} = 530\,000$$

$$J \leq 530\,000$$

The optimal size of the registry function of J . If $J = 400\,000$, threshold of the cost of typing c should be smaller than 267 euros.

An optimal size of the Registry ? The following figure derives from the results previously given. It is actually based on simulations and computations we have done. This kind of results emphasizes epidemiologists intuitions. (see Oudshoorn et alii 1997, Schiper et alii 1996, Takahashi 1989).



According to our computations, there exists obviously an optimal size of the registry existing in a population. The key parameters of the efficiency would be first the shape of the distribution of the frequencies of the types, then the number of types in the sample but also the efficiency evaluation criteria, given potential donors. This result is developed through chapter 2 when looking at the defined Welfare function (see chapter 2 section 2).

Chapitre 2

Evaluation Economique de l'Organisation d'un Registre

Ce chapitre fait l'objet d'un article soumis à la Revue d'Epidémiologie et de Santé Publique en décembre 2005. Accepté le 21 juin 2006 sous réserve de modifications, l'article soumis est actuellement en révision.

2.1 Un modèle de référence

2.1.1 Introduction

L'utilisation de greffes de cellules souches hématopoïétiques (CSH ou greffe de moelle osseuse) est une pratique courante dans le traitement des leucémies et de celui d'autres affections graves. La greffe se pratique le plus souvent à partir d'un donneur vivant compatible au sens du système HLA. On cherche en priorité un donneur familial mais en son absence un donneur volontaire est recherché dans le fichier français des donneurs et, parallèlement, dans l'ensemble des registres mondiaux (Hurley et alii 1997, Hoffman-Smith 1993). Au niveau national il existe donc un registre d'environ 130 000 donneurs dont le type HLA a été déterminé¹. Le typage HLA est coûteux et l'accroissement du registre est limité plus par des considérations budgétaires que par manque de volontaires. L'objectif de ce chapitre est d'évaluer l'efficacité du Registre à la lumière du calcul économique et d'identifier les éléments-clé les plus importants qui entrent en jeu. Certains peuvent

¹Le nombre total de donneurs inscrits au 31 décembre 2004 est de 129 042, source : www.fgm.fr

être évalués à partir d'arguments statistiques (nombre et distribution des types HLA en France, voir Gourraud 2005), d'autres peuvent évoluer en fonction de la politique de gestion du fichier et d'autres enfin relèvent d'une évaluation plus complexe liée au bénéfice attendu d'une greffe. L'intérêt de ce calcul est double : évaluer la valeur des paramètres qui rendent optimale la politique actuellement suivie ou à partir de valeurs de paramètres définis a priori calculer une gestion optimisée du registre. On pourra aussi examiner la sensibilité de ces résultats aux paramètres. La question posée est la suivante : étant donné le coût élevé de l'accueil, du typage des donneurs et de la gestion du registre, est-il socialement et économiquement efficace d'avoir un Registre de donneurs ? et quelle taille doit-il avoir ? On modélise ici la situation actuelle où d'une part, le financement du registre est assuré en grande partie par des fonds publics et où d'autre part, le typage des donneurs est réalisé par des laboratoires hospitaliers ou appartenant à l'établissement français du sang. Le système actuel est simple: les laboratoires hospitaliers réalisent les typages par petites quantités au fur et à mesure de l'arrivée des donneurs. On considère donc assez naturellement l'ensemble des laboratoires comme un monopole public et on utilise donc la théorie usuelle de type Ramsey-Boiteux. Le modèle est un modèle statique sans aléa moral ni assymétrie d'information (on considère l'information comme parfaite).

Notre modèle, basé sur la maximisation du bien-être social est schématique, mais il a le mérite de mettre en évidence les principaux ordres de grandeur de l'analyse économique de la gestion d'un registre. Soulignons que notre étude est limitée à un seul pays et que ne sont pas prises en compte les interactions entre registres : des patients français utilisent des donneurs étrangers et des patients étrangers sont greffés à partir de donneurs français. La valeur sociale du Registre devrait donc aussi être évaluée dans sa dimension internationale. Ce point fait l'objet d'un travail en cours.

2.1.2 Le modèle structurel

Fonctionnement du Registre Français

Créé en 1986, le Registre national est géré par FGM, association reconnue d'utilité

publique, sous l'égide de l'Agence de Biomédecine depuis décembre 2005. Le registre actuel de FGM se constitue en effectuant le typage HLA de classe I et II en basse résolution de tous les volontaires ayant satisfait un ensemble de conditions, d'âge et de santé notamment. Rappelons que par typage HLA de classe I, on désigne l'identification des allèles des locus A et B et par classe II celle des locus $DRB1$ et $DQB1$ (Berchery 2003). On désigne par basse résolution un niveau de typage standard, se traduisant par une nomenclature simplifiée "à deux digits" des allèles (Marsh et alii 2004). Le typage HLA de haute résolution n'est pratiqué que ponctuellement au fur et à mesure des nécessités selon les interrogations du registre pour les patients. Ces typages dépendent d'un budget différent, relevant de l'assurance maladie et sont payés par l'hôpital où est le patient. On simplifiera notre étude en supposant qu'une rencontre entre un donneur et un receveur est compatible s'il y a identité entre le type $HLA A, B, DR$ en basse résolution entre donneur et receveur. Ce n'est en effet que à ce niveau que les informations statistiques sont disponibles. Le premier appariement est fait à ce niveau de basse résolution mais la décision de faire la greffe tient compte d'un typage haute résolution qui est réalisé dans un deuxième temps. La compatibilité entre un donneur et une receveur est donc examinée à deux niveaux: on cherche d'abord dans le registre un ou plusieurs donneurs compatibles A, B, DR en basse résolution puis parmi ceux-ci on cherche un donneur pleinement compatible, en particulier à partir du typage haute résolution. On ne considèrera pas ici de possibilité de compatibilité partielle, difficile à modéliser car liée aux décisions des médecins greffeurs dans chaque cas particulier. Le nombre de donneurs nouveaux inscrits chaque année dans le registre est déterminé par une contrainte budgétaire et pour les années 2001 – 2004, au plan greffe, sur la base d'un tarif qui a été évalué et qui correspond au prix payé aux laboratoires chargés de cette prestation. Le nombre de volontaires pourrait excéder en effet le nombre d'individus pouvant être typés ou ne pas correspondre à la distribution territoriale recherchée. Dans le cadre du plan greffe, deux groupes ad-hoc de l'EFG (et FGM) ont conjointement établi un quota de distribution des donneurs à recruter annuellement entre les centres donneurs. Le niveau de diffusion des campagnes de recrutement tente de rapprocher au plus près le nombre de volontaires recherchés sur une période donnée (fonc-

tion du budget disponible) et le nombre de volontaires qui se présentent effectivement. Les 34 centres français agréés au niveau national et international (accréditation WMDA pour *World Marrow Donor Association*) responsables du recrutement des donneurs, des laboratoires chargés de leur typage HLA (accrédités généralement par l'EFI) et de leur inscription sur le registre national reçoivent un montant forfaitaire de 183-€² par typage en contrepartie. Les techniques de typage varient selon les laboratoires. L'ensemble des laboratoires est soumis à un contrôle de qualité organisé par l'AFSSAPS (Agence Française de Sécurité Sanitaire des produits de Santé).

Dans le cas d'une structure hospitalière, qui réalise les typages pour le registre français par exemple, les rendements d'échelle croissants impliquent que le coût moyen à long terme est décroissant : il y a donc des économies d'échelle et le coût marginal est toujours inférieur au coût moyen. La tarification optimale qui égalise le prix et le coût marginal conduit alors inéluctablement à un déficit de la structure hospitalière. Celui-ci doit être comblé par des attributions de fonds publics financées le plus souvent par l'impôt. Des subventions visant à résorber le déficit d'un monopole public (ce qui est le cas de l'hôpital ici) ne peuvent cependant pas toujours être mises en place (surtout à long terme), même si ce déficit est justifié par le critère d'optimalité collective que représente la tarification au coût marginal. Il est souvent plus raisonnable de supposer que le monopole public est astreint à respecter une contrainte d'équilibre budgétaire : financer les coûts de production par des recettes au moins équivalentes devient alors une contrainte qui doit être prise en compte pour définir la politique tarifaire³. On est alors conduit à définir une tarification qui maximise le surplus collectif, ou bien-être social, sous cette contrainte additionnelle que constitue l'équilibre budgétaire. Le surplus collectif constitue une mesure de l'avantage net apporté par la constitution et la gestion d'un registre de donneurs de moelle osseuse.

²Le forfait "Plan Greffe" s'élevait à 183.13-€ par typage en 2002

³Règle de Ramsey-Boiteux

Une définition du surplus pour les patients

Le surplus pour les patients associé à l'existence d'un Registre constitue une mesure de l'avantage obtenu en termes de qualité de vie (et de survie): un malade bénéficiera d'une meilleure qualité de vie s'il a la possibilité d'avoir recours à une greffe. On peut résumer la valeur pour un patient d'un appariement avec un donneur compatible par une grandeur V . Cette grandeur a une signification complexe que nous n'examinerons pas ici : en fait l'identification d'un donneur n'est bénéfique que si la greffe a bien lieu (étude plus fine de la compatibilité génétique) et dépend du succès de la greffe. Cette grandeur V est en fait une espérance des évaluations de toutes les conséquences possibles découlant de l'identification d'un donneur. Ces conséquences sont pondérées par leur probabilité d'occurrence pour le malade. On en déduit alors que si M est le nombre de patients, la valeur du registre peut être évaluée par le produit $MV\pi$ où π est la proportion de patients trouvant un donneur compatible à 4 Digits.

La politique de l'organisation en charge du registre se manifeste par la détermination de π . Cette grandeur dépend évidemment du nombre et de la distribution des types HLA dans la population, du mode de sélection des donneurs et d'une grandeur $a \in]0, 1[$ décrivant la probabilité pour qu'un donneur compatible dans le registre soit effectivement compatible pour une greffe. On verra que ce paramètre joue un rôle central dans nos calculs. La probabilité π ⁴ dépend enfin bien évidemment du nombre de donneurs inscrits dans le registre.

Le calcul de π est complexe et ne sera pas développé ici (voir chapitre 1). On note le fait qu'une borne supérieure de π peut s'écrire :

$$\pi = 1 - J \bar{p} \exp\left(\frac{-a(Q_0 + Q)}{J}\right)$$

où J est le nombre de phénotypes HLA de la population, \bar{p} est la moyenne géométrique de la fréquence des types ($\bar{p} = \exp(\frac{1}{J} \sum_{j=1}^J \ln p_j)$ où p_j est la fréquence du type j), Q_0 est le stock initial de donneurs et Q l'incrément du registre. Rappelons que $J\bar{p} \leq 1$. Cette grandeur $J\bar{p}$ sera notée α et est égale au quotient de la moyenne géométrique par

⁴Il s'agit ici d'une approximation Poissonienne de la loi binomiale, cette formule n'est donc pas valable pour de petites valeurs de Q

la moyenne arithmétique des p_j . Elle mesure l'écart de cette distribution à la loi uniforme (pour laquelle les fréquences sont toutes égales à $\frac{1}{J}$). Une distribution présentant des fréquences de phénotypes très variées aurait une valeur α faible alors que $\alpha = 1$ signifie $p_j = \frac{1}{J}$ pour tout j ⁵.

Le calcul de π précédant découle d'un mode spécifique de sélection des donneurs et de son optimisation pour l'ensemble des modes de sélection possibles. Cette probabilité n'est donc pas réalisable en ce sens qu'un mode de sélection des donneurs réalisable (fondé sur des critères objectifs de sélection des donneurs) n'atteint pas cette borne. Notre critère est donc une évaluation "optimiste" de l'efficacité du registre. Le surplus pour les patients du registre est donc:

$$\varphi(Q) = V M(1 - \alpha \exp(\frac{-a(Q_0 + Q)}{J})).$$

Remarque: Le processus de sélection d'un donneur dans un registre pour satisfaire un receveur particulier comprend deux étapes. Le registre contient un ensemble d'individus ainsi que leurs phénotypes. On cherche donc dans le registre l'ensemble des donneurs ayant le même type que le receveur.

Si aucun donneur n'est compatible le processus s'arrête. Sinon de nouveaux examens sont effectués sur les donneurs compatibles (incluant la recherche de la disponibilité effective du donneur) de manière à affiner le critère de compatibilité. Si un donneur totalement compatible est trouvé, la greffe peut avoir lieu, pour autant que l'état du patient le permette encore mais ceci n'est pas inclus dans notre analyse. On a donc deux niveaux de critères de compatibilité: un premier niveau qui est révélé par le registre et un second niveau est inconnu jusqu'au moment où le donneur est examiné pour un receveur particulier. Le partage entre les deux types de compatibilité pose une question économique pertinente qui ne sera pas traitée ici mais simplement évoquée de manière à préciser l'interprétation des paramètres. Il est clair que tous les critères de compatibilité ou de disponibilité variables au cours du temps (incompatibilité résultant du fait que le donneur lui même été

⁵un calcul élémentaire montrerait que $-\ln \alpha$ est égal à la divergence au sens de l'entropie entre la loi uniforme et la loi des types

malade ou indisponibilité du donneur due à une grossesse par exemple) ne peuvent que faire partie du second niveau (non enregistré et révélé au moment de la recherche). Par contre les critères fixes (sexe, typage HLA "fin" (à 4Digits)) peuvent être renseignés dans le registre (au prix d'un coût du typage élevé donc d'une taille plus réduite du registre à budget donné) ou ne pas être tous recherchés au moment de l'inscription. Le typage du registre définit alors une partition plus ou moins grossière qui n'est affiné qu'au moment de la décision.

Notre modèle incorpore ces deux éléments: les types $j = 1, \dots, J$ et leurs fréquences p_j formalisent les phénotypes observés car enregistrés sur le registre alors que le paramètre a décrit la probabilité pour qu'un donneur compatible au premier niveau le soit au second. Il découle de cette remarque que, en général, a devrait varier en fonction de j (le polymorphisme à 4 digits n'étant pas le même pour tous les groupes de même type à 2 digits). Notre modèle est donc une simplification permettant des calculs plus explicites. d'autre part, les connaissances actuelles ne permettent pas une calibration de a_j pour tous les types j . de plus certains éléments intervenant dans le paramètre a sont clairement indépendants de j .

Une définition du surplus pour les laboratoires

Le surplus des laboratoires est la mesure de l'avantage net associé au fait de devoir réaliser les typages. On le définit ici comme le profit issu de la production de typages, la différence entre les recettes liées aux typages PQ (P est le prix payé pour un typage) et leur coût de production $C(Q)$. Soit Q le nombre de typages HLA, on supposera pour simplifier que la fonction de coût s'écrit :

$$C(Q) = C_F + c Q$$

Le coût de production est donc la somme d'un coût fixe C_F et de coûts variables cQ : le coût marginal , c est supposé constant et il est toujours inférieur au coût moyen $c + \frac{C_F}{Q}$. On définit le surplus des laboratoires comme étant égal à :

$$PQ - C(Q)$$

Une définition du Bien-Être Social

On définit le bien-être social du modèle (BES), comme étant égal à la somme du surplus pour les patients et du surplus pour les laboratoires réalisant les typages, dont il faut retrancher le coût des fonds publics, $(1 + \lambda)PQ$. Ce coût s'explique par la nécessité de payer les typages et donc de prélever PQ d'impôts (ce qui retire PQ de bien-être des consommateurs). L'effet de distortion associé à ce prélèvement est modélisé par le coût d'opportunité des fonds publics. La constitution et la gestion d'un Registre ont un coût pour la collectivité : le coût d'opportunité des fonds publics diminue le bien-être collectif.

$$BES = \varphi(Q) + (PQ - C(Q)) - (1 + \lambda)(PQ)$$

La collectivité maximise son bien-être. Afin de tenir compte de la contrainte d'équilibre budgétaire $PQ - C(Q) = 0$, on associe un multiplicateur de Lagrange μ au modèle et on maximise :

$$\varphi(Q) + (PQ - C(Q)) - (1 + \lambda)(PQ) + \mu (PQ - C(Q))$$

Les conditions de premier ordre de la maximisation en P et Q nous donnent le système d'équations suivant :

$$\varphi'(Q) - C'(Q) - \lambda P + \mu(P - C'(Q)) = 0$$

$$-\lambda Q + \mu Q = 0$$

$$PQ = C(Q)$$

où φ' et C' sont les dérivées par rapport à Q de φ et de C , soit si $Q > 0$,

$$\mu = \lambda$$

$$\varphi'(Q) = (1 + \lambda)C'(Q)$$

$$P = \frac{C(Q)}{Q}$$

On trouve ici le résultat bien connu de l'égalité du multiplicateur de Lagrange de la contrainte de profit nul du monopole au coût d'opportunité des fonds publics. (Laffont 1988, Mas-Colell, Whinston, Green, 1995, Picard 1998). La seconde équation est la principale et s'écrit :

$$V M(1 - \alpha \exp(\frac{-a(Q_0 + Q)}{J}))' = C'(Q)(1 + \lambda)$$

ou encore :

$$c = \frac{V M \alpha \frac{a}{J} \exp(\frac{-a(Q_0 + Q)}{J})}{(1 + \lambda)}$$

On tire de cette équation la quantité optimale Q :

$$Q = \frac{J}{a} \ln \frac{V M \alpha a}{c(1 + \lambda)J} - Q_0.$$

Une autre expression de cette relation consiste à écrire :

$$\frac{V}{c(1 + \lambda)J} = \frac{1}{M \alpha a} \exp(\frac{a(Q_0 + Q)}{J})$$

où la valeur de l'identification d'un donneur compatible avec un receveur donné, exprimée en coût du typage (accru du coût d'opportunité) dépend des autres grandeurs du modèle, Q en particulier.

La troisième équation permet simplement de relier le prix aux coûts fixes.

2.1.3 La calibration du modèle

La calibration du modèle repose sur un certain nombre d'éléments de sources différentes. Le nombre J de phénotypes différents dans la population française est inconnu. On sait que sur 107 925 individus, 66 164 phénotypes différents ont été détectés et au niveau

mondial sur environ 6 millions de donneurs⁶, 400 000 phénotypes différents ont été identifiés. Dans le chapitre 3, un modèle statistique a été développé dont l'application à la population française permet de prédire un nombre de phénotypes d'environ 500 000. Bien que statistiquement pertinent ce modèle a plutôt tendance à légèrement sous évaluer le nombre de phénotypes et donc 500 000 serait plutôt une borne inférieure.⁷(voir Gourraud, 2005).

A partir de simulations réalisées dans le fichier français de donneurs, on peut supposer que pour un échantillon de taille n , le produit $\alpha = J\bar{p}$ (nombre de phénotypes observés sur cet échantillon multiplié par la moyenne géométrique des fréquences de l'échantillon) peut être estimé à 0.7. On a en effet observé que sur un fichier aléatoire de 21 360 donneurs, α vaut 0.9 puis 0.85, si le fichier se compose de 42 934 individus. Il passe à 0.827 si 64 517 donneurs sont sélectionnés et à 0.8 pour 86 149. Dans l'échantillon de 107 925, la valeur empirique de α est 0.78. Quand la taille de l'échantillon augmente J augmente et \bar{p} diminue; bien que la loi de dévolution de $J\bar{p}$ soit à notre connaissance non étudiée il semble qu'elle décroisse vers une asymptote.

Le paramètre a qui décrit la probabilité pour qu'un donneur compatible au premier niveau soit compatible au deuxième niveau et disponible est fixé à $a = \frac{1}{3}$. ce choix repose sur une opinion qualitative fournie par FGM.

Pour valider nos choix de $J = 500\,000$, $J\alpha = 0.7$ et $a = \frac{1}{3}$, considérons les deux calculs suivants. Tout d'abord en faisant $a = 1$ dans la définition de π , on doit trouver la probabilité pour qu'un receveur quelconque trouve un donneur compatible au sens $A, B, DR(2D)$. Ce calcul donne 0.45. Ce chiffre n'est pas fourni par FGM (qui définit la compatibilité sur un autre critère). Toutefois, on sait que environ 49% des nouveaux donneurs ont un phénotype déjà présent dans le Registre⁸. Si les populations de donneurs et de receveurs étaient identiques, les deux chiffres devraient être proches. On observe dans notre cas le même ordre de grandeur, la différence s'expliquant en partie par le fait

⁶Résultats obtenus et présentés dans le cadre du consortium européen MADO

⁷Ce modèle est estimé sur l'échantillon français. Au niveau mondial, il prédirait 330 000 phénotypes différents pour 6 000 000 d'individus; contre 400 000 environ observés. La population mondiale n'est pas homogène, ce qui explique cette sous prédiction

⁸cf. page 33, Rapport d'activité 2004 France Greffe de Moelle

que la population des donneurs est plus homogène que celle des receveurs. Le choix de $a = \frac{1}{3}$ peut aussi être validé par la comparaison avec des résultats internationaux. Dans Heemskerk, van Walraven, Cornelissen et alii (2005), on trouve (table 2) que le passage de la compatibilité HLA $A, B, DR(2D)$ à la compatibilité complète fait passer le pourcentage de receveurs trouvant un donneur de 84% à 59% dans la dernière période. Dans notre modèle, si $a = \frac{1}{3}$, on passe de 0.45 à 0.35, ce qui représente une diminution du même ordre de grandeur.

Les hypothèses quantitatives retenues pour la calibration de la fonction de coût sont toutes issues d'entretiens avec des professionnels du système de santé (médecins immunologistes, professeurs, chercheurs, biologistes et techniciens de laboratoire). Certaines informations, en particulier celles concernant le coût du travail ou le coût du capital, ne nous ont pas été communiquées. Or il est important de ne pas se limiter aux seules dépenses exprimées en unités monétaires (tarifs de réactifs), mais de considérer également toutes les ressources, notamment celles dont la consommation n'est pas représentée par les prix de marché, comme par exemple le travail ou le capital. A la suite d'entretiens et d'enquêtes dans différents laboratoires, nous évaluons le coût marginal du typage à 150€. Ce chiffre recouvre en fait une forte disparité due à l'utilisation de techniques différentes dans les différents laboratoires.

Le coût d'opportunité des fonds publics a fait l'objet de nombreuses études empiriques et la valeur de $\lambda = 0.5$ est couramment acceptée. Enfin nous avons fixé Q_0 à 120 000 donneurs.

Notre modèle est valide pour une période déterminée de temps. Celle-ci doit être choisie en relation avec la durée de la présence du même individu dans le registre. Nous proposons d'évaluer le registre au cours d'une période de 10 ans. Pendant cette période, $M = 10\,000$ patients sont en attente d'une greffe (on observe actuellement un peu moins de 1 000 patients en attente de greffe par an). Il ne reste donc à déterminer que V pour trouver la valeur optimale de Q , nombre de nouveaux donneurs⁹.

⁹Notre modèle suppose que Q est non contraint par le nombre de volontaires

2.1.4 Résultats

Le premier résultat consiste à déterminer V afin que la politique actuellement suivie (augmentation de $Q = 100\ 000$ donneurs en 10 ans) réalise l'optimum. Cette valeur se déduit des conditions de premier ordre et est égale à $V = 55\ 831\text{€}$. Ce premier résultat est tout à fait acceptable compte tenu du coût général du traitement des hémopathies. Un autre résultat découlant de ce modèle est la justification du tarif de 183€ qui permet de recalculer les coûts fixes des laboratoires. Un calcul élémentaire donne alors $3,312$ millions d'euros. Il ne s'agit bien sûr pas des coûts fixes des laboratoires dans leur totalité mais la part de ces coûts allouée au typage. Là encore ce chiffre est cohérent même s'il semble plutôt faible, ce qui signifierait que les autres domaines d'activité des laboratoires "subventionnent" le typage HLA. Là encore la situation des différents laboratoires est très hétérogène : certains utilisent des technologies peu coûteuses en équipement (et donc faibles en coûts fixes) alors que d'autres utilisent des technologies reposant plus sur des investissements importants.

Etant donné les paramètres retenus, il est tout d'abord important de noter que l'efficacité du registre mesurée par π , probabilité de trouver un donneur pour un receveur déterminé évolue très lentement en fonction de la quantité totale $Q_0 + Q$. π passe de 0.35 pour $Q = 0$ ($Q_0 = 120\ 000$) à 0.43 si $Q_0 + Q = 300\ 000$ et 0.64 si $Q_0 + Q = 1\ 000\ 000$ (voir graphique 1). Ce fait est bien connu empiriquement des responsables des registres et de la WMDA¹⁰ et explique en grande partie les résultats obtenus. Soulignons que la lenteur de la croissance de la probabilité n'est pas due à une mauvaise sélection des donneurs car répétons que celle-ci est supposée optimale. Une arrivée de donneurs sans sélection donnerait une croissance encore plus lente de π (voir chapitre 1 Annexe I).

¹⁰World Marrow Donor Association, <http://www.worldmarrow.org>

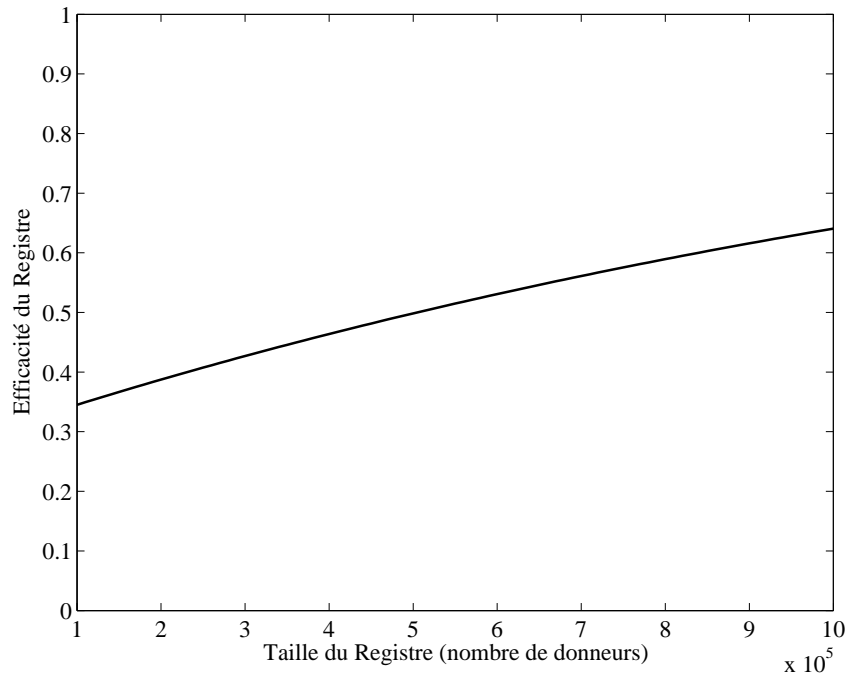


Figure 2.1: Efficacité du Registre en fonction de son accroissement

La figure 2 montre l'évolution du bien-être social (BES) en fonction de Q . Le modèle a été calibré pour que le maximum soit atteint en $Q = 100\ 000$. La principale remarque déduite de ce graphique est que la mesure du BES varie peu en fonction de Q . Maintenir le registre à sa situation actuelle ou le tripler n'a en effet que peu d'incidence sur la probabilité π et donc sur le BES.

L'étape suivante consiste en l'examen de la sensibilité des résultats en fonction des grandeurs introduites dans le modèle. La figure 3 montre la relation entre Q et BES pour différentes valeurs de J (nombre de phénotypes HLA différents dans la population). On voit en particulier que si $J = 700\ 000$ ou $J = 1\ 000\ 000$ l'optimum en Q est négatif : on n'a donc pas intérêt dans ce cas à augmenter la taille du registre actuel. Par contre si $J = 300\ 000$ un accroissement plus important du registre que celui prévu serait préférable. Rappelons toutefois que 500 000 est plutôt une hypothèse basse dans nos calculs.

La figure 4 présente les résultats de la même analyse mais pour des valeurs différentes de V (valeur estimée d'une greffe) en gardant les autres éléments du modèle à leurs valeurs de départ. Le phénomène est à peu près identique au précédent : si V baisse à 25 000 €

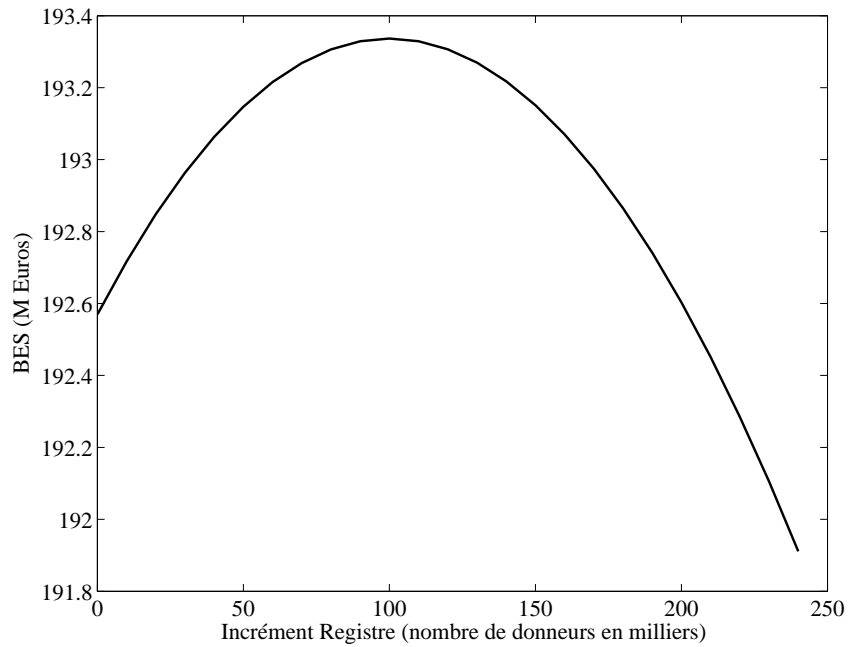


Figure 2.2: Bien-être-social fonction de l'incrément du registre

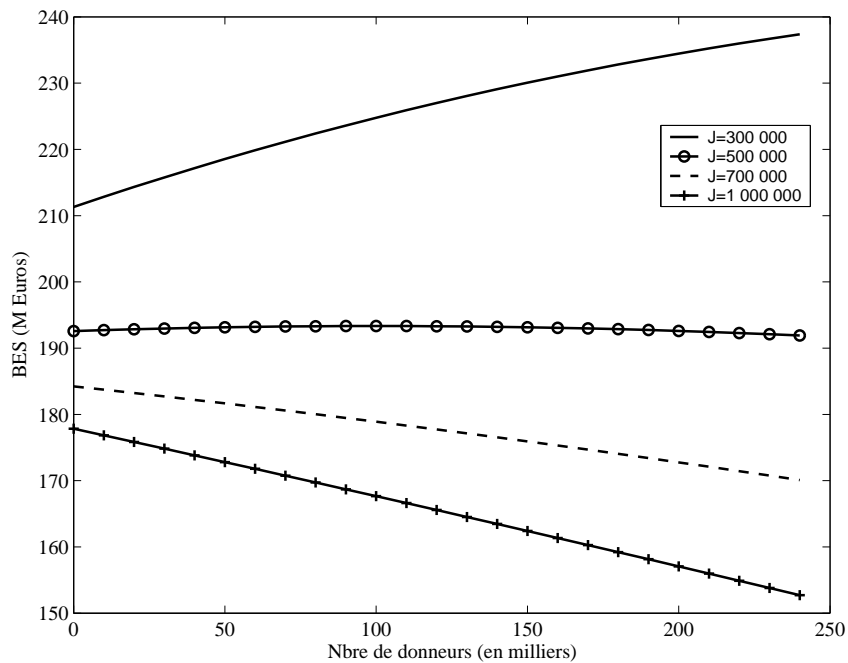


Figure 2.3: Bien-être-social et Nombre de phénotypes HLA dans la population

la politique optimale consiste à ne pas accroître le registre mais si V augmente sensiblement, il serait socialement utile d'accroître substantiellement le registre actuel. Les deux grandeurs J et V ne dépendent pas de l'action du gestionnaire du registre: J est une grandeur statistique dépendant de la structure géométrique de la population et V une grandeur "médicale" exprimée en €, dépendante en particulier de l'efficacité de la greffe et des thérapies alternatives. Les deux autres éléments du modèle peuvent au contraire être influencés par la politique de l'agence en charge du registre.

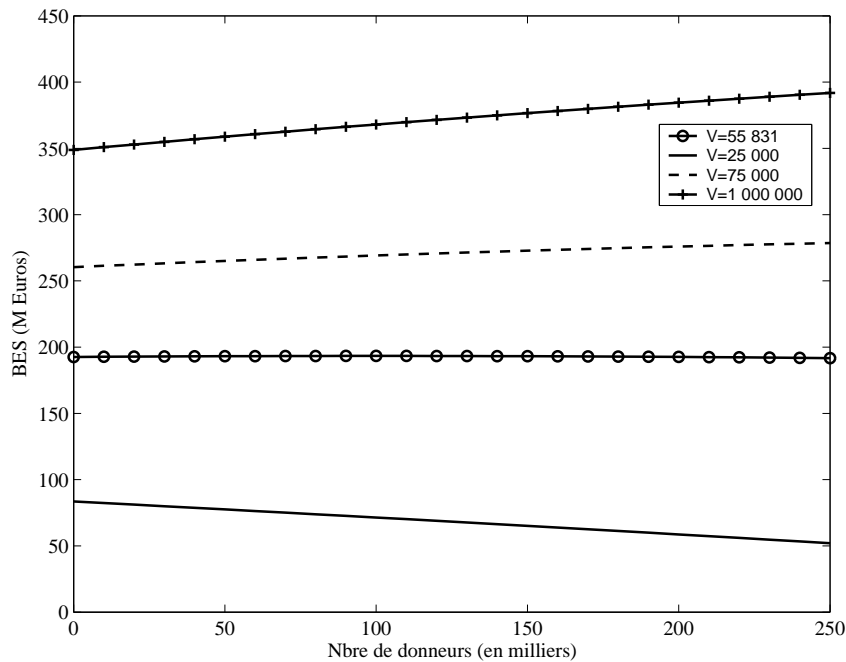


Figure 2.4: Bien-être-social et valeur implicite de l'appariement

La valeur de a modifie sensiblement le BES (voir figure 5). Si $a = \frac{1}{6}$, il n'est pas utile actuellement d'accroître le registre alors que si $a = \frac{1}{2}$, cet accroissement devrait être important. Plus encore même en gardant $Q = 100\ 000$ un accroissement de $a = \frac{1}{3}$ à $a = \frac{1}{2}$ augmenterait de près de 20% le BES grâce à l'accroissement de la probabilité de trouver un donneur qui passe dans ce cas de 0.27 à 0.30.

Certaines composantes des phénomènes décrits par le paramètre a ne sont pas modifiables par le gestionnaire du registre mais certains éléments liés à la disponibilité du donneur

pour une greffe peuvent être influencés. Par exemple, une campagne de recrutement plus informative sur le don de moelle osseuse et un meilleur suivi des donneurs présents dans le registre. Les donneurs volontaires de moelle s'inscrivent sur les registres dès lors qu'ils sont confrontés à une maladie du sang dans leur entourage proche: familles, amis. Le côté émotionnel et naturellement généreux du donneur volontaire au moment de l'inscription peut amener à minimiser l'aspect technique du prélèvement de moelle. Un donneur peut se désister à tout moment sans avoir à se justifier, et ce, même s'il est inscrit sur le Registre depuis plusieurs années. Les donneurs changent d'adresse, le Registre perd tout moyen de contacter le donneur en cas de greffe. On ne peut effectuer de prélèvement de moelle osseuse sur les femmes enceintes...

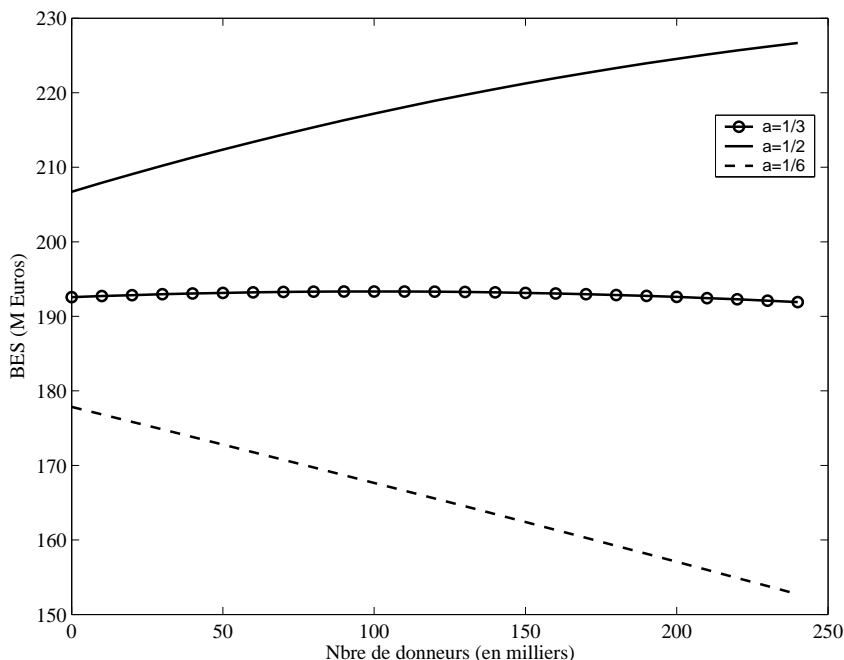


Figure 2.5: Bien-être social fonction de "a"

La figure 6 montre la relation entre le BES et le coût marginal du typage. Un accroissement de celui-ci conduirait à ne pas accroître la taille du registre. Par contre une diminution de c , même à $Q = 100\ 000$ entraînerait une amélioration de 10% environ du BES. Cet objectif peut être atteint par une meilleure organisation du typage (centralisation ?, utilisation de techniques plus efficaces ?). Une diminution des coûts marginaux justifierait

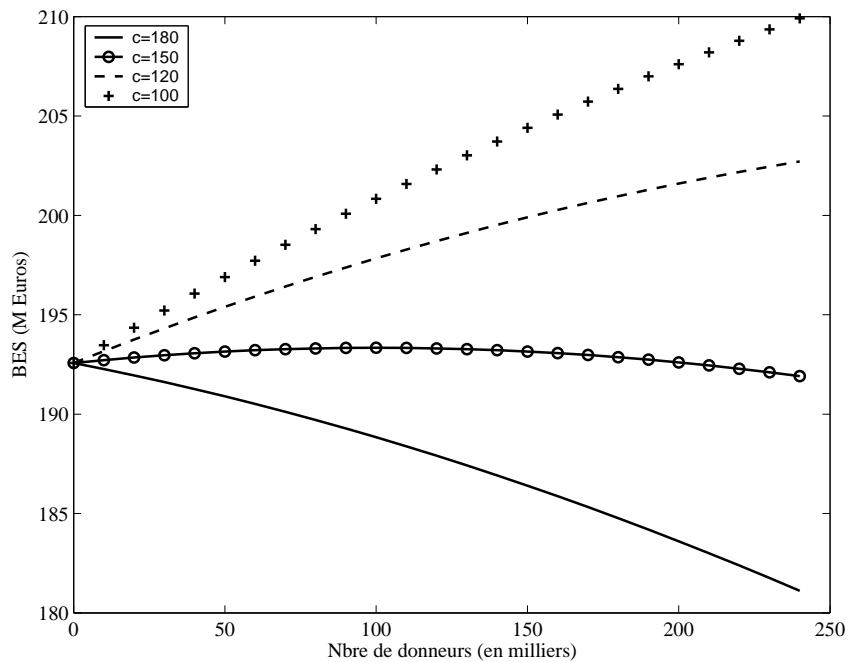


Figure 2.6: Bien-être social fonction du coût du typage

l'accroissement des registres de manière plus importante que prévue.

Discussion

Trois conclusions fondamentales se dégagent de notre étude :

- L'accroissement de la taille du registre actuel n'aura que peu d'impact sur son efficacité, donc sur le bien-être social. Malgré ce peu d'impact une valeur raisonnable des paramètres justifie le plan d'augmentation de 100 000 nouveaux donneurs sur 10 ans.
- La politique de l'agence régulatrice du système devrait être d'accroître la probabilité qu'un donneur compatible soit disponible et d'accepter la greffe (Lonjou et alii 1995, Ottinger 1994).
- De manière moins importante, toute réduction du coût marginal aurait une incidence positive sur le BES et la quantité Q optimale conduirait à une augmentation de l'efficacité du système.

2.2 Collaboration internationale entre Registres

2.2.1 Introduction

L'organisation d'un registre de donneurs de CSH au niveau d'une entité géographique (pays, région, ethnie) a une faible efficacité : la probabilité de ne pas trouver de donneur compatible avec un receveur potentiel demeure très élevée même pour des tailles de registre importantes. Ce fait a été empiriquement constaté et théoriquement démontré (cf. chapitre 1). Au niveau donc d'un seul pays on peut conclure que l'accroissement du fichier des donneurs ne se justifie pas au delà d'une certaine taille mais que d'autres mesures peuvent être seulement plus utiles (augmentation de la disponibilité des donneurs, meilleure qualité de typage, cf. paragraphe précédent). Cette analyse néglige toutefois la dimension internationale des registres : grâce à leur interconnection l'ensemble des registres peut être consulté et un donneur peut être choisi dans n'importe quel pays. Un argument en faveur de l'accroissement des registres nationaux est donc d'augmenter leur utilisation pour des receveurs étrangers. L'objectif de cette section est d'examiner cette question.

2.2.2 Le modèle

Rappelons les éléments principaux d'un modèle de registre. L'objectif d'un registre est de favoriser la mise à la disposition d'un receveur un donneur compatible disponible. Cette compatibilité est un phénomène complexe: elle comprend la compatibilité au sens du système HLA au niveau le plus fin, d'autres critères de compatibilité ainsi que la disponibilité effective du donneur. Pour atteindre cet objectif le Registre constitue une liste de donneurs dont le typage HLA est enregistré. On notera Q la taille du registre. ce typage est, en France du moins, plus grossier que celui utilisé pour la greffe (*HLA; ABDR 2 digits*). Le nombre de types enregistré au niveau 2 digits dans le registre est J et leurs fréquences dans la population des receveurs constituent le vecteur des p_j ($p_j \geq 0, \sum p_j = 1$). Un donneur compatible au sens de ces types ne le sera pas nécessairement au niveau de la décision de greffe et on notera a_j la probabilité pour qu'un

donneur compatible au premier niveau pour un receveur de type j soit pleinement compatible et disponible. En pratique on simplifiera notre modèle en supposant que a_j ne dépend pas de j et est donc égal à une valeur constante a . L'arrivée des donneurs obéit en général à un processus d'arrivée des types différents de celui des receveurs et on considèrera q_j ($q_j \geq 0$) $\sum q_j = 1$ les fréquences associées des donneurs pour chaque type. La politique du registre s'exprime en principe par grâce aux q_j (processus de sélection régionale, campagne de recrutement de donneurs auprès de minorités etc....). On a précédemment montré que:

$$\pi_0 = 1 - J\bar{p} \exp^{-\left(\frac{aQ}{J}\right)}$$

donne une borne supérieure de l'efficacité du registre.

Dans cette expression \bar{p} est la moyenne géométrique des p_j et $J\bar{p} \in [0, 1]$ mesure l'hétérogénéité des fréquences ($J\bar{p} = 1 \Leftrightarrow p_j = \frac{1}{J} \forall j$). $J\bar{p}$ faible signifie que les p_j sont tous différents entre eux. Nous considérons maintenant un modèle à 2 pays. Sans perte de généralité la liste des types sont identiques dans les deux pays et p_j^1 et p_j^2 représentent les fréquences des types dans les deux pays. ($p_j^1 = 0$ signifie que le type j n'est pas présent dans le pays 1). On notera Q_1 et Q_2 les tailles des registres et q_j^1 et q_j^2 les fréquences des donneurs. Le paramètre a sera supposé identique dans les deux pays. On supposera enfin que les proportions de la population totale sont α_1 et α_2 ($\alpha_1 + \alpha_2 = 1$) dans les deux pays et que donc la fréquence d'arrivée des malades est

$$p_j = \alpha_1 p_j^1 + \alpha_2 p_j^2$$

La probabilité pour qu'un receveur du pays 1 trouve un donneur dans son pays est

$$1 - \sum_j p_j^1 e^{-aQ_1 q_j^1}$$

et pour qu'il trouve un donneur dans l'un des deux pays est de :

$$1 - \sum_j p_j^1 e^{-a\{Q_1 q_j^1 + Q_1 q_j^2\}}$$

La probabilité pour qu'un malade quelconque trouve un donneur dans le pays 1 est:

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_1 p_j^2) e^{-a Q_1 q_j^1}$$

et celle qu'un malade quelconque trouve un donneur dans l'un des deux pays vaut:

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_1 p_j^2) e^{-a(Q_1 q_j^1 + Q_2 q_j^2)}$$

Remarque : on suppose dans notre cas que le niveau de précision des types enregistré dans les fichiers des deux pays est le même. Un modèle plus complexe serait constitué de pays ayant des niveaux de types différents (et des paramètres a différents, cf. paragraphe 2.4).

2.2.3 Les hypothèses de calibration

On considère un modèle à 1 500 types. Dans le pays 1, 1 000 types sont présents et leurs fréquences varient de 0,009 à 0,001. Pour le pays 2 on envisage trois répartitions de fréquences, appelées A, B et C représentées par le graphique 1, 2 et 3. La valeur de a est identique dans les deux pays et égale à $1/3$. On supposera le pays 1 deux fois moins peuplé que le 2 ($\alpha_1 = \frac{1}{3}$ $\alpha_2 = \frac{2}{3}$) et on supposera pour commencer que $Q_1 = 200$ et que $Q_2 = 800$ (pays à petit registre et pays à fort registre).

Les trois scénarios se distinguent par la diversité des fréquences dans les deux pays. Ils sont totalement décrits par les graphiques A, B, et C. Schématiquement, dans les deux pays, trois catégories de types existent, les fréquents, les rares et les absents. Dans le scénario A les absents du pays 1 sont les fréquents du pays 2, les rares sont identiques dans les 2 scénarii. Pour le scénario B, les fréquents sont les mêmes mais rares et absents permutent entre les 2 pays. Dans C, les fréquents du pays 1 deviennent rares pour le pays 2, les rares de 1 sont absents dans le pays 2 et les absents dans 1 deviennent fréquents

dans 2.

Le scénario B est intuitivement celui où les structures génétiques des deux populations sont les plus proches. Les deux pays partagent le même groupe de fréquents mais ne se distinguent que par leurs types rares. Le groupe C vient ensuite et le scénario A est celui du plus grand éloignement de deux pays en termes de dispersion génétique.

Figure 2.7: Scénario A: Répartition des phénotypes dans les 2 pays

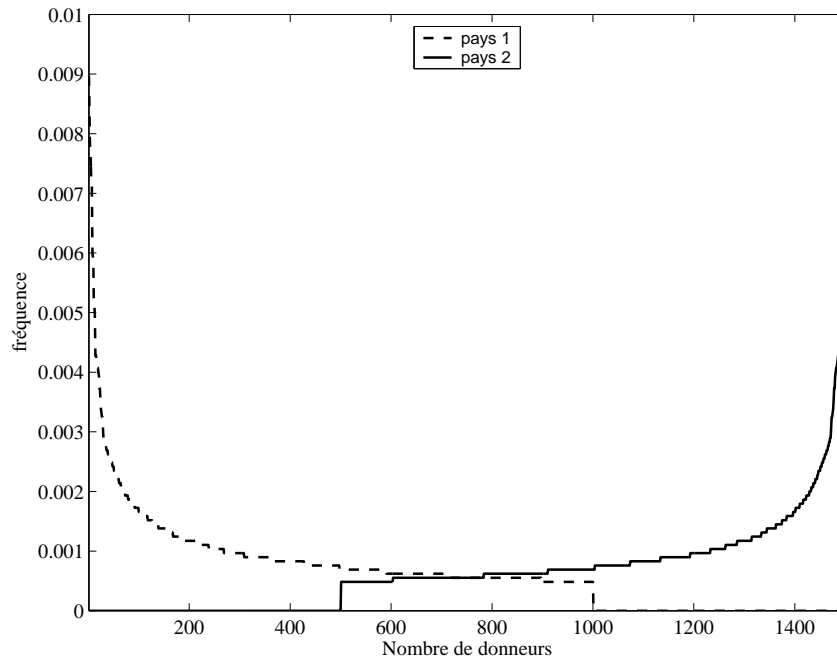


Figure 2.8: Scénario B: Répartition des phénotypes dans les 2 pays

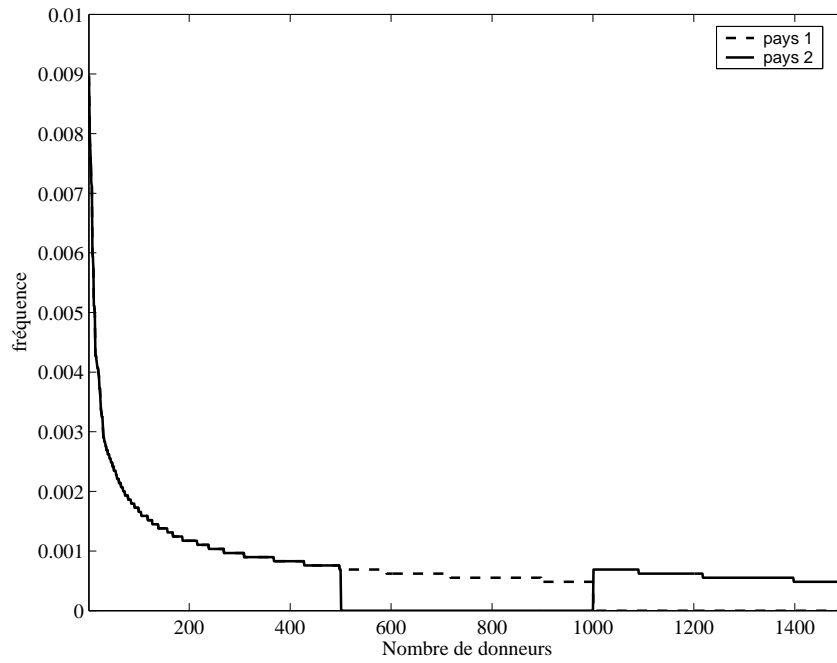
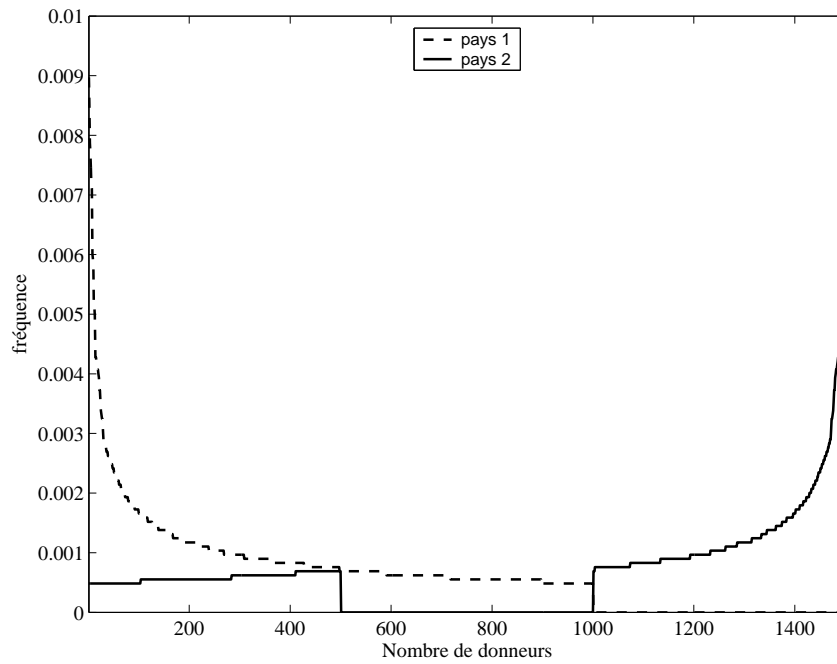


Figure 2.9: Scénario C: Répartition des phénotypes dans les 2 pays



2.2.4 Etude sans sélection des donneurs

On supposera tout d'abord que $q_j^1 = p_j^1$ et $q_j^2 = p_j^2$. On obtient les résultats suivants:

Tableau 1:

Probabilités...		$Q_1 = 200$ $Q_2 = 800$	$Q_1 = 400$ $Q_2 = 800$	$Q_1 = 200$ $Q_2 = 1000$	$Q_1 = 300$ $Q_2 = 900$
...pour qu'un receveur de 1 trouve un donneur dans 1	A	0,11	0,19	0,11	0,15
	B	0,11	0,19	0,11	0,15
	C	0,11	0,19	0,11	0,15
... pour un receveur de 1 trouve un donneur dans 1 ou 2	A	0,15	0,23	0,15	0,19
	B	0,33	0,38	0,37	0,38
	C	0,19	0,27	0,21	0,24
...pour qu'un receveur de 2 trouve un donneur dans 2	A	0,32	0,32	0,38	0,35
	B	0,32	0,32	0,38	0,35
	C	0,32	0,32	0,38	0,35
...pour qu'un receveur de 2 trouve un donneur dans 1 ou 2	A	0,33	0,34	0,38	0,36
	B	0,37	0,40	0,41	0,41
	C	0,34	0,36	0,40	0,38
...pour qu'un receveur quelconque trouve un donneur dans 1 ou 2	A	0,2704	0,304	0,308	0,307
	B	0,3554	0,395	0,398	0,397
	C	0,2930	0,331	0,333	0,333

Supposons maintenant que les deux pays "fusionnent" et que un seul fichier soit constitué de 1 000 donneurs avec $q_j = p_j$. On a:

		$Q = 1\ 000$	$Q = 1\ 200$
Probabilité pour qu'un receveur trouve un donneur dans le fichier unique	A	0,2660	0,304
	B	0,3534	0,395
	C	0,2892	0,330

Ces résultats illustrent en effet un certain nombre de propriétés valables en général au delà de nos simulations. La probabilité de trouver un donneur si les deux pays fusionnent et ne sélectionnent pas est égale à :

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_2 p_j^2) e^{-a Q(\alpha_1 p_j^1 + \alpha_2 p_j^2)} \quad (2.2.1)$$

avec $Q = Q_1 + Q_2$. L'existence de deux pays dans lesquels les donneurs arrivent sans sélection peut s'interpréter en relation avec la situation dans laquelle les deux pays fusionnent, comme un processus de filtrage. On contrôle en effet par ce moyen un critère de

sélection et on devrait idéalement optimiser, à taille donnée du fichier total, la proportion affectée à chaque pays. La solution théorique de cette optimisation n'a pas d'expression analytique simple: si Q est la taille totale du fichier, β_1 et β_2 les parts des deux registres des deux pays ($\beta_1 + \beta_2 = 1$, $\beta_1, \beta_2 \geq 0$), on devrait minimiser

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_2 p_j^2) e^{-a Q(\beta_1 p_j^1 + \beta_2 p_j^2)}$$

sous les contraintes précédentes. La résolution de la condition de premier ordre nécessite une méthode numérique.

Analysons les résultats de la simulation, dans le cas tout d'abord où $Q_1 = 200$ et $Q_2 = 800$. Pour chaque pays pris isolément, le choix du scénario n'a pas d'impact (car les types étant anonymes les 3 scénarios sont identiques quand seul le pays 2 est concerné) et le passage de 0.11 à 0.32 du pays 1 au pays 2 n'est du qu'à l'accroissement de la taille du registre. Le premier commentaire porte sur l'incidence pour un receveur de l'accès aux deux registres. Il accroît évidemment la probabilité de trouver un donneur. Il faut toutefois noter que cet accroissement est le plus important dans le cas du scénario B, puis C et A enfin. L'amélioration du système des registres est donc d'autant plus grande que les populations sont similaires. L'accès à plusieurs registres joue donc principalement par l'augmentation de la taille du registre qu'il implique plutôt que par l'accroissement de la diversité génétique qu'il permet d'obtenir.

Le second commentaire porte sur l'intérêt de maintenir deux registres plutôt qu'un seul en fusionnant les pays. La nationalité joue ici le rôle d'un filtre (non coûteux) qui même non optimisé est utile.

Le troisième commentaire répond partiellement à cette question: si on accroît l'ensemble des registres de 200 donneurs comment ceux-ci doivent-ils être répartis. Du point de vue du pays 1 il est préférable que ces donneurs augmentent le registre de ce pays et le résultat est le même pour le pays 2. Du point de vue global (probabilité pour qu'un receveur quelconque trouve un donneur dans un registre), notons que dans nos trois simulations les résultats sont très similaires entre les trois répartitions de l'accroissement. Il est très légèrement optimal de faire reporter l'accroissement uniquement sur le registre du pays 2 malgré la différence avant l'accroissement.

Notons que dans tous ces cas il est Pareto optimal de répartir l'accroissement sur les deux fichiers dans la mesure où un donneur est recherché dans les 2: dans le scénario B par exemple, passer de $Q_1 = 200$ à 400 ou augmenter Q_1 et Q_2 de 100 est équivalent pour les receveurs du pays 1 (0.38) et bénéfique pour ceux de 1.

2.2.5 Etude avec sélection des donneurs

Supposons maintenant que chaque pays pratique une sélection optimale de donneurs (voir chapitre 1) sans tenir compte de l'autre pays (maximiser la probabilité que ses résidents trouvent un donneur dans leur pays):

Tableau 2:

Probabilités...		$Q_1 = 200$ $Q_2 = 800$	$Q_1 = 400$ $Q_2 = 800$	$Q_1 = 200$ $Q_2 = 1000$	$Q_1 = 300$ $Q_2 = 900$
...pour qu'un receveur de 1 trouve un donneur dans 1	A	0,14	0,15	0,14	0,14
	B	0,14	0,15	0,14	0,14
	C	0,14	0,15	0,14	0,14
... pour un receveur de 1 trouve un donneur dans 1 ou 2	A	0,14	0,15	0,15	0,15
	B	0,39	0,40	0,41	0,41
	C	0,14	0,15	0,15	0,15
...pour qu'un receveur de 2 trouve un donneur dans 2	A	0,35	0,35	0,38	0,36
	B	0,35	0,35	0,38	0,36
	C	0,35	0,35	0,38	0,36
...pour qu'un receveur de 2 trouve un donneur dans 1 ou 2	A	0,35	0,35	0,38	0,36
	B	0,40	0,40	0,42	0,41
	C	0,38	0,39	0,41	0,40
...pour qu'un receveur quelconque trouve un donneur dans 1 ou 2	A	0,282	0,285	0,299	0,291
	B	0,396	0,402	0,416	0,408
	C	0,302	0,309	0,320	0,313

Supposons maintenant que les deux pays "fusionnent" et que un seul fichier soit constitué de 1 000 donneurs. On a:

		$Q = 1\ 000$	$Q = 1\ 200$
Probabilité pour qu'un receveur trouve un donneur dans le fichier unique	A	0,289	0,309
	B	0,382	0,462
	C	0,304	0,395

Dans le second tableau nous avons supposé que chaque pays pratique une sélection optimale de donneurs. Toutefois les q_j optimaux pouvant être négatifs nous adoptons la règle suivante: les fréquences optimales négatives sont mises à 0 et les autres fréquences sont renormalisées afin que leur somme soit égale à 1. Le passage de la table 1 à la table 2 montre l'impact de l'optimisation qui est plus grand au niveau de chaque pays isolé que globalement. Ceci découle du fait que l'optimisation a d'autant plus d'intérêt que le registre est petit. L'optimisation permet de passer de 0.11 à 0.14 pour le pays 1 seul mais seulement de 0.38 à 0.39 si les receveurs de 1 ont accès aux donneurs des 2 pays. Une première conclusion est que l'utilisation simultanée de tous les pays est un bon substitut à la selection des donneurs dans un seul pays. L'optimisation a toutefois un effet significatif si les deux pays sont proches (0.35 à 0.39 dans le scénario B) et beaucoup plus limité si les pays sont différents génétiquement. Les mêmes conclusions que précédemment s'appliquent à l'accroissement du nombre de donneurs. Dans de plus nombreux cas la répartition entre les 2 pays est Pareto optimale. Si les pays ne fusionnent pas mais ne sélectionnent pas les donneurs, on obtient:

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_2 p_j^2) e^{-a Q (\frac{Q_1}{Q} p_j^1 + \frac{Q_2}{Q} p_j^2)} \quad (2.2.2)$$

et on vérifie également qu'un choix convenable de Q_1 et de Q_2 puisse faire que (2.2.2) > (2.2.1). Si les pays optimisent indépendamment, on aura:

$$1 - \sum_j (\alpha_1 p_j^1 + \alpha_2 p_j^2) e^{-a Q (\frac{Q_1}{Q} q_j^1 + \frac{Q_2}{Q} q_j^2)} \quad (2.2.3)$$

où les q_j^1 et les q_j^2 sont les probabilités optimales de sélection de chaque pays. Enfin dans le cas fusionné et optimisé, on aura,

$$1 - \sum_i (\alpha_1 p_i^1 + \alpha_2 p_i^2) e^{-a Q q_i} \quad (2.2.4)$$

q_i minimisent la grandeur (4) qui est donc nécessairement plus grande que toutes les autres. On a clairement:

$$(2.2.4) \geq (2.2.2)$$

et

$$(2.2.4) \geq (2.2.3)$$

et

$$(2.2.2) \geq (2.2.1)$$

si Q_1 et Q_2 sont convenablement déterminés.

2.2.6 Un modèle de jeu entre plusieurs Registres

Introduction

On a examiné précédemment l'impact de la collaboration internationale sur les probabilités de trouver un donneur pour un receveur. Nous allons approfondir cette question en considérant un modèle de jeu entre les différents registres dans chaque pays ce qui permettra de mesurer l'incidence de la collaboration internationale sur les tailles optimales des registres (Speiser et alii 1994, Tiercy 2000 2003).

Le modèle

Considérons pour simplifier deux pays ayant le nombre de types J^1 et J^2 respectivement et dont les fréquences des types chez les receveurs sont p_j^1 et p_j^2 . Chaque pays choisit sa stratégie de sélection des donneurs q_j^1 et q_j^2 et la taille Q^1 et Q^2 de son registre (la taille inclut ici les stocks disponibles avant notre calcul). Le paramètre de disponibilité des donneurs compatibles est égal à a_x^1 et a_x^2 dans les deux pays. On notera M^1 et M^2 le nombre des receveurs dans la période considérée et on considère quatre valeurs V_1^1, V_1^2, V_2^1 et V_2^2 où V_a^b désigne la mise à disposition d'un donneur du pays b pour un receveur du pays a. Enfin $C_1(Q_1)$ et $C_2(Q_2)$ sont les fonctions de coût des deux pays.

Soit π_1^1 la probabilité qu'un receveur du pays 1 trouve un donneur dans le pays 1 et π_1^2 la probabilité pour qu'un receveur du pays 1 ne trouve pas de donneur dans son pays mais trouve un donneur dans le pays 2. Les nombres π_2^2 et π_2^1 sont définis de manière analogue.

On a alors:

$$\begin{aligned}\pi_1^1 &= 1 - \sum_{j \in \tau_1} p_j^1 \exp^{-a^1 Q^1 q_j^1} \\ \pi_1^2 &= \sum_{j \in \tau_1} p_j^1 \exp^{-a^1 Q^1 q_j^1} (1 - \exp^{-a^2 Q^2 q_j^2})\end{aligned}$$

où τ_1 et τ_2 sont les ensembles de types effectivement présents dans chaque pays. La valeur sociale du système des registres sera, pour le pays 1:

$$B_1^1 = M^1 V_1^1 \pi_1^1 + M^2 V_2^1 \pi_1^2 - (1 + \lambda_1) C_1(Q_1)$$

où λ_1 est le coût d'opportunité des fonds publics dans le pays 1. De manière identique, on calcule B_2 .

Plusieurs manières d'organiser la collaboration entre les pays peuvent être utilisées au niveau international. Nous nous limiterons ici à l'étude d'une solution décentralisée en cherchant l'équilibre de Nash de ce jeu. Rappelons que cette solution consiste pour chaque pays à maximiser le bien-être social de son pays en considérant les grandeurs du pays 2 données. On obtient alors un système d'équations dont la solution (si elle existe et est unique) est l'équilibre de Nash. Dans ce jeu la stratégie du pays 1 consiste dans le choix des q_j^1 et de Q_1 et celle du pays 2 dans le choix des q_j^2 et de Q_2 .

Les calculs résultent des conditions de premier ordre de l'équilibre dans ce cas sont impossibles à conduire analytiquement et leur résolution numérique est très complexe. Nous utilisons alors une approche simplifiée.

Détermination de l'équilibre en taille des registres

On a montré par ailleurs (voir Féve, Florens 2005), que dans un modèle à un seul pays l'optimisation de la probabilité de trouver un donneur est obtenue, en choisissant à Q_1 donné:

$$q_j^1 = \frac{1}{J_1} + \frac{1}{a^1 Q^1} \left\{ \ln p_j^1 - \frac{1}{J_1} \sum_l \ln p_l^1 \right\}$$

si $p_j^1 \neq 0$ et $= 0$ sinon.

Ces grandeurs sont de somme égale à 1 mais peuvent être négatives. La valeur du registre ainsi déterminée est donc une borne supérieure de la valeur effectivement réalisable.

Supposons que chacun des pays choisisse de sélectionner les donneurs suivant une règle

d'optimisation ne tenant pas compte de la collaboration internationale et adopte l'expression ci-dessus. On aura alors:

$$\pi_1^1 = 1 - H_1^1 \exp\left(\frac{-a^1 Q^1}{J_1}\right)$$

où $H_1 = J^1 \bar{p}_1$ (\bar{p}^1 =moyenne géométrique des p_j^1) est une mesure d'hétérogénéité des p_j^1 (voir Fève, Eliaou, Cambon, Florens 2005) et

$$\pi_1^2 = \sum_{j \in \tau_1 \cap \tau_2} \bar{p}^1 \exp^{-\frac{a^1 Q^1}{J_1}} (1 - \exp^{-\frac{a^2 Q^2}{J_2}} \bar{p}^2 \frac{1}{p_j^2})$$

$$\pi_1^2 = \frac{J}{J_1} H^1 \exp^{-\frac{a^1 Q^1}{J_1}} - H^1 \exp^{-\frac{a^1 Q^1}{J_1}} H^2 \exp^{-\frac{a^2 Q^2}{J_2}} \frac{1}{J_1 J_2} \sum_{j \in \tau_1 \cap \tau_2} \frac{1}{p_j^2}$$

$$\pi_1^2 = H^1 \exp^{-\frac{a^1 Q^1}{J_1}} (1 - H^2 \exp^{-\frac{a^2 Q^2}{J_2}} I_{21}) \frac{J}{J_1}$$

où J est le nombre de types communs aux deux pays et $I_{21} = \frac{1}{J_2 J} \sum_{j \in \tau_1 \cap \tau_2} \frac{1}{p_j^2}$. cet indice est le quotient de la moyenne arithmétique des p_j^2 ($\frac{1}{J_2}$ par la moyenne harmonique des P_j^2 réduite à l'intersection $\tau_1 \cap \tau_2$. $I_{21} \succeq \frac{J}{J_2}$ mesure l'inadéquation du fichier optimal du pays 2 par les receveurs du pays 1. La fonction de bien-être du pays 1 s'écrit alors:

$$B_1^1 = M_1 V_1^1 (1 - H^1 \exp^{-\frac{a^1 Q^1}{J_1}} + M_2 V_2^1 H^2 \exp^{-\frac{a^2 Q^2}{J_2}} (1 - H^1 \exp^{-\frac{a^1 Q^1}{J_1}} I_{21}) \frac{J}{J_1} - (1 + \lambda_1) C_1(Q_1)$$

. Si la fonction de coût est affine de coût marginal c_1 la condition de premier ordre pour le pays 1 s'écrit:

$$(1) \quad \frac{a^1}{J_1} M_1 V_1 X^1 + \frac{a^1}{J_1} M_2 V_2^1 I_{21} \frac{J}{J_1} X_1 X_2 = (1 + \lambda^1) c^1$$

avec $X^1 = 1 \exp^{-\frac{a^1 Q^1}{J_1}}$ et $X^2 = 1 \exp^{-\frac{a^2 Q^2}{J_2}}$. de manière symétrique la condition de premier ordre pour le pays 2 s'écrit:

$$(2) \quad \frac{a^2}{J_2} M_2 V_2 X^2 + \frac{a^2}{J_2} M_2 V_1^2 I_{12} \frac{J}{J_1} X_1 X_2 = (1 + \lambda^2) c^2$$

et l'équilibre de Nash se calcule en résolvant le système (1) et (2). Ce système est en fait du second degré mais il n'existe qu'une solution $0 \leq X_1 \leq 1$ et $0 \leq X_2 \leq 1$ d'où on tire Q_1 et Q_2 . On peut simplifier le modèle en supposant que chaque pays n'optimise son processus de sélection de donneurs que de façon aveugle en maximisant les probabilités que les receveurs nationaux trouvent un donneur. Ces calculs ne débouchent pas sur des solutions explicites. Il est donc nécessaire de rechercher des solutions numériques fonctions d'hypothèses de calibration.

2.3 Choix de la précision du typage

2.3.1 Introduction

Le Problème en l'absence de sélection des donneurs Un registre de donneurs se définit comme un ensemble de donneurs volontaires dont le type HLA est enregistré avec une certaine précision. On peut par exemple enregistrer le phénotype A,B,DR avec la précision "2 Digits" et/ou rajouter éventuellement d'autres loci (DQ, C,...). Si un receveur potentiel se présente on cherchera un donneur compatible en deux temps : on identifiera d'abord les donneurs compatibles dans le registre au niveau de la précision du fichier puis on cherchera parmi eux un donneur parfaitement compatible et disponible. La compatibilité HLA fine n'est pas suffisante pour déclencher la décision de greffe et d'autres éléments interviennent (Schuler et alii 2000, Sonnenberg et alii 1989). Pour simplifier on considèrera:

- un typage HLA "grossier" aboutissant à une compatibilité "grossière"
- un typage HLA "fin" définissant une compatibilité "fine"
- une mise à disposition d'un donneur incluant une compatibilité HLA "fine", d'autres éléments de compatibilité et la disponibilité du donneur.

On notera $j = 1, \dots, J$ les types grossiers de fréquences p_j . Un type fin sera défini par un couple (j,k) ($j = 1, \dots, J, k = 1, \dots, K$) et la fréquence de (j,k) est égale à : $p_{jk} = p_j p_{k|j}$ où $p_{k|j}$ est la probabilité conditionnelle de k sachant j . Cette écriture est très générale car $p_{k|j}$ peut être nul pour certains k et K est le nombre maximum de types fins associés à

un type grossier.

Enfin on notera D la compatibilité et disponibilité hors HLA : $D = 1$ signifie que le donneur HLA compatible et mis à disposition et $D = 0$ l'évènement contraire. On supposera que la probabilité de mise à disposition pour un receveur de type (j,k) est égale à $\alpha p_j p_{k|j}$ ce qui signifie que D est indépendant du type. Nous supposerons enfin qu'un receveur n'est identifié que par son type (j,k) .

Considérons tout d'abord le cas d'un registre constitué au niveau fin. Ainsi que nous l'avons montré précédemment, la probabilité pour qu'un receveur quelconque obtienne un donneur mis à disposition est :

$$\pi = 1 - \sum_{j=1}^J \sum_{k=1}^K p_{jk} e^{-\alpha Q p_{jk}}$$

On suppose que Q est la taille et que les donneurs sont tirés aléatoirement dans la même population que les receveurs. On n'introduit pas ici de processus de sélection des donneurs. La probabilité π ne dépend pas évidemment du choix entre un registre fin ou grossier. Dans le cas d'un registre grossier π se réécrit :

$$\pi = 1 - \sum_{j=1}^J \sum_{k=1}^K p_{jk} e^{-a_{jk} Q p_j}$$

où $a_{jk} = \alpha p_{k|j}$

On voit simplement que le passage d'un registre fin à un registre grossier fait passer du découpage α, p_{jk} (disponibilité, fréquence des types) à a_{jk}, p_j où p_j est la nouvelle fréquence des types et a_{jk} dépendant du type remplace α (cf.figure 1).

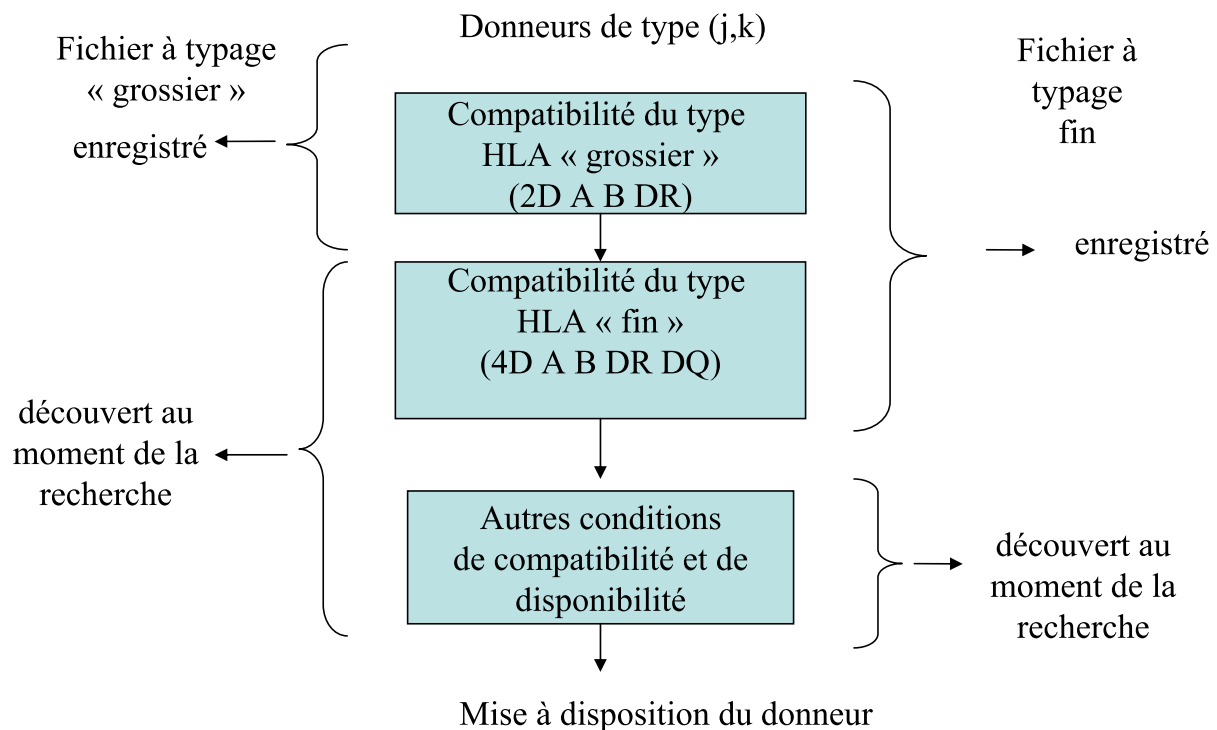


Figure 2.10: Processus d'appariement entre un donneur et un receveur

A titre d'exemple on peut considérer un cas simple (mais non réaliste) dans lequel $p_{jk} = p_j p_k$ (indépendance des 2 groupes de Digits). Dans ce cas $a_{jk} = \alpha p_k$ ne dépend que de k et :

$$\pi = 1 - \sum_{j=1}^J \sum_{k=1}^K p_j p_k e^{-a_k Q p_j}.$$

Posons $a = \sum_{k=1}^K p_k a_k$ qui mesure la probabilité pour qu'un donneur compatible à 2 Digits soit compatible à 4 (multiplié par α). Alors π peut être approché par $1 - \sum_{j=1}^J p_j p_k e^{-a Q p_j}$. En effet faisant un développement de Taylor à l'ordre 1 de $e^{-a Q p_j}$ en a, on obtient :

$$\sum_{j=1}^J \sum_{k=1}^K p_j p_k e^{-a_k Q p_j} = \sum_{j=1}^J p_j e^{-a Q p_j} + \sum_{j=1}^J \sum_{k=1}^K p_j p_k (a_k - a) e^{-a Q p_j} + \text{reste}.$$

Le second terme du membre de droite est égal à 0. Le calcul ne serait exact que si les p_k étaient identiques.

Dans la mesure où la probabilité de mise à disposition ne dépend pas de la précision du registre, les arguments de choix entre les deux organisations sont à chercher par ailleurs. Un registre fin sera plus coûteux à réaliser car le typage fin coûte plus cher que le typage grossier et donc, à budget donné un registre fin sera moins grand et donc moins efficace en

termes de d'appariements qu'un registre grossier. L'argument contre un registre grossier est la durée nécessaire pour rechercher un individu compatible. Cette durée est due aux typages nécessaires pour une éventuelle compatibilité en (j,k). Cette durée peut entraîner une dégradation de l'état de santé du receveur rendant inutile la mise à disposition du donneur. On a introduit dans le premier paragraphe du chapitre le concept de valeur de mise à disposition du donneur. Dans le cas d'un registre grossier, celle-ci est évaluée à environ 60 000-€. On considèrera donc que la valeur de la mise à disposition du donneur dans le cas d'un registre fin est plus élevé et on notera en général V_G et V_F les valeurs dans les deux cas de registre ($V_G < V_F$). Cette modélisation est naturelle car la valeur de la mise à disposition est une moyenne pondérée par toutes les valeurs des conséquences de la mise à disposition du donneur (pas de greffe, greffe effective et autres éléments de réussite). Notons que la différence $V_F - V_G$ incorpore le coût moyen de la recherche de compatibilité fine en cas de compatibilité grossière. La technologie de typage HLA a longtemps consisté en une procédure en deux temps: typage grossier d'abord plus fin ensuite. Les coûts de ces deux typages s'ajoutent dans ce cas. De manière expérimentale, un typage fin d'emblée est désormais envisagé. Pratiqué à grande échelle par des moyens optimisés il diminuerait considérablement l'écart de coût entre les deux techniques et ferait prendre la décision pour un typage fin, en cas de nouveaux entrants dans le registre.

On peut envisager le choix entre les deux registres dans le cas d'un seul pays ou en concurrence internationale. Nous testerons ici le cas d'un seul pays principalement.

2.3.2 Choix d'un niveau de typage dans un seul pays sans sélection des donneurs

Dans le cas d'un seul pays la valeur sociale d'un registre peut être évaluée (voir section 1 du chapitre) par la fonction de bien-être social (notée ci-après BES):

$$\pi(Q)MV + [PQ - C(Q)] - (1 + \lambda)PQ$$

où M est le nombre de receveurs pendant la période considérée d'utilisation du registre, P le prix d'un typage et $C(Q)$ le coût total du typage des Q donneurs. En effet πM

est le nombre espéré de mise à disposition de valeur V et πMV le bien-être pour les malades résultant du registre. $PQ - C(Q)$ est le bénéfice des laboratoires réalisant le typage et $(1 + \lambda)PQ$ représente le prélèvement réalisé sur les consommateurs pour financer le registre. Le paramètre λ représente le coûts d'opportunité des fonds publics. On suppose en général que le profit des laboratoires est nul et l'expression se ramène à $\pi MV - (1 + \lambda)C(Q)$. A partir des valeurs des différents de ce modèle, on déduit la taille optimale Q du registre.

Nous nous plaçons ici dans un cadre sans sélection de donneurs. Dans ce cas, ainsi que nous venons de le montrer, $\pi(Q)$ ne dépend pas du choix du niveau de typage. On a maintenant le choix entre deux organisations. Si le typage grossier est choisi la fonction de BES vaut :

$$\pi(Q)MV_G - (1 + \lambda)C_G(Q)$$

dont le maximum est atteint en Q_G ce qui donne un BES de

$$BES_G = \pi(Q_G)MV_G - (1 + \lambda)C_G(Q_G).$$

De la même manière si une organisation avec registre fin est choisie, on aura une fonction de BES écrivant:

$$\pi(Q)MV_F - (1 + \lambda)C_F(Q)$$

dont le maximum est atteint en Q_F qui détermine une valeur du BES de:

$$BES_F = \pi(Q_F)MV_F - (1 + \lambda)C_F(Q_F).$$

Le choix entre les deux techniques se fait donc en comparant BES_F et BES_G . Ce choix ne dépendra que de l'écart entre V_F et V_G et de l'écart entre les fonctions de coût des deux niveaux de typage. Ce problème n'a pas de solution explicite et doit faire l'objet d'une étude numérique.

2.3.3 Comparaison des registres optimaux dans les cas 2 Digits et 4 Digits par la probabilité de mise à disposition

En utilisant les notations de la section précédente, on vérifie que dans le cas d'un fichier à 4 Digits la probabilité maximale de mise à la disposition d'un donneur par un receveur

quelconque est:

$$1 - J_F \bar{p}_F \exp\left(\frac{-\alpha Q}{J_F}\right)$$

où J_F est le nombre de types (à 4D) de probabilités non nulles

$$(J_F \leq J \times K),$$

\bar{p}_F la moyenne géométrique des fréquences des types à 4D, Q la taille du Registre. Considérons maintenant un modèle à 2 Digits, le registre optimal va consister à maximiser:

$$1 - \sum_{j,k} p_{jk} \exp(-\alpha p_{k|j} Q q_j)$$

en les q_j . En d'autres termes dans un registre à 4 Digits, on optimise q_{jk} alors qu'ici on n'optimise qu'en q_j avec $q_{k|j} = p_{j|k}$. L'optimisation de cette expression ne conduit pas à une solution analytique. On peut toutefois faire un calcul approché en approchant $\alpha p_{k|j}$ par $a_j = \alpha \sum p_{k|j}^2$. On a vu que dans ce cas on réalise une approximation telle que la dérivée de l'erreur soit égale à 0. Posons $a_j = \alpha \sum p_{k|j}^2$,

$$\tilde{p} = \frac{\sum_{j=1}^J \frac{\ln p_j a_j}{a_j}}{\sum_{j=1}^J \frac{1}{a_j}}, \quad \tilde{J} = \sum_{j=1}^J \frac{1}{a_j}.$$

On montre alors que la probabilité maximale de mise à la disposition d'un donneur est:

$$1 - \tilde{J} \tilde{p} \exp\left(-\frac{\tilde{Q}}{\tilde{J}}\right).$$

Cette valeur est nécessairement inférieure (pour la même valeur de Q) à la valeur optimale du fichier à 4 Digits. On a en effet optimisé sur un ensemble plus petit dans le second cas que dans le premier.

A titre d'exemple on considère le Registre français: on ne connaît pas la distribution des types à 4 Digits et on est conduit à effectuer de fortes hypothèses. Supposons par exemple que à chaque type 2D correspond 3 types 4D de probabilité conditionnelle 0.4, 0.4, et 0.2 quel que soit le type ($p_{k|j} = 0.4$ si $k = 1$ ou 2 et 0.2 si $k = 3 \forall j = 1, \dots, J$). Dans ce cas $a_j = \alpha(0.4^2 + 0.4^2 + 0.2^2)$ ne dépend pas de j . Cela signifie que la probabilité pour qu'un donneur compatible à 4 Digits soit disponible est 0.87 et que cette probabilité tombe à 0.33 si l'on sait que la compatibilité n'est réalisée qu'à 2 Digits. L'écart entre ces

deux nombres résulte de la compatibilité 2D à 4D (36% des compatibles 2D deviennent comparables 4D. Ces chiffres sont arbitraires mais ne semblent pas contredire les intuitions des spécialistes. On a estimé par ailleurs le précédent nombre de types 2D multiplié par la moyenne géométrique des p_j à 0.7. Le nombre de types 4D est 3 fois plus grand et la moyenne géométrique des p_{jk} est égale au produit des moyennes géométriques des p_j par celle de la distribution des 2 derniers digits. On a donc $0.95 * 0.7 = J_F \bar{p}_F$ et donc $J_F \bar{p}_F = 0.664$.

Dans notre exemple simple, l'indicateur d'hétérogénéité des fréquences qui est $J\bar{p}$ passe de 0.7 à 0.664 quand le type passe de 2 à 4 Digits. On doit donc comparer:

$$1 - 0.664 \exp\left(\frac{-0.97 (Q_0 + Q)}{1\ 500\ 000}\right)$$

(sous l'hypothèse de 500 000 types à 2Digits) et

$$1 - 0.7 \exp\left(\frac{-0.33(Q_0 + Q)}{500\ 000}\right).$$

On considère le cas $Q = 220\ 000$ (120 000 dans le stock du fichier et 100 000 d'accroissement), on obtient respectivement 0.42 et 0.395. Cet écart montre l'impact maximal de la sélection optimale des donneurs à 4D et à 2D. Il est naturel que cette sélection optimale soit plus efficace au niveau 4D qu'au niveau 2D.

L'arbitrage entre 2D et 4D en matière d'organisation du fichier tient aux autres éléments que sont le coût du typage et la valeur des appariements dans les 2 cas. Ces trois éléments vont être plongés dans le modèle suivant.

2.3.4 Comparaison des registres optimaux dans les cas 2 Digits et 4 Digits par le bien-être social

Nous utilisons le modèle de la section 1 du chapitre. L'accroissement optimal du registre est:

$$Q = \frac{J}{a} \ln \frac{VMha}{c(1+\lambda)J} - Q_0$$

qui correspond à une valeur du bien-être-social de

$$VM \pi(Q + Q_0) - (1 + \lambda) C(Q),$$

c'est-à-dire à la valeur d'un appariement multiplié par leur nombre espéré moins le coût du registre pondéré par le coût d'opportunité des fonds publics.

On peut calculer ces deux grandeurs dans le cadre dans le cadre d'un typage grossier et fin. On a donc, dans le cas d'un registre à 2 Digits:

$$Q = \frac{500\,000}{1/3} \ln \frac{55\,831 * 10\,000 * 0.7 * (1/3)}{150 * 1.5 * 500\,000} - 120\,000 = 100\,000$$

$$BES = 55\,831 * 10\,000 * (1 - 0.7 \exp(-220\,000 / 1\,500\,000)) - 1.5 * 183 * 100\,000 = 194M \text{ d'€}.$$

Dans le cas du typage fin, on a déterminé dans notre exemple $h = 0.664$, $\tilde{J} = 1\,500\,000$, a devient $\alpha = 0.92$. Les deux données manquantes pour déterminer la taille optimale du fichier sont le coût marginal du typage à 4D et la valeur d'une mise à disposition dans un registre à 4D. On peut supposer tout d'abord que le coût marginal du typage à 4D est la somme du coût marginal du typage à 2D(183€) plus un incrément correspondant au coût marginal du séquençage que l'on supposera de 100€. Le coût total du typage est égal à $Q (180 + 183) + 120\,000 * 180$.

Accroissement optimal, probabilité de mise à disposition d'un donneur et bien-être-social en fonction de la valeur d'un appariement (au niveau fin)

V_F	Q	π	BES	COUT
115 000	37 388	0.40	753 092 093	35 171 793
125 000	96 429	0.42	819 604 294	56 603 889
135 000	150 925	0.44	887 261 464	76 385 666
145 000	201 524	0.45	955 943 668	94 753 209
155 000	248 747	0.47	1 025 550 854	111 895 310
165 000	293 017	0.48	1 095 998 498	127 965 308
175 000	334 682	0.50	1 167 214 403	143 089 449
185 000	374 030	0.51	1 239 136 309	157 372 915
195 000	411 307	0.52	1 311 710 065	170 904 262

Il est nécessaire d'évaluer V_F à une grandeur de l'ordre du double de V_G pour obtenir un accroissement positif du fichier. Un accroissement de 100 000 donneurs environ, typés à 4D et assorti d'un retypage des 120 000 en stock ne se justifie que si V_F est 2 fois et demi celui de V_G . Dans ce cas, le coût total pour la'Agence est de 56M€ (contre 183M€ dans le cas 2D). Le BES est alors largement supérieur au BES optimal du typage 2D.

2.3.5 Choix entre un registre à 2 Digits et un registre à 4 Digits dans un contexte de concurrence internationale

Considérons une situation réduite à 2 pays seulement:

- une distribution des types (j,k) (j=deux premiers Digits et k les deux dernières) chez les receveurs définie par ses fréquences $p_j k$
- un processus de sélection des donneurs $q_j k^2$ et une taille Q_2 .
- un paramètre α_2 égal à la probabilité pour qu'un donneur compatible à 4 Digits soit effectivement disponible pour une greffe.

Si le pays 1 choisit une stratégie de Registre à 4 Digits, ses paramètres sont analogues aux précédents et sont affectés de l'indice 1. Notons:

π_1^1 la probabilité pour un receveur de trouver un donneur en 1.

π_1^2 la probabilité pour qu'il trouve un donneur en 2 (sachant qu'il n'en a pas trouvé en 1).

π_2^1 et π_2^2 sont les probabilités pour qu'un receveur du pays 2 trouve un donneur dans les

pays 1 et 2 respectivement. On suppose que si les donneurs sont disponibles dans les deux pays, on choisit en priorité celui d'origine du receveur. On a:

$$\pi_1^1 = 1 - \sum_{j,k} p_{jk}^1 \exp^{-\alpha_1 Q_1 q_{jk}^1}$$

$$\pi_1^2 = 1 - \sum_{j,k} p_{jk}^1 \exp^{-\alpha_1 Q_1 q_{jk}^1} (1 - \exp^{-\alpha_2 Q_2 q_{jk}^2})$$

et

$$\pi_1 = \pi_1^1 + \pi_1^2$$

est la probabilité pour qu'un receveur du pays 1 trouve un donneur disponible. Des calculs analogues permettent de calculer π_2^1, π_2^2 et π_2 .

Différentes spécifications des valeurs des registres peuvent être définis. Imaginons par exemple le cas où la valeur du registre 1 s'écrit:

$$B_1 = M_1 V_1 \pi_1^1 + M_2 V_2^1 \pi_2^1 - (1 + \lambda_1) C_1(Q_1)$$

où M_1 et M_2 sont les nombres de receveurs des pays 1 et 2, V_1^1 et V_2^1 sont les valeurs pour le registre du pays 1 de trouver un donneur indigène ou étranger, $C_1(Q_1)$ le coût du registre, et λ le coût d'opportunité des fonds publics. Le registre 1 optimise B_1 en fonction des $q_j k^1$ et de Q_1 à $q_j k^2$ et Q_2 donné et le registre 2 effectuera de même. La solution de ce système donne l'équilibre de Nash du jeu et donc la valeur optimale, pour le pays 1 d'aménager un registre à 4 Digits.

Si le pays 1 décide de ne se doter que d'un registre à 2 Digits, les probabilités $\pi_1^1, \pi_1^2, \pi_2^1, \pi_2^2$ deviennent:

$$\pi_1^1 = 1 - \sum_{j,k} p_{jk}^1 \exp^{-\alpha_1 p_{k|j} Q_1 q_j^1}$$

$$\pi_1^2 = 1 - \sum_{j,k} p_{jk}^1 \exp^{-\alpha_1 p_{k|j} Q_1 q_j^1} (1 - \exp^{-\alpha_2 Q_2 q_{jk}^2}).$$

Si V_1^1 et V_1^2 sont les nouvelles valeurs de mise à disposition (V_2^2 n'est pas affecté), ($\bar{V}_1^1 < V_1^1$ $\bar{V}_1^2 < V_1^2$ $\bar{V}_2^1 < V_2^1$) et \bar{c}_1 la fonction de coût d'un registre à 2 Digits, la valeur du

registre pour le pays 1 devient:

$$\bar{B}_1 = M_1 \bar{V}_1 \bar{\pi}_1 + M_2 V_2^1 \pi_2^1 - (1 + \lambda_1) \bar{C}_1(Q_1)$$

que le pays 1 optimisera en q_j^1 et Q_1 . Le pays 2 optimisera en q_{jk}^2 et Q_2 et on trouvera l'équilibre de Nash du jeu d'où on déduira \bar{B}_1 à l'équilibre. La décision de faire un registre 2 Digits ou 4 Digits résultera enfin de la comparaison entre B_1 et \bar{B}_1 . Si $B_1 > \bar{B}_1$ le pays choisira le registre 4D et 2 D dans le cas contraire. Dans ce cas de nouveau le modèle n'admet pas de solution analytique et doit faire l'objet de simulations.

On peut imaginer d'autres modèles dans lesquels par exemple la compétition entre les 2 registres est remplacée par une autorité internationale qui coordonne leurs actions.

2.4 ANNEXE: Un cadre analytique de référence

L'objectif de cette annexe est de fournir des éléments quantitatifs observés pour la calibration des coûts du typage dans les différents modèles utilisés. En l'absence de bases de données il était impossible de procéder à une analyse économétrique des coûts telle qu'elle se pratique dans d'autres domaines (Cazals, Fève F., Fève P., Florens J.P. 2005, Fève F., Florens J.P., Roy B. 2005). La seule démarche possible consistait alors à procéder à une sorte de comptabilité analytique des coûts dans quelques laboratoires. cette démarche a bien entendu plusieurs limites:

- elle ne construit que des fonctions linéaires qui ne sont que des approximations autour de la situation actuelle: on peut ainsi extrapoler le coût du passage de 1 000 typages à 1 100 typages mais pas à 10 000 qui conduirait à un changement technique important. Par nature de telles fonctions ont des coûts marginaux constants et un coût moyen mécaniquement décroissant.
- elle se heurte à la question récurrente de la répartition des coûts au sein des laboratoires ayant des activités multiples.
- l'enquête réalisée au sein de laboratoires pose la question du biais dans l'information fournie. Ce biais est multiple et il est en particulier déterminé par l'existence du prix officiel du typage de 183.13€. L'affectation des coûts fixes que nous présentons résulte de cette contrainte et il ne faut pas s'étonner du fait que notre coût moyen se retrouve très proche de cette grandeur. La volonté de ne pas divulguer les coûts réels (et leur ignorance) conduit in fine à s'aligner sur le prix fixé par l'administration.

L'évaluation économique d'un acte de laboratoire ou plus généralement d'un programme de santé, a pour but d'allouer de façon rationnelle les ressources du système de santé. Aucune action du système de santé ne peut être menée sans conséquences financières, budgétaires et donc sans coût à supporter. Si les systèmes d'information hospitaliers connaissent le volume consommé par chaque laboratoire, ils ne fournissent pas ce type de renseignement ou très partiellement à un niveau plus désagrégé. Le volume de la consommation de réactifs par type d'examen de laboratoire reste le plus souvent une

donnée non maîtrisée : il s'avère indispensable de construire l'information nécessaire à l'évaluation économique.

On décrit ici l'étude empirique réalisée pour MADO: elle a été mise en oeuvre en association avec des techniciens, biologistes et professeurs de médecine responsables de laboratoires hospitaliers.¹¹ Si l'étude économique avait été faite en dehors de tout contexte spécifique, les quantités auraient du être estimées à partir des bases de données comptables hospitalières. Si complètes soient-elles, ces bases ne contiennent aucune information sur le temps de travail ou sur la dépréciation du capital.

Le coût d'un typage HLA dépend principalement de deux éléments, d'une part de la technique de typage retenue et d'autre part du contexte dans lequel il est mis en oeuvre (structure hospitalière ou laboratoire privé). L'objectif de l'étude empirique réalisée est d'explicitier et de ventiler le coût du typage HLA entre coûts fixes et coût marginal. On ne s'intéresse ici qu'aux laboratoires d'Immunologie et d'Histocompatibilité, appartenant à une structure hospitalière, pratiquant des typages HLA pour les DVM¹² mais aussi à des fins de transplantation et de greffes d'organes. Les options de l'économiste sont ainsi souvent limitées par la disponibilité des données (type de laboratoire, type d'information fournie). Cette étude réalisée en laboratoire nous a appris que, de par leur fiabilité et leur pouvoir de résolution, les techniques de biologie moléculaire ont progressivement remplacé, dans la pratique courante, la sérologie, tout au moins pour l'analyse du polymorphisme HLA de classe II (HLA -DR, -DQ, -DP). Le nombre croissant de kits commerciaux de typage HLA classe I et classe II disponibles sur le marché témoigne de cette tendance. Deux niveaux de résolution peuvent être atteints par les techniques de typage HLA basées sur l'analyse de l'ADN. Classiquement, dans un premier temps, la recherche de donneurs potentiels dans les Registres est réalisée en comparant le typage HLA de classe I (HLA-A, -B) et de classe II (HLA -DR et/ou -DQ) de basse ou de moyenne résolution des donneurs et des receveurs. En cas d'appariement à ce stade, le polymorphisme HLA de classe I

¹¹Toutes les données sur les quantités de ressources consommées figurent dans les cahiers d'observation tenus à jour par chaque technicien

¹²Donneurs Volontaires de Moelle

(HLA-A, -B, -C) et de classe II (HLA-DR, -DQ, -DP) est analysé de façon exhaustive aboutissant au typage HLA de haute résolution indispensable à l'appariement entre les receveurs et les donneurs de CSH sélectionnés. Les stratégies de typage HLA varient selon les laboratoires en fonction de l'expertise(l'expérience) locale et du type d'équipement disponible. Trois principales approches technologiques existent à l'heure actuelle:

1. L'amplification de l'ADN génomique par PCR (Polymerase Chain Reaction) avec des amorces oligonucléotidiques spécifiques de séquences situées dans des régions polymorphes (PCR-SSP pour *Sequence-specific primers*)
2. L'hybridation de produits de PCR avec des sondes oligonucléotidiques spécifiques de séquences (PCR-SSO pour *Sequence-specific Oligonucleotides*).

Ces deux techniques, les plus largement utilisées, un niveau de basse ou de résolution intermédiaire est atteint facilement. Cependant, des amplifications par PCR supplémentaires avec des amorces spécifiques ou des sondes additionnelles sont nécessaires pour réaliser des typages de haute résolution HLA de classe I et II. Ces approches présentent deux principaux inconvénients: la nécessité de plusieurs étapes pour aboutir à la définition optimale du polymorphisme HLA et le fait que la définition des allèles HLA est obtenue sur la base d'une information partielle, correspondant aux zones d'hybridation des sondes oligonucléotidiques à l'ADN génomique.

3. Une troisième approche technologique est basée sur le séquençage des gènes HLA amplifiés par PCR. Ce typage par séquençage (PCR-SBT pour *Sequence-Based Typing*) permet l'identification précise de tous les nucléotides de la région d'ADN amplifiée. Appliqué au typage HLA de haute résolution, le SBT peut être utilisée d'emblée pour un typage HLA complet étant donné ses potentialités techniques et ses possibilités de haut débit (d'où la nécessité de plusieurs étapes pour aboutir à la définition optimale du polymorphisme HLA). Le SBT permet l'identification de tous les nucléotides de la région d'ADN amplifiée.

2.4.1 La collecte des données

Il existe a priori trois méthodes non nécessairement exclusives l'une de l'autre¹³ : l'enregistrement rétrospectif des données à partir des dossiers du laboratoire qui effectue les typages, le protocole de typage "standardisé", l'observation directe en temps réel.

L'enregistrement rétrospectif

Il s'agit d'aller rechercher dans les dossiers tenus par les laboratoires, les informations nécessaires à l'étude économique; En général, toutes les informations sont disponibles à des endroits divers (un cahier pour l'extraction d'ADN, un cahier pour la PCR, technicienne du matin et/ou de l'après-midi, un cahier de validation du biologiste, un ordinateur PC de laboratoire,...). Bien entendu, il n'est pas prévu que tous ces cahiers puissent servir à ce type d'évaluation. Aussi, l'information disponible d'un dossier à l'autre, a fortiori d'un laboratoire à l'autre est hétérogène. On se heurte nécessairement à l'inégale qualité dans la tenue du dossier et le fait qu'il soit rempli par des intervenants multiples ne facilite pas la collecte des données. de surcroît, l'économiste est confronté à la difficulté d'exploitation des données par un "non technicien" de laboratoire. En revanche, cette méthode présente des avantages certains. Il est possible de travailler sur des populations représentatives, éventuellement obtenues par tirage au sort. On peut ainsi suivre le nombre d'actes de laboratoire sur d'assez longues périodes et la taille de l'échantillon nécessaire n'est plus un problème, car il suffit de remonter dans le temps pour obtenir autant de dossiers que nécessaire, en espérant toutefois que les techniques n'aient pas trop évolué; Inconvénient de ce type de méthode : les études réalisées par ce biais-là sont nécessairement rétrospectives. D'où la nécessité que les pratiques restent relativement constantes dans le temps.

Les protocoles de laboratoire standardisés

Ce sont les protocoles techniques (de typage HLA par exemple) définis par les laboratoires. Pour être utilisables, encore faut-il que ces protocoles existent dans chaque laboratoire et

¹³voir Drummond, 1997

qu'ils précisent les quantités consommées pour la réalisation de l'acte. Les résultats obtenus sont nécessairement "moyens" et laissent échapper toute spécificité liée au laboratoire qui emploierait une "technique maison" : certaines techniques "maison", bien que très fiables en termes de résultats, sont souvent coûteuses en temps de travail. Ni le temps de travail, ni le coût du capital ne sont envisagés par un protocole standardisé; les valeurs des besoins en quantités de facteur travail et en capital constituent dans ce cas des données manquantes.

L'observation directe en temps réel

C'est la plus adaptée à l'évaluation économique du coût du typage HLA ¹⁴. Il s'agit de mettre un observateur dans le laboratoire où s'effectue le typage HLA, la cas échéant avec un chronomètre, et de mesurer pendant une période de temps donné, tout ce qu'il se passe dans le laboratoire : aussi bien en inscrivant scrupuleusement le déroulement des actes et leur nature, mais également le temps consacré à la réalisation de l'acte. Les différentes étapes de la manipulation (préparation des paillasses, extraction d'ADN, PCR,...) suppose une disponibilité absolue de l'observateur, même si une grille de précodage des actes ¹⁵ est prévue afin de faciliter la tâche. Les structures de coût peuvent être différentes d'un établissement à l'autre; des économies d'échelle sont théoriquement possibles dans un grand établissement; par exemple, la paire de gants aura un coût unitaire très inférieur dans une petite structure à ce qu'il sera dans un grand centre hospitalo-universitaire. Le supplément de précision d'une évaluation de coût suppose une collecte d'informations non négligeable; dans le cas de la PCR, il s'agit de chronométrer le temps de manipulation de la technicienne chargée du typage (certaines "techniques" maison sont plus longues à mettre en oeuvre et donc plus coûteuse que d'autres). Ce travail d'agrégation des coûts devient difficile, dès lors que l'on est confronté à une disparité, parfois importante, des pratiques. Que dire d'une extraction d'ADN réalisée en quarante minutes par une technicienne de laboratoire chevronnée et du même acte réalisé par un interne ou même par une technicienne "plus lente" ? Que dire de la logique de calcul des coûts dans une

¹⁴l'acte de laboratoire sur lequel on a effectivement ravaillé

¹⁵voir la grille d'analyse et d'évaluation du coût du typage HLA

institution publique ou privée ? les charges de personnel sont le plus souvent inférieures dans le privé à ce qu'elles ne sont dans le public (personnel moins nombreux, alors même que le biologiste ou le Professeur d'immunologie chef de service, est en général mieux rémunéré).

La méthodologie retenue ici repose sur l'observation en temps réel des actes. Il s'agit de collecter toutes les données disponibles pour ensuite chiffrer en termes de coûts les quantités consommées. La seconde étape de l'étude consiste à appliquer le protocole retenu (le même dans chaque laboratoire visité¹⁶) à l'aide des informations empiriques collectées.

2.4.2 Le choix des inputs de la grille d'analyse et la calibration du modèle

La collecte des données nécessaires à l'évaluation du coût du typage HLA se heurte d'emblée à la suspicion du Professeur de médecine responsable du laboratoire visité. Un a priori défavorable mais compréhensible précède la collecte des informations. D'une part, le responsable d'un laboratoire d'immunologie, appartenant de surcroît, au moins en partie au secteur public, suspecte instantanément un contrôle : contrôle des autorités de tutelle (hospitalières ou non) sur la gestion des fonds attribués au laboratoire chaque année. La seconde explication relève davantage de l'éthique du domaine de la santé : "La santé n'est pas une histoire de coûts: les préoccupations économiques ne devraient jamais être soumises aux médecins". Le laboratoire ayant donné son accord pour la visite de ses installations, le protocole retenu pour l'étude (rigoureusement identique à chaque fois) est le suivant :

Ayant satisfait aux conditions d'hygiène et de sécurité (blouse blanche, gants,...) du laboratoire, le "collecteur d'informations" assiste comme observateur extérieur à la journée (ou aux journées) de travail des techniciens chargés du typage HLA. Cette phase d'observation est retranscrite au fur et à mesure: l'examen visuel de chacune des étapes du typage HLA permet de décrire par écrit la totalité de la manipulation ainsi que la collecte des infor-

¹⁶pour des raisons de confidentialité, aucun nom de laboratoire ne sera cité

mations suivantes¹⁷:

- Le nombre de typages HLA réalisés dans l'année: typages sérologiques et ou réalisés en biologie moléculaire; combien de typages alléliques ? Le laboratoire fait-il du séquençage ?
- La description de la (ou des) technique(s) utilisée(s): s'agit-il d'une technique "maison", plus commerciale (Luminex par exemple), le laboratoire utilise t-il plusieurs techniques ?
- Le nombre de prélèvements sanguins effectués dans l'année par le laboratoire ainsi que le nombre de prélèvements effectués à l'extérieur du laboratoire (le cas échéant)
- Le nombre d'extractions d'ADN réalisées dans l'année par le laboratoire (et/ou par un établissement extérieur)
- La liste de consommables et de réactifs: Le nombre exact utilisé lors de chaque manipulation ou acte¹⁸
- Le nombre de personnes réalisant les typages, ainsi que leurs qualifications (techniciens, biologistes, secrétaires), et la durée de leur temps de travail effectivement consacré au typage
- La liste et le taux d'utilisation du matériel employé. La centrifugeuse est-elle exclusivement réservée au typage HLA ? L'utilisation du séquenceur (si séquenceur il y a) est-elle uniquement dédiée au typage ?
- La maintenance du matériel: l'hôpital a t-il un contrat de maintenance avec un organisme extérieur ? Quel est son montant ? Quel % faut-il affecter à l'entretien des machines affectées au typage ?
- La superficie des locaux dédiées au typage (en m^2)

¹⁷voir en Annexe la grille d'analyse et d'évaluation du coût du typage HLA

¹⁸terme générique utilisé dans la grille: plusieurs typages sont effectués en même temps, 500 typages pour un acte dans notre exemple

- Le traitement des déchets: Quel est le montant annuel payé par l'hôpital pour le traitement des déchets ? Quelle partie faut-il attribuer au typage ?
- Prestations diverses : essentiellement les charges de ménage payées par l'hôpital et la part attribuée au laboratoire qui réalise les typages, ainsi que les charges annuelles d'électricité et de téléphone. Cette dernière catégorie d'information nous a été fournie par l'administration de l'hôpital et non par le responsable du laboratoire.

Les données retenues comme inputs pour "la calibration de la grille d'analyse" sont des variables exogènes, identiques pour chaque évaluation, permettant une comparaison directe de tous les résultats. Il s'agit des inputs suivants :

- Le coût horaire (charges sociales incluses) s'élève à 30€ pour une secrétaire, 32€ pour un technicien, 50€ pour un biologiste. Le coût horaire est évalué en divisant le coût annuel par 52 semaines et par 35 heures
- Les hypothèses de travail tiennent compte de la structure de l'établissement visité (hospitalier, Etablissement Français du Sang, Hospitalo-Universitaire). Les coûts d'investissement, plus généralement appelés coûts d'exploitation, correspondent à l'acquisition de biens en capital: il s'agit généralement d'équipements techniques (machines PCR¹⁹), de bâtiments, (laboratoires); Parfois les budgets annuels et les comptes contiennent un indicateur, l'amortissement (dépréciation). Diverses procédures comptables (linéaires, dégressive,...) sont utilisables pour évaluer la dépréciation des machines. On retiendra un amortissement du capital linéaire sur 5 ans; On considère d'autre part que l'hôpital ne se déprécie pas du tout
- Le coût d'utilisation des locaux est de 30€ par an par m^2 , soit un loyer annuel de 9€ pour un typage

Toutes ces données (empiriques: la collecte) et les hypothèses de travail retenues comme inputs, nous ont permis d'évaluer le coût d'un typage HLA pour un donneur volontaire

¹⁹Polymerase Chain Reaction

de moelle dans un laboratoire de type hospitalier, tel qu'il est réalisé aujourd'hui²⁰. Le coût unitaire indiqué dans la grille est ramené à l'acte : exemple des pipettes, on utilise 4 pipettes par typage, chaque pipette coûte 0.05€ d'où le coût en pipettes pour 1 typage qui s'élève à 0.2€. Le "temps passé sur l'acte" est le temps nécessaire à la réalisation du typage, il est exprimé en heures. Si le laboratoire emploie une technicienne dédiée au typage HLA, celle-ci a besoin de 0.8 heures pour réaliser un typage, soit²¹ un coût du travail par typage égal à 25.60€. Le laboratoire bénéficie également des services d'une secrétaire et de ceux d'un biologiste dont les salaires annuels, charges sociales incluses, sont égaux respectivement à 54 600€ et 175€. La grille indique le coût unitaire de chaque réactif employé ainsi que celui de chaque machine utilisée. Le taux d'utilisation de chaque type de matériel est également indiqué.

On calcule le coût unitaire pour un typage comme étant la somme d'un coût variable, fonction du nombre de typages réalisés, (coût marginal) et d'un coût fixe, indépendant lui par définition du nombre de typages réalisés dans l'année par le laboratoire. La quantité de consommables et le nombre d'heures de travail nécessaires au typage sont "variables" en fonction du nombre de typages. Si pour la quantité de réactifs utilisée, cela semble évident (plus on type d'individus, plus on a besoin de réactifs), on remarque ici que le nombre d'heures nécessaires au typage est proportionnel au nombre d'examen réalisés. Le modèle évalue le coût d'un typage HLA à 183€. Ces 183€ font apparaître 150€ de coût marginal. La différence est égale au montant des coûts fixes. Pour calculer ces coûts fixes, calculés par poste (maintenance, locaux, traitement des déchets, entretien, électricité et téléphone), sachant que l'hôpital a un contrat de maintenance avec une entreprise extérieure pour le matériel, on multiplie le "% de chaque poste affecté au typage" par le coût annuel du contrat et on divise le résultat obtenu par le nombre de typages réalisés dans l'année : soit, si l'on retient l'exemple du contrat de maintenance, un coût fixe annuel de 12 000€ correspondant à un coût fixe par typage de 3.60€, à ajouter par typage au coût marginal de 149.62€. A chacun des postes de coût fixe,

²⁰Lors du recrutement d'un donneur (typage + inscription sur le registre), le centre ayant effectué le recrutement reçoit un montant forfaitaire de 183 13€

²¹sachant que le coût horaire d'une technicienne s'élève à 32€ charges incluses

correspond un "% d'affectation du coût au typage HLA". On détermine ainsi un coût fixe par typage HLA égal à 33.83€. Le coût (unitaire) d'un typage HLA est égal à la somme du coût marginal 149.62€ et d'un coût fixe 33.83€, soit 183.45€. Dans notre modèle, ce coût du typage de 183€ doit s'interpréter comme la moyenne arithmétique des coûts observés dans chaque laboratoire visité. Compte tenu des différentes techniques utilisées, la disparité du coût du typage HLA est grande. On a d'abord calculé un coût direct du typage, soit 150€, et on a ensuite calibré la décomposition des coûts indirects en ventilant l'écart entre 183€ et 150€ entre différents postes de dépenses. Cette calibration "ad-hoc" nous a permis de ne pas trahir la confidentialité des données. La grille permet en outre d'évaluer le montant des coûts fixes annuels du laboratoire :

- Pour le matériel, on comptabilise comme coût fixe la "part du matériel non affectée au typage". Soit, en prenant l'exemple de la centrifugeuse, $5\,000\text{€} \times 10\% = 1\,000\text{€}$ (montant égal au coût unitaire de la centrifugeuse divisé par la durée d'amortissement retenue).
- Pour les quatre postes restants (maintenance, locaux, traitement des déchets et divers), on comptabilise comme coût fixe annuel, le coût réel pour l'hôpital (indépendamment de la part du poste affectée au typage HLA car même si le laboratoire ne réalise aucun typage, le laboratoire aura à supporter le coût annuel du contrat de maintenance souscrit (12 000€)).

Le calcul du coût du travail, du coût du capital (machines, locaux) et des charges diverses dans l'année (indépendants du typage) évalue les coûts fixes à environ 60 000€.

L'ensemble des éléments de coût et toutes les données numériques ont été traitées de façon individuelle pour chaque laboratoire visité. Pour des raisons de confidentialité, aucun nom de laboratoire ou d'hôpital, ni aucune évaluation empirique de coût ne seront communiqués. En revanche si la grille d'évaluation du coût du typage HLA a été construite ici pour la calibration de notre modèle²², elle suggère des hypothèses raisonnables pour évaluer le coût du typage HLA dans un "autre contexte".

²²183.13€, voir la première partie de ce chapitre

2.4.3 Les grilles d'analyse du coût du typage HLA

Hypothèses de travail:

- 500 typages HLA-A, B, DR/DQ réalisés dans l'année
- Le laboratoire effectue sur place et lui-même les 500 extractions d'ADN

COÛTS VARIABLES					
			Nombre nécessaire 1 typage	Coût unitaire	Coût par typage
CONSOMMABLES					
Matériel secrétariat	enveloppes timbres	-	-	-	0.80€
	étiquettes	-	-	-	0.80€
	cahiers	-	-	-	1.00€
Kits et produits labo HLA	TAQ Polymerase	30 ml	3	-	-
	Gel Agarose	2 ml	2	-	-
	pipettes	-	4	0.05€	0.20€
	éprouvettes	-	2	0.05€	0.10€
	tubes	-	3	0.04€	0.12€
	PLAQUE CLASSE I	-	0.8	35.00€	28.00€
	KITS EXTRACTION	-	0.12	150.00€	18.00€
	TESTS TOTO DQB1	-	1	26.00€	26.00€
	TESTS TOTO DRB1	-	1	26.00€	26.00€
	TEST TYP HLA A SSP	-	0.2	25.00€	5.00€
TEST TYP HLA B SSP	-	0.2	25.00€	5.00€	
					110.52€
			Temps travail 1 typage (heure)	Salaire horaire Char.soc. incluses	Coût/typage
	Secrétaire	1	0.2	30€	6.00€
	Technicienne	1	0.8	32€	25.60€
	Biologiste	1	0.1	75€	7.50€
					39.10€
Coût marginal					149.62€

COÛTS FIXES						
	% matériel affecté acte	Durée amort années	Nbre	Coût unitaire	Coût/an typage	Coût Fixe annuel
MATERIEL (y.c. amortissements)						
Microscope à fluorescence	30%	5	1	21 000€	2.52€	4 200€
Centrifugeuse type X	10%	5	1	5 000€	0.20€	1 000€
Congélateur	20%	5	1	1 700€	0.14€	340€
Extracteur ADN	25%	5	1	20 000€	2.00€	4 000€
Spectrophotomètre	50%	5	1	30 000€	6.00€	6 000€
PCR Thermocycler	75%	5	1	8 500€	2.55€	1 700€
Ordinateurs PC	25%	5	1	1 200€	0.12€	240€
					13.53€	17 480€
			Coût annuel	% affectés au typage		
MAINTENANCE						
contrat extérieur			12 000€	15%	3.60€	12 000€
LOCAUX			surface acte (m ²)	Coût/an (m ²)		
locaux HLA (surface typage HLA)			150	30€	9€	4 500€
TRAITEMENT DECHETS			Coût annuel	% affectés au typage		
			11 500€	20%	4.60€	11 500€
PRESTATIONS DIVERSES			Coût annuel	% affectés au typage		
Heures ménage			8 000€	10%	1.60€	8 000€
Charges électricité,téléphone			5 000€	15%	1.50€	5 000€
					% coût fixe affecté au typage	33.83€
COÛT UNITAIRE (coût marginal+coût fixe)					183.45€	58 480€

TYPING WITH MICROSATELLITES

1 donor = 1 typing = 16 genotypes

<i>One step typing strategy</i>				
Technical steps				
DNA extraction from peripheral blood				
RBC lysis			4 €/donor	
Proteinase K digestion				
Automatic DNA extraction			5 €/donor	
Quality controls on:	concentration			10.80 € HT /donor
	PCR sizing digestion	10% des éch.	1.5 €/donor	
Distribution in 96 wells plates			0.3 €/donor	
PCR preparation				
Mix preparation	oligos 2 dyes	manual 32 with 16 lab	1h	0.3-0.4 € HT /donor
	Taq pol		50% of the typing reagent cost	
Preparation of 384 wells PCR plates		pipetting robo		
PCR			1h30	
Dilut°/pool/purif°/96 wells plates distrib°			2h	
Quality controls		4 QC/plate		
PCR migration				
Quality controls		Usual failure < 10%		
		Plate rejected when > 10% failure		
Validation & interpretation				
			2h	
Instruments				
96 capillaries automated sequencer	5	6X96 runs/d		1h30/96 runs
	no use	↳20% of the time		
PCR instruments	11			
Staff				
10 persons	2 eng	except QC		
	6 tech			
Flow				
2000 genotypes/d/person				
28,125 non validated typings/person/year				

Stratégie de typage : analyse locus par locus				
HLA-A	1snp	A,C,T	A: rare C: 6 snps T: 8 snps	2 temps
HLA-B	20 snps			1 temps
HLA-DR	20 snps			1 temps
Extraction d'ADN sur le sang				
Lyse des GR				4 €/éch
Digestion protéin. K				
Extraction/automate				5 €/éch
Contrôle de qualité	concentration	10% des éch.		1,5 €/éch
	PCR			
	sizing			
	digestion			
Mise en plaques 96 puits				0,3 €/éch
Préparation de la PCR :				0,18 € HT
	Préparation plaques PCR 384 puits	automatique		/généotype
	Contrôle de qualité	4 contrôles/plaque		
MasSpec				
	Contrôle de qualité	Taux habituel d'échec : ? Rejet de la plaque si >10% échec		
Matériels				5% du coût
	Spectromètre de masse	1	10-15000 snps/j	(5 sem
		non utilisation	?	util°)
	Appareils PCR	?		
Personnels				10% du
	2 personnes pendant 5 semaines	50 snps (moy)/individu		coût
	hors préparation & mise en plaque A 15000 individus			(5 sem de
				travail)
Débit				
	13-15000 génotypes/j/personne			
	Non rentable au dessous de ce seuil			

Cost Evaluation of Sequenced Base Typing									
10 000 individuals 30 000 typings									
Equipments									
A sequencer (16 capillaries) : Cost : 182 000 €									
Maintenance :	10 000 €/an								
Depreciation :	26 000 €/an								
Cost (*)	36 000 €/an								
4 months necessary for 10 000 individuals.....12 000 € (*) corresponding to the 100 % activity of the equipment									
2 PCR machines.....	2 500 €								
Kits/reagents									
12 sequences / Individual Cost : 3,2 € / each..... 384 000 €									
Labour costs									
Laboratory technician for the sequencer									
Labour cost : 3 000 €/an									
4 months.....	12 000 €								
Interpretation of the results									
8 months.....	24 000 €								
	434 500 €								
Overheads									
50 %.....	217 250 €								
TOTAL COST	651 750 €								
+ 25 % "ambiguities"									
	<table style="border: none;"> <tr> <td style="font-size: 2em; vertical-align: middle;">}</td> <td style="padding-left: 5px;">Ambiguity rate</td> </tr> <tr> <td></td> <td>HLA-A 8 %</td> </tr> <tr> <td></td> <td>HLA-B 15 %</td> </tr> <tr> <td></td> <td>HLA-DR 5 %</td> </tr> </table>	}	Ambiguity rate		HLA-A 8 %		HLA-B 15 %		HLA-DR 5 %
}	Ambiguity rate								
	HLA-A 8 %								
	HLA-B 15 %								
	HLA-DR 5 %								
Cost : 40 €/individual									
40 x 2 500.....	100 000 €								
	751 750 €								
75€ /indiv									

Figure 2.11: Exercice de construction du coût du Séquencage selon un protocole théorique de typage en grandes quantités

Chapitre 3

Contributions Statistiques à l'organisation d'un Registre de donneurs

3.1 Introduction

Haplotype information is essential for many analyses of genetics data, for example, in disease mapping (Risch and Merikangas 1996) or in interpreting data generated through DNA pooling (Wang, Kidd, Zhao 2003). Haplotype estimation is an important issue, both in population genetics (Single, Merger et al. 2002) and in the identification of complex disease genes (Niu, Qin et al. 2002). For example, associations between markers and disease loci that are not evident with a single marker locus may be identified in multi-locus analyses using estimated haplotype frequencies. Current genotyping methods do not provide phase information. This can be obtained, partially through genotyping of additional family members (Duldbridge, Kolleman et al. 2000).

If no information is available from family members, a statistical method may be used. From genotype observations, the joint distribution of haplotypes is estimated and the knowledge of this distribution and of the genotype of an individual may be used to infer the phase.

The first part of the chapter addresses the question of estimation of the joint distribution of the haplotypes using genotype data and proposes a new simple estimation method. The use of this distribution to infer the phase for an individual does not depend on the

estimation procedure and is not explicitly considered in this paper. Most of algorithms are based on searching for the haplotype vector that maximizes a likelihood computation. The two most popular existing methods are maximum likelihood, implemented via the EM Algorithm (Excoffier and Slatkin, 1995), and a parsimony method created by Clark (1990). A third method was proposed by Stephens, Smith and Donnely (2001). Their phase reconstruction method (a bayesian one) uses Gibb's sampling, a type of MCMC algorithm. We present a new statistical method, based on Hardy-Weinberg Equilibrium to infer the phase of genotype, which is neither an EM Algorithm nor a bayesian MCMC approach. The EM algorithm starts with an initial guess of haplotype frequencies and iteratively updates the frequency estimates, to maximize the log-likelihood function. Niu et al. (2002) introduce a bayesian procedure that uses a statistical model used in the EM algorithm. In their model each individual's haplotype pair is treated as two random draws from a pool of haplotypes with unknown population frequencies. Here we take a different approach and place the problem of estimating haplotype frequencies by introducing an algorithm based on Hardy-Weinberg equilibrium. This method is a moment estimation, based on the fitting between theoretical and empirical moments and provides estimation of the haplotype distribution using genotype data. Even if moment estimation does not reach asymptotic efficiency bound like maximum likelihood, it has similar advantages. This estimator is more easy to compute. It does not depend on a stopping rule and simulations show that it can perform better than maximum likelihood in small sample analysis.

In section 2 we present our moment estimation method: latent model and observables, likelihood, asymptotic distribution, a Monte Carlo simulation. In section 3 we will describe the algorithm for a generic model with five loci, with each locus possessing two alleles, as an example of implementation of our algorithm. The last section tries to estimate the number of types in the whole population.

3.2 A moment estimation of the haplotypes distribution using genotypes data

3.2.1 The latent model and the observables

The specification of the statistical model starts by assumptions on a set of latent observations. These assumptions will be completed by the description of the observation scheme and the distribution of the observables is deduced from these two parts of the specification.

The latent variables are the observations at the haplotype level. They are constituted by a sequence :

$$(\xi_1^q(i), \xi_2^q(i)) \quad q = 1, \dots, Q \quad i = 1, \dots, n \quad (3.2.1)$$

where i denotes the individual and q the locus. For each individual and each locus, $(\xi_1^q(i), \xi_2^q(i))$ represents the values of the alleles on the paternal chromosome (1 in our notation) and on the maternal chromosome (indexed by 2).

The number of possible alleles for locus q is r^q and the set of these alleles is $J^q = \{1, \dots, r^q\}$.

The parameter of interest is the joint distribution of the alleles on different loci on each phase, namely the numbers :

$$\begin{aligned} p(j^1, \dots, j^Q) &= P(\xi_1^1(i) = j^1, \dots, \xi_1^Q(i) = j^Q) \\ &= P(\xi_2^1(i) = j^1, \dots, \xi_2^Q(i) = j^Q) \end{aligned} \quad (3.2.2)$$

This notation implicitly assumes that the distribution of the alleles is identical for each individual and each chromosome. We assume more over that individuals and chromosomes are independent. These assumptions are implicitly based on the Hardy Weinberg equilibrium of the population. In other words the latent model describes a sample of $2n$ observations of a Q varied discrete random vector.

Unfortunately, the location of alleles on the two chromosomes of a given individual is not observable. Different ways exist for modelling this partial observability.

The more common way is to transform each pair $(\xi_1^q(i), \xi_2^q(i))$ into the rank statistic and the order statistic and to analyse the case where the order statistic only is observable. The difficulty of this method is the possibility of ties (homozygous individual) which requires a specific analysis. We prefer to adopt an other strategy based on the idea of a random (non observable) permutation of the data.

This presentation of the model and of the lack of observation is done using Rubin (1976) principle (“missing as random”). This principle is used in the paper by Li et al (2003) who deduce for this formalization an efficient way to implement the EM algorithm (see Dempster, Laird, Rubin 1977).

Let us extend the latent model by considering a vector

$$(\delta^q(i)) \quad q = 1, \dots, Q \quad i = 1, \dots, n$$

where $\delta^q(i) \in \{0, 1\}$. The new latent observations are now

$$(\xi_1^q(i), \xi_2^q(i), \delta^q(i)) \quad q = 1, \dots, Q \quad i = 1, \dots, n$$

and we define

$$X_1^q(i) = \delta^q(i) \xi_1^q(1) + (1 - \delta^q(i)) \xi_2^q(i)$$

$$X_2^q(i) = (1 - \delta^q(i)) \xi_1^q(i) + \delta^q(i) \xi_2^q(i)$$

Equivalently $X_1^q(i)$ is equal to $\xi_1^q(i)$ if $\delta^q(i) = 1$ and equal to $\xi_2^q(i)$ else. The variable $\delta^q(i)$ may be interpreted as the indicator of the allele which is observed first.

The probabilistic specification of the model is completed by the distribution of the $\delta^q(i)$. They are assumed to be *i.i.d* between individuals and between locus and

$$P(\delta^q(i) = 1) = P(\delta^q(i) = 0) = \frac{1}{2}.$$

Moreover the $\delta^q(i)$ are independent of the $(\xi_1^q(i), \xi_2^q(i))_{q,i}$. Let us underline that the distribution of the $\delta^q(i)$ is the marginal *i.i.d.* distribution. As we will see later, this distribution of the $\delta^q(i)$ given the $(X_1^q(i), X_2^q(i))$ is different.

There obviously exists a one to one transformation between the $(\xi_1^q(i), \xi_2^q(i), \delta^q(i))$ and $(X_1^q(i), X_2^q(i), \delta^q(i))$. Indeed

$$\xi_1^q(i) = \delta^q(i) X_1^q(i) + (1 - \delta^q(i)) X_2^q(i)$$

$$\xi_2^q(i) = (1 - \delta^q(i)) X_1^q(i) + \delta^q(i) X_2^q(i)$$

The vector $\delta(i) = (\delta^1(i), \dots, \delta^Q(i))$ may be called the phase configuration for an individual i .

Finally the observed sample is characterized by the

$$(X_1^q(i), X_2^q(i)) \quad q = 1, \dots, Q \quad i = 1, \dots, n.$$

or equivalently the $(\delta^q(i)) \quad q = 1, \dots, Q \quad i = 1, \dots, n$ are not observed. The actual observations are called the genotypes.

Remark: It should be stressed that the informational contain of our observed sample is equivalent to the order statistic. Actually the knowledge of the $X_1^q(i), X_2^q(i)$ implies the knowledge of the order statistic

$$\min(X_1^q(i), X_2^q(i)) = \min(\xi_1^q(i), \xi_2^q(i))$$

and

$$\max(X_1^q(i), X_2^q(i)) = \max(\xi_1^q(i), \xi_2^q(i))$$

Reciprocally given the order statistic, one can draw a vector of $(\delta^q(i))_{q,i}$ and construct the $X_1^q(i), X_2^q(i)$ from the order statistic. In that case the likelihood of the $X_1^q(i), X_2^q(i)$ construct from the latent observations or from the order are identical.

3.2.2 Likelihood

The sampling distribution of the latent variables $(\xi_1^q(i), \xi_2^q(i), \delta^q(i))$ is equal to:

$$L^* = \prod_{i=1}^n \left\{ \frac{1}{2^Q} p(\xi_1^1(i), \dots, \xi_1^Q(i)) p((\xi_2^1(i), \dots, \xi_2^Q(i))) \right\} \quad (3.2.1)$$

or using the one to one transformation between the $(\xi_1^q(i), \xi_2^q(i), \delta^q(i))$ and $(X_1^q(i), X_2^q(i), \delta^q(i))$ this likelihood is equal to:

$$L^* = \prod_{i=1}^n \left\{ \frac{1}{2^Q} p(\delta^1(i)X_1^1(i) + (1 - \delta^1(i))X_2^1(i), \dots, \delta^Q(i)X_1^Q(i) + (1 - \delta^Q(i))X_2^Q(i)) \times p((1 - \delta^1(i))X_1^1(i) + \delta^1(i)X_2^1(i), \dots, (1 - \delta^Q(i))X_1^Q(i) + \delta^Q(i)X_2^Q(i)) \right\} \quad (3.2.2)$$

and the marginal likelihood of the observable data is obtained by summing up the $(\delta^q(i))_q = \delta(i)$ in the set of all possible $\delta(i)$, namely $\{0,1\}^Q$. Then the likelihood of the genotype data is:

$$L = \prod_{i=1}^n \left\{ \frac{1}{2^Q} \sum_{\delta(i) \in \{0,1\}^Q} p(\delta^1(i)X_1^1(i) + (1 - \delta^1(i))X_2^1(i), \dots, \delta^Q(i)X_1^Q(i) + (1 - \delta^Q(i))X_2^Q(i)) \times p((1 - \delta^1(i))X_1^1(i) + \delta^1(i)X_2^1(i), \dots, (1 - \delta^Q(i))X_1^Q(i) + \delta^Q(i)X_2^Q(i)) \right\} \quad (3.2.3)$$

This marginal likelihood involves a sum over 2^Q terms and is untractable but the EM algorithm provides an efficient way to compute numerically the value of $p(j^1, \dots, j^Q)$ which maximises this likelihood. This method is based on two features of this model :

- Given the parameters $p(j^1, \dots, j^Q)$ the probability of $\delta(i)$ given

$(X_1(i), X_2(i)) = (X_1^q(i), X_2^q(i))_{q=1, \dots, Q}$ is easily deduced from the Bayes rule:

$$P(\delta(i)|X_1(i), X_2(i), p) = \frac{P(\delta(i)|p)P(X_1(i), X_2(i)|\delta(i), p)}{\sum_{\delta(i) \in \{0,1\}^Q} P(\delta(i)|p)P(X_1(i), X_2(i)|\delta(i), p)} \quad (3.2.4)$$

The term $P(\delta(i)|p) = \frac{1}{2^Q}$ can be simplified and using an elementary vectoriel notation:

$$P(\delta(i)|X_1(i), X_2(i), p) = \frac{p(\delta(i)X_1(i) + (1 - \delta(i))X_2(i))p((1 - \delta(i))X_1(i) + \delta(i)X_2(i))}{\sum_{\delta(i) \in \{0,1\}^Q} p(\delta(i)X_1(i) + (1 - \delta(i))X_2(i))p((1 - \delta(i))X_1(i) + \delta(i)X_2(i))} \quad (3.2.5)$$

- Given the $\delta(i)$ the loglikelihood of $(X_1(i), X_2(i))$ may be rewritten on the form :

$$\sum_{\substack{\bar{j} = (j^1, \dots, j^Q) \\ \bar{k} = (k^1, \dots, k^Q)}} \alpha(\bar{j}, \bar{k}) \{ \ln p(j^1, \dots, j^Q) + \ln p(k^1, \dots, k^Q) \} \quad (3.2.6)$$

where $\alpha(\bar{j}, \bar{k}) = P(\delta(i) | X_1(i), X_2(i), p)$

Then given α the $\alpha(\bar{j}, \bar{k})$ this likelihood may be easily maximized (M step) and given p the $\alpha(\bar{j}, \bar{k})$ may easily be computed (E step). The EM algorithm is based on a recursive application of these two steps under the convergence.

A bayesian analogous of the EM algorithm is provided by an MCMC treatment of the posterior distribution of the vector p . If the prior probability on p is a Dirichlet distribution, its posterior given $(X_1(i), X_2(i), \delta(i)) i = 1, \dots, n$ is also a Dirichlet distribution. Then, samples from p given $(X_1(i), X_2(i), \delta(i))$ are easily generated. Using the previous argument, $(\delta(i)) i = 1, \dots, n$ given p and $(X_1(i), X_2(i))$ are easily generated and a recursive use of the Gibbs sampling algorithm will provided draws, after convergence, from $(p, (\delta(i)) i = 1, \dots, n)$ given the actual data.

3.2.3 A moment estimation: an introductory case

Let us consider a simple case where Q , the number of loci, is equal to 2. The goal of the procedure is to estimate the joint distribution of the alleles on these two loci, described by the

$$p(j^1, j^2) \quad j^1 = 1, \dots, r^1 \quad j^2 = 1, \dots, r^2$$

where $p(j^1, j^2) \geq 0$ and $\sum_{j^1, j^2} p(j^1, j^2) = 1$. We denote by $p(j^1, \cdot)$ and $p(\cdot, j^2)$ the marginal probabilities, ie:

$$\begin{aligned} p(j^1, \cdot) &= \sum_{j^2=1}^{r^2} p(j^1, j^2) = P(\xi_1^1(i) = j^1) \\ &= P(\xi_2^1(i) = j^1) \end{aligned} \quad (3.2.1)$$

and

$$\begin{aligned} p(\cdot, j^2) &= \sum_{j^1=1}^{r^1} p(j^1, j^2) = P(\xi_1^2(i) = j^2) \\ &= P(\xi_2^2(i) = j^2) \end{aligned} \tag{3.2.2}$$

It is well known that the lack of observation of the phase configuration does not raise any problem for the estimation of these marginal probabilities.

Indeed:

$$\hat{p}(j^1, \cdot) = \frac{1}{2n} \sum_{i=1}^n \{ \mathbb{I}(X_1^1(i) = j^1) + \mathbb{I}(X_2^1(i) = j^1) \} \quad (3.2.3)$$

and

$$\hat{p}(\cdot, j^2) = \frac{1}{2n} \sum_{i=1}^n \{ \mathbb{I}(X_1^2(i) = j^2) + \mathbb{I}(X_2^2(i) = j^2) \} \quad (3.2.4)$$

where $\mathbb{I}(X_1^1(i) = j^1)$ equal 1 if $X_1^1(i) = j^1$ and 0 else provide consistent estimators if $p(j^1, \cdot)$ and $p(\cdot, j^2)$. This consistency follows from the strong law of large number.

Consider now the statistic:

$$\begin{aligned} \hat{A}(j^1, j^2) = \frac{1}{4n} \sum_{i=1}^n & \mathbb{I}(X_1^1(i) = j^1, X_1^2(i) = j^2) \\ & + \mathbb{I}(X_1^1(i) = j^1, X_2^2(i) = j^2) \\ & + \mathbb{I}(X_2^1(i) = j^1, X_1^2(i) = j^2) \\ & + \mathbb{I}(X_2^1(i) = j^1, X_2^2(i) = j^2) \end{aligned} \quad (3.2.5)$$

which count all the possible pairs of alleles on the two loci equal to (j^1, j^2) .

It will easily show that the expectation of each terms of the sum is equal to

$$\frac{1}{2}(p(j^1, j^2) + p(j^1, \cdot)p(\cdot, j^2)) \quad (3.2.6)$$

Let us take for example the expectation of the first term:

$$\begin{aligned}
 E & (\mathbb{I}(X_1^1(i) = j^1, X_1^2(i) = j^2)) \\
 = & \sum_{\delta(i) \in \{0,1\}^2} P(\delta(i)) P(X_1^1(i) = j^1, X_1^2(i) = j^2 | \delta(i)) \\
 = & \frac{1}{4} \{ P(X_1^1(i) = j^1, X_1^2(i) = j^2 | \delta^1(i) = 0, \delta^2(i) = 0) \\
 & + P(X_1^1(i) = j^1, X_1^2(i) = j^2 | \delta^1(i) = 1, \delta^2(i) = 0) \\
 & + P(X_1^1(i) = j^1, X_1^2(i) = j^2 | \delta^1(i) = 0, \delta^2(i) = 1) \\
 & + P(X_1^1(i) = j^1, X_1^2(i) = j^2 | \delta^1(i) = 1, \delta^2(i) = 1) \} \\
 = & \frac{1}{4} \{ P(\xi_1^1(i) = j^1, \xi_1^2(i) = j^2) \\
 & + P(\xi_1^1(i) = j^1, \xi_2^2(i) = j^2) \\
 & + P(\xi_2^1(i) = j^1, \xi_1^2(i) = j^2) \\
 & + P(\xi_2^1(i) = j^1, \xi_2^2(i) = j^2) \} \\
 = & \frac{1}{4} \{ p(j^1, j^2) + p(j^1, \cdot) p(\cdot, j^2) + p(j^1, \cdot) p(\cdot, j^2) + p(j^1, j^2) \} \quad (3.2.7)
 \end{aligned}$$

Then, using the strong law of large numbers:

$$\hat{A}(j^1, j^2) \xrightarrow{a.s} \frac{1}{2} (p(j^1, j^2) + p(j^1, \cdot) p(\cdot, j^2)) \quad (3.2.8)$$

The intuition behind this result is that a pair of two observed alleles on two loci has a (marginal) probability $\frac{1}{2}$ to be on the same locus and then to be generated with a probability $p(j^1, j^2)$ and a (marginal) probability $\frac{1}{2}$ to be on different chromosomes and then to be independently generated.

Remark: A more tedious computation would show that the behavior of $\hat{A}(j^1, j^2)$ is identical if the observed data are constructed using the order statistics on each locus. Actually one can remark that \hat{A} is chosen such that its value is invariant by permutation of the data between the two phases.

Following (4.4) and (4.8), a consistent estimation of $p(j^1, j^2)$ for any value of j^1, j^2 is given by:

$$\hat{p}(j^1, j^2) = 2\hat{A}(j^1, j^2) - \hat{p}(j^1, \cdot)\hat{p}(\cdot, j^2) \quad (3.2.9)$$

This argument may be extended to three loci. Let us now consider $\hat{A}(j^1, j^2, j^3)$ equal to the total number of possible triplets of alleles j^1, j^2 and j^3 observed for each individual, divided by $8n$. Using an equivalent argument to the two loci case (a general presentation will be given in the next section) we can check that:

$$\begin{aligned} \hat{A}(j^1, j^2, j^3) \rightarrow \frac{1}{4} \{ & j^1, j^2, j^3 \} + p(j^1, j^2, \cdot)p(\cdot, \cdot, j^3) \\ & + p(j^1, \cdot, j^3)p(\cdot, j^2, \cdot) + p(j^1, \cdot, \cdot)p(\cdot, j^2, j^3) \} \end{aligned} \quad (3.2.10)$$

where e.g. $p(j^1, j^2, \cdot)$ is the marginal distribution on the two first loci.

Using (4.3), (4.4) and (4.9) the marginal probabilities on a single locus or on two loci can be estimated and we obtain a consistent estimator of $p(j^1, j^2, j^3)$ by:

$$\begin{aligned} \hat{p}(j^1, j^2, j^3) = & 4\hat{A}(j^1, j^2, j^3) - \hat{p}(j^1, j^2, \cdot)p(\cdot, \cdot, j^3) \\ & - \hat{p}(j^1, \cdot, j^3)\hat{p}(\cdot, j^2, \cdot) - \hat{p}(j^1, \cdot, \cdot)\hat{p}(\cdot, j^2, j^3) \end{aligned} \quad (3.2.11)$$

3.2.4 Moment estimation : the general case

For any individual we consider the following random element:

$$Z_{j_1, \dots, j_Q}(i) = \frac{1}{2^Q} \sum_{\tau \in T} \mathbb{I}(X_{\tau(1)}^1(i) = j^1, \dots, X_{\tau(Q)}^Q(i) = j^Q)$$

where T is the set of all functions from $\{1, \dots, Q\}$ into $\{1, 2\}$ whose cardinality is 2^Q . Intuitively $Z_{j_1, \dots, j_Q}(i)$ counts the number of Q -uple equal to j^1, \dots, j^Q obtained by all the possible selections of one element in each pair of alleles for an individual i . Thanks to the strong law of large number, the empirical mean converges to the theoretical mean, i.e. :

$$\hat{A}(j_1, \dots, j_Q) = \frac{1}{n} \sum_{i=1}^n Z_{j^1, \dots, j^Q}(i) \xrightarrow{a.s.} E(Z_{j^1, \dots, j^Q}(i)) \quad (3.2.1)$$

We now compute the theoretical mean, which does not depend on the individual and we drop out for simplicity the index i .

$$\begin{aligned} E(Z_{j^1, \dots, j^Q}) &= \frac{1}{2^Q} \sum_{\tau \in T} E(\mathbb{I}(X_{\tau(1)}^1 = j^1, \dots, X_{\tau(Q)}^Q = j^Q)) \\ &= \frac{1}{2^Q} \sum_{\tau \in T} \sum_{\delta \in D} P(X_{\tau(1)}^1 = j^1, \dots, X_{\tau(Q)}^Q = j^Q | \delta) \end{aligned} \quad (3.2.2)$$

where $\delta = (\delta^1, \dots, \delta^Q)$ may be any element of the set D of all the function from $\{1, \dots, Q\}$ to $\{0, 1\}$ ($= \{0, 1\}^Q$).

Then:

$$E(Z_{j^1, \dots, j^Q}) = \frac{1}{2^Q} \sum_{\tau \in T} \sum_{\delta \in D} \left(\sum_{\substack{j^1, \dots, j^Q \\ s.t. \tau(q)-1 \neq \delta^q}} p(j^1, \dots, j^Q) \right) \left(\sum_{\substack{j^1, \dots, j^Q \\ s.t. \tau(q)-1 \neq \delta^q}} p(j^1, \dots, j^Q) \right) \quad (3.2.3)$$

In the first parenthesis the sum of $p(j^1, \dots, j^Q)$ is computed with the index j^q only if $\tau(q) - 1 \neq \delta^q$.

By regrouping equal terms in the sum, we get

$$E(Z_{j^1, \dots, j^Q}) = \frac{1}{2^Q} \sum_{\delta \in D} \left(\sum_{\substack{j^1, \dots, j^Q \\ \delta^q=0}} p(j^1, \dots, j^Q) \right) \left(\sum_{\substack{j^1, \dots, j^Q \\ \delta^q=1}} p(j^1, \dots, j^Q) \right)$$

This expression will be denote by $\lambda_{j^1, \dots, j^Q}(p)$ where p is the vector of probabilities to be estimated.

Formulae (4.6) and (4.10) give particular cases of this results for $Q = 2$ and 3 . For the case $Q = 4$ we obtain:

$$\begin{aligned}
E(Zj^1, j^2, j^3, j^4) = \frac{1}{16} & \left\{ 2 \left(p(j^1, j^2, j^3, j^4) + p(j^1, j^2, j^3, \cdot) p(\cdot, \cdot, \cdot, j^4) \right. \right. \\
& + p(j^1, j^2, \cdot, j^4) p(\cdot, \cdot, j^3, \cdot) + p(j^1, \cdot, j^3, j^4) p(\cdot, j^2, \cdot, \cdot) \\
& + p(\cdot, j^2, j^3, j^4) p(j^1, \cdot, \cdot, \cdot) + p(j^1, j^2, \cdot, \cdot) p(\cdot, \cdot, j^3, j^4) \\
& \left. \left. + p(j^1, \cdot, \cdot, j^4) p(\cdot, j^2, j^3, \cdot) + p(j^1, \cdot, j^3, \cdot) p(\cdot, j^2, \cdot, j^4) \right) \right\}
\end{aligned}$$

Then a recursive estimation method may be easily implemented to solve the set of moment conditions (see section 2).

3.2.5 Asymptotic distribution

This approach belongs to the family of estimation using estimating function or moment estimation. The estimation of p is constructed in order to match the empirical moments $\hat{A}(j^1, \dots, j^Q)$ and the theoretical moments $\lambda_{j^1, \dots, j^Q}(p)$ depending on the parameters of interest. The number of conditions is $r^1 \times \dots \times r^Q$ (the number of possible combinations of alleles on the Q locus on a chromosome), but one condition may be dropped out because the $\lambda_{j^1, \dots, j^Q}(p)$ are themselves probabilities and sum to one. In other words, if we stack the $Zj^1, \dots, j^q(i)$ in a $(r^1 \times \dots \times r^Q) - 1$ vector $Z(i)$ and the $\lambda_{j^1, \dots, j^Q}(p)$ in a $(r^1 \times \dots \times r^Q) - 1$ vector $\lambda(p)$, the estimation of p we propose is obtained by solving:

$$\frac{1}{n} \sum_{i=1}^n Z(i) = \lambda(p) \quad (3.2.1)$$

Let us underline that the recursive method we have proposed is only a resolution method of this system and the equations corresponding to subsets of the Q locus are contained in the set (6.1). In a very compact way, our estimator may be summarized by

$$\hat{p} = \lambda^{-1} \left(\frac{1}{n} \sum Z(i) \right) \quad (3.2.2)$$

and the existence of the inverse follows from the recursive solution we have introduced.

The functions λ and λ^{-1} are continuously differentiable. Then:

i) $\hat{p} \xrightarrow{a.s} p$ solution of $E(Z(i)) = \lambda(p)$

ii) $\sqrt{n}(\hat{p} - p) \implies \mathcal{N}(0, \Sigma)$

where $\Sigma = \left(\frac{\partial \lambda}{\partial p'}\right)^{-1} Var Z(i) \left(\frac{\partial \lambda}{\partial p}\right)^{-1}$

(see Serfling (1980)).

The variance Σ can be easily estimated : $Var Z(i)$ may be estimated by the empirical variance of the $Z(i)$ and $\frac{\partial \lambda}{\partial p}$ the matrix of partial derivatives of λ is a matrix of functions of p . The general computation of this matrix is very tedious (but may be immediately realized by computer software as Mapple or Mathematica) and we just illustrate this asymptotic distribution by an example.

Consider an example with two locus and two alleles on each locus. The probabilities of interest are defined by:

	locus 1	allele 1	allele 2
	\	1	2
locus 2			
allele 1		p_{11}	p_{12}
allele 2		p_{21}	p_{22}

Actually this case only involves three parameters because $p_{11} + p_{12} + p_{21} + p_{22} = 1$ and this constraint is introduced by replacing p_{22} by $1 - (p_{11} + p_{12} + p_{21})$. In that case we have $\lambda(p) = (\lambda_{11}(p), \lambda_{12}(p), \lambda_{21}(p))$ where:

$$\begin{aligned}
 \lambda_{11}(p) &= \frac{1}{2}(p_{11} + (p_{11} + p_{12})(p_{11} + p_{21})) \\
 \lambda_{12}(p) &= \frac{1}{2}(p_{12} + (p_{11} + p_{12})(p_{12} + p_{22})) \\
 \lambda_{21}(p) &= \frac{1}{2}(p_{21} + (p_{11} + p_{21})(p_{21} + p_{22}))
 \end{aligned} \tag{3.2.3}$$

and the matrix of partial derivatives is:

$$\frac{\partial \lambda}{\partial p'} = \frac{1}{2} \begin{pmatrix} 1 + (p_{11} + p_{12}) + (p_{11} + p_{21}) & p_{11} + p_{21} & p_{11} + p_{12} \\ p_{22} - p_{11} & 1 + (p_{21} + p_{22}) & -(p_{11} + p_{12}) \\ p_{22} - p_{11} & -(p_{11} + p_{21}) & 1 + (p_{21} + p_{22}) \end{pmatrix}$$

This matrix is estimated by replacing the unknown values of p by their estimated values.

In practice, bootstrap confidence interval may be used in order to improve asymptotic distribution. This approach will be discussed in section 9.

3.2.6 Correction for negative probabilities

The asymptotic analysis done in section 6 is implicitly based on the assumption that the true value p of the parameter is an interior point of the set of all possible values of

this parameter, namely the simplex of probabilities of $(r^1 \times \dots \times r^Q) - 1$ dimensions. If this assumption is satisfied (i.e. $p(j^1, \dots, j^Q) > 0 \forall j^1, \dots, j^Q$) the estimator $\hat{p}(j^1, \dots, j^Q)$ is necessarily positive for n sufficiently large because this estimator is consistent.

In practice our estimation method constructs estimations $\hat{p}(j^1, \dots, j^Q)$ which verifies

$$\sum_{j^1, \dots, j^Q} \hat{p}(j^1, \dots, j^Q) = 1$$

but which may fail to be positive. Remember that, in the two loci cases we have (see 4.8)

$$\hat{p}(j^1, j^2) = 2A(j^1, j^2) - \hat{p}(j^1, \cdot) \hat{p}(\cdot, j^2)$$

and the $2A(j^1, j^2)$ may be equal to zero (if the pair (j^1, j^2) is never observed) or smaller to the product of the estimated marginal probabilities. In that case, we suggest to transform our estimator by the following rule :

- i) put equal to 0 any probability estimated by a negative number
- ii) renormalized the positive probabilities by dividing their sum.

As noted below, this modification does not affect the asymptotic behavior of the estimator if the true probabilities are positive. In that case, the correction is only a small sample improvement of the estimator. However, if some of the true probabilities are zero, the asymptotic distribution of our estimator, as well as the distribution of the maximum likelihood estimator, is definitely more complex and will not be considered in this paper (see Andrews 1999).

3.2.7 A Monte Carlo simulation

In order to evaluate the small sample performance of our estimator and to compare with other approaches we have done a Monte Carlo simulation using the following design.

1) We consider two loci and two alleles for each locus and the simulation are generated using the values:

	locus 1	allele 1	allele 2
locus 2 \			
allele 1		0, 1	0, 3
allele 2		0, 2	0, 4

2) For different sample sizes (100, 500, 1000 and 10000) a sample of pairs of chromosome is generated and four estimations are performed

- $\hat{q}(j^1, j^2)$ is the maximum likelihood estimation using haplotypes which represents the "best" estimation (inaccessible in practice but accessible using simulation)
- $\hat{p}(j^1, j^2)$ is the estimation consider in the paper (see section 4). Negative probability are never obtained by the simulation.
- $\hat{p}_{EM}(j^1, j^2)$ and $\hat{p}_{Max}(j^1, j^2)$ denotes two evaluations of the maximum likelihood estimation : the first one is based on the EM algorithm (using as as shopping role the variation of the four parameter is lower than 10^{-6}). The second is a direct resolution of the first order condition of the likelihood maximization by the procedure "solve" of matlab. This two methods only differ by the numerical computation of the maximum likelihood estimator.

3) This experiment has been reproduced 100 times and results are summarized by the root mean square errors of each parameter for each sample size.

The results are summarized in table I. For almost all the cases (except p_{11} for a sample size of 100) our estimator is superior to the two numerical evaluations of the maximum likelihood, even for a sample size as large as 10000. For large sample size the difference between our estimator and the "best" possible estimator \hat{q} is very low.

3.2.8 Application to the relation between the microsatellite MOGc and gene HLA-A

The original motivation of this research was to analyse the capacity of a set of microsatellites to predict groups of HLA types in the framework of optimizing HLA typing policies of Bone Marrow Donor Registries . In this paper we just present a preliminary step of this study concentrated on a single microsatellite MOGc and the A locus of the HLA system. We consider a sample (of size 2117)¹ of genotypes used for the estimation of the joint distribution of size 2117 of MOGc and A on a single chromosom.

The result of our estimation is given in table II where "0" denotes pairs of alleles of MOGc/A never observed. Probability values are rounded off.

The precision of this estimation result is analysed by a non parametric bootstrap. From the original sample we have contracted 1000 samples by random drawing with replacement. Each sample is used for a new estimation of the joint probability (see Efron 1982, Hall (1999)).

We just illustrate the power of this analysis by two examples. We have constructed the bootstrap distribution of two measures of the linkage disequilibrium. The first one is the entropy measure defined by

$$I = \sum_{j,k} p_{jk} \ln \frac{p_{jk}}{p_j \cdot p_k} \quad (3.2.1)$$

where j is the index of possible alleles of MOGc, k is the index of possible alleles of A, p_{jk} is the joint probability and p_j and p_k the marginal probabilities.

The estimated value of I (9.1) is 0,9701. The bootstrap mean is 0,9676 and a confidence interval of I at 95 % is [0,9215; 1,0173]. The distribution of I is given by the histogram in table III.

It is well know that entropy has some unclearable features and a better association measure

¹This sample was randomly extracted from the France Greffe de Moelle Registry. In this data set missing data are reconstructed by answering homozygoty

is provided by Hellinger distance between the joint distribution and the product of its marginals, i.e.

$$H = \frac{1}{\sqrt{2}} \left[\sum_{j,k} (\sqrt{p_{jk}} - \sqrt{p_j \cdot p_k})^2 \right]^{\frac{1}{2}}$$

In particular, by construction, it is normalized in order to be between 0 and 1 where 0 is equivalent to independence. The actual estimated value of H is 0,4270. The bootstrap mean is 0,4271 and a confidence interval is [0,4033; 0,4520]. Histograms of bootstrap distribution of this, linkage disequilibrium measure is given in tables III and IV.

Conclusion

This chapter presents a moment estimation of the joint distribution of the alleles on several loci on a chromosome using genotype data. This estimator is not constructed as the limit of a recursive algorithm (dependent on starting point and on stopping rule) but is immediately computable. This estimator is strongly consistent and asymptotically normal and although it does not reach the efficiency bound as maximum likelihood, Monte-Carlo simulations show that it performs better in some small sample cases. Moreover a bootstrap analysis of the distribution of the estimator is possible thank to its efficiency in term of computation time. We have illustrated the power of our methodology by an empirical analysis of linkage disequilibrium between MOGc and gene HLA-A. Two extensions of this approach are in project. First the computation of the estimator in case of numerous loci may be improved by an optimisation of the numerous countings required for the estimation, and second asymptotic properties of Maximum Likelihood Estimation and of our estimator should be studied in case where the true joint probability has elements exactly equal to 0. Then a major hypothesis of MLE is not satisfied (namely the true parameter is an interior point of the parametric space) and optimality of MLE is no longer warranted .

Table I

RMSE of different estimators

	Sample size	100	500	1 000	10 000
MLI using Haplotype Data	q11	0,0183	0,0093	0,0059	0,0023
Moment Estimator using Phenotype Data	q11	0,0255	0,0111	0,0089	0,0026
EM Algorithm computation of MLI using Phenotype Data	q11	0,0311	0,013	0,011	0,0053
Direct MLI estimator using Phenotype Data (*)	q11	0,0237	0,0104	0,0087	0,005
MLI using Haplotype Data	q12	0,0332	0,0145	0,0102	0,003
Moment Estimator using Phenotype Data	q12	0,0358	0,0166	0,0121	0,0032
EM Algorithm computation of MLI using Phenotype Data	q12	0,0423	0,0176	0,0136	0,0033
Direct MLI estimator using Phenotype Data	q12	0,0396	0,0158	0,0124	0,0031
MLI using Haplotype Data	q21	0,0296	0,0124	0,0081	0,0028
Moment Estimator using Phenotype Data	q21	0,036	0,014	0,01	0,0034
EM Algorithm computation of MLI using Phenotype Data	q21	0,049	0,0172	0,0126	0,0052
Direct MLI estimator using Phenotype Data	q21	0,0419	0,0152	0,0111	0,005
MLI using Haplotype Data	q22	0,0335	0,0163	0,011	0,0037
Moment Estimator using Phenotype Data	q22	0,0374	0,0184	0,0124	0,004
EM Algorithm computation of MLI using Phenotype Data	q22	0,0451	0,022	0,0155	0,0094
Direct MLI estimator using Phenotype Data	q22	0,042	0,0205	0,0147	0,0094

2 characters et 2 modes
 0,1 0,3
 0,2 0,4

(*)MATLAB procedure : Chris Smith 1997 revised by F. Collard (1999)

Table II

Joint Distribution of MOGC and HLA-A on a single chromosome

MOGCHLA-A	1	2	3	9	10	11	23	24	25	26	28	29	30	31	32	33	34	36	43	66	68	69	74	80	
121	0,076	0,003	0	0	0,001	0,003	0,015	0,016	0,035	0,001	0,052	0,002	0,004	0,004	0,002	0,010	0	0	0,002	0,014	0	0	0	0	0,24
123	0	0	0,001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
125	0	0	0	0	0	0	0	0	0	0	0	0	0	0,001	0	0	0	0	0	0	0	0	0	0	0,00
127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
129	0,005	0,016	0,117	0	0,002	0,002	0,022	0	0,005	0	0,001	0,014	0,003	0,001	0	0,001	0	0	0,001	0	0	0	0	0,001	0,19
131	0,005	0,169	0,003	0,001	0,004	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,013	0,001	0,001	0,013	0,001	0	0	0,012	0	0	0	0	0,21
133	0,002	0,002	0	0	0	0	0	0,001	0,011	0,016	0,001	0,001	0,016	0,001	0,001	0	0	0	0	0	0	0	0	0	0,03
135	0,002	0	0,051	0,020	0,001	0,001	0,001	0,001	0,001	0	0,001	0	0	0,001	0	0	0	0	0	0	0	0	0	0	0,08
137	0,003	0	0	0,002	0,001	0,024	0	0,001	0	0,001	0	0,016	0,004	0,001	0,001	0	0	0	0,010	0	0	0	0	0	0,06
139	0,001	0,001	0	0	0	0	0	0	0	0	0	0,002	0	0	0	0	0	0	0	0	0	0	0	0	0,00
141	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
143	0,052	0,003	0	0	0,002	0,002	0,002	0	0,001	0,001	0	0,001	0,001	0	0,001	0	0	0	0	0	0	0	0	0	0,06
145	0,001	0,001	0	0	0,003	0	0,003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,01
147	0,061	0,006	0,002	0,001	0	0,030	0	0	0	0	0	0	0	0,001	0	0	0	0	0,001	0	0	0	0	0	0,10
149	0	0	0	0	0,001	0,001	0,001	0	0	0	0	0	0	0	0	0	0	0	0	0,001	0,001	0	0	0	0,00
151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,00
153	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,001	0	0	0	0,00

0,13 0,28 0,13 0,00 0,00 0,06 0,03 0,10 0,02 0,04 0,00 0,06 0,04 0,03 0,03 0,02 0,00 0,00 0,00 0,00 0,04 0,00 0,00 0,00

Table III

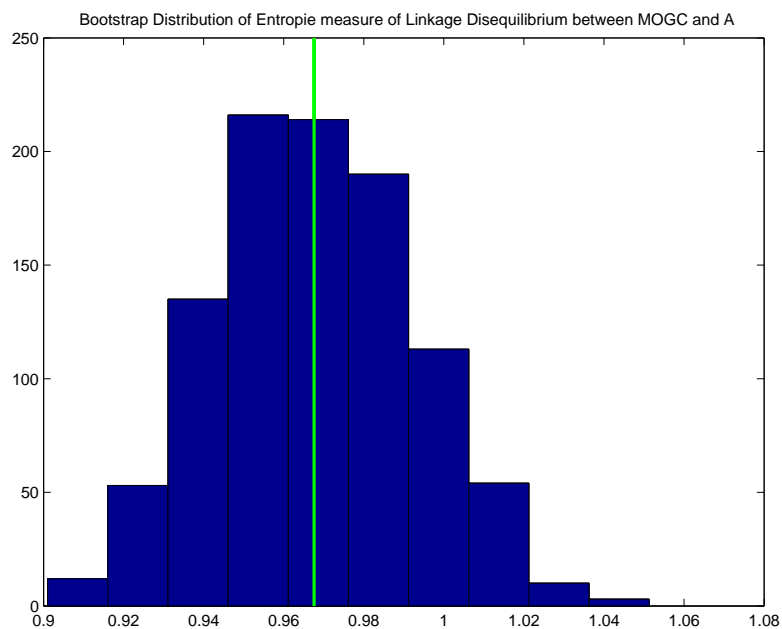
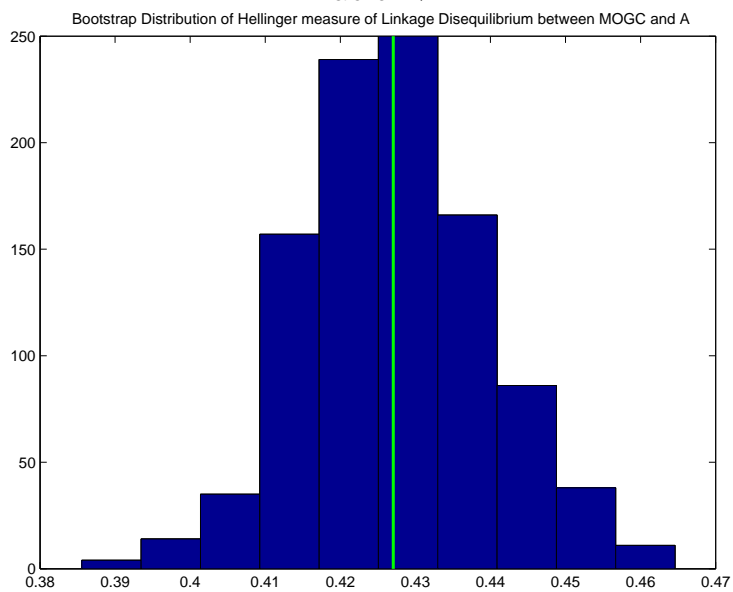


Table IV



3.3 Implementation of the estimation algorithm

For simplicity we consider first a case where only two loci are observed. At the haplotype level the variables are denoted by:

$$((\xi_{11}(i), \xi_{12}(i)), (\xi_{21}(i), \xi_{22}(i))) \quad i = 1, \dots, n$$

where i represents the individual and where the first pair denotes the alleles on the first locus of the paternal and maternal chromosome and the second pair denotes the alleles on the second locus. In our notation greek letters represents latent (inobservable) variables, the first index correspond to the locus and the second characterizes paternal (1) and maternal (2) chromosome.

On the first locus J_1 alleles are possible and J_2 on the second locus. The probabilities of interest constitute the sequence:

$$\begin{aligned} p(j_1, j_2) &= P(\xi_{11}(i) = j_1, \xi_{21}(i) = j_2) \\ &= P(\xi_{12}(i) = j_1, \xi_{22}(i) = j_2) \end{aligned}$$

This notation implicitly assumes that the distribution of the alleles is identical for each individual and each chromosome. We assume more over that individuals and chromosomes are independent. These assumptions are implicitly based on the Hardy Weinberg equilibrium of the population. In other words the latent model describes a sample of $2n$ observations of a Q varied discrete random vector.

Unfortunately, the repartition of alleles on the two chromosomes (phase) of a given individual is not observable and we denote by

$$(X_{11}(i), X_{12}(i)), (X_{21}(i), X_{22}(i)) \quad i = 1, \dots, n$$

the observed data (genotype data). Different ways exists in order to formalize the absence

of the phase. For example $X_{11}(i)$ is the smallest value of $\xi_{11}(i)$ and $\xi_{12}(i)$ and $X_{12}(i)$ is the larger value of these two elements. In case of ties (homozygous observations) we reproduce the observation. We don't consider here missing data, i.e. single observation of an allele means unambiguously homozygous data.

Let us denote by $p(j_1, \cdot)$ and $p(\cdot, j_2)$ the marginal probabilities. For example :

$$p(j_1, \cdot) = \sum_{j_2=1}^{J_2} p(j_1, j_2) = P(\xi_{11}(i) = j_1) = P(\xi_{12}(i) = j_1).$$

It is well known that the lack of observation of the phase configuration does not raise any problem for the estimation of these marginal probabilities.

Indeed:

$$\hat{p}(j_1, \cdot) = \frac{1}{2n} \sum_{i=1}^n \{ \mathbb{I}(X_{11}(i) = j_1) + \mathbb{I}(X_{12}(i) = j_1) \} \quad (3.3.1)$$

and

$$\hat{p}(\cdot, j_2) = \frac{1}{2n} \sum_{i=1}^n \{ \mathbb{I}(X_{21}(i) = j_2) + \mathbb{I}(X_{22}(i) = j_2) \} \quad (3.3.2)$$

(where $\mathbb{I}(X_{11}(i) = j_1)$ equal 1 if $X_{11}(i) = j_1$ and 0 else) provide consistent estimators of $p(j_1, \cdot)$ and $p(\cdot, j_2)$.

Consider now the statistic:

$$\begin{aligned} \hat{A}(j_1, j_2) = \frac{1}{4n} \sum_{i=1}^n & \mathbb{I}(X_{11}(i) = j_1, X_{21}(i) = j_2) \\ & + \mathbb{I}(X_{11}(i) = j_1, X_{22}(i) = j_2) \\ & + \mathbb{I}(X_{12}(i) = j_1, X_{21}(i) = j_2) \\ & + \mathbb{I}(X_{12}(i) = j_1, X_{22}(i) = j_2), \end{aligned} \quad (3.3.3)$$

which count all the possible pairs of alleles on the two loci equal to (j_1, j_2) . This value is independent if the phase and then may be compute using genotype data only.

Equivalently for each individual observation we create four possible chromosomes and we compute the frequencies if the pairs of alleles. These frequencies do not estimate the

probabilities of interest but it may be easily shown that the expectation of each terms of the sum is equal to

$$\frac{1}{2}(p(j_1, j_2) + p(j_1, \cdot)p(\cdot, j_2)). \quad (3.3.4)$$

Then, using the strong law of large numbers,

$$\hat{A}(j_1, j_2) \xrightarrow{a.s} \frac{1}{2}(p(j_1, j_2) + p(j_1, \cdot)p(\cdot, j_2)). \quad (3.3.5)$$

The intuition behind this result is that a pair of two observed alleles on two loci has a (marginal) probability $\frac{1}{2}$ to be on the same locus and then to be generated with a probability $p(j_1, j_2)$ and a (marginal) probability $\frac{1}{2}$ to be on different chromosomes and then to be independently generated.

Following (2.1), (2.2) and (2.5), a consistent estimation of $p(j_1, j_2)$ for any value of j_1, j_2 is given by:

$$\hat{p}(j_1, j_2) = 2\hat{A}(j_1, j_2) - \hat{p}(j_1, \cdot)\hat{p}(\cdot, j_2). \quad (3.3.6)$$

This argument may be extended to three loci. Let us now consider $\hat{A}(j_1, j_2, j_3)$ equal to the total number of possible triplets of alleles j_1, j_2 and j_3 observed for each individual, divided by $8n$. Using an equivalent argument to the two loci case (a general presentation will be given in the section 4) we can check that:

$$\begin{aligned} \hat{A}(j_1, j_2, j_3) \rightarrow \frac{1}{4} \{ & p(j_1, j_2, j_3) + p(j_1, j_2, \cdot)p(\cdot, \cdot, j_3), \\ & + p(j_1, \cdot, j_3)p(\cdot, j_2, \cdot) + p(j_1, \cdot, \cdot)p(\cdot, j_2, j_3) \} \end{aligned} \quad (3.3.7)$$

where e.g. $p(j_1, j_2, \cdot)$ is the marginal distribution on the two first loci.

Using (2.1),(2.3) and (2.7) the marginal probabilities on a single locus or on two loci can be estimated and we obtain a consistent estimator of $p(j_1, j_2, j_3)$ by:

$$\begin{aligned} \hat{p}(j_1, j_2, j_3) &= 4\hat{A}(j_1, j_2, j_3) - \hat{p}(j_1, j_2, \cdot)p(\cdot, \cdot, j_3) \\ &\quad - \hat{p}(j_1, \cdot, j_3)\hat{p}(\cdot, j_2, \cdot) - \hat{p}(j_1, \cdot, \cdot)\hat{p}(\cdot, j_2, j_3). \end{aligned} \quad (3.3.8)$$

3.3.1 A five Loci simulation

In order to illustrate our estimation algorithm and to analyse its properties by simulation we have considered a five loci example where each locus has two possible alleles only (denoted by 0 and 1). A chromosome is then characterized by a sequence of five 0 and 1 generated by the following distribution: the first element has a probability $\frac{1}{2}$ to be equal to 0 and 1 and the probability of a loci to be identical to the previous one is $\alpha \in [0, 1]$. For example $P(0, 0, 1, 1, 0) = \frac{1}{2} \times \alpha \times (1 - \alpha) \times \alpha \times (1 - \alpha)$. Equivalently the sequence of loci is a stationary markov chain with transition probabilities:

$$Prob(0/0) = Prob(1/1) = \alpha. \quad (3.3.1)$$

If $\alpha = \frac{1}{2}$ all the loci are independent and the linkage disequilibrium increases if α goes to 1 or 0. Simulations are realized with $\alpha = 0,8$. An interest of this example is to produce a large variation of haplotype's probabilities (between 0,2 and 0,0008). Haplotype data for an individual has the form:

$$[(1, 1, 1, 0, 0), (0, 1, 0, 0, 1)].$$

Genotypes data are then constituted by five couples of observations:

$$[(0, 1)(1, 1)(0, 1)(0, 0), (0, 1)].$$

For each individual we create 32 possible sequences of alleles on 5 loci (actually 32 phases are possible generating 64 chromosomes but in that case each chromosome is repeated twice).

Computation are done under *Stata*² and the command for generating this set of chromosome is "append using". Then for n individual we generate 32 n possible sequences of 5 loci.

The next step consists to count alleles on the marginal, the pairs, the triplets and the quadruplets. In practice each sequence (a, b, c, d, e) is transformed into a row:

$$(a, b, c, d, e, ab, ac, ae, bc, \dots, abc, abd, \dots, abcd, abce, \dots, abcde)$$

of 31 elements (5 loci, 10 pairs, 10 triplets, 5 quadruplets and 1 quintuplet). This matrix is created using "egen" and "merge" instructions of *Stata*.

The number of possible recurrences of each column of this matrix is obtained by the instruction "contract".

Finally the estimation of each of the 32 probabilities are obtained by the formulae:

$$\begin{aligned} \hat{p}(j_1, j_2, j_3, j_4, j_5) &= \frac{1}{2^n} N(j_1, j_2, j_3, j_4, j_5) \\ &- \{ \hat{p}(j_1, \dots, \dots) \hat{p}(\dots, j_2, j_3, j_4, j_5) + \hat{p}(\dots, j_2, \dots, \dots) \hat{p}(j_1, \dots, j_3, j_4, j_5) \\ &+ \hat{p}(\dots, \dots, j_3, \dots) \hat{p}(j_1, j_2, \dots, j_4, j_5) + \hat{p}(\dots, \dots, \dots, j_4, \dots) \hat{p}(j_1, j_2, j_3, \dots, j_5) \\ &+ \hat{p}(\dots, \dots, \dots, j_5) \hat{p}(j_1, j_2, j_3, j_4, \dots) + \hat{p}(j_1, j_2, \dots, \dots) \hat{p}(\dots, \dots, j_3, j_4, j_5) \\ &+ \hat{p}(j_1, \dots, j_3, \dots) \hat{p}(\dots, j_2, \dots, j_4, j_5) + \hat{p}(j_1, \dots, \dots, j_4, \dots) \hat{p}(\dots, j_2, j_3, \dots, j_5) \\ &+ \hat{p}(j_1, \dots, \dots, j_5) \hat{p}(\dots, j_2, j_3, j_4, \dots) + \hat{p}(\dots, j_2, j_3, \dots) \hat{p}(j_1, \dots, \dots, j_4, j_5) \\ &+ \hat{p}(\dots, j_2, \dots, j_4, \dots) \hat{p}(j_1, \dots, j_3, \dots, j_5) + \hat{p}(\dots, j_2, \dots, \dots, j_5) \hat{p}(j_1, \dots, j_3, j_4, \dots) \\ &\hat{p}(\dots, \dots, j_3, j_4, \dots) \hat{p}(j_1, j_2, \dots, \dots, j_5) + \hat{p}(\dots, \dots, j_3, \dots, \dots, j_5) \\ &\hat{p}(j_1, j_2, \dots, j_4, \dots) + \hat{p}(\dots, \dots, \dots, j_4, j_5) \hat{p}(j_1, j_2, j_3, \dots, \dots) \end{aligned} \quad (3.3.2)$$

where $N(j_1, j_2, j_3, j_4, j_5)$ is the number of sequence of $(j_1, j_2, j_3, j_4, j_5)$ in the 32n created sequences.

The intermediate marginal probabilities are recursively computed used formulae (2.1), (2.5) and (2.8).

Some results are given in tables I and II. In table I three estimations are performed for three different sample size. In each case have computed our estimator (where the negative probabilities are transformed into 0) and for comparison, the estimator derived from the haplotypes' observation (FH).

²Copy of the program is available upon request at feve@cict.fr

Table II gives more precisions in case of a sample size $n = 1000$. We have performed a monte carlo simulation of size 100 and we compare the properties of estimation derived from haplotypes' observation and of our estimator (non corrected or corrected for negative probabilities). The comparaison shows the performance of our estimator relative to the (unfeasible) case where haplotypes are observable. It should be underlined that corrected estimator gives probabilities which do not sum to 1. In that case the mean sum is 1.15. This may be corrected by renormalizing the estimated probabilities.

Finally let us underlined that moment estimation is computed extremely fast (less then 30 second for $n = 1000$ and less than 6 minutes for $n = 10000$ using *Stata* through *windows* on a Dell/Pentium 3 (1.13 gigahertz) computer). Moreover the computation does not require any specification of a starting value of the parameters.

Table I : Three Estimations

Haplotype	probability	n=1 000		n = 5 000		n = 10 000	
		Hapl Freq	Moment Est	Hapl Freq	Moment Est	Hapl Freq	Moment Est
0 0 0 0	.2048	.206	.2088298	.1989	.2011688	.20255	.2112688
0 0 0 1	.0512	.052	.0636018	.0529	.0632893	.05055	.0583022
0 0 1 0	.0128	.0145	0	.0119	0	.0135	0
0 0 1 1	.0512	.0475	.0162172	.051	.0091953	.0527	.0122803
0 1 0 0	.0128	.012	.0149978	.0114	.0056206	.0123	.0043247
0 1 0 1	.0032	.005	0	.0035	0	.0031	0
0 1 1 0	.0128	.0115	.0292689	.0126	.0399666	.01305	.0352466
0 1 1 1	.0512	.0515	.0826214	.0507	.0889362	.05005	.0912542
1 0 0 0	.0128	.013	.0175201	.0145	.0255962	.01295	.0207061
1 0 0 1	.0032	.004	.01137618	.0031	.0096301	.00305	.0096966
1 0 1 0	.0008	.0015	0	.0007	0	.0012	0
1 0 1 1	.0032	.0025	0	.0028	0	.0035	0
1 1 0 0	.0128	.0165	.0060716	.0116	.0004915	.013	.0043296
1 1 0 1	.0032	.0035	0	.0031	0	.0035	0
1 1 1 0	.0128	.013	.0461852	.0123	.0479367	.01355	.0520412
1 1 1 1	.0512	.0495	.0783266	.0512	.0724294	.05345	.0726751
1 0 0 0	.0512	.048	.0706784	.0519	.0839282	.0495	.0712941
1 0 0 1	.0128	.0175	.0509093	.0119	.0463588	.0127	.0476921
1 0 1 0	.0032	.0055	0	.0029	0	.00315	0
1 0 1 1	.0128	.0175	.0119039	.013	.0059473	.0129	.0063592
1 1 0 0	.0032	.0025	0	.0034	0	.003	0
1 1 0 1	.0008	0	0	.0007	0	.00065	0
1 0 1 0	.0032	.0025	.015952	.0029	.0093975	.00325	.0077471
1 0 1 1	.0128	.0155	.0244887	.0149	.0209064	.0116	.0213855
1 1 0 0	.0512	.046	.0826132	.0496	.0841322	.05305	.0914693
1 1 0 1	.0128	.018	.0395336	.012	.0378055	.01275	.0367616
1 1 1 0	.0032	.0045	.0024004	.0028	0	.00315	0
1 1 1 1	.0128	.014	.0015409	.0135	.0045156	.01205	.0043965
1 1 1 0	.0512	.038	.0119291	.0525	.0154248	.051	.0120378
1 1 1 1	.0128	.012	0	.0144	0	.0123	0
1 1 1 0	.0512	.0595	.0639496	.0551	.0640807	.05185	.0607033
1 1 1 1	.2048	.195	.2001555	.2063	.2194556	.2051	.2109377

Table II : A Monte Carlo simulation

Haplotype	probability	Haplotype Frequency			Moment Estimator			Corrected Moment Estimator			MSE
		mean	bias	variance	mean	bias	variance	mean	bias	variance	
0 0 0 0	0,2048	0,206525	0,001725	0,0000895	0,0000925	0,0000227	0,0000227	0,205974	0,001174	0,0002256	0,000227
0 0 0 1	0,0512	0,051245	0,000045	0,0000238	0,0000238	0,0000238	0,0000238	0,0634829	0,0122829	0,000107	0,0002578
0 0 1 0	0,0128	0,01465	0,00185	5,23E-06	8,66E-06	0,0004697	0,000413	0,000487	-0,012313	3,74E-06	0,0001553
0 0 1 1	0,0512	0,047445	-0,003755	0,0000209	0,000035	0,0000929	0,0000929	0,0159665	-0,035235	0,0000879	0,0013293
0 1 0 0	0,0128	0,01183	-0,00097	4,95E-06	5,89E-06	0,000485	0,0000645	0,0152697	0,0024697	0,0000479	0,000054
0 1 0 1	0,0032	0,0046465	0,0014465	2,44E-06	4,53E-06	0,0001596	0,0001596	0,0003701	-0,0028299	1,59E-06	9,60E-06
0 1 1 0	0,0128	0,01155	-0,00125	5,54E-06	7,11E-06	0,000323	0,0002901	0,0288541	0,0160541	0,0000323	0,0002901
0 1 1 1	0,0512	0,05093	-0,00027	0,0000249	0,0000249	0,0000323	0,0002901	0,0288541	0,0160541	0,0000323	0,0002901
1 0 0 0	0,0128	0,012885	0,000065	6,80E-06	6,81E-06	0,0000954	0,0010666	0,0823638	0,0311638	0,0000954	0,0010666
1 0 0 1	0,0032	0,0042368	0,0010368	1,84E-06	2,92E-06	0,000481	0,0000731	0,0177907	0,0049907	0,0000481	0,0000731
1 0 1 0	0,0008	0,0019417	0,0011417	6,96E-07	2,00E-06	0,0000188	0,0001466	0,0177907	0,0049907	0,0000481	0,0000731
1 0 1 1	0,0032	0,0026812	-0,0005187	1,06E-06	1,33E-06	0,000251	0,0001995	0,0145054	0,0113054	0,0000188	0,0001466
1 1 0 0	0,0128	0,016765	0,003965	6,87E-06	0,0000226	0,0000274	0,0006078	0	-0,0008	0	6,40E-07
1 1 0 1	0,0032	0,0034521	0,0002521	1,60E-06	1,66E-06	0,0000274	0,0006078	0,00736	-0,0052	0	0,0000102
1 1 1 0	0,0128	0,01303	0,00023	5,17E-06	5,22E-06	0,000033	0,0000664	0,00736	-0,00544	0,0000258	0,0000554
1 1 1 1	0,0512	0,049535	-0,001665	0,0000236	0,0000236	0,000033	0,0000664	0,004763	-0,0027237	1,94E-06	9,35E-06
1 0 0 0	0,0512	0,047755	-0,003445	0,0000287	0,0000405	0,0000523	0,0001497	0,0459265	0,0331265	0,0000523	0,0001497
1 0 0 1	0,0128	0,016925	0,004125	8,92E-06	0,0000259	0,0001174	0,0008754	0,0787319	0,0275319	0,0001174	0,0008754
1 0 1 0	0,0032	0,00526	0,00206	2,53E-06	6,78E-06	0,000168	0,0000935	0,0718873	0,0206873	0,0001274	0,0005554
1 0 1 1	0,0128	0,01736	0,00456	6,88E-06	0,000277	0,000428	0,001458	0,0505185	0,0377185	0,0000353	0,001458
1 0 1 0	0,0032	0,0029096	-0,0002904	1,14E-06	1,23E-06	0,0000275	0,0000454	0,0000858	-0,0031142	1,37E-07	9,83E-06
1 0 1 1	0,0008	0	-0,0008	0	0,0000064	0,0000254	0,00010321	0,0113397	-0,0014603	0,0000383	0,0000404
1 0 1 0	0,0032	0,0028141	-0,0003859	1,08E-06	1,23E-06	0,0000275	0,0008812	0	-0,0032	0	0,0000102
1 0 1 1	0,0128	0,01524	0,00244	6,17E-06	0,0000121	0,0000275	0,0001787	0,0154983	0,0122983	0,0000275	0,0001787
1 1 0 0	0,0512	0,04627	-0,00493	0,0000235	0,0000478	0,0000704	0,0002075	0,0245094	0,0117094	0,0000704	0,0002075
1 1 0 1	0,0128	0,01772	0,00492	6,07E-06	0,0000303	0,0000321	0,0007334	0,0826416	0,0314416	0,0000726	0,0010611
1 1 0 1	0,0032	0,0043838	0,0011838	1,85E-06	3,25E-06	0,000039	0,000039	0,039283	0,026483	0,0000321	0,0007334
1 1 0 1	0,0128	0,01438	0,00158	6,12E-06	8,62E-06	0,000041	0,0001581	0,0037508	0,0005308	0,0000193	0,0000196
1 1 1 0	0,0512	0,037695	-0,013505	0,0000154	0,0001978	0,0000975	0,0016811	0,0033879	-0,0094121	0,000024	0,0001126
1 1 1 1	0,0128	0,012425	-0,000375	5,60E-06	5,74E-06	0,0000383	0,0009316	0,0118887	-0,0393113	0,0000821	0,0016275
1 1 1 0	0,0512	0,060145	0,008945	0,0000385	0,0001185	0,0001555	0,0003402	0,0647904	0,0135904	0,0001555	0,0003402
1 1 1 1	0,2048	0,196315	-0,008485	0,0000971	0,0001691	0,0002437	0,0002572	0,2011142	-0,0036859	0,0002437	0,0002572

3.3.2 An application to HLA haplotype frequencies and comparison with EM algorithm

Using FGM basis of 96000 individuals, we have reconstructed the frequencies of the haplotypes A,B,DR (at the 2 Digits precision). Using results from Gourraud (2005) and personal communication, we have compare our approach and EM results.

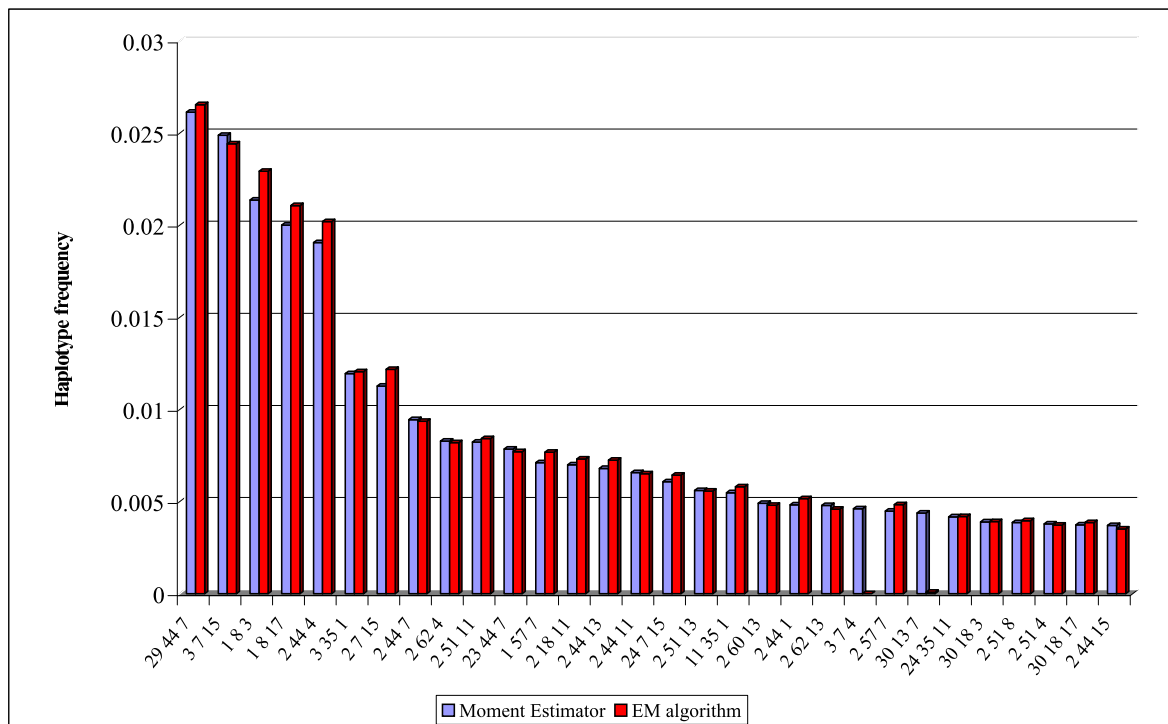


Figure 3.1: Comparison between Moment Estimator and EM algorithm

Haplotype	Moment Estimator	EM algorithm
29 44 7	0.026143	0.026550
3 7 15	0.024884	0.024420
1 8 3	0.021372	0.022940
1 8 17	0.020016	0.021074
2 44 4	0.019068	0.020201
3 35 1	0.011949	0.012058
2 7 15	0.011276	0.012179
2 44 7	0.009455	0.009360
2 62 4	0.008293	0.008202
2 51 11	0.008232	0.008430
23 44 7	0.007853	0.007713
1 57 7	0.007117	0.007691
2 18 11	0.006998	0.007315
2 44 13	0.006806	0.007250
2 44 11	0.006578	0.006519
24 7 15	0.006087	0.006441
2 51 13	0.005606	0.005573
11 35 1	0.005487	0.005809
2 60 13	0.004907	0.004806
2 44 1	0.004826	0.005163
2 62 13	0.004797	0.004595
3 7 4	0.004616	0.000000
2 57 7	0.004494	0.004838
30 13 7	0.004379	0.000090
24 35 11	0.004178	0.004189
30 18 3	0.003904	0.003922
2 51 8	0.003863	0.003967
2 51 4	0.003793	0.003728
30 18 17	0.003738	0.003865
2 44 15	0.003713	0.003519
1 8 15	0.003523	0.003512
2 13 7	0.003511	0.003482
2 7 4	0.003510	0.003824
2 50 7	0.003468	0.003659
31 60 4	0.003436	0.003509
2 62 11	0.003390	0.003604
24 18 11	0.003351	0.003776
25 18 15	0.003336	0.000436
2 27 1	0.003230	0.002226
26 38 13	0.003220	0.003054
3 7 1	0.003198	0.002939
2 8 3	0.003155	0.003550
3 35 11	0.003150	0.003278
2 51 1	0.003014	0.008430
2 35 11	0.002972	0.002924
24 44 7	0.002935	0.002906
1 8 13	0.002830	0.003000
2 51 7	0.002816	0.002873
1 7 15	0.002770	0.002608
1 8 4	0.002761	0.002511

Figure 3.2: The "50s" most frequent haplotypes

3.4 How many different HLA genotypes exist in a population?

3.4.1 The problem

We consider a population of N individuals where each individual has an unknown "type" (namely the HLA genotype (A, B, DR two digits)). We observe a sample of n individuals. In this sample J_n different types are observed. The question is how to deduce from this information an estimation of J_N , the number of different types in the complete population of size N . We present in this paper a statistical model which lead to a number of different types proportional to the logarithm of the population size. If multiples observations are available the model introduces overidentification constraints which may be tested.

Actually two approaches may be adopt to analyse the problem of the number of types. The first one is a finite population approach : we assume a large population where individuals have types and we consider the observed population as a sample obtained through a random survey. In that case the only stochastic element of the data generating process is the survey mechanism (see references in the survey by Haas et al. 1995). The statistical problem is then to infer a characteristic of the total population (namely the number of different types) from the observed sample. The second one, which is adopted in this note consists in the specification of a model generations of types which allows ties. The sample is then used in order to estimate (and to test the model). Finally the estimated model is used for prediction for the whole population.

3.4.2 A Polya urn scheme for types' generation

We first describe a simple one parameter statistical model explaining types' generation. This model is purely descriptive because it does not consider the haplotypes'level. We consider a continuous (non atomic) probability measure on \mathbb{R}^+ characterized by its cdf F_0 . This distribution describes the latent generation of all possible types. As F_0 is non atomic the probability of ties in a sample from F_0 is zero. This assumption is only an

approximation : the distribution of all possible genotypes is actually discrete (The number of DNA sequences is finite) but extremely much higher than the size of the population. Then the ties in the population cannot come from the discreteness of F_0 and should be modeled explicitly.

A sequential generation of types may be expressed by this Polya urn scheme :

1. the first individual has a type generated randomly by F_0 .
2. let us assume that i individuals ($i \geq 1$) has been generated and that F_i represents the empirical distribution of the observed types. Then the $i + 1$ the individual will have a type generated by F_0 with a probability $\frac{n_0}{n_0+i}$ and by F_i with a probability $\frac{i}{n_0+i}$. Then the probability of a new type is $\frac{n_0}{n_0+i}$ and n_0 is an unknown parameter.

Despite this sequential description the joint distribution of the types across the individuals is exchangeable : the distribution is invariant by individual permutations (individuals are anonymous). This exchangeability property may be derived from the interpretation of this distribution as the predictive probability of a bayesian non parametric model. Let F is an unknown distribution function generated by a Dirichlet process parametrized by n_0 and F_0 and assume that x_1, \dots, x_n are generated given F by an iid sampling process. Then the unconditional distribution of (x_1, \dots, x_n) follows the preceding model.

An elementary computation shows that the probability to observe in a sample J different types where each type j is repeated k_j times is:

$$\frac{(n_0 - 1)!}{(n_0 + n - 1)!} n_0^J \prod_{j=1}^J (k_j - 1)! = (n_0 - 1)! n_0^J \frac{\prod_{j=1}^J \Gamma(k_j)}{\Gamma(n_0 + n)}$$

where $1 \leq J \leq n, 1 \leq k_j \leq n - J$ and $\sum_{j=1}^J k_j = n$.

3.4.3 Exact and approximate number of types

The marginal distribution on J is difficult to derive for the previous formulae. Actually it can be deduce from the following recursive argument. If we call \tilde{J}_n the random variable

generating the number of different types in a sample of size n we have:

$$P(\tilde{J}_n = k) = \frac{n_0}{n_0 + n - 1} P(\tilde{J}_{n-1} = k - 1) + \frac{n - 1}{n_0 + n - 1} P(\tilde{J}_{n-1} = k) \quad k = 1, \dots, n$$

This recursive relation generates a probability distribution:

$$P(\tilde{J}_n = k) = \frac{(n_0 - 1)!}{(n_0 + n - 1)!} n_0^k s_{n,k}$$

where $s_{n,k}$ does not depend on n_0 and are defined by the property:

$$\frac{(n_0 + n - 1)!}{(n_0 - 1)!} = \sum_{k=1}^n n_0^k s_{n,k}$$

Following Rolin (1993) we may reconstruct \tilde{J}_n by the following way. Let u_i a random element in $\{0, 1\}$ which takes the value 1 if the individual i has a new type and 0 else.

Then:

$$\tilde{J}_n = \sum_{i=1}^n u_i$$

The u_i are independent and

$$P(u_i = 1) = \frac{n_0}{n_0 + i - 1}$$

Then

$$E(\tilde{J}_n) = \sum_{i=1}^n \frac{n_0}{n_0 + i - 1} = n_0 \{ \psi(n_0 + n) - \psi(n_0) \}$$

where $\psi(t)$ is the digamma function $\psi(t) = \frac{d}{dt} \ln \Gamma(t)$.

Thank to the approximation $\frac{\psi(t)}{\ln t} \rightarrow 1$ if t large, we have :

$$E(\tilde{J}_n) \simeq n_0[\ln(n_0 + n) - \ln n_0]$$

The variance of \tilde{J}_n is obtained by :

$$V(\tilde{J}_n) = \sum_{i=1}^n V(u_i) = \sum_{i=1}^n \frac{n_0 (i-1)}{(n_0 + i - 1)^2}$$

and

$$V(\tilde{J}_n) \leq n_0(\psi(n_0 + n) - \psi(n_0))$$

thanks to

$$\frac{i-1}{n_0 + i - 1} \leq 1$$

.

Actually, an application of Lindeberg Feller theorem shows that (see Rolin(1993)) :

- i) $[n_0(\ln(n_0 + n) - \ln(n_0))]^{\frac{1}{2}}[\tilde{J}_n - n_0(\ln(n_0 + n) - \ln(n_0))] \rightarrow \mathcal{N}(0, 1)$
- ii) $\tilde{J}_N - \tilde{J}_n$ is independent in distribution to \tilde{J}_n and, asymptotically

$$\tilde{J}_N | \tilde{J}_n \sim \mathcal{N}(\tilde{J}_n + n_0(\ln(N + n_0) - \ln(n + n_0)), n_0[\ln(N + n_0) - \ln(n + n_0)])$$

3.4.4 Estimation of n_0 and prediction of the number of types

Let us consider first the case where a single sample of size n is available. We observe in the sample J_n distinct types. The log likelihood is then equal to :

$$\ln \ell = J \ln n_0 + \ln \Gamma(n_0) - \ln \Gamma(n_0 + n) + \text{constant}$$

and the first order conditions reduces to :

$$n_0 = \frac{J_n}{\psi(n_0 + n) - \psi(n_0)}$$

which has no solution in a closer form.

Using the approximation $\psi(t)/\ln t \rightarrow 1$ this equation simplifies into:

$$J_n = n_0[\ln(n_0 + n) - \ln n_0]$$

This estimator may also be view as the solution of the moment condition

$$E(\tilde{J}_n) = n_0[\ln(n_0 + n) - \ln n_0]$$

where the expectation is replaced by the observed value. Indeed, let us compute first an approximation of the variance of \hat{n}_0 . Using a Taylor expression we derive from

$$J_n = \hat{n}_0[\ln(\hat{n}_0 + n) - \ln \hat{n}_0]$$

the relation :

$$V(\hat{n}_0) \simeq [\ln(\hat{n}_0 + n) - \ln \hat{n}_0 + \frac{n_0}{n_0 + n} - 1]^{-2} V(J_n)$$

and we get

$$\hat{n}_0 \simeq N(n_0, \frac{\hat{n}_0(\ln(\hat{n}_0 + n) - \ln \hat{n}_0)}{[\ln(\hat{n}_0 + n) - \ln \hat{n}_0 + \frac{\hat{n}_0}{(\hat{n}_0 + n)} - 1]^2})$$

The speed of convergence of \hat{n}_0 to n_0 is actually of order $\ln n$.

In the case of the prediction of J_N we can use the predictor :

$$\hat{J}_N = J_n + \hat{n}_0 [\ln(N + \hat{n}_0) - \ln(n + \hat{n}_0)]$$

$$V(J_N - \hat{J}_N | J_n) = V[\hat{n}_0 [\ln(N + \hat{n}_0) - \ln(n + \hat{n}_0)]]$$

$$V(J_N - \hat{J}_N | J_n) \simeq [\ln(N + \hat{n}_0) - \ln(n + \hat{n}_0) + \hat{n}_0 \left(\frac{1}{N + \hat{n}_0} - \frac{1}{n + \hat{n}_0} \right)]^2 V(\hat{n}_0)$$

Example: In the France Greffe de Moelle registry we have observe 107 925 individuals and 66 164 different types are present. The estimator \hat{n}_0 is then equal to $\hat{n}_0 = 72 703$ with

a standard deviation equal to 823. A confidence interval at 95 % is then $[72\ 703 \pm 1\ 613] = [71\ 090, 74\ 316]$.

The forecast of the number of types for a population of 60 000 000 individuals is :

$$\widehat{J}_n = 66\ 164 + 72\ 703[\ln [60\ 000\ 000 + 72\ 703] - \ln [107\ 925 + 72\ 703]] = 488\ 340$$

with a standard deviation equal to 44 880.

Finally the probability that a new donor does not have a previously observed type is equal to : 0.40

3.4.5 A repeated observations case and a test of the model

We now assume that we observe an increasing sample of individuals at different dates t_1, \dots, t_p . At t_i the file contain n_i individual and J_{n_i} different types. Using previous properties we have the asymptotic approximation.

$$\begin{aligned} J_{n_1} &\sim N(n_0 \ln n_1, n_0 \ln n_1) \\ J_{n_2} - J_{n_1} &\sim N(n_0(\ln n_2 - \ln n_1), n_0(\ln n_2 - \ln n_1)) \\ &\vdots \\ J_{n_p} - J_{n_{p-1}} &\sim N(n_0(\ln n_p - \ln n_{p-1}), n_0(\ln n_p - \ln n_{p-1})) \end{aligned}$$

and all these distributions are independent. We then propose a weighted mean square estimation of n_0 :

$$\widehat{n}_0 = \arg \min \sum_{i=2}^p \frac{[(J_{n_i} - J_{n_{i-1}}) - n_0(\ln(n_0 + n_i) - \ln(n_0 + n_{i-1}))]^2}{n_0(\ln(n_0 + n_i) - \ln(n_0 + n_{i-1}))}$$

Unfortunately this optimisation problems conduct to the estimator \widehat{n}_0 solution of $J_{n_p} = n_0 (\ln(n_0 + n_p) - \ln n_0)$ which only uses the last observation. However the statistics

$$\sum_{i=2}^p \frac{[(J_{n_i} - J_{n_{i-1}}) - \widehat{n}_0(\ln(\widehat{n}_0 + n_i) - \ln(\widehat{n}_0 + n_{i-1}))]^2}{\widehat{n}_0(\ln(\widehat{n}_0 + n_i) - \ln(\widehat{n}_0 + n_{i-1}))}$$

may be used as a test statistic of the model validity. Under the assumption of correct specification this statistic follows a χ_{p-1}^2 distribution and if this value is statistically large the model may be rejected.

Example:

Let us consider the FGM file of 107 925 individuals in 2004. We have not the history of the file but in order to test our model we propose the following strategy: for the original file we may draw uniformly without replacement) increasing subfiles.

For example we have done the next simulation:

n	J_n	Estimated J_n	χ^2
107 925	66 164	66 164	-
86 427	55 617	56 951	193.2
64 789	44 139	46 325	68.3
43 170	31 529	33 888	2.4
21 551	17 328	18 874	44.0
			$\Sigma = 307.9$

The column "estimated J_n " compute the number of types deduced from our model with $n_0 = 77\ 203$. Even if the model performs correctly it is rejected by the test (307.9 is greater than the χ_p^2 threshold at 5% which is equal to 9.48). A conclusion of this simulation is that our model under estimate the number of types of larger samples than the actual one. This point is confirmed by an observation at the world level: the BMDW ³says that around 400 000 have been observed for a sample of 6 000 000 individuals and our model predicts 330 000 types only.

These results and the test are motivations for extensions of our model. A possible direction to extend our model is to consider heterogenous populations. Even if this is eventually a topic for future researches, let us consider the following example. Let us assume that the population of size N is separated into two subpopulations of sizes N_1 and N_2 and the preceding Polya urne scheme applied to each population with two different parameters n_{01} and n_{02} . We assume moreover that the two populations are independently generated. Finally the two populations has no common types.

³Bone Marrow Donors Worldwide

Then the expected number of types will be:

$$E(\tilde{J}_N) = E(\tilde{J}_{N_1}) + E(\tilde{J}_{N_2})$$

where

$$E(\tilde{J}_{N_1}) = n_{01} (\ln(N_1 + n_{01}) - \ln n_{01})$$

and

$$E(\tilde{J}_{N_2}) = n_{02} (\ln(N_1 + n_{02}) - \ln n_{02})$$

using repeated observations (or resampling in a file) the three parameters of the model (n_{01} , n_{02} and α , the proportion of the two subpopulations) may be estimated using at least three observations. The implementation and the properties of this approach will be developed in the future.

Conclusions et Extensions

Ce travail de recherche utilise le calcul économique et les outils statistiques pour modéliser le mode de régulation des registres de donneurs de CSH. On utilise les arguments de la théorie de la décision pour déterminer comment optimiser l'organisation du système actuel de don de cellules en vue de greffes. On retient comme critère d'optimisation la probabilité pour un receveur quelconque de trouver un donneur. On met en évidence la valeur du Registre si donneurs et receveurs sont tirés d'une même population du point de vue des fréquences des groupes HLA (pas de sélection) ainsi que la valeur du Registre liée à la sélection optimale des donneurs.

Quel que soit le critère de sélection retenu, notre travail souligne la faible efficacité de l'organisation d'un registre de CSH au niveau d' "une seule" entité géographique (pays, région, ethnie): la probabilité de ne pas trouver de donneur compatible avec un receveur potentiel demeure très élevée même pour des tailles de registres importantes si l'on néglige la dimension internationale des registres. Notre étude pourrait être transposée au niveau mondial en supposant que tous les Registres fonctionnent comme un Registre "intégré". On pourrait également y intégrer différentes sources de CSH: les donneurs volontaires et le sang placentaire . La modélisation de ces deux sources de CSH fait l'objet de travaux en cours.

Par ailleurs, les responsables du Registre français soulignent le fait que l'accroissement du registre est davantage limité par des considérations budgétaires que par manque de volontaires. Notre travail évalue l'efficacité d'un registre de donneurs de CSH à la lumière du calcul économique et permet d'identifier les éléments-clé qui entrent en jeu. Certains peuvent être évalués à partir d'arguments statistiques (nombre et distribution des types HLA en France), d'autres peuvent évoluer en fonction de la politique de gestion du fichier et d'autres enfin relèvent d'une évaluation plus complexe liée au bénéfice attendu d'une

greffe. L'organisation des registres pour de très nombreux pays et au niveau mondial est une organisation complexe et coûteuse. On modélise le bien-être social apporté par cette organisation et propose une stratégie d'optimisation de l'évolution des registres. Le modèle calibré proposé montre l'incidence relativement faible de l'accroissement des registres avec typages HLA à faible résolution comparé à l'accroissement de la disponibilité effective des donneurs typés HLA à haute résolution et à la diminution des coûts. On a montré que la probabilité de ne pas trouver de donneur compatible avec un receveur potentiel demeure très élevée même pour des tailles de registre importantes. Au niveau d'un seul pays, on peut conclure que l'accroissement du fichier des donneurs ne se justifie pas au delà d'une certaine taille mais que d'autres mesures peuvent être plus utiles (augmentation de la disponibilité des donneurs pleinement compatibles, meilleure qualité de typage).

Dans un troisième temps, on apporte une contribution statistique à la sélection des donneurs : on pose le problème de la distribution jointe de plusieurs variables discrètes (les haplotypes) connaissant les phénotypes des individus (variables observables). On propose une méthode simple d'estimation que l'on compare par simulations à celle de la méthode utilisée habituellement par les biostatisticiens. Cette question statistique est un élément du problème décisionnel relatif à l'optimisation du fichier de donneurs de CSH défini au début de cette thèse. Deux extensions de la partie statistique de notre travail sont en cours: la première étant l'extension de cette méthode d'estimation à un nombre toujours plus important d'allèles et de locus (optimisation de la méthode de comptage), la seconde étant l'étude du cas où les probabilités jointes des allèles sont égales à 0 strictement. La dernière partie de ce travail de thèse suggère un petit modèle statistique qui montre le lien de proportionnalité existant entre le nombre de phénotypes HLA (différents) dans une population et le logarithme de la taille de cette population. Ce résultat est basé sur la modélisation du nombre de types d'une population à partir de schémas d'urnes de Polya. Ce modèle est actuellement l'objet d'extensions et d'un traitement systématique des problèmes d'estimation.

Au total ce travail a permis:

1. une modélisation théorique des registres qui n'avait jamais été entreprise et sa validation sur un Registre national
2. la définition des paramètres utiles pour une optimisation dynamique des registres
3. la proposition d'une nouvelle méthode d'estimation des fréquences haplotypiques populationnelles qui trouvera d'autres terrains d'application en génétique épidémiologique

Les travaux qui vont se poursuivre à la suite de cette thèse vont s'appuyer sur un consortium international et sceller une collaboration à plus long terme entre les deux équipes de recherche qui ont permis la réalisation des travaux présentés ici.

Bibliographie

Andrews, D. (1999), Estimation when a Parameter is on the Boundary, *Econometrica*, 67, 1341-1383.

Arrow, K.J., Uncertainty and the Welfare Economics of Medical Care, *the American Economic Review*, 1963, vol.53, (5)941 – 973

Berchery D. (2002), " Le Registre français des donneurs volontaires de moelle osseuse: évaluation médico-économique de l'introduction des microsattellites, Thèse de Doctorat de Médecine, Université Toulouse III

Berchery, D., Molinier L., Baouz A., Raffoux C., Cambon-Thomsen A., Cost-effectiveness analysis of two strategies for typing unrelated donors for bone marrow transplantation in France, *Eur J Health Econom* 2003.4:130-137

Cazals C., Fève F., Fève P., Florens J.P., Simple Structural Econometrics of Price Elasticity, *Economics Letters*, vol.86, n°1, janvier 2005, p 1-6.

Clark, A. (1990), Inference of Haplotypes from PCR-amplified Samples of Diploid Populations, *Mol. Biol. Evol.*, 7(2): 111-122.

Dempster A.P., Laird N.M., Rubin D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society Series B*, 39, 1-38.

Drummond Michael F., O'Brien Bernie J., Stoddart Greg L., Torrance George W., méthodes d'évaluation économique des programmes de santé, *Economica*, 1997.

Dudbridge, F. B. Koeleman, J. Todd, D. Clayton (2000) Unbiased Applications of the

Transmission/Disequilibrium Test to Multilocus Haplotypes, *Am. J. Hum. Genet*; 66: 2009-2012.

Efron B., (1982), *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.

Excoffier, L. M. Slatkin (1995), Maximum Likelihood Estimation of Molecules Haplotype Frequencies in a Diploid Population, *Mol. Biol. Evol.*, 12(5): 921-927.

Fève F., Florens J.P., Roy B., *Delivery Costs II: Back to Parametric Models, Regulatory and Economic Challenges in the Postal and Delivery Sector*, sous la direction de M. A. Crew and P. R. Kleindorfer, Kluwer, Boston, 2005.

Fève F., Florens J.P. (2003), "A Moment Estimation of the Haplotypes Distribution using Phenotypes Data", working paper, University of Toulouse.

Fève F., Florens J.P. (2003), "Matching Models and Optimal Registry for Voluntary Organ Donation Registries", working paper, University of Toulouse.

Fève F., Florens J.P. (2005), "How many different HLA genotypes exist in a population?", working paper, University of Toulouse.

Fève F., Cambon A., Eliaou J.F., Gourraud P.A., Raffoux C., Florens J.P., (2005) "Evaluation économique de l'organisation d'un Registre de donneurs de Cellules Souches Hématopoïétiques", working paper, University of Toulouse.

France Greffe De Moelle, Fichier National de donneurs de Cellules Souches Hématopoïétiques, Rapport d'activité 2004, Hôpital Saint-Louis, Paris, France.

Hall, P. (1999), *The Bootstrap and Edgeworth Expression*, Springer Verlag, New York.

Gourraud P.A., Fève F., Florens J.P., Cambon-Thomsen A., *Models of donors registries*

optimal HLA Composition, Genes and Immunity, vol.5, n°1, 2004.

Gourraud P.A. (2005), "Utilisation d'analyses d'haplotypes dans l'organisation de la transplantation de cellules souches Hématopoïétiques", Thèse de Doctorat de Biologie, Université Toulouse III

Haas P.J., Naughton J.F., Seshadri S., Stokes L., *Sampling-Based Estimation of the Number of Distinct Values of an Attribute*, in VLDB'95, Proceedings of 21th International Conference on very Large data Bases, sept 11-15, 1995, Zurich, Zwitzerland, 311-322, Morgan Kaufman

Heemskerk MBA, Van Walraven SM, Cornelissen JJ, Barge TMY et alii (2005), How to improve the search for an unrelated haematopoietic stem cell donor. Faster is better than more!, Bone Marrow Transplantation, 35, 645-652.

Hurley CK, Schreuder GM, Marsh SG, Lau M, Middleton D, Noreen H. (1997), The search for HLA-matched donors: a summary of HLA-A,-B,-DRB1/3/4/5 alleles and their association with serologically defined HLA-A,-B,-DR antigens. Tissue Antigens. Oct;50(4):401-18.

Hoffman-Smith C. (1993), Matching marrow donors & recipients: downsized system helps save more lives. Health Inform. Sep;10(9):18, 20.

Laffont, J.J, Fondements de l'Economie Publique, Economica, 1988

Lonjou C, Clayton J, Cambon-Thomsen A, Raffoux C. (1995), HLA-A,-B,-DR haplotype frequencies in France implications for recruitment of potential bone marrow donors. Transplantation. Aug 27;60(4):375-83.

Li, S., Carlson C., Zhao LP. (2003), Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms, Biostatistics, 4,4: 513-522.

Mas-Colell, A., Whinston M.D., Green J.R, Microeconomic Theory, Oxford University Press, 1995

Marsh SG, Albert ED, Bodmer WF, Bontrop RE, Dupont B, Erlich HA, Geraghty DE, Hansen JA, Hurley CK, Mach B, Mayr WR, Parham P, Petersdorf EW, Sasazuki T, Schreuder GM, Strominger JL, Svejgaard A, Terasaki PI, Trowsdale J., Nomenclature for factors of the HLA system, (2004). *Tissue Antigens*. 2005 Apr;65(4):301-69

Niu, T. Z. Qin, X. Xu, J. Liu (2002), Bayesian Haplotype Inference for Multiple Linked Single. Nucleotide Polymorphisms, *Am.J. Hum. Genet.*, 70: 157-169.

Ottinger H, Grosse-Wilde M, Schmitz A, Grosse-Wilde H. (1994), Immunogenetic marrow donor search for 1012 patients: a retrospective analysis of strategies, outcome and costs. *Bone Marrow Transplant*. 1994-14 Suppl 4:S34-8.

Oudshoorn M, Cornelissen JJ, Fibbe WE, de Graeff-Meeder ER, Lie JL, Schreuder GM, Sintnicolaas K, Willemze R, Vossen JM, van Rood JJ. (1997), Problems and possible solutions in finding an unrelated bone marrow donor. Results of consecutive searches for 240 Dutch patients. *Bone Marrow Transplant*. Dec;20(12):1011-7.

Oudshorn M., Van der Zanden H.G.M., Bakker J.N.A., Van Rood J.J., Bone Marrow Donors Worldwide - the present and the perspectives in facilitating the donor-recipient matching in hematopoietic stem cell transplantation., *textit Focus on immunology research* (in press).

Picard, P., *Elements de Microeconomie*, Montchestien, 1998.

Rendine S, Barbanti M, Borelli I, Dall'Omo AM, Roggero S, Sacchi N, Curtioni ES. (1999), The Italian Registry of Bone Marrow Donors: genetic structure and recruitment strategy - *Ann Ist Super Sanita*.35(1):21-34. Italian.

Rolin, J.M., On the distribution of jumps of the Dirichlet process, Institut de Statistique DP 9302, Universit  Catholique de Louvain, Louvain-la-Neuve, Belgium, 1993

Risch, N., K. Merikangas (1996), The Future of Genetic Studies of Complex Human Dis-

eases, *Science*, New series, 273, 5281, 1516-1517.

Rubin D.B. (1976), *Inference and Missing Data*, *Biometrika*, vol. 63, pp.581 – 592, 13

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Schipper RF, Oudshoorn M, D’Amaro J, van der Zanden HG, de Lange P, Bakker JT, Bakker J, van Rood JJ. (1996), Validation of large data sets, an essential prerequisite for data analysis: an analytical survey of the Bone Marrow Donors Worldwide. *Tissue Antigens*. Mar;47(3):169-78.

Schuler U, Rutt C, Baier D, Keller JV, Stahr A, Grathwohl A, Ehninger G.(2000), Approaches to managing volunteer marrow donor registry HLA data. Algorithms for directing donor center-initiated HLA-DR typing of selected donors. *Rev Immunogenet*. 2000-2(4):541-6.

Single, R.M., D. Meyer, J. Hollenback, M.P. Nelson, J. Noble, H. Erlich, G. Thomson (2002), Haplotype Frequency Estimation in Patient Populations: The effect of Departures from Hardy-Weinberg Proportions and Collapsing over a Locus in the HLA Region, *Genetic Epidemiology* 22: 186-195.

Speiser DE, Tiercy JM, Rufer N, Chapuis B, Morell A, Kern M, Gmur J, Gratwohl A, Roosnek E, Jeannet M.(1994),Relation between the resolution of HLA-typing and the chance of finding an unrelated bone marrow donor. *Bone Marrow Transplant*. 1994 Jun;13(6):805-9.

Sonnenberg FA, Eckman MH, Pauker SG. (1989),Bone marrow donor registries: the relation between registry size and probability of finding complete and partial matches. *Blood*. 1989 Nov 15;74(7):2569-78.

Stephens, M. N. Smith, P. Donnelly (2001), A New Statistical Method for Haplotype Reconstruction from Population Data, *Am. J. Hum. Genet*. 68:978-989.

Takahashi K, Juji T, Miyazaki H. (1989), Determination of an appropriate size of unrelated donor pool to be registered for HLA-matched bone marrow transplantation. *Transfusion*. 1989 May;29(4):311-6.

Tiercy JM, Bujan-Lose M, Chapuis B, Gratwohl A, Gmur J, Seger R, Kern M, Morell A, Roosnek E. (2000), Bone marrow transplantation with unrelated donors: what is the probability of identifying an HLA-A/B/Cw/DRB1/B3/B5/DQB1-matched donor? *Bone Marrow Transplant*. 2000 Aug;26(4):437-41.

Tiercy JM, Stadelmann S, Chapuis B, Gratwohl A, Schanz U, Seger RA, Faveri GN, Kern M, Morell A, Schwabe R, Schneider P. (2003), Quality control of a national bone marrow donor registry: results of a pilot study and proposal for a standardized approach. *Bone Marrow Transplant*. 2003 Sep;32(6):623-7.

Wang, S. K.K. Kidd, H. Zhao (2003), On the use of DNA Pooling to Estimate Haplotype Frequencies, *Genetic Epidemiology*, 24: 74-82.

Résumé en anglais

AUTHOR: Frédérique Fève

TITLE: Economics and Statistics of Hematopoietic Stem Cell Gift: Contributions to the Optimization of Voluntary Donors Registries

PHD SUPERVISORS:

Anne Cambon Thomsen, DR CNRS Toulouse

Jean-Pierre Florens, Professor Mathematics Toulouse I

DATE: March, 27th, 2006, Toulouse

SUMMARY

This thesis uses economical calculation in order to model the organization of the mechanism of haematopoietic stem cells transplantation. The financing of French Public Health system tends to allocate financial duties to the health sector. This choice implies a required development of the measure of efficiency and costs of medical care. The first chapter proposes a measure of efficiency for a voluntary donors registry. We underline the value of the registry if donors arrive with the same frequency as receivers (no selection) and the value of the registry in case of optimal selection of donors. Given the high cost of welcome, the cost of typing and the cost of management registry, is it socially and economically efficient to have a registry ? In chapter 2 we show that it is possible to evaluate a certain number of key-elements of the registry system from statistical arguments (number and distribution of HLA types in France), some may change according to the registry management policy and some others deal with a more complex valuation related to a desired profit from a graft. The last part provides statistical contributions to the selection of donors : it raises the problem of the joint distribution of several discrete variables, haplotypes, knowing phenotypes of individuals and we define a simple estimation method. We study the properties of our method and we compare its efficiency by simulations to the method used by epidemiologists. Finally we propose a statistical method in order to evaluate the number of phenotypes in the population.

KEYWORDS:

Social welfare, Biostatistics, Economical Valuation, Haplotypes, Models, Optimization, Registry, Surplus, Haematopoietic Stem Cells Transplantation, HLA Typing

PhD in Economics

Université Toulouse III

Résumé en français

AUTEUR : Frédérique Fève

TITRE : Economie et Statistique du Don de Cellules Souches Hématopoïétiques : Contributions à la Gestion de Registres de Donneurs Volontaires

DIRECTEURS DE THÈSE :

Anne Cambon Thomsen, DR CNRS Toulouse

Jean-Pierre Florens, Professeur de Mathématiques Toulouse I

PRÉSENTÉE ET SOUTENUE LE : 27 mars 2006, à Toulouse

RÉSUMÉ

Cette thèse utilise le calcul économique pour modéliser l'organisation de la filière du don de Cellules Souches Hématopoïétiques en vue de greffes. Le financement du système de santé en France évolue vers l'attribution d'enveloppes affectées aux différents secteurs de la santé. Ce choix implique le nécessaire développement de la mesure et de l'efficacité et du coût des soins dispensés. Le premier chapitre propose une mesure d'efficacité d'un Registre de donneurs volontaires. On met en évidence la valeur du Registre si les donneurs arrivent à la même fréquence que les receveurs (pas de sélection) ainsi que la valeur du Registre liée à la sélection optimale des donneurs. Etant donné le coût élevé de l'accueil, du typage des donneurs et de gestion du registre, est-il socialement et économiquement efficace d'avoir un registre de donneurs ? Le chapitre 2 montre que certains éléments-clé du fonctionnement d'un Registre peuvent être évalués à partir d'arguments statistiques (nombre et distribution des types HLA en France), d'autres peuvent évoluer en fonction de la politique de gestion du fichier et d'autres enfin relèvent d'une évaluation plus complexe liée au bénéfice attendu d'une greffe. Le troisième chapitre apporte des contributions statistiques à la sélection des donneurs : il pose le problème de la distribution jointe de plusieurs variables discrètes, les haplotypes, connaissant les phénotypes des individus et propose une méthode simple d'estimation. On étudie les propriétés de cette méthode et l'on compare par simulations son efficacité à celle de la méthode utilisée par les statisticiens épidémiologistes. Enfin une procédure statistique d'évaluation du nombre de phénotypes dans la population est proposée.

MOTS CLÉS :

Bien-être-social, Biostatistique, Evaluation Economique, Haplotypes, Modélisation, Optimisation, Registre, Surplus, Transplantation de Cellules Souches Hématopoïétiques, Typage HLA

Discipline : Sciences Économiques

Doctorat de l'Université Toulouse III