

⑫

DEMANDE DE BREVET D'INVENTION

A1

②2 Date de dépôt : 01.08.08.

③0 Priorité :

④3 Date de mise à la disposition du public de la
demande : 05.02.10 Bulletin 10/05.

⑤6 Liste des documents cités dans le rapport de
recherche préliminaire : *Se reporter à la fin du
présent fascicule*

⑥0 Références à d'autres documents nationaux
apparentés :

⑦1 Demandeur(s) : COMMISSARIAT A L'ENERGIE ATO-
MIQUE Etablissement public à caractère industriel et
commercial — FR et CENTRE DE RECHERCHE
POUR LES PATHOLOGIES PROSTATIQUES -
CEREPP Association loi de 1901 — FR.

⑦2 Inventeur(s) : AURIBAUT KARINE, MULLER JEAN
DENIS, CANCEL TASSIN GERALDINE, CUSSENOT
OLIVIER, GAZUT STEPHANE, GILARDI NICOLAS,
MERCIER DAVID, POLI JEAN PHILIPPE, RAMASSO
EMMANUEL et SUARD FREDERIC.

⑦3 Titulaire(s) : COMMISSARIAT A L'ENERGIE ATOMI-
QUE Etablissement public à caractère industriel et com-
mercial, CENTRE DE RECHERCHE POUR LES
PATHOLOGIES PROSTATIQUES - CEREPP Associa-
tion loi de 1901.

⑦4 Mandataire(s) : MARKS & CLERK FRANCE.

⑤4 PROCEDE DE PREDICTION POUR LE PRONOSTIC OU LE DIAGNOSTIC OU LA REPONSE THERAPEUTIQUE
D'UNE MALADIE ET NOTAMMENT DU CANCER DE LA PROSTATE ET DISPOSITIF PERMETTANT LA MISE
EN OEUVRE DU PROCEDE.

⑤7 L'invention a pour objet un procédé de prédiction indi-
viduelle pour le dépistage ou le diagnostic ou la prise en
charge thérapeutique ou le pronostic d'une maladie com-
portant le recueil de données individuelles (x_i) d'entrée et la
fourniture d'une information prédictive de risque (y) liée à un
type de maladie, caractérisé en ce que les données d'entrée
comportent au moins une variable ou une combinaison de
variables type génétique telles que l'identification de mar-
queurs de polymorphismes génétiques considérés comme
étant liés au développement de la maladie.

L'invention a aussi pour objet un dispositif de prédiction
individuelle pour le dépistage ou le diagnostic ou la prise en
charge thérapeutique ou le pronostic d'une maladie com-
portant des premiers moyens de saisie de données d'infor-
mation individuelles par un utilisateur, au moins une
première interface logicielle sur laquelle opère les dits pre-
miers moyens caractérisé en ce qu'il comporte en outre un
logiciel mettant en oeuvre le procédé de l'invention et four-
nissant une information prédictive de risque lié à une mala-
die.



Procédé de prédiction pour le pronostic ou le diagnostic ou la réponse thérapeutique d'une maladie et notamment du cancer de la prostate et dispositif permettant la mise en œuvre du procédé.

Le domaine de l'invention est celui des méthodes de prédiction individuelle pour le dépistage, le diagnostic, le pronostic, la réponse thérapeutique de maladies et les effets secondaires des médicaments dans
5 le cas de maladies complexes et multifactorielles telles que des cancers et notamment le cancer de la prostate.

La présente invention propose un procédé et un outil conduisant à une décision fiable pour instaurer un dépistage ou un diagnostic précoce et ce, en associant de nombreuses données d'entrée cliniques et/ou génétiques
10 associées de manière complexe.

Aujourd'hui il existe des formes de cancers et notamment le cancer de la prostate fortement répandues chez l'homme dans les pays industrialisés et dont l'incidence a augmenté de manière importante ces dernières années.

15 L'étiologie relative aux causes et à l'évolution des cancers de la prostate est complexe et résulte d'interactions stochastiques multiples entre les facteurs génétiques constitutionnels, des facteurs tissulaires acquis et des facteurs environnementaux.

Le diagnostic et les traitements proposés nécessitent la mise en
20 œuvre de procédures invasives et coûteuses. Les méthodes actuelles développées pour déterminer les populations à risque ou les stratégies de prise en charge proposent des valeurs prédictives positives ou négatives (cancer/non cancer) en fonction de tests (marqueurs tumoraux, signatures moléculaires ...) ou de résultats d'algorithmes linéaires encore appelés
25 nomogrammes, mais leur fiabilité est inférieure à 80% et les résultats sont rarement reproductibles à l'échelle individuelle.

Actuellement, il est proposé d'évaluer un risque de cancer de la prostate par dosage sanguin de l'antigène prostatique spécifique (PSA) qui est le marqueur de référence pour décider d'une procédure invasive de type
30 biopsie pour la confirmation histologique d'un cancer de la prostate, typiquement en cas de détection d'un taux mesuré supérieur à 4 ng/ml voire 2,5 ng/ml dans certains protocoles.

Au-delà de 4 ng/ml de taux sanguin de PSA la sensibilité est de 30%, ce qui signifie que parmi les personnes qui ont un taux de PSA total supérieur à 4 ng/ml, de l'ordre de seulement 3 sur 10 ont un cancer de la prostate.

- 5 Au seuil de 4 ng/ml la spécificité du test PSA est de l'ordre de 80%, ce qui signifie que lorsque le seuil de PSA est inférieur à 4 ng/ml, l'absence de cancer de la prostate est réelle dans 8 cas sur 10.

Des outils d'évaluation du risque de type nomogrammes intégrant plusieurs paramètres ont été développés pour répondre aux questions
10 individuelles. Les nomogrammes sont des outils statistiques destinés à la prise de décision, qui contiennent de l'information provenant de centaines d'observations concrètes sur des cas prouvés de cancer de la prostate. Ces outils aident les patients et les médecins à la prise de décision. Ils fournissent des prévisions calculées à partir de données cliniques variées
15 issues de cancers de la prostate précédemment traités. Il s'agit de règles de calcul ou d'abaques construits sur la base de régressions logistiques multivariées. Ces nomogrammes ont un taux moyen d'exactitude de 80 % ce qui demeure insuffisant. Les patients en tirent cependant d'indéniables avantages parce qu'ils sont dépourvus de la partialité et de la subjectivité
20 que l'on retrouve chez les divers cliniciens et soignants. A titre d'exemple 12 questions et les outils associés prédictifs sont proposés par la Fondation de Recherche Canadienne sur le Cancer de la Prostate.

Les solutions existantes utilisées dans ce type d'outils prédictifs reposent la plupart du temps sur le recueil de données cliniques et
25 d'évaluation utilisant des méthodes de modélisation linéaires par rapport aux paramètres. Les méthodes développées sont insuffisantes en terme de fiabilité et ne permettent pas de réaliser des prédictions hiérarchisées : risque de maladie, risque de maladie rapidement évolutive, risque de maladie répondant mieux ou mal à tel traitement.

30 Dans ce contexte, la présente invention propose un procédé de prédiction individuelle pour le pronostic ou le diagnostic ou la réponse thérapeutique d'une maladie et notamment de cancer et plus particulièrement bien adapté au cancer de la prostate, basé sur le recueil de très nombreuses données cliniques et/ou génétiques et comportant
35 l'élaboration d'un modèle évolué permettant de délivrer une valeur de risque

pouvant avantageusement être soumise de plus à une procédure de validation.

Plus précisément la présente invention a pour objet un procédé de prédiction individuelle pour le pronostic ou le diagnostic ou la réponse thérapeutique d'une maladie, comportant le recueil de données individuelles (x_i) d'entrée et la fourniture d'une information prédictive de risque (y) liée à un type de maladie, caractérisé en ce que les données d'entrée comportent au moins une variable ou une combinaison de variables de type SNP (Single Nucleotide Polymorphism ou Polymorphisme Nucléotidique Simple) ou Polymorphisme de longueur ou Polymorphisme de répétition parmi les suivantes :

- variable définissant le génotype lié au SNP **rs2174183** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673 à 128447913 du chromosome 4 ;
- variable définissant le génotype lié au SNP **rs7576160** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 37855761-38126567 du chromosome 2 ;
- variable définissant le génotype lié au SNP **rs2012385** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 241767109-242119399 du chromosome 2 ;
- variable définissant le génotype lié au SNP **rs888298** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 63815611 à 64165896 du chromosome 17 ;
- variable définissant le génotype lié au SNP **rs8110935** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 62026584-62294837 du chromosome 19.
- variable définissant le génotype lié au SNP **rs2190453** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 17464539 à 17757162 du chromosome 11 ;
- variable définissant le génotype lié au SNP **rs2788140** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 ;
- variable définissant le génotype lié au SNP **rs3828054** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 ;

- variable définissant le génotype lié au SNP **rs1499955** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 116302446-117011700 du chromosome 3.

5 - variable définissant le génotype lié au SNP **rs4855539** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 69049525 - 69153397 du chromosome 3 ;

- variable définissant le génotype lié au SNP **rs4242382** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8 ;

10 - variable définissant le génotype lié au SNP **rs11526176** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7

- variable définissant le génotype lié au SNP **rs7934514** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 99092040-99333419 du chromosome 11

- variable définissant le génotype lié au SNP **rs6681102** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1

20 - variable définissant le génotype lié au SNP **rs6492998** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 38991207-39584443 du chromosome 15.

Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP **rs2174183** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle
25 127602673-128447913 du chromosome 4.

Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP **rs2174183** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le
30 génotype lié au SNP **rs7576160** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 37855761-38126567 du chromosome 2 et/ou d'une variable définissant le génotype lié au SNP **rs2012385** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 241767109-242119399 du chromosome 2.

Selon une variante de l'invention, les données d'entrée
35 correspondent à la combinaison d'une variable définissant le génotype lié au

SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et/ou d'une variable définissant le génotype lié au SNP rs2190453 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 17464539-17757162 du chromosome 11 et/ou d'une variable
5 définissant le génotype lié au SNP rs888298 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 63815611-64165896 du chromosome 17.

Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle
10 127602673-128447913 du chromosome 4 et/ou d'une variable définissant le génotype lié au SNP rs2788140 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 et/ou d'une variable définissant le génotype lié au SNP rs7934514 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 99092040-99333419 du
15 chromosome 11.

Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et/ou d'une variable définissant le
20 génotype lié au SNP rs3828054 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 et/ou d'une variable définissant le génotype lié au SNP rs1499955 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 116302446-117011700 du chromosome 3.

25 Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le génotype lié au SNP rs8110935 et/ou à l'un ou à plusieurs de ses voisins
30 dans l'intervalle 62026584-62294837 du chromosome 19.

Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le
35 génotype lié au SNP rs4855539 et/ou à l'un ou à plusieurs de ses voisins

dans l'intervalle 69049525-69153397 du chromosome 3 et/ou d'une variable définissant le génotype lié au SNP rs4242382 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8.

5 Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le génotype lié au SNP rs11526176 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7.

10 Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs6492998 ou à l'un de ses voisins dans l'intervalle 38991207-39584443 du chromosome 15 et d'une variable définissant le génotype lié au SNP rs11526176 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7 et d'une variable définissant le génotype le génotype lié au SNP rs6681102 ou à l'un de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1.

20 Selon une variante de l'invention, les données d'entrée correspondent à la combinaison d'une variable liée à l'âge et à des données cliniques et/ou à des données d'anamnèse personnelles et familiales.

Selon une variante de l'invention, la maladie étant de type cancer, les données d'anamnèse correspondent à des variables d'antécédents personnels ou familiaux de cancer, à une variable liée à l'âge ou à toute combinaison de ces variables.

25 Selon une variante de l'invention, les données d'anamnèse incluent la combinaison de quatre variables d'antécédents de cancer et d'une variable de catégorie d'âge, lesdites variables d'antécédents concernant respectivement des antécédents familiaux d'un cancer du sein, des antécédents familiaux de cancer de la prostate, des antécédents personnels de cancer, des antécédents familiaux d'autres cancers.

Selon une variante de l'invention, la maladie est un cancer pouvant notamment être de type cancer urologique ou de type hormonodépendant ou de type cancer de la prostate.

35 Selon une variante de l'invention, le procédé de prédiction individuelle concerne le dépistage, le diagnostic, le pronostic ou la réponse

thérapeutique d'un cancer de la prostate, les données étant de type clinique telles que des données individuelles concernant l'âge du patient, son poids, sa taille, les antécédents personnels et familiaux de cancer, de type biologique avec par exemple le taux de PSA, et de type génétique telles que
 5 l'identification de marqueurs de polymorphismes génétiques considérés comme étant liés au développement de la maladie.

Selon une variante de l'invention le procédé de l'invention comprend :

- la constitution d'une base de données d'exemples (Bex)
 10 constitués de données d'entrées (x_{mi}) et de résultats avérés (y_m^*) ;
- la construction d'au moins un modèle optimal par apprentissage statistique comportant les étapes suivantes :
 - le choix d'une famille (F) de fonctions multivariées ($f_1, \dots, f_i, \dots, f_N$) ;
 - 15 • pour une fonction donnée f_i , l'élaboration d'un modèle défini par ajustement de paramètres θ_j tels que l'estimation délivrée par le modèle $y_m = f_i(x_{mi}, \theta_j)$ soit la plus proche possible de celle du résultat avéré y_m^* ;
 - la comparaison des différentes estimations de manière à
 20 définir une fonction f_i optimisée f_{iop} permettant de définir un modèle optimal ;
- l'exploitation dudit modèle optimal à partir desdites données individuelles (x_i) de manière à fournir ladite information prédictive (y) de risque liée à une maladie.

25 Selon une variante de l'invention, le procédé comprend la construction en parallèle d'un ensemble de modèles optimaux, chaque modèle étant élaboré à partir d'une famille (F_k) de fonctions, l'information prédictive de risque lié à une maladie résultant de l'exploitation de l'ensemble des modèles optimaux.

30 Selon une variante de l'invention, le procédé comprend :

- la création d'une base d'apprentissage (BA) et d'une base de validation (BV) à partir de la base d'exemples ;
- un processus de validation du résultat prédictif (y^*) par
 comparaison entre ledit résultat prédictif obtenu avec un modèle construit
 35 avec l'ensemble de données d'entrée appartenant à la base d'apprentissage,

et le résultat avéré obtenu à partir d'un ensemble de données d'entrée similaires appartenant à la base de validation.

Selon une variante de l'invention, le procédé comprend pour une base de données comportant N données, la construction de la base d'apprentissage réalisée par le tirage aléatoire (sans remise) de M données appartenant à la base d'exemples, N-M données restantes constituant la base de validation.

Selon une variante de l'invention, la famille de fonctions est de type des MLP (Multi Layer Perceptron), sous-ensemble de la famille des réseaux de neurones ou de type des Support Vector Machines (SVM) ou de type des Relevant Vector Machines (RVM) ou de type des modèles fréquentistes s'apparentant à la méthode des plus proches voisins.

Selon une variante de l'invention, l'estimation délivrée par le modèle $y_m = f_i(x_{mi}, \theta_j)$ est comparée au résultat avéré y_m^* avec une fonction de coût de type score d'entropie croisée dans le cas de la discrimination : $-[y^* \log(f(x, \theta)) + (1 - y^*) \log(1 - f(x, \theta))]$ ou de type critère de log vraisemblance notée $-\log(P(y | x, \theta))$ et correspondant à la probabilité d'obtenir y avec à partir des paramètres x et θ ou de type écart quadratique dans le cas de la régression : $(f(x, \theta) - y^*)^2$.

Selon une variante de l'invention, la comparaison entre ledit résultat prédictif obtenu avec un modèle construit avec l'ensemble de données d'entrée appartenant à la base d'apprentissage, et le résultat avéré obtenu à partir d'un ensemble de données d'entrée appartenant à la base de validation est effectuée avec une fonction de coût similaire à celle utilisée dans la comparaison entre l'estimation délivrée par le modèle et le résultat avéré y^* .

Selon une variante de l'invention, la maladie étant de type cancer de la prostate, les données cliniques individuelles correspondent à la combinaison de quatre variables d'antécédents de cancer et d'une variable de catégorie d'âge, lesdites variables d'antécédents concernant respectivement des antécédents familiaux de cancer du sein, des antécédents de cancer de la prostate, des antécédents personnels de cancer, des antécédents familiaux d'autres cancers.

L'invention a aussi pour objet un dispositif de prédiction individuelle pour le pronostic ou le diagnostic ou la réponse thérapeutique

d'une maladie comportant des premiers moyens de saisie de données d'information individuelles par un utilisateur, au moins une première interface logicielle sur laquelle opère les dits premiers moyens caractérisé en ce qu'il comporte en outre un logiciel mettant en œuvre le procédé selon l'invention
5 et fournissant une information prédictive de risque lié à une maladie.

Selon une variante de l'invention, ladite information prédictive de risque est restituée à l'utilisateur via ladite interface logicielle.

Selon une variante de l'invention, le dispositif comporte en outre des moyens de communication entre les premiers moyens de saisie et le
10 logiciel, permettant la transmission des données d'information et celle de l'information prédictive.

Selon une variante de l'invention, le dispositif comporte en outre des seconds moyens de saisie de données d'information individuelles et une seconde interface logicielle, les premiers moyens de saisie concernant la
15 saisie d'informations de type clinique, les seconds moyens concernant la saisie d'informations issues de prélèvement sur l'individu.

L'invention sera mieux comprise et d'autres avantages apparaîtront à la lecture de la description qui va suivre donnée à titre non limitatif et grâce
20 aux figures annexées parmi lesquelles :

- la figure 1 illustre un schéma qui résume les interactions entre la base d'exemples, les résultats réels et les résultats prédictifs ;

- la figure 2 illustre une représentation d'un type de réseau de neurones ;

- 25 - la figure 3 illustre un premier exemple d'utilisation dans laquelle l'outil logiciel est implanté chez le praticien ;

- la figure 4 illustre un second exemple d'utilisation dans laquelle l'outil logiciel est centralisé chez un professionnel fournisseur de résultats prédictifs ;

- 30 - les figures 5a à 5e illustrent respectivement les performances d'algorithmes de type *Multi-Layer Perceptron* vis à vis de la discrimination malades atteints de cancers de la prostate des témoins avec en variables d'entrée la catégorie d'âge et respectivement le génotype associé au SNP rs2969612 ou à ses voisins, le génotype associé au SNP rs1167190 ou à
35 ses voisins, le génotype associé au SNP rs1314813 ou à ses voisins, le

génotype associé au SNP rs2174183 ou à ses voisins, le génotype associé au SNP rs1604724 ou à ses voisins.

L'intérêt de la présente invention réside notamment dans le fait de
5 mettre à la disposition des médecins un outil d'aide à la décision pour une prise en charge personnalisée de leurs patients. Son originalité réside dans l'association d'une base de données exclusive et d'analyses statistiques multidimensionnelles. L'utilisateur peut ainsi bénéficier d'un savoir issu de recherches pluridisciplinaires en médecine, biologie, génétique,
10 mathématiques et de résultats objectifs. L'impact médical de ce système expert est également économique car il permet aux praticiens de mieux détecter les stades précoces et curables de la maladie, de réduire les coûts et les effets secondaires liés aux méthodes diagnostiques et thérapeutiques invasives. Enfin pour le patient il s'agit d'obtenir une prise en charge optimale
15 de sa pathologie, une diminution du risque de surtraitement, une augmentation de son espérance de vie et une amélioration de sa qualité de vie.

Selon l'invention, l'outil de prédiction est réalisé grâce à la construction en amont de modèles par apprentissage statistique. Nous allons
20 en décrire ci-après le principe de construction.

Un modèle, construit dans le cadre de la théorie de l'apprentissage statistique, est généralement une fonction mathématique paramétrée f qui contient des paramètres ajustables θ et appartenant à une famille de fonctions plus large F .

25 Cette fonction permet de délivrer une estimation y en fonction d'un certain nombre d'entrées x qui sont les variables d'entrées du problème.

Dans le cas de la présente invention :

- les entrées x sont les informations génétiques et/ou les résultats codés d'informations cliniques pouvant notamment
30 être issues d'un questionnaire du patient ; lorsque les entrées x sont des variables qualitatives (ou catégorielles), l'encodage de ces variables en valeurs numériques est nécessaire afin de les rendre directement utilisables par les modèles dans le cadre de leur construction et de leur utilisation en tant qu'estimateur. À
35 titre d'exemple, pour l'information sur les antécédents familiaux

de cancer de la prostate, l'encodage peut consister à coder la variable qualitative « mon grand père » avec la valeur «1» qui regroupera tous les parents du second degré. L'encodage ne doit ni masquer, ni brouiller l'information, il doit être pertinent.

5 Dans l'exemple précédent, on peut affiner le codage si l'on souhaite distinguer ou non l'atteinte du grand-père maternel de l'atteinte du grand-père paternel. L'encodage des données peut être inventif, sa qualité (exhaustivité, pertinence) détermine pour partie les possibilités de résolution du problème de

10 discrimination posé. L'encodage n'est pas obligatoirement binaire, le nombre de catégories (donc de valeurs numériques possibles) dépend du nombre d'états de la variables qualitative. Pour un SNP donné, s'il y a deux allèles A et B dans la population, un individu peut être de génotype AA BB ou AB,

15 l'encodage est ici ternaire. Rajoutons un allèle C dans la population, les combinaisons qui s'ajoutent sont CC CA CB donc un encodage à 6 catégories.

- l'estimation y , délivrée par le modèle, est la classe du patient (cancer / pas cancer) ou le risque d'avoir le cancer.
- 20 Cette estimation y peut être considérée comme étant une fonction f dépendante des entrées x et des paramètres θ .

Toute la difficulté de la création du modèle réside dans l'ajustement des paramètres θ . Ces paramètres θ sont ajustés dans une phase dite d'apprentissage qui nécessite des exemples et la mise en œuvre

25 d'algorithmes dédiés.

De manière générale, tous les modèles construits par apprentissage statistique nécessitent des exemples. En effet, en tant que système capable d'apprendre, ces modèles mettent en œuvre le principe de l'induction, c'est-à-dire l'apprentissage par l'expérience. La base d'exemples est constituée

30 d'un ensemble de N couples (x, y^*) représentatifs du processus étudié que l'on souhaite modéliser.

La variable x est, comme précédemment, une valeur parmi un ensemble de valeurs d'entrées et y^* est la sortie réelle associée à ces entrées considérées comme la vérité que l'on souhaite estimer (le diagnostic

35 cancer / pas cancer délivré par un spécialiste par exemple). On représente

cette base de données sous la forme d'un tableau de N lignes, où chaque ligne représente un exemple (les valeurs d'entrée pour un individu et sa classe associée). Le but de l'apprentissage est de construire un modèle, à partir de ces N exemples, afin d'estimer *in fine* la réponse qu'aurait donné le spécialiste sur un nouveau cas jamais rencontré. On parle dans ce cas de capacité de généralisation. Dans la procédure de création de modèles, on choisira celui qui délivrera la meilleure capacité de généralisation.

La représentativité des données est une notion très importante puisqu'elle conditionne la qualité du modèle construit et que l'information que peut apprendre le modèle est contenue dans la base au travers des N exemples. On entend par représentativité le caractère exhaustif des cas contenus dans la base. C'est-à-dire que l'on doit s'assurer que le modèle a rencontré un ensemble de cas similaires à ceux qu'il rencontrera dans son utilisation future en tant qu'estimateur. La phase de constitution de la base d'apprentissage est donc une étape clé et doit être conduite avec beaucoup de rigueur.

Le paragraphe suivant décrit comment l'algorithme d'apprentissage ajuste les paramètres du modèle en fonction des éléments constituant la base d'apprentissage.

La figure 1 illustre un schéma qui résume les interactions entre la base d'exemples Bex, les résultats réels et les résultats prédictifs.

Durant la phase d'apprentissage, l'algorithme modifie les paramètres ajustables θ du modèle pour que l'estimation y soit la plus proche possible de celle du résultat avéré encore dénommé « superviseur » y^* . Le critère que l'on souhaite donc minimiser en agissant sur les paramètres θ est l'écart entre la réponse du modèle et la réponse du superviseur sur les cas dont on dispose. Cet écart peut s'obtenir de différentes façons suivant le problème traité et est appelé « fonction de coût » :

Typiquement la « fonction de coût » que l'on cherche à minimiser peut être par exemple une des fonctions suivantes :

- le score d'entropie croisée dans le cas de la discrimination (cela revient à estimer l'appartenance à une classe donnée) : $-[y^* \log(f(x, \theta)) + (1 - y^*) \log(1 - f(x, \theta))]$;

- le critère de log vraisemblance notée $-\log(P(y|x, \theta))$ et correspondant à la probabilité d'obtenir y à partir des entrées x et des paramètres θ ;
- l'écart quadratique dans le cas de la régression : $(f(x, \theta) - y^*)^2$.

La phase d'apprentissage consiste donc à trouver un jeu de
 5 paramètres θ , pour une fonction f_i de la famille F de fonctions qui minimise la fonction de coût sur l'ensemble des exemples, à l'aide d'algorithmes d'optimisation.

Cependant, un modèle capable de prédire des informations déjà
 10 connues n'a que peu d'intérêt. Il faut s'assurer qu'il est capable de prédire correctement des cas non présents mais représentés dans la base d'apprentissage, et qui suivent les mêmes lois que ceux ayant servi à l'apprentissage. C'est pourquoi la base d'exemple est généralement scindée en une base d'apprentissage BA, pour ajuster les paramètres du modèle, et
 15 une base de validation BV encore appelée base de test, pour tester le modèle choisi et vérifier sa robustesse.

L'important pour les deux ensembles est d'être le plus représentatif possible d'une part de la base d'exemples totale et d'autre part du problème traité. Si la base d'apprentissage ne l'est pas, on risque de ne pas modéliser
 20 le phénomène que l'on cherche. Si la base de validation ne l'est pas, les scores de validation risquent de donner une fausse idée des performances des modèles, si la base d'exemple n'est pas représentative des cas réels aucune application pratique ne pourra en découler.

Lorsque l'on dispose de suffisamment de données, les deux
 25 ensembles (base d'apprentissage et base de test) sont construits en tirant aléatoirement les éléments de la base d'exemples. Ainsi, sur une base de N éléments, on en sélectionne au hasard M qu'on utilisera pour l'entraînement, et les $(N - M)$ qui restent serviront pour la validation.

Pour que le score de validation ne soit pas dépendant du tirage
 30 particulier d'une seule partition de la base totale en base d'apprentissage et base de validation, la procédure est répétée un certain nombre de fois.

Ainsi nous allons décrire plus en détails le processus proposé dans la présente invention.

Dans une première étape on choisit une famille F de fonctions, le choix dépendant du problème posé et de la connaissance *a priori* que l'on en a. Typiquement dans le cadre de l'invention, le problème rencontré entre dans la catégorie des problèmes de discrimination, c'est-à-dire que l'on
 5 cherche à classer de nouveaux individus en deux groupes : malades ou témoins.

Dans une seconde étape on choisit un type de fonction f_i appartenant à la famille F .

Dans une troisième étape, on construit un modèle optimal $f_i(\mathbf{x}, \theta)$
 10 par la procédure d'apprentissage en ajustant les paramètres θ .

On réitère, cette construction de modèle avec $n-1$ fonctions de manière à tester suffisamment de type de fonctions $f_1, f_2, \dots, f_n$, on compare les qualités respectives de leurs modèles optimaux.

Dans une quatrième étape, on sélectionne la fonction f_i conduisant
 15 au modèle optimal ayant le meilleur score de validation, déterminant ainsi la fonction f_i dite qui « généralise le mieux ».

Dans une cinquième étape, on évalue, avec l'ensemble des exemples de la base d'apprentissage, les paramètres θ de la fonction retenue à l'étape précédente. On obtient ainsi le modèle optimal $f_{iop}(\mathbf{x}, \theta)$ qui
 20 à partir de données individuelles d'entrée x_i sera en mesure de fournir le résultat prédictif y .

Parmi les nombreuses familles de fonctions disponibles, on peut notamment citer les familles :

- 25 • des MLP (*Multi Layer Perceptron*), sous-ensemble de la famille des réseaux de neurones,
- la régression logistique (sous-ensemble de la famille des MLP) ;
- des *Support Vector Machines* (SVM) ;
- des *Relevance Vector Machines* (RVM) ;
- 30 • des modèles fréquentistes s'apparentant à la méthode des plus proches voisins.

La plupart de ces types de fonctions sont notamment décrit dans l'ouvrage de référence "Réseaux de Neurones, Méthodologie et Applications" de G. Dreyfus et al., Editions Eyrolles ou dans "Pattern Recognition and
 35 Machine Learning" de C.M. Bishop, Springer 2006. Les *Relevance Vector*

Machines sont décrits dans "Sparse Bayesian learning and the relevance vector machine", Tipping, M. E. (2001), Journal of Machine Learning Research 1, 211-244.

5 Le principal apport des modèles décrits précédemment, par rapport aux modèles déjà utilisés pour évaluer les risques, réside dans la non-linéarité des modèles d'apprentissage statistique. En effet, les modèles généralement employés sont dits linéaires par rapport aux paramètres ce qui induit une plus grande facilité de mise en œuvre au prix généralement d'un
10 plus faible pouvoir prédictif. Dans le cas de modèles précédemment décrits, non linéaires par rapport aux paramètres, la mise en œuvre est plus délicate mais permet :

- d'obtenir en général, de meilleures performances du modèle ;
- de détecter les synergies entre variables d'entrées.

15 La possibilité d'exploiter les synergies entre les variables d'entrées est un aspect essentiel du caractère inventif de l'objet de la présente invention. Il constitue l'apport principal de la collaboration des mathématiciens aux découvertes biologiques et médicales dans ces travaux. En effet les outils mathématiques et statistiques à la disposition des médecins et des
20 biologistes ne permettent pas en général de détecter ces synergies.

 De plus, ces algorithmes ayant des capacités d'apprentissage élevées, il est très important de pouvoir mesurer leurs performances afin de vérifier qu'ils ne se sur-ajustent pas aux exemples d'entraînement (on parle alors d'apprentissage « par cœur » ou de « sur-apprentissage »). Les
25 méthodologies de l'apprentissage statistique permettent, notamment grâce à l'usage des exemples de validation, de résoudre ce problème et de s'assurer que le modèle obtenu représente un phénomène général et non pas un cas particulier des exemples d'entraînement. Ceci permet de modéliser des phénomènes pour lesquels on a peu ou pas de connaissance *a priori*.

30

 L'invention va être décrite ci-après dans le cadre du diagnostic du cancer de la prostate, mais pourrait tout à fait aussi bien être décrite dans le cadre d'autres pathologies pour lesquelles on a identifié un certain nombre de facteurs génétiques et biologiques (par exemple cancer de la vessie,
35 cancer du rein, ...).

Exemple d'application pour le diagnostic du cancer de la prostate

5 Dans cet exemple d'application, la problématique est de déterminer les personnes dont le risque de développer un cancer de la prostate est plus élevé que la normale. Pour cela, une base de données de 1315 individus dont la moitié a été diagnostiquée comme étant atteinte du cancer en question (catégorie des Malades) est réalisée, et l'autre moitié est
10 considérée comme saine (catégorie des Témoins). On dispose d'un certain nombre de variables explicatives et d'un diagnostic : positif (malades) ou négatif (témoins).

Selon la présente invention, un modèle est constitué, capable à partir des variables explicatives, de prédire une réponse interprétée comme une
15 probabilité d'être malade ou témoin.

Il convient dans un premier temps de choisir une famille F de fonctions de modèle :

Le présent problème entre dans la catégorie des problèmes de
20 discrimination, c'est-à-dire que l'on cherche à classer de nouveaux individus en deux groupes : malades ou témoins.

De nombreuses familles de fonctions sont adaptées à la résolution de ces problèmes. Certaines sont très simples à mettre en œuvre mais ne permettent pas de tenir compte des synergies entre les variables. Or, on ne
25 sait pas *a priori* si de telles relations existent ou non. Il convient donc de choisir une famille de fonctions capable d'en tenir compte si elles existent.

Une famille simple à décrire et généralement efficace est celle des Perceptrons Multi-Couches ou MLP (pour *Multi Layer Perceptron*). Il s'agit d'un type de réseau de neurones que l'on représente généralement selon le
30 schéma illustré en figure 2.

La formule mathématique est de la forme suivante :

$$f(\mathbf{x}, \boldsymbol{\theta}) = L \left(\theta_0 + \sum_{i=1}^n \theta_i S_i \left(\theta_{i0} + \sum_{j=1}^p \theta_{ij} x_j \right) \right)$$

Où L est la fonction « Logistique », les S_i sont des fonctions de type « Sigmoides » (comme par exemple la fonction « tangente hyperbolique »), n est le nombre de neurones cachés, p le nombre de variables d'entrée et θ désigne le vecteur de paramètres constitué des composantes θ_i et θ_{ij} pour $1 \leq i \leq n$ et $1 \leq j \leq p$.

Etant donné que le nombre m de variables est dicté par le problème traité, seul le nombre n de neurones cachés peut être choisi dans la phase de modélisation. C'est pourquoi les fonctions composant la famille des MLP pour le problème traité se différencient uniquement par leur nombre de « neurones cachés », chacun d'eux représentant en réalité une fonction Sigmoidale. Par exemple, la fonction « régression logistique », bien connue dans le domaine médical, appartient à cette famille. C'est en effet un cas particulier de MLP n'ayant aucun neurone caché. Dans ce cas, le modèle est linéaire par rapport aux paramètres et la construction du modèle met alors en œuvre des techniques d'apprentissage différentes de celles utilisées dans le cadre des MLP.

Dans une seconde étape il convient de valider des fonctions :

Plus un MLP a de neurones cachés, plus il est capable de modéliser des phénomènes complexes. Il a en effet été démontré que toute fonction continue pouvait être approximée par un MLP ayant suffisamment de neurones cachés.

Cependant, dans le cas présent, seule est prise en compte la modélisation de comportements « généraux », et pas la particularité des individus présents dans la base de données. Il convient donc de trouver un MLP avec un nombre optimal de neurones cachés afin de construire le modèle le plus général possible. Pour cela, on peut décider *a priori* de tester 5 MLP, chacun ayant de 1 à 5 neurones cachés, construire pour chacun un modèle optimal qui sera évalué sur des données de validation. Le MLP ayant le meilleur pouvoir de généralisation est alors retenu.

Dans une troisième étape, on détermine une méthode de validation :

Compte tenu du nombre d'exemples à disposition, il est possible de procéder à une simple construction aléatoire des ensembles de validation et

d'entraînement. Cependant, les données contenant beaucoup d'informations inutiles, on ne peut se contenter d'un unique couple entraînement/validation car il y a risque de construire un modèle adapté à un sous-problème, et de le valider sur autre chose. Pour cela, on évalue les modèles par une procédure de validation croisée. Le principe est le suivant :

- 1) On sépare aléatoirement la base d'exemples en cinq sous-ensembles numérotés de 1 à 5.
- 2) On prend le sous-ensemble 1 comme ensemble de validation, et on construit l'ensemble d'entraînement avec le sous-ensemble composé de l'union des sous-ensembles 2 à 5.
- 3) On entraîne le modèle numéro 1 et on calcul son score de validation numéro 1.
- 4) On prend le sous-ensemble 2 comme ensemble de validation, et on construit l'ensemble d'entraînement avec les sous-ensembles 1, 3, 4 et 5.
- 5) On entraîne le modèle numéro 2 et on calcul son score de validation numéro 2.
- 6) On poursuit la procédure jusqu'à ce que chaque sous-ensemble ait été utilisé en validation. On a donc cinq scores de validation. Le score de validation final est la moyenne de ces cinq scores.

Grâce à cette procédure, on utilise toutes les données pour calculer le score de validation, ce qui permet d'éviter de se focaliser sur des cas particuliers.

Dans une quatrième étape on procède au choix d'une fonction de coût d'entraînement :

La fonction de coût utilisée pour l'entraînement est en partie dictée par le problème posé (discrimination) et la famille de fonction (MLP). Dans le cas présent on peut avantageusement utiliser l'entropie croisée.

Dans une cinquième étape on procède au choix de la fonction de calcul du score de validation :

Le score de validation correspond à une mesure d'évaluation de la qualité du modèle. Ce score peut correspondre à son taux de bonne

classification, soit la somme du nombre de malades et de témoins correctement identifiés, divisé par le nombre total d'individus dans la base de validation. Ce score est simple à calculer et facilement interprétable et utilisable, bien qu'il occulte les performances classe par classe (il peut en effet arriver que l'une des classes soit mieux identifiée que l'autre). Ce score
5 peut également être l'AUC (Area Under Curve), c'est-à-dire l'aire sous la courbe ROC (*Receiver Operating Characteristic*) comme illustré en figures 5a, 5b, 5c, 5d et 5e.

L'ensemble des choix précédents étant effectué, la procédure de
10 sélection de la fonction MLP « idéale » peut être lancée. Celle qui permet d'obtenir le meilleur score de validation est sélectionnée pour construire le modèle final.

Dans une sixième étape on procède à la construction du modèle final
15 dit optimal.

Pour le modèle final dit optimal, c'est-à-dire celui qui est effectivement utilisé pour le calcul de risque, on lance une procédure d'entraînement sur la fonction « idéale » identifiée. L'ensemble d'entraînement utilisé est cette fois la totalité de la base d'exemple car
20 aucune validation n'est plus nécessaire.

Selon une variante plus élaborée de l'invention, il est également possible pour différentes familles de fonctions F, d'élaborer un modèle optimal conduisant ainsi à la détermination d'un ensemble de modèles optimaux, destinés à gérer en cours d'utilisation, des données individuelles
25 d'entrée pour fournir un résultat prédictif.

Selon une variante plus élaborée de l'invention, il est également possible pour différentes familles de fonctions F, d'élaborer un modèle optimal résultant d'une fusion de décision d'autres modèles optimaux construits à partir de tout ou partie des variables d'entrées. Cette étape
30 conduisant à une variante plus élaborée de l'invention entre dans le cadre de la septième étape décrite ci-dessous.

Dans une septième étape on procède à la fusion d'informations de
35 modèles optimaux.

La fusion d'informations a pour objectif d'améliorer la prise de décision en terme de robustesse et de fiabilité à partir de la combinaison, via un opérateur mathématique, des décisions ou des scores fournis par la famille de fonctions [I. Bloch. *Fusion d'informations numériques : panorama méthodologique*. In *Journées Nationales de la Recherche en Robotique*, Guidel, Morbihan, Octobre 2005.. Ces opérateurs doivent à la fois tirer parti des complémentarités entre les différentes fonctions en entrée de la fusion mais aussi prendre en compte leurs redondances. Les opérateurs de fusion sont nombreux [Ludmila I. Kuncheva, James C. Bezdek, and Robert P.W. Duin. *Decision templates for multiple classifier fusion : an experimental comparison*. *Pattern Recognition*, 34 :299–314, 2001] et peuvent être basés sur différents formalismes mathématiques tels que la théorie des probabilités, la théorie des fonctions de croyance ou les mesures floues [G.J. Klir and M.J. Wierman. *Uncertainty-based information. Elements of generalized information theory*, 2nd edition. *Studies in fuzzyness and soft computing*. Physica-Verlag, 1999].

Des algorithmes d'apprentissage statistiques ou automatiques peuvent par ailleurs être utilisés pour une fusion paramétrique mais ils requièrent généralement plus d'informations *a priori* pour l'estimation de l'opérateur de fusion.

Quel que soit le formalisme utilisé, les opérateurs de fusion peuvent prendre la forme de table de règles, de règles de combinaison de type "ET/OU logiques", de produit de scores avec ou sans *a priori* pouvant être conditionnel ou non comme dans le cas de la fusion basée sur le théorème de Bayes généralisé ou non [Ph. Smets. *Beliefs functions: The Disjunctive Rule of Combination and the Generalized Bayesian Theorem*. *Int. Jour. of Approximate Reasoning*, 9 :1–35, 1993], de distances à des modèles prédéfinis par apprentissage ou expertise, de somme pondérées avec ou sans prise en compte des interactions entre les entrées de la fusion...

Le pouvoir explicatif et l'interprétation des résultats, qui sont des critères importants pour les applications médicales et industrielles, sont généralement beaucoup plus aisés via l'utilisation d'opérateurs spécifiques de fusion au lieu des algorithmes d'apprentissage statistiques ou automatiques.

Ainsi et selon l'invention, lorsque le procédé de prédiction est construit, il est possible de proposer à l'utilisateur typiquement le médecin ou tout autre entité type laboratoire la mise à disposition d'un outil d'aide à la décision à la fois impartial, fiable et permettant une utilisation personnalisée à différents stades du parcours du patient, permettant par la même avec un outil unique de réaliser des prédictions hiérarchisées, comportant des entrées de type données cliniques et ou données génétiques ledit outil procurant en sortie une information de type évaluation d'un risque ou degré d'avancée de maladie détectée.

10 Avec un tel outil, il devient possible de réaliser une identification précoce et non invasive de risque de développer un cancer de la prostate avec évaluation de la gravité (y compris de cancer en fonction d'exposition professionnelles à des carcinogènes, les variantes génétiques déterminant une plus ou moins grande sensibilité à ces agents).

15 Il est également possible d'évaluer le risque de récurrence des cancers selon le traitement y compris la validation d'essais cliniques pour l'industrie pharmaceutique, sous la forme d'une activité de service de « fouille de données » ou de biostatistique.

20 Il est également possible d'évaluer les risques de complication de la radiothérapie ou curiethérapie (ou d'exposition aux radiations ionisantes en général), les risques pour d'autres pathologies urologiques (hypertrophie bénigne prostatique, incontinence urinaire).

Travailler sur le génotype de patients permet d'avoir accès à des éléments pouvant être très déterminants dans l'apparition d'une pathologie et facile à collecter. Un simple prélèvement de salive permet en effet de travailler facilement sur l'ADN constitutionnel invariant. Le matériel génétique est informatif car il est susceptible par l'identification du profil génétique de déterminer le risque de développer la maladie mais aussi le risque qu'elle soit agressive.

30

Exemple d'application installée chez le praticien :

Selon un exemple d'utilisation, l'application est installée chez le praticien qui saisit les informations dont il dispose pour son patient, comme par exemple le taux sanguin de PSA total ou de PSA libre, l'âge, le poids, la taille, les antécédents familiaux et personnels, le résultat d'examens de type

35

toucher rectal et les génotypes d'intérêt. Il sélectionne les questions pertinentes et l'application interroge le ou les différents modèles statistiques à sa disposition. L'outil donne des réponses personnalisées et hiérarchisées avec par exemple pour le cancer de la prostate, le risque de développer un

5 cancer agressif à un âge donné, le risque de développer des métastases ou une récurrence de la tumeur après traitement initial (à un âge donné). La figure 3 illustre une telle configuration dans laquelle des données individuelles x_i sont saisies par un utilisateur U_0 grâce à des premiers moyens au niveau d'une interface 1, ladite interface assurant la liaison avec le logiciel 2 mettant

10 en œuvre le procédé de l'invention. L'information prédictive y est restituée au niveau de l'interface à l'utilisateur U_0 , en l'occurrence le praticien.

Exemple d'application installée chez un professionnel fournisseur de résultats.

15 Dans ce cas, les informations de type cliniques sont expédiées par un patient ou par un praticien chez le professionnel fournisseur de résultats, via des réseaux de communication pouvant être de type internet.

En parallèle des informations issues de prélèvements de type sanguin et/ou salivaires analysés en laboratoire sont également expédiées

20 chez le professionnel de résultat prédictif, l'ensemble des informations est traité par le ou les modèles préalablement élaborés de manière à fournir un résultat prédictif, ledit résultat étant réexpédié chez un professionnel de santé qui est ainsi en mesure d'en informer le patient.

La figure 4 schématise ce type de configuration. Un premier

25 utilisateur U_1 saisit un certain nombre de données individuelles x_{1i} pouvant être de type données cliniques au niveau d'une première interface 10 et les envoie via une liaison à distance de type internet par exemple à un professionnel fournisseur de résultats FRP, chez qui est installé le logiciel de prédiction 2.

30 En parallèle, un second utilisateur pouvant être un laboratoire d'analyses envoie un autre flux d'informations issues de prélèvements sanguins ou salivaires x_{2i} et saisies au niveau d'une seconde interface 11 et également expédié au fournisseur FRP via une liaison à distance. Après traitement de l'ensemble des données reçues via une interface 12 installée

35 chez le fournisseur FRP, ce dernier expédie le résultat y , à un troisième

utilisateur U_3 habilité à informer le patient concerné. Typiquement lorsque l'utilisateur U_1 est le praticien, il peut n'y avoir que deux utilisateurs U_1 et U_2 . Par contre si un patient a la possibilité d'adresser directement des informations au professionnel FRP, le résultat y ne pourra pas lui être
5 communiqué directement par FRP.

Le professionnel fournisseur de résultat peut à tout moment venir enrichir ses bases de données d'exemples par les nouveaux cas traités de manière à fournir des résultats prédictifs plus performants.

Pour les soumissions de cas à distance il est prévu une protection
10 des données personnelles de chaque patient, compatible avec les règles de sécurité et d'éthique en usage.

Nous allons décrire ci-après des exemples de combinaisons de données d'entrée ou variables, particulièrement adaptées au calcul de risque
15 de survenue de cancer de la prostate.

Une première variable est dénommée «antécédents familiaux de cancer de la prostate» les valeurs de cette variable permettent de définir le contexte familial de survenue du cancer de la prostate d'un patient. Les
20 valeurs affectées à chaque individu dépendent de l'âge et/ou du degré de parenté et/ou du nombre de cas de survenue de cancer de la prostate dans sa famille.

Une deuxième variable est dénommée «antécédents familiaux de cancer du sein» les valeurs de cette variable permettent de définir le
25 contexte familial de survenue de cancer sein d'un patient. Les valeurs affectées à chaque individu dépendent de l'âge et /ou du degré de parenté et/ou du nombre de cas de survenue de cancers du sein dans sa famille.

Une troisième variable est dénommée «antécédents personnels de cancer» elle permet de distinguer les patients ayant déjà eu un cancer, quel
30 qu'il soit.

Une quatrième variable est dénommée «antécédents familiaux d'autres cancers» les valeurs de cette variable définissent le contexte familial de survenue de cancer (autres que le cancer du sein ou de la prostate) et dépendent de l'âge et/ou du degré de parenté et/ou du nombre de cas de
35 survenue d'autres formes de cancer pour patient donné.

Une cinquième variable est l'âge encodé sous la forme de catégories d'âges.

Ces variables sont utilisables en combinaison ou seules comme variables d'entrées d'algorithmes pertinents pour obtenir un calcul de risque de survenue de cancer de la prostate ou déterminer la susceptibilité au cancer de la prostate.

La valeur prédictive de ces variables est renforcée par leur utilisation en combinaison avec des marqueurs de la variabilité biologique individuelle comme par exemple des polymorphismes génétiques ponctuels aussi appelés SNP (Single Nucleotide Polymorphisms). Une propriété essentielle des marqueurs génétiques dont les SNP font partie est leur capacité à être transmis en déséquilibre de liaison avec des marqueurs de leur voisinage défini au sens de la localisation chromosomique. On parle de distance génétique entre deux marqueurs ou SNP. On considère que deux marqueurs sont ainsi génétiquement liés lorsque la fréquence des recombinaisons entre eux est rare. L'existence de ces liaisons génétiques fait que les SNP au voisinage d'un SNP d'intérêt sont susceptibles d'apporter la même information ou une part d'information sur un caractère de susceptibilité. Comme pour chaque SNP, on dispose de la pertinence de différents SNP présents dans son voisinage, on peut obtenir pour chaque SNP d'intérêt majeur, la liste des SNP voisins pouvant apporter de l'information sur la susceptibilité au cancer de la prostate. La définition d'un tel intervalle est d'un intérêt majeur d'un point de vue pratique puisqu'il permet de choisir des marqueurs apportant une information pertinente parmi une liste selon des critères pratiques de disponibilité commerciale des réactifs et expérimentaux par exemple.

Dans la présente demande, est mentionnée l'utilisation d'un SNP d'intérêt et/ou d'un ou plusieurs de ses voisins. En effet, chacun des SNP génétiquement liés au SNP d'intérêt est susceptible d'apporter tout ou partie de l'information apportée par le SNP d'intérêt. La liaison génétique dépendant de la distance physique entre deux éléments génétiques (en général exprimée en nucléotides) et de la fréquence des recombinaisons entre ces deux éléments. Le SNP d'intérêt peut être lui-même l'agent causal de la susceptibilité que l'on cherche à prédire, il peut également simplement lui être génétiquement lié. Par un effet de transitivité un SNP génétiquement

lié au SNP d'intérêt sera également génétiquement lié au facteur génétique de susceptibilité. Cette possibilité explique la nécessité d'introduire un premier « ou ». Le « et » provient lui aussi de la propriété donnée par les liaisons génétiques. Si le facteur de susceptibilité est positionné entre deux

5 SNP génétiquement liés le fait de connaître chez un individu les allèles présents pour chaque SNP permet de compléter l'information sur la probabilité de présence de l'agent causal d'une susceptibilité. L'ensemble de ces propriétés nous a semblé représenté au mieux par la formulation utilisée dans les revendications.

10

Liste des SNP liés à une susceptibilité au cancer de la prostate et intervalles chromosomiques correspondants:

SNP rs2174183 localisé en 4q28.1 sur le chromosome 4 entre les

15 positions 127907634-127908134 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs2174183 : nucléotide polymorphe en gras.

ACCAAATTGTTGCTACCAATCAGTCAATCCTAGGCACATTTACCTTCCCAGTTGAACAATCAATTATT

20 TACACTTCCTACTTCACTGTATCTTTAGATTATCAATATTTTCTTCAATCTTTTAGTTATTTAATGTC

ATATGACTACCCTCAATAATAGTATATATGAATGTTTGTGTTTGGTGATGGGAGGTCAATCAGAT (**G/T**)

GTTCCAGATAACCACTGCCTTCTTACCTTGCCTAAATAGGTATTTACATATTTCTTCCCTTAAAAA

CTGACATAggtcaggcacggtggctgacgcctgtaatcccagcactttgggaggccgaggcaggtgga

tcacttgaggtcgggagtttgagaccagcccgaccaacatggagaaaccccgctctctactaaaaatac

25 aaaattagccaggtgtggtggcacatgcctgtaatcccagctactggggaggctgagacaggagaatt

gcttgaactcaggaggcagaggttgagtgagccaagatcaagccattgcactcaagcttgggcaaca

agagcaaaactccatctcaagaaacaaaaaaaaaacaagacaaaaC AAAAGAACCTGACATAGTTGT

TTATCTGCTGAGAGTACAAGTTATTGTGATAACAAATGGCATTGCAATTGGTCATCCTTTTCTAATGG

TATATTTGCATTTTAATAACTGTATTGAAAACT

30

Les SNP voisins du SNP **rs2174183** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans une base de données selon le tableau suivant et sont positionnés dans l'intervalle 127602673-128447913 du chromosome 4 ou encore entre les SNP

35 rs12651126 et rs13122922 sur le chromosome 4:

SNP	Chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs12651126	4	-304961	chr4:127602673-127603173
rs2969612	4	-41669	chr4:127865965-127866465
rs1167190	4	-32365	chr4:127875269-127875769
rs13148138	4	-10633	chr4:127897001-127897501
rs2174183	4	0	chr4:127907634-127908134
rs1604724	4	21908	chr4:127929542-127930042
rs13122922	4	539779	chr4:128447413-128447913

La pertinence des SNP associés et du SNP d'intérêt pour la discrimination de malades atteints de cancers de la prostate et de témoins peut être démontrée par l'établissement de courbes ROC (correspondant à une variable relative à la sensibilité à un test encore dénommée « *Receiver Operating Characteristic* ») comme illustré en figure 5 qui montrent les performances d'algorithmes de type *Multi-Layer Perceptron* vis-à-vis de la discrimination des malades atteints de cancers de la prostate et des témoins en utilisant comme variables d'entrée la catégorie d'âge et le génotype associé au SNP rs2174183 ou à ses voisins. Les SNP intermédiaires non mentionnés sont donc susceptibles d'être porteurs d'information. Les AUC correspondantes (*Area Under Curve* ou aires sous la courbe ici courbe ROC) sont susceptibles d'être renforcées par l'utilisation des variables d'antécédents en entrée.

SNP rs7576160 localisé en 2p22.2 sur le chromosome 2 entre les positions 37957978-37958478 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs7576160 : nucléotide polymorphe en gras.

GTCAGATATATGTGAGTTTTTTGTCAACTAAATTCATAGTTGTCTTAATATTCATCCCTT
GCTAAAATTAAGGTGCAGAAATAAATCTGTCTAATAGAGAAATATAAATCCATCTTTTGTCTGGATA
ATCAAATTTTACTATATTTTGT'TTTAATCCTGAGAATGAAATTTTACAAATAGCTCAGGAGGTTTTTC
CTAGAGTTCCAAATAAAAGTGTGTGGATCATATACGTTCTGCTTAATCACATGACGGTTCCAAATT
TTTAATTTCAATCCTTCATTACGATGAAAAATTTTG (**C/T**) GTTTTTTTTCCACCAGCTCTTTGTTTT
GTTTTTCAATGGCTCAGGAAAGGAGAGGGGTGTGGGAGACTCTGTCTCTTTTGACAAATCACCAGCGCC

ATCTACTGTCAAGAAATAAAATCGTGACTCATTGTTAACGCGTCAATGAACATTAGGGCTTAAAGAGG
 GAAAGACAATTTTATACCCAGTACTTACTGATAAAATATAAGTTCATGTACACATATTTTATCTTAT
 ATTATTGTATTCTTAAGCAGCCTATAGGGAGAATACAATGAACCTAATATATAATCATTTATGTAATT
 C

5

Les SNP voisins du SNP **rs7576160** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 37855761-38126567 du chromosome 2 ou entre les SNP
 10 rs7562836 et rs17021897 du chromosome 2.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs7562836	2	-102217	chr2:37855761-37856261
rs4670780	2	-56053	chr2:37901925-37902425
rs4670222	2	-50101	chr2:37907877-37908377
rs10206788	2	-48321	chr2:37909657-37910157
rs7598641	2	-38008	chr2:37919970-37920470
rs9967771	2	-12100	chr2:37945878-37946378
rs879321	2	-3587	chr2:37954391-37954891
rs2565640	2	-3285	chr2:37954693-37955193
rs2278320	2	-414	chr2:37957564-37958064
rs7576160	2	0	chr2:37957978-37958478
rs2707223	2	5806	chr2:37963784-37964284
rs4670788	2	7502	chr2:37965480-37965980
rs17021897	2	168089	chr2:38126067-38126567

SNP rs2012385 localisé en 2q38.1 sur le chromosome 2 entre les
 15 positions 242070828 et 242071328 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs2012385 : nucléotide polymorphe en gras.

CTGGCGGATGCACTAGCCGGGCTGAGGGTCAGGAATAGCCTTGTGGCCGCTTGTGCTCCTCT
 20 GGCTCCTCCCAATGAGGGTCCTCTAGTGGAGCCTCCCAATGGGGCTCCTCTACCCCTCAGCAGTGCCTT
 TGGTCACCAGGTCTGTCTTGGTGCCAACAAATTCAGTTCTCAAACCATCTACTGAGCACCTGCTCTG
 GGCTAGGAGCCCTGGAGCCCTGATACAACCAAGAGGTAGAGCCCGGAGTATTGTTCTTGTCTGAGGAGA
 AGCTTCTGGAAGGTTTCAGCCACAAAGATGTCATCTGAGATCAGCTTTGAAAACATTGGACAGGAGCAG
 GTTCGAGAATGGGAGGAGGAAAGGAGGGTTCTCCTAAGTATTCAAATTAGCACCAGGAGCAGGTTTCGA
 25 GAATGGGAGGAGGAAAGGAGGGTTCTCC (**C/T**)GAGTATTCAAATTAGCACCAAGAGCAGGTTTCGAGA

ATGGGAGGAGGAAAGGAGGGTTCTCCTAAGTATTCAAATTAGCACCACTCGTCCACCACAGGGCGTT
 AGATAAGAAAAAGAATCCTGCCAGTATCAGACACCTGCGCAGATAGGGTAAGCGAGAGTCCTGGGAG
 CCCCTCAGATTCTTAACCTGGACTGCTCTGGAGCCCTTCCACCATCTGTTCTTTTCAGACAACAGGAG
 GAGCAGCAGGTGTCCGGAGAATGTGCTAGGGGCCTCCTAGTATGAGCAGTCCCACATACTGCGTGAGC
 5 AGAAGGAGGAGCCACTCACGAATATCCTCACAGAACGCAGATGAAAAACAAGCCAAACAGAAACGTCA
 CCCACACATGAAGAAGGTGGTCATATGGATG

Les SNP voisins du SNP **rs2012385** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle
 10 241767109-242119399 du chromosome 2 ou entre les SNP rs1540528 et rs7567892 du chromosome 2.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs1540528	2	-303719	241767109-241767609
rs16843438	2	-284703	241786125-241786625
rs2074840	2	-280686	241790142-241790642
rs2055566	2	-71468	241999360-241999860
rs2012385	2	0	242070828-242071328
rs7567892	2	48071	242118899-242119399

15 **SNP rs2190453** localisé en 11p15.1 sur le chromosome 11 entre les positions 17489723-17490223 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs2190453 : nucléotide polymorphe en gras.

20 AGCCGCAGACCATACTCTAAGTAGCCTCAGAGCCACACCTGAGATGGAGAGGCCAGCCT
 TAGACTCTGGTGGGGTAGAGTGAAGAGGACAGACTCAAATCTCTAAGCCAGGTGTATCAAAGGCTAAC
 CTGAGACCTACCATCTGGTCAGAAAGGCTAACCTCAGACTCACACCCCCGACCAAGGAGGCTAGTTT
 CAATTCCAAAGCCAGGAGCAAGACTCACACCCCCAAGCAAGGAGATTAGTTTCAATTCTTAAGCCAGG
 AGCTAACCTCAGATGGCCCTGGGCAGGTGGCATGATCTCTCTCTCCAGGCTGGGGAGCAGGAAAGGGC
 25 TCACTCCACCCTTGTATGCCATTTGAGGAGAACAACCTCAGCTGGTCTCTGGGAGCACATGGAGAAC
 (A/G) ACCACATTGTGTCCAGGGTTGCTTGCCCTGGCCTGCAGGCAGGACACATACCTCCTGGGCCAG
 CCGGTTGATCTTTAGCTGCTTTTCTTCTCCAGCATTTCTCTTTCTCTTTGTAAAGCTTTTGCTCAA
 ACTCCAGTTCTTTCTTATTCTTTCTCAAGTCCTGCAGGCTGCCATACTTGGCTTTCTTCTTATCTTTT
 CCTTTCTGAGTAGATGTGGCATTGTTTATATGACAAAGGTTAGAAATAGTGTGACAGCACAGCACAC

GGGGCATCCAGTCCTCACATAACACAACCATCCCATGGTGAGCCCCTCCCCAGCTCTCTCACCACTC
TGGACATCAGACCTCAGGTTTAGGACAGGAAGGCCACTGCTACCTACTGCAGAGTGGGAGACACA

Les SNP voisins du SNP **rs2190453** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 17464539-17757162 du chromosome 11 ou entre les SNP rs12278956 et rs1003921 du chromosome 11.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs12278956	11	-25184	17464539-17465039
rs1006099	11	-2934	17486789-17487289
rs2190453	11	0	17489723-17490223
rs2190454	11	238	17489961-17490461
rs7119071	11	39005	17528728-17529228
rs1003921	11	266939	17756662-17757162

10

SNP rs888298 localisé en 17q24.2 sur le chromosome 17 entre les positions 63955680 à 63956180 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs888298 : nucléotide polymorphe en gras.

CTTAGAAAAAAGGGATTTGGggccaggtgcggtgggtcacacctgtaatccctgcacttt
gggaggccgaggtgggtggatcacgaggtcaggagatcgagaacatcctggctaacatggtgaaaccc
catctctactaaaaatacaaaaacattagccgggctggtggcaggtgcttgtagtcccagctacttg
ggaggggtgaggcaggagaattgcttgaacacgggaggtagaggttgtggtgagctgagactgcactcc
agcctgggcaacagagtgaactctatctcaaaaaaaaaaaaaaaaaaaaaagataaaaGGGATTTTG
GATCCTTATAACACCTTATCCAAATCTTTAACTTTTTCCTGTTTTTCAAAAAAGAACTGTGCTGTCT
GAAGGCCTGAGGAAGTAGCAGACTGAGTGCTACAGAATAGAACAGGACACACTCCCCTTGGGCCTTTA
TCATTTCCCCAGAGTGGGCAGTCCTCCCGGACACC (**A/G**) CAGAATCCCTACCTGGCAAGAGAGGCTG
CAGCAGCTGAGTTGCTTAAACCAAAATTTAAGTCCCAAACCTGAAAGTTTAAAGAAAAGCAAACCCCC
AATACTTCCCAGACCTGTTTCAAATCATTTCTTGTCGGAGAAGAAATGTAAAGGAAGGGAGAACTCTTA
GATATTGGTTCCAATGAACCGATGCTCATCTTGGTT

25

Les SNP voisins du SNP **rs888298** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de

données selon le tableau suivant et sont positionné dans l'intervalle 63815611-64165896 du chromosome 17:

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs7211107	17	-140069	chr17:63815611-63816111
rs888298	17	0	chr17:63955680-63956180
rs887281	17	209716	chr17:64165396-64165896

5

SNP rs8110935 localisé en 19q13.43 sur le chromosome 19 entre les positions 62239851-62240351 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

10 **Séquence génomique au voisinage de rs8110935 : nucléotide polymorphe en gras.**

TTTAAAAACAATTTTTGTTCTCTCGGTAAGTGTGGTCTCCATTCATCCCAGTGTGTTTC
 CCTGAAAGCAGAGATCcttctccaaattcatgttgaagtcctaaacccagtacctcagaatgagatt
 gtatthttgagatgggcctttacagaggttaattaagggttaaataatgatattatcagggttaggcctaattcc
 aatatggctgggtgtccttatagaaggagattaggacacagacacacacaggggatgaccacgtga
 15 ggagaggagggaagacgggccaatacagagccaagcagagacaccttagcagaaaccaacccctgcccac
 accttgatgttgacctgcagcctccagaactgtgaaaattttctgttacatgagccaccagtcctgtg
 gtactttattatggctgccagagcagactaagacaGTCACCCATTTAAGGGGAAAAAAGGAAGTTC
 AGGTTGAAGAAACAGGAAACATTCTGAAAACATGCATATAATCAACAAGAAAACAAAGAATTATTAG
 20 CATATTAGAAATGGAAAAAAGTccggggcgcatggctcatgcaggtaatcccagcacttcggggagggc
 tgaggcaggcagatcacctgaggtcaggagttcgagaccagcctggccaatatggtg (**A/C**) atcccc
 gtctagaatatgaagcaggcagaagaacgtgaaaaactagactggcttagcctcccagcccacatctt
 tctcccatgctggatgctccctgccattaaacatcagactccaagttcttcagttttgggactcggac
 tggctctccttgctcctcagcttgagatggcctattgtgggaccttgatgatcatgtgagttaatatt
 25 taataaaactccctaataatatacctatcagttctgtccctctagagaacactgactaatacaCCCAGACT
 TGCAGAAATCACCTCACCTTCAACACCAGCATTTCTGGCCTGGGGGCTGGACATGCAGGCTGGCCTGTT
 CCTTTGCAATCATCCCAGCATCACAGAGGCCACTGTGGCTGCATGGACCTATCACTCCTGACCTGTTG
 TTACTCCCTCTCCTCATCTTCCCTGTCCTGCCCTTGAGACggctccacttcctgaactcccaaatc
 caacttccacattccatcttcattgctaacaccctggaccagggcactgagatctctaccctacaaga
 30 ccacggcaccctcctcatggggctccccacctccacaccaggccctgggtcctccaccttcccaacag
 gagccagagggagagctttaagtcataaaacagatgatgttgccctctccttgccattcggacttaciaa
 ctttccagtgggcctccaatgaacctacaatgaaatccaaaatccCCAGCATAAGAGTAT

Les SNP voisins du SNP **rs8110935** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 62026584-62294837 du chromosome 19 ou entre les SNP rs1860565 et rs1565944 du chromosome 19.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs1860565	19	-213267	chr19:62026584-62027084
rs8110935	19	0	chr19:62239851-62240351
rs1565944	19	54486	chr19:62294337-62294837

SNP **rs2788140** localisé en 1q32.3 sur le chromosome 1 entre les positions 210171227-210171727 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs2788140 : nucléotide polymorphe en gras.

CCAATACAGTGCACATTCTTCAATATATCATTGAAGATCCTCCACAATTAGACACAGGCC
 15 TAGCAGCCAGACCTCTCttttcttttttttttttttttgagacggagtcctcgctctgtcgcccaggctgga
 gtgcagtgggcgcagtcctcggtcaccgcaagctcgcctcccggttcattgccattctcctgcctcag
 cctcccgagtagctgggactacaggcgctgccaccacgcccggctaattttttgtatttttagtaga
 gacgggggtttcaccgtgttagccaggatgggtctcgatctcctgacctcgtgatctgcccgcctcgggc
 tcccaaagtgtgggattacaggcgtgagccactgcacccggccCAGACCTCTCTTTTCTACGGCCCT
 20 CTGTGTGTATCCAGCCCGCAGTAAACTGGCACCTGGGCATTCCATGAGCTCAGTTTGCACTATCT
 TACCTTTGTGGCTTTGCTCATATTTTCCCTCT (**A/G**) TCTGAACACTCTTCCCTCCATCCGTGAAAAA
 CCTGTTTCGTCCTCCATGTCCTGATTTCTAGCCAGACACAATACTCAGTATTCCTCCATAGCCCGTAT
 CCCAATCCATCTGTGTGAAGCAGTCTAGCTGCATGGCCCTGGGGTCGGAGGCACTGTAGACAAATGGA
 GGCTAATGTTACCATGTCCTGCCAGGAGCAGCCAGCTCCCTCCACTGCCCCATGCCTCCCATCAGCTC
 25 CCTGGCTATT

Les SNP voisins du SNP **rs2788140** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 210157195-210446272 du chromosome 1 ou entre les SNP rs12135924 et rs7546833 du chromosome 1.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs12135924	1	-14032	chr1:210157195-210157695
rs2788140	1	0	chr1:210171227-210171727
rs7546833	1	274545	chr1:210445772-210446272

- 5 **SNP rs7934514** localisé en 11q22.1 sur le chromosome 11 entre les positions 99214118-99214618 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs7934514 : nucléotide polymorphe en gras.

10 GTAACCAAGCTAAGACTGGATATAGATCCCACAGATATTTTGGAAATGATGCCTGAAATGAATCGTT
 CTTCTTCCAGTTCTGAAAGCTTATGGCCCTATGATAGCATAAAAAATCAAACATCTATCAAGTATTTT
 ATTTTCTCCAGTATCACTCTTTGTAAATGATACTTCTATCTCTTATTTTTGTTTTTTCATCttttat
 ttttaaaataattttCT **(C/T)** ACAATTAATATAGGGAGAGGAAAAATGGTTtatttagttacctattc
 ctatatttaaaaaatcctcaaaacttagcaatttaaaacaacaatcaagcattttctcttcaagtctg
 aaatctgagtagcttagctgggaggttctggtcttaggtctttcatgaggctgcagtcagtcgtgctg
 15 ttatagctccattctcatttgaaaactttacaaagggaggatccacttaacaattcacctatgtgatt
 gttgttaggctcagtttcttctgctgccttttgccaagccaggtatttcagttccttaccatgtcgcc
 ctctccacagcctgaaaaaatttcttggatagcaatggtcttcttcttgaggagtgacccacga
 ggaaagtgtacccagaaggaagttgcattacttagtattagaagtaatatagtagtccttttgcttt
 tagctagaataagtcattaagtcagctgacactcacggggaaagaaattaagctcaactccttgaa
 20 gggaggggttatcaaaaaagttgtggacatatcttttaactaACCCAAGTAGGTTGGAAAAATTCTT
 CACAAGTAGGTTTGGAAAAATCTTCACAAGTTAATTGGTCTAAAGATGATATAAAAGGCATGTTTAC
 TTTATATCATTATTTTGAAATACAATTAAAACAAACAAGATTAAAAAGGAGGCATGAAAAGGTTACTT
 TCATTGAA

- 25 Les SNP voisins du SNP **rs7934514** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 99092040-99333419 du chromosome 11 ou entre les SNP rs605559 et rs12574821 du chromosome 11.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs605559	11	-122078	chr11:99092040-99092540
rs1441381	11	-88366	chr11:99125752-99126252
rs10750395	11	-78780	chr11:99135338-99135838
rs2583150	11	-58325	chr11:99155793-99156293
rs7934514	11	0	chr11:99214118-99214618
rs12574821	11	118801	chr11:99332919-99333419

SNP rs3828054 localisé en 1q21.3 sur le chromosome 1 entre les positions 149779269-149779769 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

5 **Séquence génomique au voisinage de rs3828054 : nucléotide polymorphe en gras.**

TGAGACCCGCGGCCCAAGCACGGGCTCGCCGGCGCCGAGTCCAGGCAGGAGCCGCAGTGTG
 CTACCAAAGGGCAGGGACGCCCCGAACCTCCAGCCTCAAAGGAGTCTTCACCCGCGACTCCCACTG
 10 CCCGTCGCAGGCAAAAGAATAAAAAGAGAGAAGCGCCGCGCAGGGCTGACCGCGCGAGCCGGGCACCA
 GGTGATGTCAGCCAACACGGCGCGGGGCACGGAAGGGGCGGACTTAGAAACCGGGAATACAAAACGGA
 GAAGACAGCGAGAGCGCTTTTTCTTACCGCCGCC (**C/T**)GGTCCTCTGGGTGCACGTCCACCAGGGTA
 CACCAGTTCCGCGTCCCGTTTCATCTTCCCTCGGGGTCGCAGCACACACGCCACTTGTCCACCCCGCTG
 TCTGGCTCCAACCTGGGCGGGCGCGCGGAACCGCCCCCTTGTATAGGCCCATCAGGGGCGGGGCTGA
 15 AGATAGGCCGCGCCCCCAGTTCGCGGTTTCGCAGAGAACTAACGATAGGCGAGGAGGTGAGGTGGGCG
 GAGCCAATGGGTCTGGGACATGCCCCATCGGTGCTCGCATAGATTTACACAAAGGTGGGGCTTGGGA

Les SNP voisins du SNP **rs3828054** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle
 20 149382371-149874970 du chromosome 1 ou entre les SNP rs11807526 et rs6702842 du chromosome 1.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs11807526	1	-396898	chr1:149382371-149382871
rs3828054	1	0	chr1:149779269-149779769
rs6702842	1	95201	chr1:149874470-149874970

SNP rs1499955 localisé en 3q13.31 sur le chromosome 3 entre les positions 116719413-116719913 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs1499955 : nucléotide polymorphe en gras.

CCTCTATTACAGATGTCTAGAATAACAAGCAAATTTAACCCTATCACCTACGGCACAAA
 CTTGCAAAAGCTGTCCACACCATTTTTTCTTTCTTGCTTGCTTTAATTGTCAGGCTGCCCATTCCTCC
 CACTTCTGTCTATTTTCTTAAAGCACACGAGTTCCTAGTTGATAGTATGGTGGAGAAGAGTAGAAA
 10 CAGCATGGTCTATTTATTTTATTTTAAATTCACCTAGTATTCACAAATAAGAAACGGGTATTTGTAGA
 AAAAATATATCATATATAAAAAAGTAGATAAGTCCCA (**G/T**) GCAGGCCATTTTTTAGCTGATATTTAC
 TTATTGCAGATTCATACAAGGGTTAAATTAGATAAAACACTTTGCGTGCTGCTAATAAACAATATAAA
 TGTAAAAATACAATTCTGTAGACGTTAAAGTACAAATGGAATAGTATTTACATTTCAAAGGAAC'TT
 GGGTTCAGTCAGCCTTTATAGGTATAAGAAATGATGTAACAGAACTATCACTGGACTAGCAGTAAGGA
 15 AACCTGGGCTCCAACCTTGCTTTATCACAGTCTCTAAATGACTGTGATATTAGAAAAGTCACTCATTT
 T

Les SNP voisins du SNP **rs1499955** pouvant apporter de
 l'information sur la susceptibilité au cancer de la prostate sont définis dans
 20 notre base de données selon le tableau suivant et sont positionnés dans
 l'intervalle 116302446-117011700 du chromosome 3 ou entre les SNP
 rs9289008 et rs2289271 du chromosome 3

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs9289008	3	-416967	chr3:116302446-116302946
rs17755786	3	-296763	chr3:116422650-116423150
rs7428182	3	-118281	chr3:116601132-116601632
rs7650434	3	-92831	chr3:116626582-116627082
rs1353909	3	-75480	chr3:116643933-116644433
rs1499954	3	-75317	chr3:116644096-116644596
rs1499955	3	0	chr3:116719413-116719913
rs2289271	3	291787	chr3:117011200-117011700

25 **SNP rs4855539** localisé en 3p14.1 sur le chromosome 3 entre les
 positions 69108069-69108569 suivant la localisation déterminée par l'UCSC
 genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs4855539 : nucléotide polymorphe en gras.

AAGTCAATGTCTTTAGTTTGTTTTTTCTGGTCTTACTTTTCACAGGGAAAAATTCCTCT
 TCATGAGGCTAATTTGAAGTTTTGAAATTAAAGACTGGAATACTTTCATGCTGACAGAGGTAGACGC
 ACACGCACTGGTATATGCAGTTACAAATACTCGCATAAAATGGAAACCATTATTTTCATATATAAATTA
 5 ATTAATCACAAATGCTCTCCATGGCTAAGAAGGAATCAGTGGAAACCAGACAGAAGGTATGCAAGACA
 GTCCTACAGAATGTTCTAATTTGCTTTTATCACATG (C/T) AGTTGCTACATTTTAGGAAAACATGAT
 TTAAATATGAAACATGTAATATAAATTAATATAGTGGCATGATTTATTACAGGTCTCTCGATGCATATAA
 CCTGGAGGTGACTAAACGCTGATCTATAACATGGTCTTATAGCTTGGTACTGAGAATCACAACTCTGC
 GTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTATGTTTTGCATGTTTTCTTTCTTACCA
 10 CAAACAGTGTTATAACCAGATTATGGCAAATAAAAGAACAGTTGTAAATTTACCCAAATATATCATAA
 A

Les SNP voisins du SNP **rs4855539** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 69049525-69153397 du chromosome 3;

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs6768792	3	-58544	chr3:69049525-69050025
rs6785239	3	-24227	chr3:69083842-69084342
rs4855539	3	0	chr3:69108069-69108569
rs1745	3	44828	chr3:69152897-69153397

SNP rs4242382 localisé en **8q24.21** sur le chromosome 8 entre les positions 128586505-128587005 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

Séquence génomique au voisinage de rs4242382 : nucléotide polymorphe en gras.

25 CTTACAGCATACCCGAAAGCATTGGTGAGGACACAAAACTACAGATAAGAATCAGATTC
TAAAAAGACAATTCTCTTTTCCATTCTCTGTCCTCTCCCCTGCAACTTCCCAATCCCTCACCTCTAATT
AACCCGCCCCACCCCTTCACTAGCTTCTGATTTTCAGGCAACGTCCAGTACTTGTTCACCTTTCTCTCT
GACCAGCCATCAAGAAGATCTTGTATGTTTCTCCTACACACCCCTGCCCTGGACCCAGGAATTCTTC
CATTTTTTCCATATTTGGGCTATATTAAGTAATAAGCCACATGCTTTTCTGTTGAGAAAATACAAAAG
30 ATGTTTCCCTCTGTCATAAAGAAAAAGAGGTAACCCAGGGAACATTTTGTCCCTCTAGTTATCTTCCC
(A/G) CAGGCCCATCAAGAATCAGGCAGTAGGTGAAAAAGAAACACAGAGAACCCTAGGAACACAATAG

GAAGACCACCATGGGCCCTTAGGGAGTCAGCGAAGGCTTATGATGCAAAAAGAAGGTCCCAGGTACCT
TAAAAACTCCACTTCCCTCTCTAGGATCCCCAAGAGAGCTTGACAGCGTCCCTCTATGCAGATGTTCA
TAAATCAGGCATATGTAACCTCTGCGGTTTCTTGCACATAATTGATCACAGTTGAGCTGCTCAGACATT
AAATCCAAAGGACATCAGAGAAGGACGAGTTCAGTAAAGAACACTGAGAAAGAAGTGGACCCCTGAGCA
5 TAGATCTTGGCATAACATGCGTGGGAAATGGCCTCTCAAGGGGTCATTATCCATTCAATTACACAC

Les SNP voisins du SNP **rs4242382** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans
10 l'intervalle 128539973-128619555 du chromosome 8 ou entre les SNP rs7830412 et rs4407842 du chromosome 8.

SNP	chromosome	distance (pb) au SNP principal	localisation UCSC genome browser assemblage mars 2006
rs7830412	8	-46532	chr8:128539973-128540473
rs1447293	8	-45253	chr8:128541252-128541752
rs921146	8	-42388	chr8:128544117-128544617
rs4871799	8	-34931	chr8:128551574-128552074
rs1447295	8	-32535	chr8:128553970-128554470
rs9297758	8	-30985	chr8:128555520-128556020
rs7831028	8	-25544	chr8:128560961-128561461
rs11775749	8	-22907	chr8:128563598-128564098
rs16902169	8	-21067	chr8:128565438-128565938
rs13253127	8	-20982	chr8:128565523-128566023
rs6985504	8	-20797	chr8:128565708-128566208
rs7831150	8	-18135	chr8:128568370-128568870
rs723555	8	-17474	chr8:128569031-128569531
rs16902173	8	-13574	chr8:128572931-128573431
rs17766217	8	-13076	chr8:128573429-128573929
rs12155672	8	-10549	chr8:128575956-128576456
rs1562432	8	-9971	chr8:128576534-128577034
rs4871808	8	-4028	chr8:128582477-128582977
rs4242382	8	0	chr8:128586505-128587005
rs4242384	8	981	chr8:128587486-128587986
rs7017300	8	7695	chr8:128594200-128594700
rs11988857	8	14300	chr8:128600805-128601305
rs9656816	8	17081	chr8:128603586-128604086
rs12542685	8	20010	chr8:128606515-128607015
rs7837688	8	21787	chr8:128608292-128608792
rs6991990	8	27810	chr8:128614315-128614815
rs13258742	8	31105	chr8:128617610-128618110
rs4407842	8	32550	chr8:128619055-128619555

SNP rs11526176 localisé en **7p15.2** sur le chromosome 7 entre les positions 27546048-27546548 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

5 **Séquence génomique au voisinage de rs11526176 : nucléotide polymorphe en gras.**

```

CATACTTCTAAATGAAAGTTACTTGCTTTTCAAGAAAAATTTGAAGTCCATGGGTTATTG
CTGCGTGATTGTACTACAAATAGAGAGGACTATGGCAAGTACAGTTGACCCTTGAATGATGAGGGGT
10 TAGGGGTGCCAACCCCCAGTGCAGTCAAAAACCCATGTATAACTTTTGACTCTCCAAAACTTAACTA
CTAATAGCCCACTGTTGACTGGAAGCCTCGTCAATAACATAAACAGTTGATTAACACATATTTTGTAT
ATGTATTATATATTGTATTCTTATGGTAAAGCAAGCTAGAGAAAAAATGTTACTAAGGGAATCATTA
AGGAAGATAAAAATATATTTATTATTCATTAAGTGGAGTGGATCATCATAAAGGTCTTCAATCCCATC
ATCTTAATAATGAGTAGGCTGAGGAGGAAGAGGAGGGTTGCTCTTCGCTGTCTCGGGGTGACAGAGG
15 CAGAAGAGGTGGAGGTGGTAGAAGGGGAGGCAGAAGGGGCAGGCACACTCCGGATAACTTTATGGAAA
TTGTAATTTCTATCTGATGTTTTTGTCTTTTCATTTCTCTAAAAACGTTTTTGTATGGTACCAATC (C
/T)GTCTTCCACTGTTTGCTTTATTTTTCAGTGTCTGTATCAGAGAAGGGTCCATGTTGTAAGAAGT
TGAAAGGAGTCTTGAATAATCAGAACCGTTCTGCCATACTGTCTAATGTCAATTTGTTTCTGGCACT
GCTTTTGGTACATCTTCTTCCCTCATCATCTGGTACTGTTTCAGAAGCACTCATCTCCATCAAGCCTCTT
20 CTGTTAATTACTCTGCTGTGGTGTCTATTAGCTCTTGAATTAATCCAAGATCCATATCTTGAAGCCT
TCATACACTCCCCACCTTTTTTGGCATATGCACAATCTCTTTAGTGATTTCTTGGATTGGCCCTGCCA
TAAATCCTGTGAAGTCTTGCACAACATCTGGACAGTTTTTTTCCAGCAGGAATTTACTGTTAGGGGCTT
GATGGCCTTCAAGGCGTTTTTCCACAATAACAATGGCATCTTCAATGGTGTAATCTTTCCAGATTTTCA
TGTTCTATCAGGGTTTTCTTCCACAGTGACAATCCTTCCCATAGAGTACCATGTGTAATGAGCCTTAA
25 AGGTCCTTATGATCCCCTACTCTAGAGGCTGAATTAGGGGCGTTATGTTTAGGGGCAAGTTGGCCCCCT
TGGACACCTTCAGTGTGAACTCATGTTATTCTGGGTGGCCAGGGGTACTGTCCAATATCAAAATAAC
TTTAAAAGTCAGTCCCTTACTGGCAAGATATTGCCTGACTCCAGAGACAAAGCCATTGATGGAAACAA
TCCAGAAACAGGGTTCTCATCGTCCAGGCCCTTCTTGCTGTACAACCAAAGACAGGCAGCTGGTATTT
ATCTTTTCACTTAAAGCCTCAGAAGTTAGCAACTTTATAGATAAGGGCAGTCCTGATTTTCAACCCAA
30 CTGCATTTGTACAAAACAGTAGAGTTAGCCTATCCTTTTCTGCCTTAAATCCTGGTGCTGCTTGCTTC
TCTTCCTAATAAATGTCCTTCGAGCATCCTTTTTTTTTTTTTTCTCCGTAATAGGGCACTTCTGTCT
TGCAATAAAACTCATTCAGGCAGATATACTTTCTCTTCAATGATTTTTTCTTAATGGCGCCTGGGAA
CTGTCTGCTGTCTCTTGGTTGGCAGAAGCTACTTCGCCTATTTCTTGACATTTTTTAAGCAAACCTCT
TCCTAAAATTATCAAACCATCCTTTGCTGGCATTAATTCTCCAGCTTTAGATCCTTCACTTTCTTTT
35 TGCTTTAAGTTGTCATATTTTCTTGAATCATATTAGATGTAAGTATGCCTTTCTACAGCAATCCTGC

```

ATCTACATAAAAGCTGCATTTTCAATGTGAGATAAAAAGATGTTCTGCAAAAAGTGCAAGCCTGCTGG
 AGTAGCTGCAGTGATGGGTTTCATGACTATTCTTTTCTTTGTTTACAATGGTCCTTACATTGGATTTGT
 TTATCTTGAAATGGAGGGCAAACGCAGCCGCAGACCTCAATCCATGGTATGTATCAGGCAATTCAACT
 TTTTCTTGTAATGTCATGACTTTTCTCAGCTTCTTAGGAGCACTTCCAGCATCACTAGTGCGCACTTTG
 5 TATGGGTCCCATGGTGTCAATTCAAGGTTTATGGTATTGCACTAAACATGATAAAAAATACAAGAGAA
 TTCCAAGAGATCAATTTTTACTATGATACACAATTTACTAAAGAGATGAACCACTCACACAAAAGATGA
 TTAGTGTCACATGACATTTTATGCTCAATACTTGTAACTTGTAGTTCACTGCAATAGCAACAGGTGG
 CCACAAAATTATTACAGTAGTACAGTATTACTAGAGTTAATTTTATGCCATTATGATTTAATGCATCT
 TTACATTTCTTTACATTTCTCTCAACTGTAAATGGTGCCATGTATGGTCTATAAATATTTGTAAACTT
 10 TGATAAATTTTAACTCTTTATAACAGATTTGTGCATATTTATAAACTAGTATCTATCTACATATATTT
 TATGCGTTCACGACATATCTAACTTTTTCTT

Les SNP voisins du SNP **rs11526176** pouvant apporter de
 l'information sur le risque de survenue de cancer de la prostate sont définis
 15 dans notre base de données selon le tableau suivant et sont positionnés
 dans l'intervalle 27414591-27808301 du chromosome 7 ou entre les SNP
 rs11761572 et rs2237344.

SNP	chromosome	distance pb	localisation UCSC genome browser assemblage mars 2006
rs11761572	7	-131457	chr7:27414591-27415091
rs11526176	7	0	chr7:27546048-27546548
rs10447552	7	103525	chr7:27649573-27650073
rs42088	7	122088	chr7:27668136-27668636
rs2237344	7	261753	chr7:27807801-27808301

20 **SNP rs6492998** localisé en **15q15.1** sur le chromosome 15 entre les
 positions 39,333,673-39,334,173 suivant la localisation déterminée par
 l'UCSC genome browser, assemblage de mars 2006.

**Séquence génomique au voisinage de rs6492998 : nucléotide
 polymorphe en gras.**

25 ACCTCCTTATTGAGACTGAAGTTCAGGCTAGGTTGTGCATCACCCTTGATACTAGACTTGGTATTTA
 AACTGCCTTTTCTCAGCTAAAGTTTCTTAAGCTTGTTAGACATTAACTGAAGTATGTAGCCATGCAA
 TTCAAATCAGCCTTAGTCTTAATTTAAAAGTGAGTAGTTATTGTTTCTTGACCTCTGTCAGACA (**A/G**
 30) GAGGAGCTACATTTTGATGATAGTGTAGACTTTGTATTACAGAACAAATTATGTAATAAAGCTTAG
 TACATGTTTGTGTAATTAATAATCAGGACCTCGGTAATTTTCTCTTTCATCATCTTAAAGCAATCCAG
 TTATCTTATGAATGACTTCTTCTGGTTCATGCATTGATATAAAATTATTACACTAAATGGTCAAG

Les SNP voisins du SNP **rs6492998** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 38991207-39584443 du chromosome 15 :

5

SNP	chromosome	distance pb	localisation USC genome browser assemblage mars 2006
rs12592197	15	-337006	chr15:38991207-38991707
rs6492997	15	-5460	chr15:39328213-39328713
rs6492998	15	0	chr15:39333673-39334173
rs170296	15	250270	chr15:39583943-39584443

SNP rs6681102 localisé en **1q43** sur le chromosome 1 entre les positions 236,853,987-236,854,487 suivant la localisation déterminée par l'UCSC genome browser, assemblage de mars 2006.

10 **Séquence génomique au voisinage de rs6681102 : nucléotide polymorphe en gras.**

AAGGACTGAAAAC TGCAATAGAGTTACCAGAGATGCCATTCTTTTAAAATTCAGCAACGTTTCATTTCC
 ATTGTGCTTAAAGTTTTGTATTTCTCTTTTAGCAACATAGGTTTGAAGACTATTTTACAATATTGT
 15 ATAGAATATAAACTTCAAAGTACATATTTCTATGTAAAGTCACATGCTGTATAATGACATTTcagt
 ggtcccataagattataatggagctggaaaattcctattgcctcgtatttacaatactatatatttttac
 tgttatttttagagtgtaccccgacttattaaaaaaatcaaacaagttaactataatacagcctcagg
 ctgtcttcacgagggcatccagaagaaggtattgttatcataggagatgacacctctatgcttggtatt
 gcccctgaataccttccagtgggacaagaggtggaggtggaaaacagtgatattgatgatcctgactt
 20 gtgcaggcctaggctaattgtatgtgtctgtgtcttaatttttaccaaagttttaaaggtaaaaaatt
 gggaaaaagcttattgaataaggatataaagaatatgttttgtacagctctgcgatattgttttaaact
 acgttattactaaagagtcaaaaagccttaaaaaacttaaaaaattatttaataaaaaagttacagtat
 gctaagggttaattttattattgaagaaaaaattaacaagtttagtattgtctgatttgtaaatgctcat
 aaagtctatagtagtattagtaatatcctaggccttcacatacactccccattcactctgactcacc
 25 cagagcaacttccagtcctgcaagctccattcatggtaagtgacactgtacaggtgtcccatggctgga
 aaccatcattctcagcaactaacacaggaacagaaaaccaaaccgcagtggtctcactcataaatg
 ggagttgcacaatgagaacgcatggacacaaggagggggaatatcacacactggggcctgtcgtgggt
 ggggggctaggggagggtatagcattagaagaaatacctaattgtatgatgacgggttaattgggtgcagca
 aaccaccatggcagctgtatacctatgtaacaaacctgcacgttctgcacatgtatcccagaacttaa
 30 agtataataaagaaagtaaaaaaaaaaattcttttatacttttttactgcgcttttctatgtttaga
 tagacacatacttactgttgtgttataactgcctacagtatatagtagtaacatgctacacaggtt
 ttagagccaggagcaataggctatactatataggctagggtgtgtggttagactatgatatactaaattg
 tacactctatgatgttcacacaatgatggaatcacctaacttatcaggacgtatccc (**c/t**) ggtg
 ttaagcaacacatgattTTGTTATACTAACAATTCTCTTAGAGATTATTGGGGAAAAATTTAATAAGA
 35 TATTTCTTACGTTTGTAAATAGACCATCAGTGGTGACGCTCTAACAAGCTGTCATGAAGATGGCCATAC
 ACAACAATTCTGCGTGTTTCTTTTGCTATTTAAGAGTGCTCTGTTTGGGAACCTGACTTATAAACC
 GTGGTTCTGGCCA

Les SNP voisins du SNP **rs6492998** pouvant apporter de l'information sur la susceptibilité au cancer de la prostate sont définis dans notre base de données selon le tableau suivant et sont positionnés dans l'intervalle 236815776-236998150 du chromosome 1 :

40

SNP	chromosome	distance pb	localisation UCSC genome browser
rs652252	1	-38211	chr1:236815776-236816276
rs10754645	1	-1597	chr1:236852390-236852890
rs6681102	1	0	chr1:236853987-236854487
rs7547641	1	34418	chr1:236888405-236888905
rs2174076	1	50252	chr1:236904239-236904739
rs2689128	1	143663	chr1:236997650-236998150

Ainsi les 4 variables dites d'antécédents ainsi que la variable de
 5 catégorie d'âge peuvent être combinées aux SNP mentionnés ci-dessus
 comme variables d'entrée d'algorithmes de type régression logistique MLP
 SVM RVM ou encore une fusion de ces 4 algorithmes ou encore un autre
 type d'algorithme d'apprentissage statistique pour réaliser un calcul de risque
 de cancer de la prostate.

10 Parmi toutes les combinaisons possibles de variables d'entrée ont
 été retenues comme particulièrement pertinentes (toutes les localisations
 nucléotidiques citées correspondent à celle définie par UCSC genome
 browser, assemblage de mars 2006) :

- la combinaison des quatre variables d'antécédents de cancer et
 15 d'une variable de catégorie d'âge ;

- la combinaison des quatre variables d'antécédents de cancer,
 d'une variable de catégorie d'âge et d'une variable définissant le génotype lié
 au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-
 128447913 du chromosome 4;

20 - la combinaison d'une variable de catégorie d'âge et d'une variable
 définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans
 l'intervalle 127602673-128447913 du chromosome 4;

- la combinaison des quatre variables d'antécédents de cancer,
 d'une variable de catégorie d'âge, d'une variable définissant le génotype lié
 25 au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-
 128447913 du chromosome 4, d'une variable définissant le génotype lié au
 SNP rs7576160 ou à l'un de ses voisins dans l'intervalle 37855761-
 38126567 du chromosome 2 et d'une variable définissant le génotype lié au
 SNP rs2012385 ou à l'un de ses voisins dans l'intervalle 241767109-
 30 242119399 du chromosome 2 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs7576160 ou à l'un de ses voisins dans l'intervalle 37855761-38126567 du chromosome 2 et d'une variable définissant le génotype lié au SNP rs2012385 ou à l'un de ses voisins dans l'intervalle 241767109-242119399 du chromosome 2 ;

- la combinaison des quatre variables d'antécédents de cancer, d'une variable de catégorie d'âge d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673 à 128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs2190453 ou à l'un de ses voisins dans l'intervalle 17464539-17757162 du chromosome 11 et d'une variable définissant le génotype lié au SNP rs888298 ou à l'un de ses voisins dans l'intervalle 63815611-64165896 du chromosome 17 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs2190453 ou à l'un de ses voisins dans l'intervalle 17464539-17757162 du chromosome 11 et d'une variable définissant le génotype lié au SNP rs888298 ou à l'un de ses voisins dans l'intervalle 63815611-64165896 du chromosome 17 ;

- la combinaison des quatre variables d'antécédents de cancer, d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs2788140 ou à l'un de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 et d'une variable définissant le génotype lié au SNP rs7934514 ou à l'un de ses voisins dans l'intervalle 99092040-99333419 du chromosome 11. ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs2788140 ou à l'un de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 et d'une variable

définissant le génotype lié au SNP rs7934514 ou à l'un de ses voisins dans l'intervalle 99092040-99333419 du chromosome 11 ;

- la combinaison des quatre variables d'antécédents de cancer, d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs3828054 ou à l'un de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 et d'une variable définissant le génotype lié au SNP rs1499955 ou à l'un de ses voisins dans l'intervalle 116302446 - 117011700 du chromosome 3 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs3828054 ou à l'un de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 et d'une variable définissant le génotype lié au SNP rs1499955 ou à l'un de ses voisins dans l'intervalle 116302446 - 117011700 du chromosome 3 ;

- la combinaison des quatre variables d'antécédents de cancer d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs8110935 ou à l'un de ses voisins dans l'intervalle 62026584-62294837 du chromosome 19 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs8110935 ou à l'un de ses voisins dans l'intervalle 62026584-62294837 du chromosome 19 ;

- la combinaison des quatre variables d'antécédents de cancer, d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs4855539 ou à l'un de ses voisins dans l'intervalle 69 049 525-69 153 397 du chromosome 3 et d'une variable définissant le génotype lié

au SNP rs4242382 ou à l'un de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans
5 l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs4855539 ou à l'un de ses voisins dans l'intervalle 69 049 525-69 153 397 du chromosome 3 et d'une variable définissant le génotype lié au SNP rs4242382 ou à l'un de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8 ;

10 - la combinaison des quatre variables d'antécédents de cancer, d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 selon, d'une variable définissant le génotype lié au SNP rs11526176 ou à l'un de ses voisins dans l'intervalle 27414591-
15 27808301 du chromosome 7 ;

- la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs2174183 ou à l'un de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4, d'une variable définissant le génotype lié au SNP rs11526176 ou à l'un de ses voisins dans
20 l'intervalle 27414591-27808301 du chromosome 7.

- la combinaison des quatre variables d'antécédents de cancer, la combinaison d'une variable de catégorie d'âge, d'une variable définissant le génotype lié au SNP rs6492998 ou à l'un de ses voisins dans l'intervalle 38991207- 39584443 du chromosome 15, d'une variable définissant le
25 génotype SNP rs6681102 ou à l'un de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1, d'une variable définissant le génotype lié au SNP rs11526176 ou à l'un de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7.

- la combinaison d'une variable de catégorie d'âge, d'une variable
30 définissant le génotype lié au SNP rs6492998 ou à l'un de ses voisins dans l'intervalle 38991207-39584443 du chromosome 15, d'une variable définissant le génotype SNP rs6681102 ou à l'un de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1, d'une variable définissant le génotype lié au SNP rs11526176 ou à l'un de ses voisins dans
35 l'intervalle 27414591-27808301 du chromosome 7.

REVENDECATIONS

1. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie comportant le recueil de données individuelles (x_i) d'entrée et la fourniture d'une information prédictive de risque (y) liée à un type de maladie, caractérisé en ce que :
 - on recueille des informations représentatives qui sont des informations génétiques et/ou des résultats d'informations cliniques d'un patient pour obtenir lesdites données individuelles ;
 - on saisit les données individuelles (x_i) à l'aide de moyens de saisie de données ;
 - on réalise un outil de prédiction en construisant au moins un modèle par apprentissage statistique, les variables d'entrée de ce modèle étant lesdites informations représentatives ;
 les informations génétiques d'entrée comportant au moins une variable ou une combinaison de variables parmi les suivantes (toutes les localisations nucléotidiques citées correspondent à celles définies par « UCSC genome browser », assemblage de mars 2006) :
 - variable définissant le génotype lié au SNP **rs2174183** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673 à 128447913 du chromosome 4 ;
 - variable définissant le génotype lié au SNP **rs7576160** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 37855761-38126567 du chromosome 2 ;
 - variable définissant le génotype lié au SNP **rs2012385** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 241767109-242119399 du chromosome 2 ;
 - variable définissant le génotype lié au SNP **rs888298** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 63815611 à 64165896 du chromosome 17 ;
 - variable définissant le génotype lié au SNP **rs8110935** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 62026584-62294837 du chromosome 19.
 - variable définissant le génotype lié au SNP **rs2190453** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 17464539 à 17757162 du chromosome 11 ;
 - variable définissant le génotype lié au SNP **rs2788140** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 ;
 - variable définissant le génotype lié au SNP **rs3828054** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 ;
 - variable définissant le génotype lié au SNP **rs1499955** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 116302446-117011700 du chromosome 3.

- variable définissant le génotype lié au SNP **rs4855539** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 69049525 - 69153397 du chromosome 3 ;
- variable définissant le génotype lié au SNP **rs4242382** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8 ;
- 5 - variable définissant le génotype lié au SNP **rs11526176** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7
- variable définissant le génotype lié au SNP **rs7934514** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 99092040-99333419 du chromosome 11
- variable définissant le génotype lié au SNP **rs6681102** et/ou à l'un ou à
- 10 plusieurs de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1
- variable définissant le génotype lié au SNP **rs6492998** et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 38991207-39584443 du chromosome 15.

2. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou
 15 la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4.

3. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou
 la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une
 25 variable définissant le génotype lié au SNP rs7576160 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 37855761-38126567 du chromosome 2 et/ou d'une variable définissant le génotype lié au SNP rs2012385 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 241767109-242119399 du chromosome 2.

4. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou
 la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et/ou d'une
 35 variable définissant le génotype lié au SNP rs2190453 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 17464539-17757162 du chromosome 11 et/ou d'une variable définissant le génotype lié au SNP rs888298 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 63815611-64165896 du chromosome 17.

5. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et/ou d'une variable définissant le génotype lié au SNP rs2788140 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 210157195-210446272 du chromosome 1 et/ou d'une variable définissant le génotype lié au SNP rs7934514 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 99092040-99333419 du chromosome 11.

6. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et/ou d'une variable définissant le génotype lié au SNP rs3828054 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 149382371-149874970 du chromosome 1 et/ou d'une variable définissant le génotype lié au SNP rs1499955 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 116302446-117011700 du chromosome 3.

7. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le génotype lié au SNP rs8110935 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 62026584-62294837 du chromosome 19.

8. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le génotype lié au SNP rs4855539 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 69049525-69153397 du chromosome 3 et/ou d'une variable définissant le génotype lié au SNP rs4242382 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 128539973-128619555 du chromosome 8.

9. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une
5 variable définissant le génotype lié au SNP rs2174183 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 127602673-128447913 du chromosome 4 et d'une variable définissant le génotype lié au SNP rs11526176 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7.
- 10 10. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 1, caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable définissant le génotype lié au SNP rs6492998 ou à l'un de ses voisins dans l'intervalle 38991207-39584443 du chromosome 15 et d'une
15 variable définissant le génotype lié au SNP rs11526176 et/ou à l'un ou à plusieurs de ses voisins dans l'intervalle 27414591-27808301 du chromosome 7 et d'une variable définissant le génotype le génotype lié au SNP rs6681102 ou à l'un de ses voisins dans l'intervalle 236815776-236998150 du chromosome 1.
- 20 11. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 10 caractérisé en ce que les données d'entrée correspondent à la combinaison d'une variable liée à l'âge et à des données cliniques et/ou à des données d'anamnèse personnelles et familiales.
- 25 12. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 11, caractérisé en ce que la maladie étant de type cancer, les données d'anamnèse correspondent à des variables d'antécédents personnels ou familiaux de cancer, à
30 une variable liée à l'âge ou à toute combinaison de ces variables.
13. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'un cancer selon la revendication 12, caractérisé en ce que les données d'anamnèse incluent la combinaison de
35 quatre variables d'antécédents de cancer et d'une variable de catégorie d'âge, lesdites variables d'antécédents concernant respectivement des antécédents familiaux d'un cancer du sein, des antécédents familiaux de cancer de la prostate, des antécédents personnels de cancer, des antécédents familiaux d'autres cancers.

14. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 13, caractérisé en ce que la maladie est un cancer

5

15. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 13, caractérisé en ce que la maladie est un cancer urologique.

10

16. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 13, caractérisé en ce que la maladie est un cancer hormonodépendant.

15

17. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 13, caractérisé en ce que la maladie est un cancer de la prostate.

18. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 1 à 17, caractérisé en ce qu'il comprend :

- la constitution d'une base de données d'exemples (Bex) constitués de données d'entrées (x_{mi}) et de résultats avérés (y_m^*) ;
 - la construction d'au moins un modèle optimal par apprentissage statistique comportant les étapes suivantes :

- le choix d'une famille (F) de fonctions multivariées ($f_1, \dots, f_i, \dots, f_N$);
- pour une fonction donnée f_i , l'élaboration d'un modèle défini par ajustement de paramètres θ_j tels que l'estimation délivrée par le modèle $y_m = f_i(x_{mi}, \theta_j)$ soit la plus proche possible de celle du résultat avéré y_m^* ;
- la comparaison des différentes estimations de manière à définir une fonction f_i optimisée f_{iop} permettant de définir un modèle optimal ;
- l'exploitation dudit modèle optimal à partir desdites données individuelles (x_i) de manière à fournir ladite information prédictive (y) de risque liée à une maladie.

30

35

19. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication

18, caractérisé en ce qu'il comprend la construction en parallèle d'un ensemble de modèles optimaux, chaque modèle étant élaboré à partir d'une famille (Fk) de fonctions, l'information prédictive de risque lié à une maladie résultant de la fusion de l'ensemble des modèles optimaux.

5

20. Procédé de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon l'une des revendications 18 à 19, caractérisé en ce que la famille de fonctions est de type des MLP (Multi Layer Perceptron), sous-ensemble de la famille des réseaux de neurones
10 ou de type des Support Vector Machines (SVM) ou de type des Relevant Vector Machines (RVM) ou de type des modèles fréquentistes s'apparentant à la méthode des plus proches voisins.

21. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou
15 la prise en charge thérapeutique ou le pronostic d'une maladie comportant des premiers moyens de saisie de données d'information individuelles (1, 10, 11) par un utilisateur, au moins une première interface logicielle sur laquelle opèrent lesdits premiers moyens, caractérisé en ce qu'il comporte en outre des moyens (2) exploitant un logiciel mettant en œuvre le procédé selon l'une des revendications 1 à
20 20 et fournissant une information prédictive de risque lié à une maladie.

22. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 21 caractérisé en ce que ladite information prédictive de risque est restituée à
25 l'utilisateur via ladite interface logicielle.

23. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 22, caractérisé en ce qu'il comporte en outre des moyens de
30 communication entre les premiers moyens de saisie et le logiciel, permettant la transmission des données d'information et celle de l'information prédictive.

24. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la
35 revendication 23, caractérisé en ce qu'il comporte en outre des seconds moyens de saisie de données d'information individuelles et une seconde interface logicielle, les premiers moyens de saisie concernant la saisie d'informations de type clinique, les

seconds moyens concernant la saisie d'informations issues de prélèvement sur l'individu .

23. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie selon la revendication 22, caractérisé en ce qu'il comporte en outre des moyens de communication entre les premiers moyens de saisie et le logiciel ,
5 permettant la transmission des données d'information et celle de l'information prédictive.

24. Dispositif de prédiction individuelle pour le dépistage ou le diagnostic ou la prise en charge thérapeutique ou le pronostic d'une maladie
10 selon la revendication 23, caractérisé en ce qu' il comporte en outre des seconds moyens de saisie de données d'information individuelles et une seconde interface logicielle, les premiers moyens de saisie concernant la saisie d'informations de type clinique, les seconds moyens concernant la saisie d'informations issues de prélèvement sur l'individu .

15

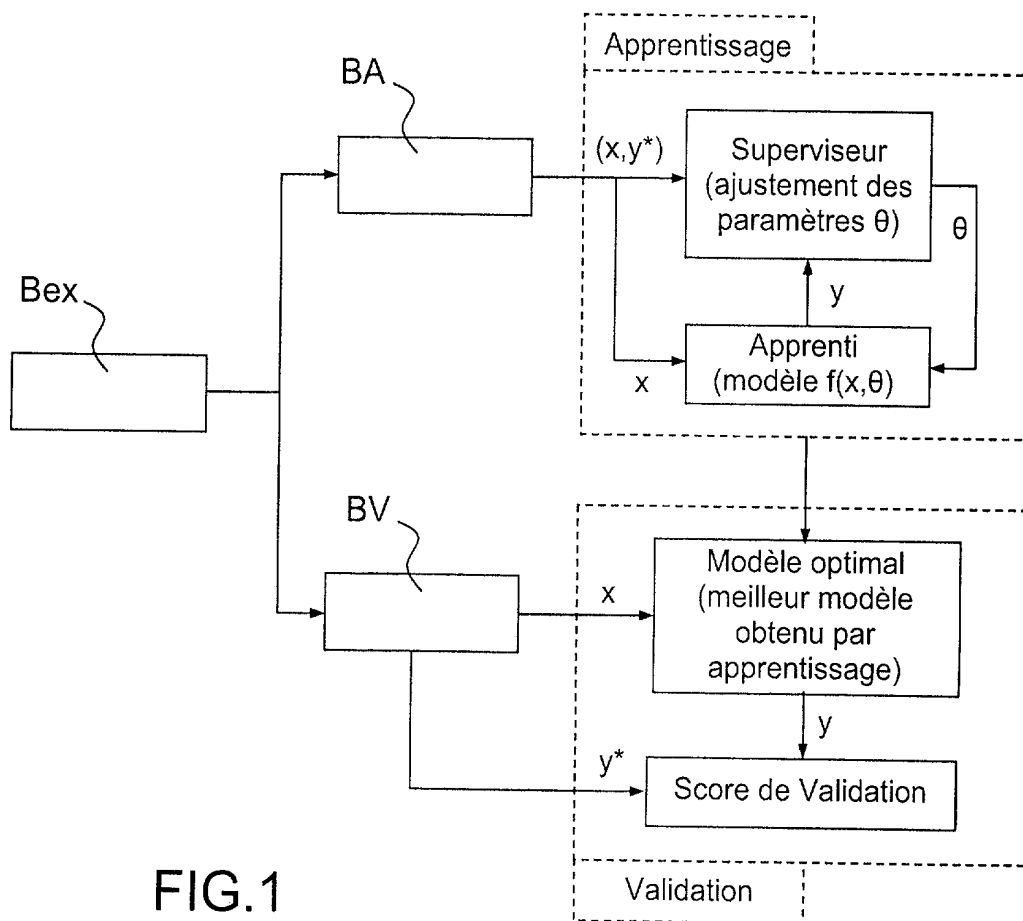


FIG.1

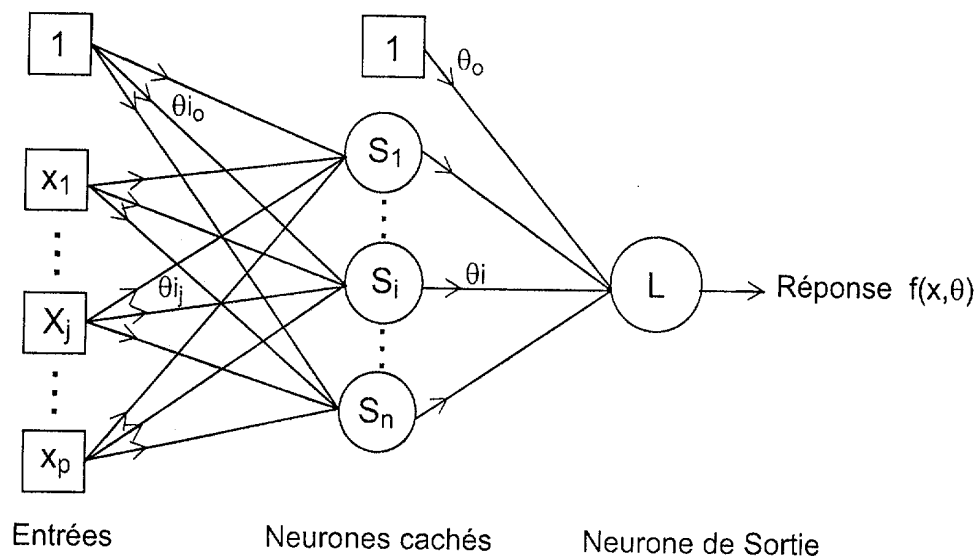


FIG.2

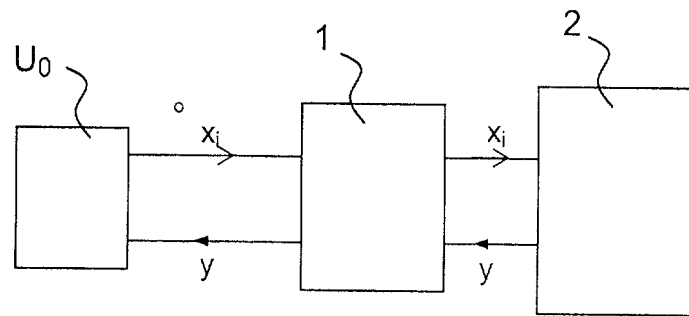


FIG.3

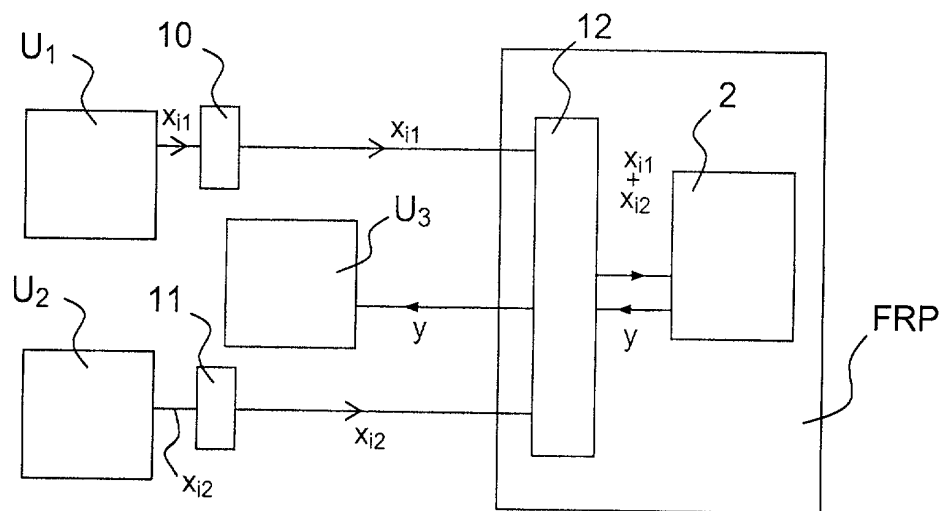


FIG.4

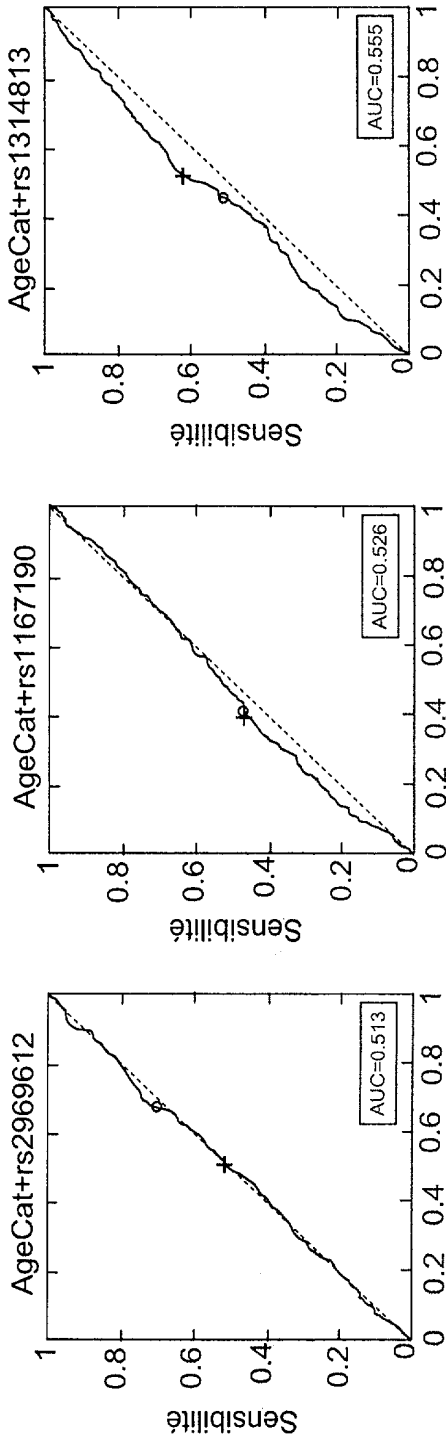


FIG.5a

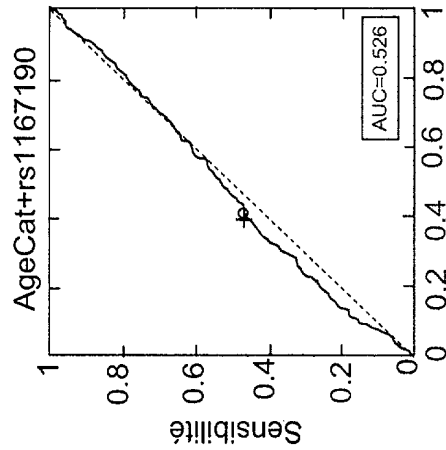


FIG.5b

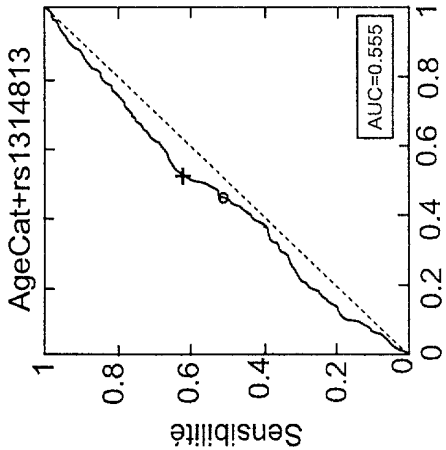


FIG.5c

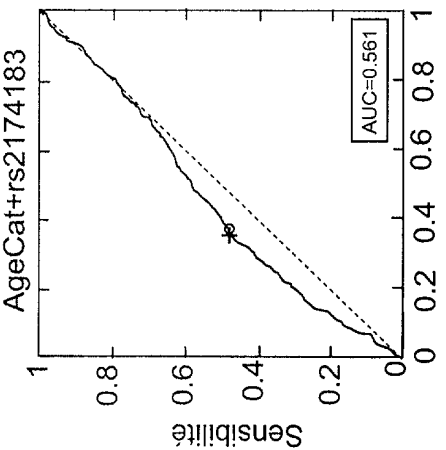


FIG.5d

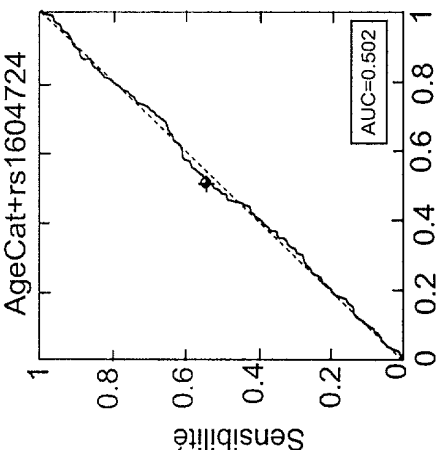


FIG.5e



RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications
déposées avant le commencement de la recherche

N° d'enregistrement
national

FA 713378
FR 0804414

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
X	US 2007/092888 A1 (DIAMOND CORNELIUS [US] ET AL) 26 avril 2007 (2007-04-26) * abrégé; revendications 1-3,7-10,27; figures 1-4 * * alinéa [0040] * * alinéa [0081] - alinéa [0084] * * alinéa [0099] - alinéa [0103] * * alinéa [0130] - alinéa [0143] * * alinéa [0157] - alinéa [0174] * * alinéa [0183] - alinéa [0193] * * alinéa [0209] - alinéa [0212] * * alinéa [0320] - alinéa [0328] *	1-24	G06N3/00 C12Q1/68
X	WO 2007/109571 A (PROMETHEUS LAB INC [US]; HARVEY JEANNE [US]; NERI BRUCE [US]; SINGH SH) 27 septembre 2007 (2007-09-27) * abrégé; figure 3 * * alinéa [0056] - alinéa [0063] * * alinéa [0191] *	1-24	
X	GB 2 444 410 A (NAVIGENICS INC [US]; FILIPPONE MELISSA FLOREN [US]; CARGILL MICHELE [U]) 4 juin 2008 (2008-06-04) * abrégé; revendications 102,125; figures 1-4,16 * * alinéa [0194] - alinéa [0217] * * alinéa [0035] - alinéa [0044] * * alinéa [0090] - alinéa [0098] * * alinéa [0171] - alinéa [0183] *	1-17	DOMAINES TECHNIQUES RECHERCHÉS (IPC) G06F C12Q
A	US 2008/025591 A1 (BHANOT GYAN VATSARAJA [US] ET AL) 31 janvier 2008 (2008-01-31) * abrégé; revendications 1-4; figure 1 * * alinéa [0005] - alinéa [0009] * * alinéa [0022] - alinéa [0030] * * alinéa [0049] - alinéa [0058] *	1-24	
		-/--	
Date d'achèvement de la recherche		Examineur	
8 avril 2009		Türkeli, Yasemin	
CATÉGORIE DES DOCUMENTS CITÉS X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant			



RAPPORT DE RECHERCHE PRÉLIMINAIRE

établi sur la base des dernières revendications
déposées avant le commencement de la recherche

N° d'enregistrement
national

FA 713378
FR 0804414

DOCUMENTS CONSIDÉRÉS COMME PERTINENTS		Revendication(s) concernée(s)	Classement attribué à l'invention par l'INPI
Catégorie	Citation du document avec indication, en cas de besoin, des parties pertinentes		
D,A	TIPPING M E: "Sparse Bayesian learning and the relevance vector machine" JOURNAL OF MACHINE LEARNING RESEARCH, MIT PRESS, CAMBRIDGE, MA, US, vol. 1, no. 3, 1 juin 2001 (2001-06-01), pages 211-244, XP008099985 ISSN: 1532-4435 * le document en entier * -----	1-24	DOMAINES TECHNIQUES RECHERCHÉS (IPC)
Date d'achèvement de la recherche		Examineur	
8 avril 2009		Türkeli, Yasemin	
CATÉGORIE DES DOCUMENTS CITÉS X : particulièrement pertinent à lui seul Y : particulièrement pertinent en combinaison avec un autre document de la même catégorie A : arrière-plan technologique O : divulgation non-écrite P : document intercalaire		T : théorie ou principe à la base de l'invention E : document de brevet bénéficiant d'une date antérieure à la date de dépôt et qui n'a été publié qu'à cette date de dépôt ou qu'à une date postérieure. D : cité dans la demande L : cité pour d'autres raisons & : membre de la même famille, document correspondant	

**ANNEXE AU RAPPORT DE RECHERCHE PRÉLIMINAIRE
RELATIF A LA DEMANDE DE BREVET FRANÇAIS NO. FR 0804414 FA 713378**

La présente annexe indique les membres de la famille de brevets relatifs aux documents brevets cités dans le rapport de recherche préliminaire visé ci-dessus.

Les dits membres sont contenus au fichier informatique de l'Office européen des brevets à la date du **08-04-2009**

Les renseignements fournis sont donnés à titre indicatif et n'engagent pas la responsabilité de l'Office européen des brevets, ni de l'Administration française

Document brevet cité au rapport de recherche	Date de publication	Membre(s) de la famille de brevet(s)	Date de publication
US 2007092888 A1	26-04-2007	AUCUN	

WO 2007109571 A	27-09-2007	AUCUN	

GB 2444410 A	04-06-2008	WO 2008067551 A2	05-06-2008

US 2008025591 A1	31-01-2008	US 2008187207 A1	07-08-2008
